昇腾全栈AI软硬件平台

HUAWEI

昇腾AI 基础软硬件平台

极简易用　极致性能　构筑智能世界的基石

超强算力
超强AI算力，更优能效比

端边云协同
一次开发，多次部署

全栈开放
软硬件开放，使能伙伴

使能应用
业务快速上线

行业应用

智慧城市、制造、能源、交通、金融、运营商、教育等更多行业

应用使能

ModelArts　HiAI Service　第三方平台

MindX 昇腾应用使能

- **MindX DL** 深度学习使能
- **MindX Edge** 智能边缘使能
- **ModelZoo** 优选模型库
- **MindX SDK** 行业SDK

AI框架

[M]S 昇思 MindSpore
最佳匹配昇腾AI处理器算力的全场景AI框架

TensorFlow / PyTorch等第三方框架
可基于第三方框架开发的模型进行二次开发、训练和推理

异构计算架构

CANN
统一异构计算架构，释放昇腾硬件澎湃算力

系列硬件

端　边　云

全流程开发工具链 MindStudio

管理运维工具 FusionDirector / SmartKit

昇腾社区 hiascend.com

昇腾社区　MindSpore官网　MindSpore掌中宝

Atlas 200 AI加速模块

Atlas 200 AI加速模块（型号：3000）可以在端侧实现目标识别、图像分类等，广泛用于智能摄像机、机器人、无人机等端侧AI场景。

极致性能

- 半张信用卡大小即可提供22 TOPS INT8算力，支持20路高清视频实时分析（1080P 25FPS）
- 多级算力配置，支持22/16/8 TOPS三级算力

超低功耗

- 支持毫瓦级休眠、毫秒级唤醒，典型功耗仅6.5 W，使能边缘AI应用

应用场景

嵌入边缘设备，使能智能边缘

摄像机　机器人

无人机　工控机

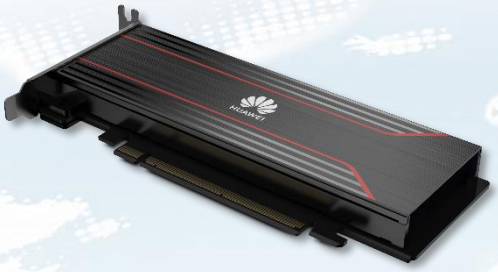
图像分析　视频分析

图像分割　物体识别

产品规格

AI算力	22/16/8 TOPS INT8 11/8/4 TFLOPS FP16
内存规格	LPDDR4X，8 GB/4 GB，总带宽51.2 GB/s
编解码能力	• 支持H.264硬件解码，20路1080P 25 FPS（2路3840*2160 60 FPS） • 支持H.265硬件解码，20路1080P 25 FPS（2路3840*2160 60 FPS） • 支持H.264硬件编码，1路1080P 30 FPS • 支持H.265硬件编码，1路1080P 30 FPS • JPEG解码能力1080P 256 FPS，编码能力1080P 64 FPS，最大分辨率：8192*4320 • PNG解码能力1080P 24 FPS，最大分辨率：4096*2160
接口	• PCIe ×4 Gen3.0 • ×1 USB2.0 / USB3.0 • ×1 RGMII
串行总线	UART / I2C / SPI
接口规格	144 pin BTB连接器
典型功耗	4GB：6.5W/8GB: 9.5W
工作环境温度	-25℃ ~ 80℃（-13℉ ~ 176℉）
重量	30 g
结构尺寸	52.6 mm * 38.5 mm * 8.5 mm

Atlas 300V 视频解析卡



Atlas 300V 视频解析卡融合“通用处理器、AI Core、编解码”于一体，提供超强AI推理、视频图片编解码等功能，具有超大视频解析路数、高性能特征检索、安全启动等优势，支持100路高清视频实时分析，可广泛应用于智慧城市、智慧交通、智慧园区、智慧金融等诸多AI行业场景。

超强算力

- 单卡最大提供100 TOPS INT8算力，为数据中心推理提供更强大支持
- 支持 8 core * 1.9 GHz CPU计算能力

超大视频解析路数

- 支持100路高清视频实时分析
- 支持JPEG和视频硬件编解码，提升图片和视频类应用性能

安全启动

- 设备启动链完整，启动初始状态确定，防止后端植入

应用场景

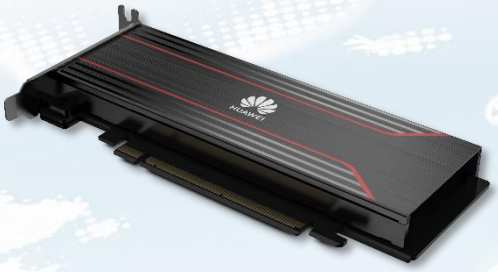
集成于服务器中，进行AI推理

智慧城市　智慧交通　智慧金融　智慧园区

产品规格

形态	半高半长PCIe卡
AI算力	100 TOPS INT8 50 TFLOPS FP16
内存规格	LPDDR4X 24GB，总带宽 204.8 GB/s
编解码能力	• 支持H.264/H.265 硬件解码，100路 1080P 25 FPS / 80路 1080P 30FPS / 10路 4K 60FPS • 支持H.264/H.265 硬件编码，30路 1080P 25 FPS / 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：8192x8192
PCIe	x16 Lanes，兼容x8/x4/x2 PCIe Gen4.0，兼容3.0/2.0/1.0
功耗	72 W
结构尺寸	169.5mm（L） x 68.9mm（H） x 18.45mm（W）
工作环境温度	0℃ ~ 55℃（32℉ ~ 131℉）

Atlas 300I Pro 推理卡



Atlas 300I Pro 推理卡融合“通用处理器、AI Core、编解码”于一体，提供超强AI推理、目标检索等功能，具有超强算力、超高能效、高性能特征检索、安全启动等优势，可广泛应用于OCR识别、语音分析、搜索推荐、内容审核等诸多AI应用场景。

超强算力

- 单卡最大提供140 TOPS INT8算力，为数据中心推理提供更强大支持
- 支持 8 core * 1.9 GHz CPU计算能力

超高能效

- 提供 2 TOPS/W 超高能效比，达到业界2.1倍

安全启动

- 设备启动链完整，启动初始状态确定，防止后端植入

应用场景

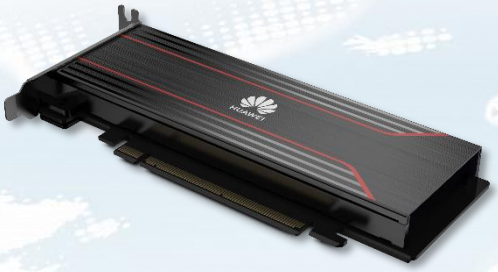
集成于服务器中，进行AI推理

OCR 识别　语音分析　搜索推荐　内容审核

产品规格

形态	半高半长PCIe卡
AI算力	140 TOPS INT8 70 TFLOPS FP16
CPU算力	8 core * 1.9 GHz
内存规格	LPDDR4X 24 GB，总带宽 204.8 GB/s
编解码能力	• H.264、H.265视频编解码 • JPEG图片编解码
PCIe	PCIe x16 Gen4.0
功耗	最大72 W
工作环境温度	0℃ ~ 55℃（32℉ ~ 131℉）
结构尺寸	169.5 mm * 68.9 mm

Atlas 300V Pro 视频解析卡



Atlas 300V Pro 视频解析卡融合“通用处理器、AI Core、编解码”于一体，提供超强AI推理、视频图片编解码等功能，具有超大视频解析路数、高性能特征检索、安全启动等优势，支持128路高清视频实时分析，可广泛应用于智慧城市、智慧交通、智慧园区、智慧金融等诸多AI行业场景。

超大视频解析路数

- 支持128路高清视频实时分析
- 支持JPEG和视频硬件编解码，提升图片和视频类应用性能

安全启动

- 设备启动链完整，启动初始状态确定，防止后端植入

产品规格

形态	半高半长PCIe卡
AI算力	140 TOPS INT8 70 TFLOPS FP16
内存规格	LPDDR4X 48 GB，总带宽 204.8 GB/s
编解码能力	• 支持H.264硬件解码，128路 1080P 30 FPS（16路 3840*2160 60 FPS） • 支持H.265硬件解码，128路 1080P 30 FPS（16路 3840*2160 60 FPS） • 支持H.264硬件编码，24路 1080P 30 FPS（3路 4K 60 FPS） • 支持H.265硬件编码，24路 1080P 30 FPS（3路 4K 60 FPS） • JPEG解码能力4K 384 FPS，编码能力4K 192 FPS，最大分辨率：8192*8192
PCIe	PCIe x16 Gen4.0
功耗	最大72 W
工作环境温度	0℃ ~ 55℃（32℉ ~ 131℉）
结构尺寸	169.5 mm * 68.9 mm

应用场景

集成于服务器中，进行AI推理

智慧城市　智慧交通　智慧金融　智慧园区

Atlas 300I Duo 推理卡

Atlas 300I Duo 推理卡融合“通用处理器、AI Core、编解码” 于一体，提供AI推理、视频分析等功能，具有超强算力、超高能效、高性能视频分析等优势，可广泛应用于互联网、智慧城市、智慧交通等多场景，支持检索聚类、内容审核、OCR识别、语音分析、视频分析等多应用。

超强算力

- 单卡最大提供280 TOPS INT8算力，为中心推理提供更强大算力支持
- 支持 16 core * 1.9 GHz CPU计算能力

超高能效

- 提供 1.86 TOPS/W 超高能效比，业界领先

高性能视频分析

- 支持256路高清视频实时分析
- 支持JPEG和视频硬件编解码，提升图片和视频类应用性能

应用场景

集成于服务器中，进行AI推理

互联网　智慧城市　智慧交通　智慧园区

检索聚类　内容审核　视频分析　语音分析

产品规格

形态	单槽位全高全长（10.5英寸）
AI算力	280 TOPS INT8 140 TFLOPS FP16
CPU算力	16 core * 1.9 GHz
内存规格	LPDDR4X 48GB，总带宽408 GB/s；支持ECC
编解码能力	• 支持H.264硬件解码，256路 1080P 30 FPS（32路 3840*2160 60 FPS） • 支持H.265硬件解码，256路 1080P 30 FPS（32路 3840*2160 60 FPS） • 支持H.264硬件编码，48路 1080P 30 FPS（6路 4K 60 FPS） • 支持H.265硬件编码，48路 1080P 30 FPS（6路 4K 60 FPS） • JPEG解码能力4K 768 FPS，编码能力4K 384 FPS，最大分辨率：8192*8192
• PCIe接口	• x16 Lances，兼容x8/x4/x2 • PCIe Gen4.0，兼容3.0/2.0/1.0
功耗	150W
尺寸	266.7mm（长）×111.15mm（高）×18.46mm（宽）
工作温度	0℃～55℃（32℉～131℉）

Atlas 300T Pro 训练卡

华为Atlas 300T Pro训练卡配合服务器，为数据中心提供强劲算力的AI加速卡，单卡可提供最高280 TFLOPS FP16算力，加快深度学习训练进程。Atlas 300T Pro具有超强算力、高度集成、高速带宽等特点，满足互联网、运营商、金融等需要人工智能训练以及高性能计算领域的算力需求。

超强算力

- 内置30个达芬奇AI Core
- 提供业界领先的280 TFLOPS FP16算力

高度集成

- AI算力、通用算力、I/O能力三合一
- 处理器集成30个华为达芬奇AI Core + 16个TaiShan核 + 1 * 100GE RoCE v2网卡

高速带宽

- 支持PCIe 4.0和 1*100G RoCE高速接口，出口总带宽56.5 Gb/s
- 无需外置网卡，训练数据和梯度同步效率提升10%-70%

产品规格

形态	全高全长，双槽位
AI算力	280 TFLOPS FP16
编解码能力	• 支持16 channel 4K（或64 channel 1080P） 60 FPS H.264/H.265 • JPEG解码能力 1080P 2048 FPS, 或等价的解码能力, 最高分辨率为8192*4320 • PNG解码能力 1080P 240 FPS, 或等价的解码能力, 最高分辨率为4096*2160 • JPEG编码能力 1080P 256 FPS, 或等价的编码能力, 最高分辨率为8192*4320
内存规格	• 32 GB HBM • 16 GB DDR4
网络	1*100GE QSFP-DD接口
PCIe	PCIe x16 Gen4.0
功耗	最大300W①
散热方式	被动风冷
工作温度	5℃~45℃（41℉ ~ 113℉）

① 持续调优中，数值根据调优结果动态更新

应用场景

模型训练　HPC

智慧城市　智慧交通

智能制造　智慧金融

构建万物互联的智能世界

HUAWEI

Atlas 500 智能小站

Atlas 500智能小站（型号：3000）是面向边缘应用的产品，具有超强计算性能、体积小、环境适应性强、易于维护和支持云边协同等特点，可以在边缘环境广泛部署，满足在安防、交通、社区、园区、商场、超市等复杂环境区域的应用需求。

智能边缘

- 业界领先的集成AI处理能力的边缘产品
- 无风扇散热，支持-40℃至70℃室外工作

小身材大能量

- 机顶盒大小即支持22 TOPS INT8算力
- 支持20路高清视频处理（1080P 25FPS）

边云协同

- 支持LTE无线传输
- 云边协同，模型实时更新
- 可在云端统一进行设备管理和固件升级

应用场景

边缘侧独立部署，使能智能边缘

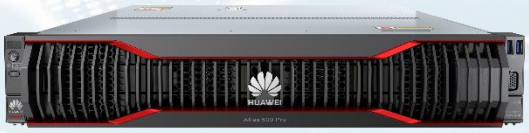
智慧变电站　智慧交通　智慧社区　环境监控

智能制造　智慧营业厅　无人零售　智能楼宇

产品规格

AI算力	22/16 TOPS INT8 11/8 TFLOPS FP16
内存规格	LPDDR4X，8 GB / 4 GB，最大51.2 GB/s
编解码能力	支持H.264硬件解码，20路1080P 30 FPS（2路3840*2160 60 FPS） 支持H.265硬件解码，20路1080P 30 FPS（2路3840*2160 60 FPS） 支持H.264硬件编码，1路1080P 30 FPS 支持H.265硬件编码，1路1080P 30 FPS JPEG解码能力1080P 256 FPS，编码能力1080P 64 FPS，最大分辨率：8192*4320 PNG解码能力1080P 24 FPS，最大分辨率：4096*2160
接口	网络：2个GE RJ45 其他I/O： 1个HDMI接口，1对3.5 mm立体声输入输出接口；2个外部和1个内部USB2.0接口（Type-A）
典型功耗	无盘配置：25 W 有盘配置：40 W
环境条件	无盘配置：-40°C~70°C 有盘配置：-40°C~60°C
结构尺寸	无盘配置：45 mm * 235 mm * 220 mm 有盘配置：45 mm * 355 mm * 220mm

Atlas 500 Pro 智能边缘服务器



Atlas 500 Pro 智能边缘服务器（型号：3000）是面向边缘应用的产品，具有超强计算性能、高环境适应性、易于部署维护和支持云边协同等特点。可以在边缘场景中广泛部署，满足在安防、交通、社区、园区、商场、超市等复杂环境区域的应用需求。

超强算力

- 最大支持3张Atlas 300VI/V Pro推理卡，满足多场景推理需求；整机可提供384路高清视频实时分析（1080P 30 FPS）
- 搭载鲲鹏920处理器，高效加速应用

超高能效

- 发挥鲲鹏架构多核、低功耗优势，为推理场景构建高效能、低功耗的AI计算平台
- Atlas 300I Pro单卡功耗仅72W，为AI服务器算力加速同时提供更优的能效比

应用场景

边缘侧独立部署，使能智能边缘

平安城市　智慧交通　智慧社区　环境监控

智能制造　智慧营业厅　无人零售　智能楼宇

产品规格

形态	2U AI服务器
CPU	1 * 鲲鹏920
CPU内存	4个DDR4内存插槽，最高3200 MT/s
AI加速卡	最大支持3张 Atlas 300V 视频解析卡 Atlas 300I Pro 推理卡 Atlas 300V Pro 视频解析卡
AI算力	最大420 TOPS INT8
本地存储	(8~12)*3.5 SAS/SATA
RAID支持	RAID 1/5/6/10等
PCIe	最多4个PCIe 4.0 x8标准扩展槽位
板载网卡	4*10GE/25GE(光口)+2*GE(电口)
电源	• 2个550W或900W交流热插拔电源，支持AC 220V/DC 240V或者2个1200W直流热插拔电源，支持DC -48V • 支持1+1冗余
风扇	4个热拔插风扇，支持N+1冗余
工作环境温度	• 长期：5°C~ 50°C • 短期：0°C~ 55°C
结构尺寸	86.1 mm * 447 mm * 475 mm

Atlas 800 推理服务器

型号：3000

Atlas 800 推理服务器（型号：3000）最大可支持8个Atlas 300I/V Pro，提供强大的实时推理能力和视频分析能力，广泛应用于中心侧AI推理场景。

超强算力

- 支持8张Atlas 300I/V Pro 推理卡，满足多场景推理需求；整机可提供1024路高清视频实时分析（1080P 30FPS）
- 搭载超强算力的鲲鹏920处理器，高效加速应用

超高能效

- 发挥鲲鹏架构多核、低功耗优势，为推理场景构建高效能、低功耗的AI计算平台
- Atlas 300I/V Pro 单卡功耗仅72W，为AI服务器算力加速同时提供更优的能效比

产品规格

形态	2U AI服务器
CPU	2 * 鲲鹏920
CPU内存	32个DDR4内存插槽，最高3200 MT/s
AI加速卡	最大支持8张 Atlas 300V 视频解析卡 Atlas 300I Pro 推理卡 Atlas 300V Pro 视频解析卡
AI算力	最大1120 TOPS INT8
本地存储	25*2.5 SAS/SATA 12*3.5 SAS/SATA 8*2.5 SAS/SATA+12x2.5 NVMe
RAID支持	RAID 0/1/10/5/50/6/60等
PCIe	最多支持9个PCIe4.0 PCIe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
电源	2个热插拔900 W或2000 W交流电源模块，支持1+1冗余备份
风扇	4个热拔插风扇，支持N+1冗余备份
工作环境温度	5°C ~ 40°C（41°F ~ 104°F）
结构尺寸	447 mm * 790 mm * 86.1 mm

应用场景

部署在数据中心机房，使能中心推理

精准营销　医疗影像分析　视频分析　OCR

智慧零售　智慧医疗　智慧城市　智慧金融

Atlas 800 推理服务器

型号：3010

Atlas 800 推理服务器 （型号：3010）是基于Intel处理器的推理服务器，最多可支持7个Atlas 300I/V Pro，支持896路高清视频实时分析，广泛应用于中心侧AI推理场景。

灵活配置，适配多项负载

- 支持SAS/SATA/NVMe/M.2 SSD硬盘多种组合灵活配置
- 支持板载网卡和灵活I/O卡，提供丰富多样的网络接口

智能视频分析

- 最大支持7张Atlas 300I/V Pro，支持896路高清视频实时分析（1080P 30FPS）

产品规格

形态	2U AI服务器
CPU	1/2个Intel® Xeon® SP Skylake 或 Cascade Lake处理器，最高205W
CPU内存	24个DDR4内存插槽，最高3200 MT/s
AI加速卡	最大支持7张 Atlas 300V 视频解析卡 Atlas 300I Pro 推理卡 Atlas 300V Pro 视频解析卡
AI算力	最大980 TOPS INT8
本地存储	8*2.5 SAS/SATA 12*3.5 SAS/SATA 8* 2.5 SAS/SATA+12*2.5 NVMe 24*2.5 SAS/SATA 24*2.5 NVMe 25*2.5 SAS/SATA
RAID支持	RAID 0/1/5/6/10/1E/50/60等
PCIe	10个PCIe Gen3.0接口 (含1个RAID控制卡+1个灵活LOM)
电源	可配置2个冗余热插拔电源，支持1+1冗余备份，选择规格如下： 550 W AC 白金电源、900 W AC 白金/钛金电源、1500 W AC 白金电源 1500 W 380 V 高压直流电源、1200 W -48 V ~ -60 V 直流电源
风扇	4个热拔插风扇，支持N+1冗余备份
工作环境温度	5℃ ~ 45℃（41℉ ~ 113℉）
结构尺寸	3.5英寸硬盘机箱尺寸： 86.1 mm * 447 mm * 748 mm 2.5英寸硬盘机箱尺寸： 86.1 mm * 447 mm * 708 mm

应用场景

部署在数据中心机房，使能中心推理

精准营销　医疗影像分析　视频分析　OCR

智慧零售　智慧医疗　智慧城市　智慧金融

构建万物互联的智能世界

HUAWEI

Atlas 800 训练服务器
型号：9000

Atlas 800 训练服务器（型号：9000）具有更高算力密度、极致能效比与高速网络带宽等特点。该服务器广泛应用于深度学习模型开发和训练，适用于智慧城市、智慧医疗、天文探索、石油勘探等需要大算力的行业领域。

更高算力密度
- 4U高度提供 2.24 PFLOPS FP16超强算力

极致能效比
- 提供2.24 PFLOPS/5.6 kW①超高能效比

高速网络带宽
- 8*100G RoCE v2高速接口
- 处理器间跨服务器互联时延缩短10~70%

产品规格

形态	4U AI服务器
CPU	4 * 鲲鹏920
CPU内存	• 最多32个DDR4内存插槽，支持RDIMM • 内存速率最高3200 MT/s • 单根内存条容量支持32 GB/64 GB
HBM	8 * 32 GB
AI算力	1.76 ~ 2.24 PFLOPS FP16
本地存储	• 2 * 2.5 SAS/SATA+3 * 2.5 NVMe • 2 * 2.5 SATA+3 * 2.5 NVMe • 2 * 2.5 SAS/SATA+6 * 2.5 NVMe • 2 * 2.5 SATA+6 * 2.5 NVMe • 2 * 2.5 SATA+8 * 2.5 SAS/SATA
RAID支持	支持 RAID 0/1/10/5/50/6/60
网络	8 * 100GE+ 4 * 25GE/2 * 100GE
PCIe扩展	最多支持2个PCIe 4.0扩展插槽
电源	4个热插拔2 kW或3 kW交流电源模块，支持2+2冗余
供电	• 200 ~ 240 V AC • 240 V DC
功耗	最大功耗5.6 kW①
散热方式	风冷
风扇	支持8个热拔插风扇模组，支持N+1冗余
工作温度	5℃ ~ 35℃（41℉ ~ 95℉）
结构尺寸	175 mm * 447 mm * 790 mm

应用场景

部署在数据中心机房，使能中心训练

- 模型训练
- HPC
- 智慧城市
- 智慧医疗
- 天文探索
- 石油勘探

构建万物互联的智能世界

① 持续调优中，数值根据调优结果动态更新

HUAWEI

Atlas 800 训练服务器

型号：9010

Atlas 800 训练服务器（型号：9010）具有更高算力密度、高速网络带宽等特点。该服务器广泛应用于深度学习模型开发和训练，适用于智慧城市、智慧医疗、天文探索、石油勘探等需要大算力的行业领域。

更高算力密度

- 4U高度提供 2.24 PFLOPS FP16超强算力

极致能效比

- 提供2.24 PFLOPS/5.6 kW①超高能效比

高速网络带宽

- 8*100G RoCE v2高速接口
- 处理器间跨服务器互联时延缩短10~70%

应用场景

部署在数据中心机房，使能中心训练

- 模型训练
- HPC
- 智慧城市
- 智慧医疗
- 天文探索
- 石油勘探

产品规格

形态	4U AI服务器
CPU	2 * Intel V5 Cascade Lake处理器
CPU内存	最多24个DDR4内存插槽，支持RDIMM
HBM	8 * 32 GB
AI算力	1.76 ~ 2.24 PFLOPS FP16
本地存储	• 2 * 2.5 SATA+8 * 2.5 SAS/SATA • 2 * 2.5 SAS/SATA+6 * 2.5 NVMe
RAID支持	支持 RAID 0/1/10/5/50/6/60
网络	8 * 100GE 1 * OCP NIC 3.0标卡，支持2 * 25GE
PCIe扩展	最多支持2个PCIe 3.0 x16和4个PCIe 3.0 x8扩展插槽
电源	4个热插拔2 kW或3 kW交流电源模块 支持2+2冗余
供电	• 200 ~ 240 V AC • 240 V DC
功耗	最大功耗5.6 kW①
散热方式	风冷
风扇	支持8个热拔插风扇模组，支持N+1冗余
工作温度	5℃ ~ 35℃（41℉ ~ 95℉）
结构尺寸	175 mm * 447 mm * 790 mm

① 持续调优中，数值根据调优结果动态更新

Atlas 900 PoD

Atlas 900 PoD（型号：9000）AI训练集群基础单元，具有超强AI算力、更优AI能效、极佳AI拓展等特点。该基础单元广泛应用于深度学习模型开发和训练，适用于智慧城市、智慧医疗、天文探索、石油勘探等需要大AI算力的领域。

超强AI算力

- 47U 高度提供最高20.4 PFLOPS FP16超强AI算力

更优AI能效

- 最大可提供20.4 PFLOPS/46 kW超高能效比

极佳AI拓展

- 支持机柜单元扩展，最大可拓展至4096颗昇腾910处理器集群，总算力达1.3 EFLOPS FP16

应用场景

模型训练　HPC

智慧城市　智慧医疗

天文探索　石油勘探

产品规格

形态	47U 机柜
CPU	32 * 鲲鹏920
CPU内存	• 最多256个DDR4内存插槽，支持RDIMM • 单根内存条容量支持32 GB/64 GB
HBM	2048 GB
AI算力规格	14.0 ~ 20.4 PFLOPS @FP16
AI算力扩展	最大可扩展至1.3 EFLOPS FP16
本地存储	最大支持 80 * 2.5英寸硬盘
RAID支持	支持 RAID 0/1
供电	• 交流：6路3+3，电源：380V/32A • 直流：4路2+2，电源：380V/32A
功耗	最大功耗46 kW*
散热方式	液冷
温度	• 工作温度：5℃ ~ 40℃ （符合ASHRAE Class A2/A3/A4）
结构尺寸（H*W*D）	• 2250mm×600mm×1500mm（全液冷，含液冷门及造型门） • 2250mm×600mm×1350mm（全液冷，含液冷门） • 2250mm×600mm×1200mm（半液冷，无风液换热器）

*功耗随配置&负载变化

异构计算架构-CANN

CANN（Compute Architecture for Neural Networks）是专为深度学习所设计的异构计算架构，通过各核心组件充分释放昇腾处理器澎湃算力，支持用户快速构建基于昇腾平台的AI应用和业务，主要包含AscendCL、DVPP、HCCL等组件：昇腾统一编程接口AscendCL实现软硬件解耦；华为通信集合库HCCL在分布式训练中为不同昇腾AI处理器之间提供高效的数据传输能力；DVPP实现硬件加速，提升图像预处理并行能力。

CANN

异构计算架构

使能全场景

向下支持14+操作系统；
底层支持10+端边云设备形态；
向上能够适配多种AI框架

使能极致性能

亲和昇腾的极致图编译技术；
丰富的高性能算子

使能极简开发

统一API适配全系列硬件；
四大开放性设计：Plugin适配、图融合接口、Ascend-IR、算子库

操作系统

硬件设备

全场景AI框架-昇思MindSpore

昇思MindSpore是新一代全场景AI框架，最佳匹配昇腾AI处理器算力，支持端、边、云全场景灵活部署，开创全新的AI编程范式，降低AI开发门槛，旨在实现开发友好、运行高效、部署灵活三大目标，推动人工智能软硬件应用生态繁荣发展。

全流程极简

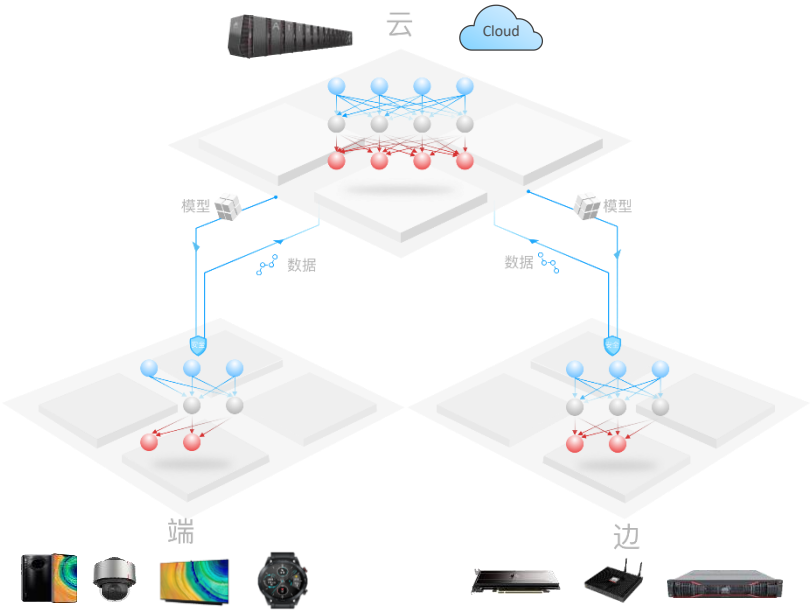
- 模型开发套件，“即开即用”
- 模型调优套件，“所见即所得”
- 第三方支持套件，“一键式转换”

全自动并行

- 一行代码，串行算法并行化
- 张量自动切分，最大化并行效率

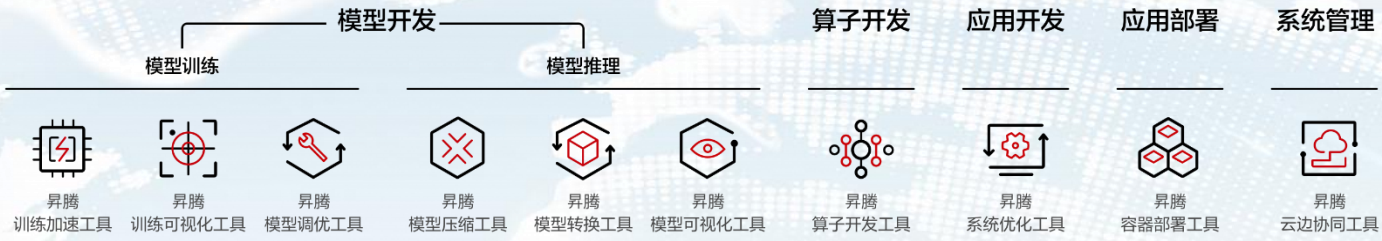
全场景协同

- 全场景自适应部署，跨异构硬件执行，无需模型转换
- 端侧轻量学习，模型“私人订制”



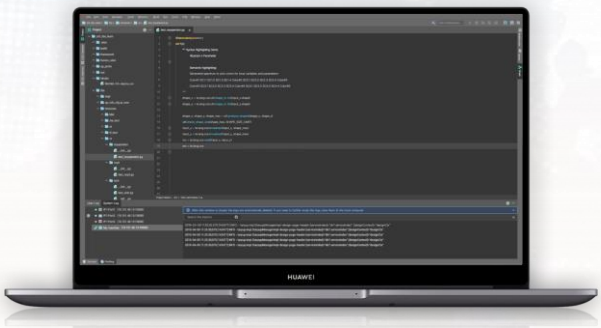
全流程开发工具链-MindStudio

MindStudio提供了AI开发所需的一站式开发环境，支持模型开发、算子开发及应用开发的全流程任务。依靠模型可视化、算力测试、IDE本地仿真调试等功能，帮助开发者高效便捷的完成AI开发。



模型开发

MindStudio包含了模型开发所涉及的推理、训练全系列工具，同时支持随时调用ModelZoo中提供的大量预训练好的AI模型、模型训练脚本以及模型开发案例，让开发者能够更高效的完成AI模型的开发。



算子开发

MindStudio兼顾算子开发的易用性与灵活性，提供了DSL和TIK两种算子开发方式，在算子开发过程中，还提供了性能调优与精度比对等功能。

TBE-DSL
更优开发效率

- 自动实现数据切分和调度，只需关注计算表达
- 覆盖70%算子，算子开发时间较业界降低70%

TBE-TIK
更佳算子性能

- 提供指令级编程和调优能力，需关注指令集调用过程和数据切分及编排
- 覆盖全部算子，可以发挥处理器的极致性能

应用开发

通过AscendCL接口，进行系统级调优、调试传输等AI应用开发，提供模型/算子加载与执行、多种C++的API接口等功能。

应用部署

通过连接IP地址统一管理调试设备，实现远程管理、调试及应用推送，无缝兼容不同形态的设备。

系统管理

由FusionDirector及SmartKit组成的昇腾云边协同工具，可以让开发者对系统进行实时地设备管理、模型部署等操作。

昇腾应用使能-MindX

昇腾应用使能MindX为行业应用开发者而设计，快速使能开发者进行各行业AI应用开发。MindX包含“2+1+X”，深度学习使能 MindX DL 、智能边缘使能 MindX Edge、1个优选模型库 ModelZoo和X个行业SDK

深度学习使能 - MindX DL

数据中心计算资源统一管理与调度，使能合作伙伴快速开发深度学习系统

4Paradigm　AAPULIS 依瞳科技　中科弘云　第三方深度学习系统　|　ModelArts　第三方云平台

计算资源更优调度支持

NPU设备发现、集合通信优化、大批量数据群组调度

边云协同参考设计

支持中心训练模型发布、更新、推送至边缘进行推理，形成模型的完整闭环

智能边缘使能 - MindX Edge

轻量化的边缘计算资源管理运维，使能行业客户快速搭建边云协同推理平台

边云协同设计

云端模型推送至边缘快速部署
边缘数据支持上传云端持续训练

多样化硬件形态支持

摄像头、工控机、机器人、无人机、边缘推理服务器...

轻量化部署

极致轻量化，平台内存开销仅256MB， CPU占用率仅3%

优选模型库 - ModelZoo

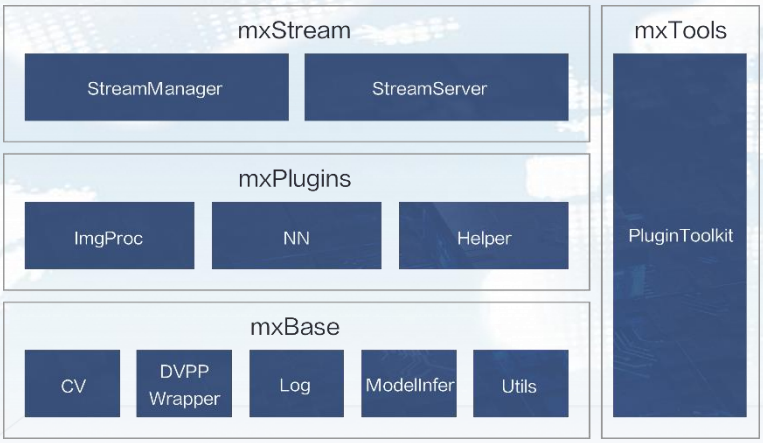
为开发者提供丰富的场景化优选预训练模型，为开发者解决了模型的选型难、训练难、优化难等问题

易获取	ascend.huawei.com > ModelZoo	**多种框架**	MindSpore、TensorFlow、PyTorch、Caffe等
多场景	OCR、图像检测、图像分类、图像分割、推荐类、NLP、机器翻译、语音生成、增强学习等	**高性能**	模型提前调优并保障精度性能

昇腾应用使能-MindX

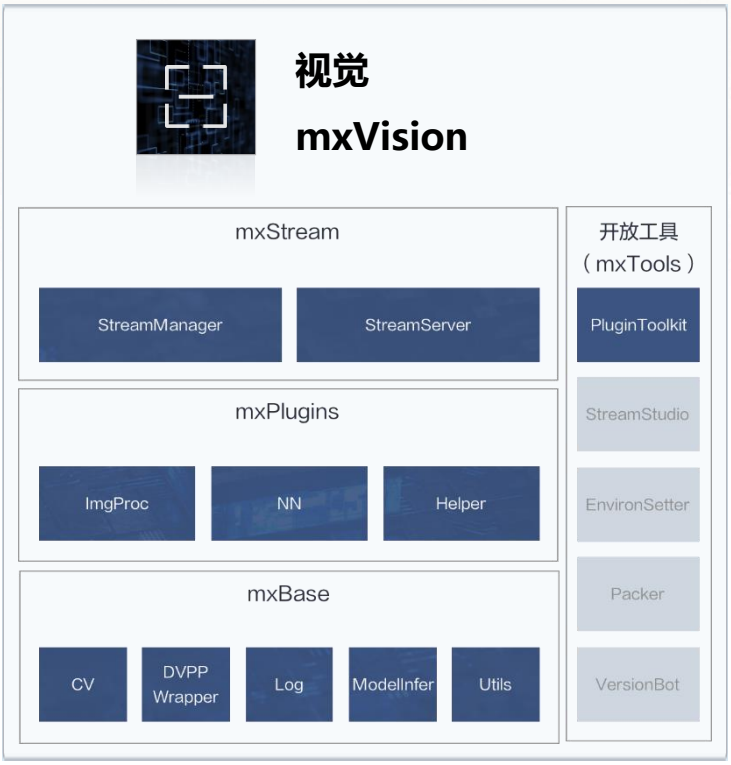
行业应用开发套件 - MindX SDK

面向行业场景的完整AI开发套件，提供极简易用的API及图形界面，使能开发者以极少代码快速开发行业AI应用



mxStream	管理业务流和生成推理服务。StreamManager负责业务流的构建和销毁，同时提供发送数据和获取结果的接口；StreamServer将Stream构建成推理服务，对外提供restful接口接收推理请求。
mxPlugins	基础功能插件，例如模型推理插件、图像解码插件
mxBase	昇腾处理器功能基础库：包括图像编解码、裁剪缩放，模型推理
mxTools	SDK提供的开发工具。其中PluginToolkit提供插件开发的API

已上线 **昇腾社区**



请访问昇腾社区获取更多支持

ascend.huawei.com

昇腾计算产业生态

昇腾计算产业生态包括围绕着昇腾计算技术和产品体系所开展的学术、技术、公益及商业活动，产生的知识和产品以及各种合作伙伴，主要包括硬件合作伙伴、软件算法合作伙伴、初创公司、高校和业界开发者。以上共同构成了昇腾产业的合作伙伴生态体系，不同的角色相互配合，共同促进AI赋能千行百业。

昇腾计算产业

硬件开放、软件开源、使能伙伴、发展人才

1 个创新发展的AI计算产业
昇腾使能千行百业的智能化转型

医疗 | 运营商 | 能源 | 交通 | 互联网 | AI计算中心 | 制造 | 金融 | 机器人 | 平安城市

2 大商业扶植计划
助力昇腾合作伙伴商业成功

- 昇腾万里 ISV合作伙伴 发展计划 → 实现联合方案的商业成功
- 昇腾万里 初创伙伴 加速计划 → 加速初创企业的创新和成长

3 大人才培养措施
为昇腾产业长期发展培育核心人才

- 昇腾 高校教学 合作计划
- MindSpore 论文+模型开发 激励计划
- 昇腾 开发人员 成长计划

2 大昇腾生态发展的基础平台

昇腾生态创新中心 | 昇腾开发者社区

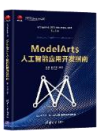
与业界TOP ISV联合创新

制造	交通	能源	金融	互联网	医疗
工业质检（烟草、半导体、PCB、线管材、镜筒）	自由流收费 高速云联网 车辆稽核	输电线路 智能运检 智能变电站 智能营业厅 智能加油站	智慧网点 金融OCR	精准推荐 内容审核	肺炎诊断 骨龄检测

昇腾系列教材
让昇腾成为
高校学生必备技能


昇腾AI处理器架构与编程


深度学习与MindSpore实践


ModelArts人工智能应用开发指南

高校合作
“智能基座”产教融合
协同育人基地

构建万物互联的智能世界

HUAWEI

把数字世界带入每个人、每个家庭、
每个组织，构建万物互联的智能世界。

修订时间：2022年9月

华为技术有限公司
深圳龙岗区坂田华为基地
电话：+86 755 28780808
邮编：518129
www.huawei.com