昇腾全栈AI软硬件平台

HUAWEI

昇腾全栈AI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

应用使能

ModelArts | HiAI Service | 第三方平台

MindX

行业SDK：mxVision / mxManufacture　　优选模型库：ModelZoo

智能边缘使能：MindX Edge　　深度学习使能：MindX DL

框架：MindSpore / TensorFlow / PyTorch / …

芯片使能：CANN


Atlas 200
AI加速模块


Atlas 300I
推理卡


Atlas 300T 训练卡


Atlas 900 PoD


Atlas 500
智能小站


Atlas 500 Pro
智能边缘服务器


Atlas 800
推理服务器


Atlas 800
训练服务器


Atlas 900
AI集群

全流程开发工具链　MindStudio

管理运维工具　FusionDirector / SmartKit

昇腾社区　ascend.huawei.com

Atlas 200 AI加速模块

型号：3000

Atlas 200 AI加速模块（型号：3000）集成了昇腾310 AI处理器，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5.5 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芯片	昇腾310
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，16路1080P 30 FPS（2路3840*2160 60 FPS） • 支持H.265硬件解码，16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	• 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	4 GB: 5.5 W / 8 GB: 8 W
工作环境温度	-25℃ ~ 80℃
重量	30 g
结构尺寸	52.6 mm * 38.5 mm * 8.5 mm

Atlas 200 DK 开发者套件

型号：3000

Atlas 200 DK 开发者套件（型号：3000）是一款高性能AI应用开发板，集成了昇腾310 AI处理器，方便用户快速开发、快速验证，可广泛应用于开发者方案验证、高校教育、科学研究等场景。

高集成度

- 基于华为昇腾310 AI处理器，集成丰富的外设接口和MindStudio开发环境，方便开发者简捷接入、快速开发

易用的软件环境

- MindStudio编程界面友好，具备图形化调试能力，可全自动管理离线模型并提供仿真环境

应用场景

开发者方案验证

模型验证
方案验证

高校教育

AI入门教育
人才培养

科学研究

应用研究
算法研究

产品规格

AI芯片	昇腾310
AI算力	22/16/8 TOPS INT8 11/8/4 TFLOPS FP16
内存规格	LPDDR4X，8 GB，总带宽51.2 GB/s
编解码能力	• 支持H.264 硬件解码，16路1080P 30 FPS（2路3840*2160 60 FPS） • 支持H.265 硬件解码，16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：1个GE RJ45 • USB： 1个USB2.0 / USB3.0 • Camera：2个51 pin 相机连接器 • 其他：1个40 pin IO连接器
电源	5~28 V DC，默认配置12 V / 3 A适配器
功耗	典型功耗20 W
工作环境温度	0℃ ~ 35℃
结构尺寸	137.8 mm * 93.0 mm * 32.9 mm

Atlas 300I 推理卡

型号：3000/3010

Atlas 300I 推理卡（型号：3000/3010）基于昇腾310 AI处理器，提供超强AI推理性能，单卡算力可达88 TOPS INT8，支持80路高清视频实时分析，可广泛应用于智慧城市、智慧交通、智慧金融等场景。

超强算力

- 单卡提供88 TOPS INT8算力，支持80路高清视频实时分析（1080P 25FPS），为边缘推理提供更强大支持

硬件编解码

- 支持JPEG和视频硬件编解码，提升图片和视频类应用性能

低时延

- 内存容量大、带宽高，满足特征匹配场景内存需求，降低应用时延

应用场景

集成于服务器、工控机中，进行AI推理

智慧城市　智慧交通　智慧社区　智慧营业厅

智能制造　无人零售　智能楼宇　智慧金融

产品规格

形态	半高半长PCIe卡
AI芯片	昇腾310
AI算力	88 TOPS INT8 44 TFLOPS FP16
内存规格	LPDDR4X 32 GB，总带宽204.8 GB/s
编解码能力	• 支持H.264硬件解码，64路1080P 30 FPS（8路 3840*2160 60 FPS） • 支持H.265硬件解码，64路1080P 30 FPS（8路 3840*2160 60 FPS） • 支持H.264硬件编码，4路1080P 30 FPS • 支持H.265硬件编码，4路1080P 30 FPS • JPEG解码能力4*1080P 256 FPS，编码能力4*1080P 64 FPS，最大分辨率：8192*4320 • PNG解码能力4*1080P 48 FPS，最大分辨率：4096*2160
PCIe	PCIe x16 Gen3.0（型号：3010） PCIe x8 Gen3.0（型号：3000）
功耗	最大67 W
工作环境温度	0℃ ~ 55℃
结构尺寸	169.5 mm * 68.9 mm

Atlas 300T 训练卡

型号：9000

华为Atlas 300T 训练卡（型号：9000）是基于昇腾910 AI芯片，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具有最强算力、最高集成、最快带宽等特点，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。

最强算力

- 内置32个达芬奇AI Core
- 提供业界领先的280 TFLOPS FP16算力

最高集成

- AI算力、通用算力、I/O能力三合一
- 芯片集成32个华为达芬奇AI Core + 16个TaiShan核 + 1 * 100GE RoCE v2网卡

最快带宽

- 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%

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

产品规格

形态	全高，3/4长，双槽位
AI处理器	昇腾 910
AI算力	280 TFLOPS FP16 (Pro) 256 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℃

① 持续调优中，数值根据调优结果动态更新

Atlas 500 智能小站

型号：3000

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

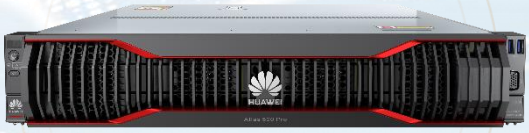
智能制造　智慧营业厅　无人零售　智能楼宇

产品规格

AI芯片	昇腾310
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，16路1080P 30 FPS（2路3840*2160 60 FPS） 支持H.265硬件解码，16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：-40ºC ~ 70ºC 有盘配置：-40ºC ~ 60ºC
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

型号：3000



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4张Atlas 300I 推理卡，满足多场景推理需求；整机可提供320路高清视频实时分析（1080P 25 FPS）
- 搭载64核架构、具有超强算力的鲲鹏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 单卡功耗仅67W，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

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

形态	2U AI服务器
CPU	1 * 鲲鹏920
CPU内存	4个DDR4内存插槽，最高2933 MT/s
AI加速卡	最大支持4个Atlas 300I AI推理卡
AI算力	最大352 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）是基于昇腾310芯片的推理服务器，最大可支持8个Atlas 300I 推理卡，提供强大的实时推理能力，广泛应用于中心侧AI推理场景。

超强算力

- 支持8张Atlas 300I 推理卡，满足多场景推理需求；整机可提供640路高清视频实时分析（1080P 25FPS）
- 搭载64核架构、具有超强算力的鲲鹏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 单卡功耗仅67W，为AI服务器算力加速同时提供更优的能效比

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

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2933 MT/s
AI加速卡	最大支持8个Atlas 300I 推理卡
AI算力	最大704 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
结构尺寸	447 mm * 790 mm * 86.1 mm

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 推理加速卡，支持560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 推理卡，支持560路高清视频实时分析（1080P 25FPS）

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

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2933 MT/s
AI加速卡	最大支持7个Atlas 300I 推理卡
AI算力	最大616 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℃
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

Atlas 800 训练服务器
型号：9000

Atlas 800 训练服务器（型号：9000）是基于华为鲲鹏920+昇腾910处理器的AI训练服务器，具有最强算力密度、超高能效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。

最强算力密度

- 4U高度提供 2.56 PFLOPS FP16超强算力
- 算力密度达到业界1.7倍

超高能效

- 单机支持风冷和液冷两种散热方式
- 提供2.56 PFLOPS/5.6 kW①超高能效比，达到业界1.3倍

高速网络带宽

- 8*100G RoCE v2高速接口
- 芯片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	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2933 MT/s • 单根内存条容量支持32 GB/64 GB
AI处理器	8 * 昇腾910
HBM	8 * 32 GB
AI算力	2.56 / 2.24 / 2 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/液冷
风扇	支持8个热拔插风扇模组，支持N+1冗余
工作温度	5℃ ~ 40℃（液冷） 5℃ ~ 35℃（风冷）
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 800 训练服务器

型号：9010

Atlas 800 训练服务器（型号：9010）是基于Intel处理器+华为昇腾910芯片的AI训练服务器，具有最强算力密度、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。

最强算力密度

- 4U高度提供最高2.24 PFLOPS FP16超强算力
- 算力密度达到业界1.5倍

高速网络带宽

- 8*100G RoCE v2高速接口，芯片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
AI处理器	8 * 昇腾910
HBM	8 * 32 GB
AI算力	2.24 PFLOPS FP16 2 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℃
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
型号：9000

Atlas 900 PoD（型号：9000）是基于华为昇腾910 + 鲲鹏920 处理器的AI训练集群基础单元，具有超强AI算力、更优AI能效、最佳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20.48 PFLOPS FP16超强AI算力

更优AI能效

- 提供20.48 PFLOPS/43 kW超高能效比

最佳AI拓展

- 支持机柜单元扩展，最大可拓展至4096颗昇腾910芯片集群，总算力达1 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
AI处理器	64 * 昇腾910
HBM	2048 GB
AI算力	最高20.48 PFLOPS FP16
AI算力扩展	最大可扩展至1 EFLOPS FP16
本地存储	最大支持 64 * 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43 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×1200mm（半液冷，无风液换热器） • 2250mm×600mm×1250mm（半液冷，安装液冷前后门） • 2250mm×600mm×1350mm（全液冷，无风液换热器） • 2250mm×600mm×1375mm（全液冷，安装液冷前门）

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。

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

全场景AI计算框架-MindSpore

MindSpore是新一代全场景AI计算框架，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，旨在实现开发态友好、运行态高效、部署态灵活三大目标，推动人工智能软硬件应用生态繁荣发展。

全流程极简

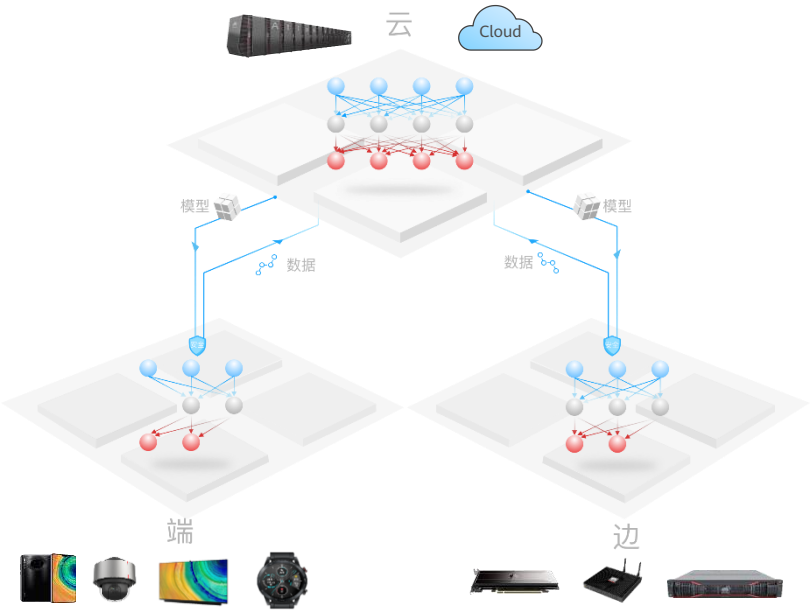
- 模型开发套件，“即开即用”
- 模型调优套件，“所见即所得”
- 第三方支持套件，“一键式转换”

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

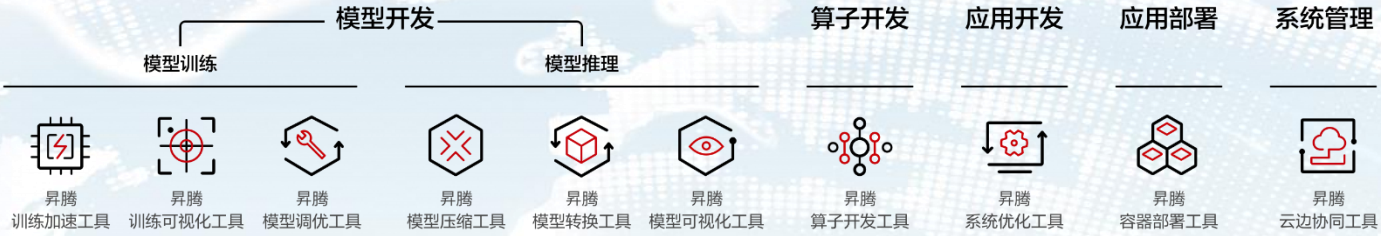
全场景协同

- 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”



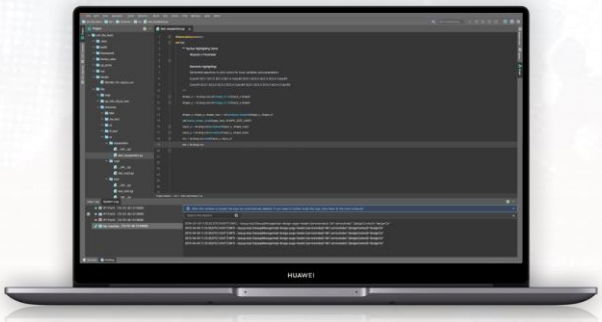
全流程开发工具链-MindStudio

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，让开发者能够最高效的完成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
最优开发效率

- 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
最佳算子性能

- 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，可以发挥芯片的极致性能

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

计算资源最优化调度支持

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

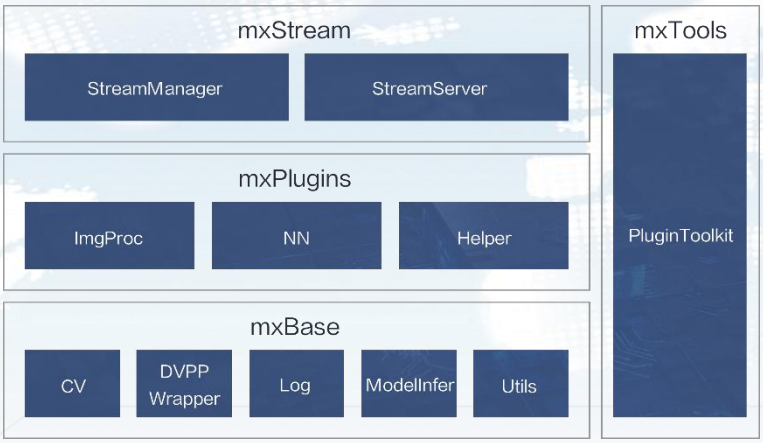
多场景 OCR、图像检测、图像分类、图像分割、推荐类、NLP、机器翻译、语音生成、增强学习等

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

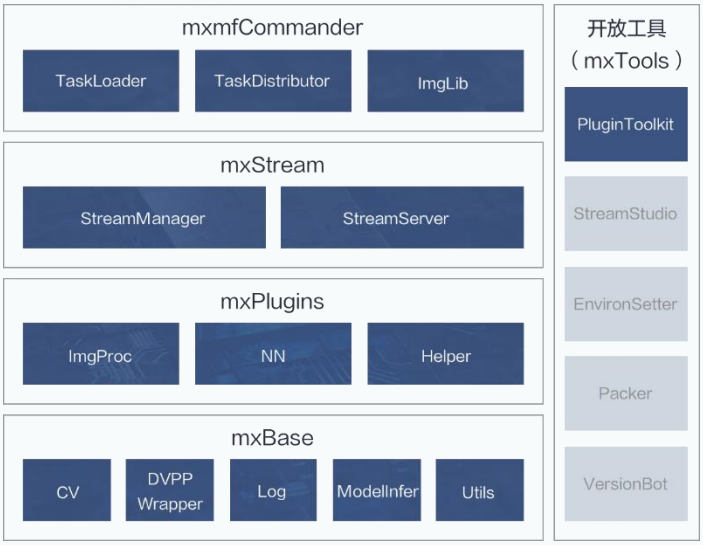
面向行业场景的完整AI开发套件，提供极简易用的API及图形界面，使能开发者以极少代码快速开发行业AI应用



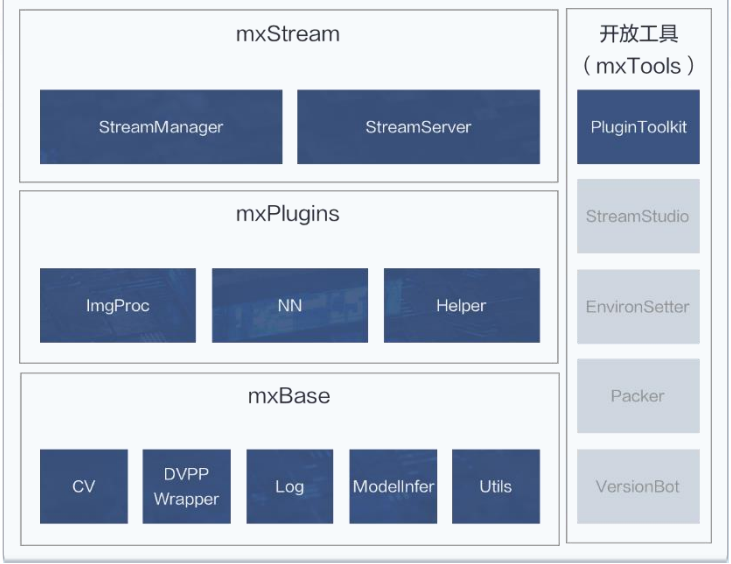
mxStream	管理业务流和生成推理服务。StreamManager负责业务流的构建和销毁，同时提供发送数据和获取结果的接口；StreamServer将Stream构建成推理服务，对外提供restful接口接收推理请求。
mxPlugins	基础功能插件，例如模型推理插件、图像解码插件
mxBase	昇腾芯片功能基础库：包括图像编解码、裁剪缩放，模型推理
mxTools	SDK提供的开发工具。其中PluginToolkit提供插件开发的API

已上线 **昇腾社区**

制造 mxManufacture



视觉 mxVision



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，使能合作伙伴

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 高速视频云联网 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

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：2020 年 12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