Atlas AI Computing Solution

HUAWEI

Atlas 200 AI Accelerator Module

Model: 3000

The Atlas 200 AI accelerator module (model: 3000) integrates the Ascend 310 AI processor to implement video analysis and image classification on the device side. Atlas 200 is widely used in AI scenarios such as intelligent cameras, robots, and drones.

Ultimate performance

- 22 TOPS INT8 in the size of half a credit card, supporting real-time analysis of 20-channel HD videos (1080p 25FPS)
- Multi-level computing power configuration: 22/16/8 TOPS

Ultra-low consumption

- Hibernation at milliwatts and wakeup in milliseconds, typical power consumption of 7.5 W, enabling edge AI applications

Application Scenarios

Embedded in edge intelligence

- Cameras
- Robots
- Drones
- Industrial computers
- Image analytics
- Video analytics
- Image segmentation
- Object recognition

Specifications

AI Processor	Ascend 310
AI Computing Power	22/16/8 TOPS INT8 11/8/4 TFLOPS FP16
Memory	LPDDR4X, 8 GB/4 GB, total bandwidth 51.2 GB/s
Encoding/ Decoding	• H.264 hardware decoding, 16-channel 1080p 30 FPS (2-channel 3840 x 2160 @ 60 FPS) • H.265 hardware decoding, 16-channel 1080p 30 FPS (2-channel 3840 x 2160 @ 60 FPS) • H.264 hardware encoding, 1-channel 1080p 30 FPS • H.265 hardware encoding, 1-channel 1080p 30 FPS • JPEG decoding: 1080p 256 FPS; encoding: 1080p 64 FPS; maximum resolution: 8192 x 4320 • PNG decoding: 1080p 24 FPS; maximum resolution: 4096 x 2160
Port	• PCIe x4 Gen3.0 • 1 USB 2.0/USB 3.0 • 1 RGMII
Serial Bus	UART/I2C/SPI
Interface Specifications	144-pin BTB connector
Typical Power Consumption	4 GB: 7.5 W 8 GB: 11 W
Operating Temperature	-25°C to +80°C
Weight	30 g
Dimensions (H x W x D)	8.5 mm x 52.6 mm x 38.5 mm

Building a Fully Connected, Intelligent World

HUAWEI

Atlas 200 DK AI Developer Kit

Model: 3000

The Atlas 200 DK (model: 3000) is a high-performance AI application developer board that integrates the Ascend 310 AI processor to facilitate quick development and verification. It has been widely used in scenarios such as developer solution verification, higher education, and scientific research.

High integration

- Powered by the Huawei Ascend 310 AI processor, and integrates various peripheral interfaces and the Mind Studio, facilitating access to the development environment and enabling quick development

Easy-to-use software environment

- Mind Studio provides a user-friendly programming interface and GUI-based debugging, allowing automatic management of offline models with a simulation environment

Application Scenarios

Developer solution verification

Model verification
Solution verification

Higher education

Entry-level AI education
Talent cultivation

Scientific research

Application research
Algorithm research

Specifications

AI Processor	Ascend 310
AI Computing Power	22/16/8TOPS INT8 11/8/4 TFLOPS FP16
Memory	LPDDR4X, 8 GB/4 GB, total bandwidth 51.2 GB/s
Encoding/ Decoding	• H.264 hardware decoding, 16-channel 1080p 30 FPS (2-channel 3840 x 2160 @ 60 FPS) • H.265 hardware decoding, 16-channel 1080p 30 FPS (2-channel 3840 x 2160 @ 60 FPS) • H.264 hardware encoding, 1-channel 1080p 30 FPS • H.265 hardware encoding, 1-channel 1080p 30 FPS • JPEG decoding: 1080p 256 FPS; encoding: 1080p 64 FPS; maximum resolution: 8192 x 4320 • PNG decoding: 1080p 24 FPS; maximum resolution: 4096 x 2160
Port	• Network: 1 GE RJ45 port • USB: 1 USB 2.0/USB 3.0 port • Camera: 2 51-pin connector • Others: 1 40-pin I/O connector
Power Supply	12 V DC
Power Consumption	Typical: 20 W
Operating Temperature	0°C to 35°C
Dimensions (H x W x D)	32.9 mm x 137.8 mm x 93.0 mm

Atlas 300I Inference Card

Model: 3000/3010

Powered by the Ascend 310 AI processor, the Atlas 300I inference card (model: 3000/3010) unlocks superior AI inference performance. A single card provides up to 88 TOPS INT8 computing power and supports 80-channel real-time HD video analytics, making it an ideal option for intelligent scenarios such as smart city, transportation, and finance.

Superior computing

- A single card provides 88 TOPS INT8 computing power and supports 80-channel HD video real-time analytics (1080p 25 FPS), providing powerful support for edge inference

Hardware encoding/decoding

- Supports JPEG and video hardware codecs, improving image and video application performance

Low latency

- Supports large-capacity and high-bandwidth memory for feature matching scenarios, reducing application latency

Application Scenarios

Integrated in servers and industrial computers for AI inference

Smart city | Smart transportation | Smart community | Smart customer service center

Smart manufacturing | Unmanned retail | Smart building | Smart finance

Specifications

Form Factor	Half-height half-length PCIe standard card
AI Processor	Ascend 310
AI Computing Power	88 TOPS INT8 44 TFLOPS FP16
Memory	LPDDR4X, 32 GB, total bandwidth 204.8 GB/s
Encoding/ Decoding	• H.264 hardware decoding, 64-channel 1080p 30 FPS (8-channel 3840 x 2160 @ 60 FPS) • H.265 hardware decoding, 64-channel 1080p 30 FPS (8-channel 3840 x 2160 @ 60 FPS) • H.264 hardware encoding, 4-channel 1080p 30 FPS • H.265 hardware encoding, 4-channel 1080p 30 FPS • JPEG decoding: 4-channel 1080p 256 FPS; encoding: 4-channel 1080p 64 FPS; maximum resolution: 8192 x 4320 • PNG decoding: 4-channel 1080p 48 FPS; maximum resolution: 4096 x 2160
PCIe	PCIe x8 Gen3.0 (Model: 3000) PCIe x16 Gen3.0 (Model: 3010)
Power Consumption	Maximum: 67 W
Operating Temperature	0°C to 55°C
Dimensions (W x D)	169.5 mm x 68.9 mm

Atlas 300T Training Card

Model: 9000

The Huawei Atlas 300T training card (model: 9000) is based on the Ascend 910 AI processor and works with servers to provide powerful computing for data centers. A single card provides 280 TFLOPS FP16 computing power, accelerating deep learning and training. Atlas 300T features the highest computing power, integration, and bandwidth, meeting the AI training and high-performance computing requirements of the Internet, carriers, and finance.

Ultimate computing power

- 32 built-in Da Vinci AI Cores
- Industry-leading 280 TFLOPS FP16 computing power

Highest integration

- AI computing, general computing, and I/O 3-in-1
- Integrates 32 Huawei Da Vinci AI Cores, 16 TaiShan Cores, and 1 100GE RoCE v2 NICs

Highest bandwidth

- Supports PCIe 4.0 and 1 100 Gbit/s RoCE high-speed ports, with a total egress bandwidth of 56.5 Gbit/s
- Boosts the efficiency of data training and gradient synchronization by 10–70% without the need for external NICs

Application Scenarios

- **Model training**
- **HPC**
- **Smart city**
- **Smart transportation**
- **Smart manufacturing**
- **Smart finance**

Specifications

Form Factor	Full height 3/4 length, dual-slot
AI Processor	Ascend 910
AI Computing Power	280 TFLOPS FP16 (Pro) 256 TFLOPS FP16
Encoding/ Decoding	• 16-channel 4K (or 64-channel 1080p) 60 FPS H.264/H.265 • JPEG decoding: 1080p 2048 FPS, or equivalent decoding capability; maximum resolution: 8192 x 4320 • PNG decoding: 1080p 240 FPS, or equivalent decoding capability; maximum resolution: 4096 x 2160 • JPEG encoding: 1080p 256 FPS, or equivalent encoding capability; maximum resolution: 8192 x 4320
Memory	• 32 GB HBM • 16 GB DDR4
Network	1 100GE QSFP-DD ports
PCIe	PCIe x16 Gen4.0
Power Consumption	Maximum: 300 W[1]
Cooling Mode	Passive air cooling
Operating Temperature	5°C to 45°C

1. This specification item is in continuous optimization. The value is dynamically updated based on the optimization result.

Building a Fully Connected, Intelligent World

HUAWEI

Atlas 500 AI Edge Station

Model: 3000

The Atlas 500 AI edge station (model: 3000) is designed for edge applications. It features superb computing performance in a compact size, strong environmental adaptability, easy maintenance, and cloud-edge collaboration, and can be widely deployed at the edge. The Atlas 500 AI edge station meets complex requirements in scenarios such as security, transportation, community, campus, shopping malls, and supermarkets.

Intelligent edge

- State-of-the-art edge product with AI processing capabilities
- Fan-free heat dissipation, stable outdoor at –40°C to +70°C

Superb capacity in a compact size

- 22 TOPS INT8 computing power in the size of an STB
- 20-channel HD video processing (1080p 25 FPS)

Edge-cloud collaboration

- LTE wireless transmission
- Cloud-edge collaboration for real-time model update
- Unified device management and firmware update on the cloud

Application Scenarios

Independently deployed for edge intelligence

Smart city

Smart transportation

Smart community

Environment monitoring

Smart manufacturing

Smart customer service center

Unmanned retail

Smart building

Specifications

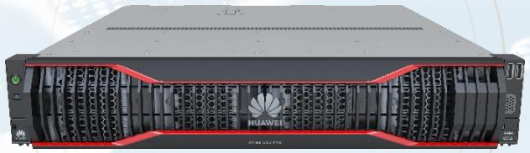
AI Processor	Ascend 310
AI Computing Power	22/16 TOPS INT8 11/8 TFLOPS FP16
Memory	LPDDR4X, 8 GB/4 GB, up to 51.2 GB/s
Encoding/ Decoding	H.264 hardware decoding, 16-channel 1080p 30 FPS (2-channel 3840 x 2160 @ 60 FPS) H.265 hardware decoding, 16-channel 1080p 30 FPS (2-channel 3840 x 2160 @ 60 FPS) H.264 hardware encoding, 1-channel 1080p 30 FPS H.265 hardware encoding, 1-channel 1080p 30 FPS JPEG decoding: 1080p 256 FPS; encoding: 1080p 64 FPS; maximum resolution: 8192 x 4320 PNG decoding: 1080p 24 FPS; maximum resolution: 4096 x 2160
Port	Network: 2 GE RJ45 ports Other I/O ports: 1 HDMI port 1 input and 1 output (stereo), 3.5 mm audio connector 2 external USB 2.0 ports and 1 internal USB 2.0 port (Type-A)
Typical Power Consumption	Without disks: 25 W With disks: 40 W
Environment Conditions	Without disks: –40°C to +70°C With disks: –40°C to +60°C
Dimensions (H x W x D)	Without disks: 220 mm x 45 mm x 235 mm With disks: 220 mm x 45 mm x 355

Building a Fully Connected, Intelligent World

HUAWEI

Atlas 500 Pro AI Edge Server

Model: 3000



The Atlas 500 Pro AI edge server (model: 3000) is designed for edge applications. It features superb computing performance, strong environmental adaptability, easy maintenance, and cloud-edge collaboration. It can be widely deployed at the edge to meet application requirements in complex scenarios and environments such as security, transportation, communities, campuses, shopping malls, and supermarkets.

Superior computing

- Supports up to 4 Atlas 300I inference cards to meet the inference requirements in multiple scenarios; 320-channel real-time HD video analytics (1080p 25 FPS)
- Runs on the 64-core Kunpeng 920 processors to unlock powerful computing for application acceleration

Superior perf./watt

- Provides an AI computing platform with high efficiency, low power for inference scenarios, fully leveraging the multi-core, low-consumption advantages of Kunpeng
- Atlas 300I runs at only 67 W, fueling the AI server with faster computing and higher performance per watt

Application Scenarios

Independently deployed for edge intelligence

- Smart city
- Smart transportation
- Smart community
- Environment monitoring
- Smart manufacturing
- Smart customer service center
- Unmanned retail
- Smart building

Specifications

Form Factor	2U AI server
Processor	1 Kunpeng 920 processor
Processor Memory	4 DDR4 DIMM slots, up to 2933 MT/s
AI Accelerator Card	Up to 4 Atlas 300I AI inference cards
AI Computing Power	Up to 352 TOPS INT8
Local Storage	8–12 x 3.5" SAS/SATA drives
RAID	RAID 1, 5, 6, or 10
PCIe	Up to 4 PCIe 4.0 x8 standard slots
LOM	4 10GE/25GE (optical ports) + 2 GE (electrical ports)
Power Supply	• 2 hot-swappable 550 W or 900 W AC PSUs, supporting 220 V AC or 240 V DC; or 2 hot-swappable 1200 W DC PSUs, supporting –48 V DC • 1+1 redundancy
Fan Modules	4 hot-swappable fan modules, supporting N+1 redundancy
Operating Temperature	• Long term: 5°C to 50°C • Short-term: 0°C to 55°C
Dimensions (H x W x D)	475 mm x 86.1 mm x 447 mm

Building a Fully Connected, Intelligent World

HUAWEI

Atlas 800 Inference Server

Model: 3000

Powered by the Ascend 310 processor, the Atlas 800 inference server (model: 3000) supports up to 8 Atlas 300I inference cards to provide powerful real-time inference. It is widely used for AI inference in data centers.

Superior computing

- Supports 8 Atlas 300I inference cards to meet the inference requirements in multiple scenarios; 640-channel real-time HD video analytics (1080p 25 FPS)
- Runs on the 64-core Kunpeng 920 processors to unlock powerful computing for application acceleration

Superior perf./watt

- Provides an AI computing platform with high efficiency, low power for inference scenarios, fully leveraging the multi-core, low-consumption advantages of Kunpeng
- Atlas 300I runs at only 67 W, fueling the AI server with faster computing and higher performance per watt

Application Scenarios

Deployed in data centers to enable AI inference

Precision marketing | **Medical image analytics** | **Video analytics** | **OCR**

Smart retail | **Smart healthcare** | **Smart city** | **Smart finance**

Specifications

Form Factor	2U AI server
Processor	2 Kunpeng 920 processors
Processor Memory	32 DDR4 DIMM slots, up to 2933 MT/s
AI Accelerator Card	Up to 8 Atlas 300I inference cards
AI Computing Power	Up to 704 TOPS INT8
Local Storage	25 x 2.5" SAS/SATA drives 12 x 3.5" SAS/SATA drives 8 x 2.5" SAS/SATA + 12 x 2.5" NVMe
RAID	RAID 0, 1, 10, 5, 50, 6, or 60
PCIe	Up to 9 PCIe 4.0 PCIe ports, among which one is a PCIe slot dedicated for screw-in RAID controller card, and the other 8 are for plug-in PCIe RAID controller cards
Power Supply	2 hot-swappable 900 W or 2000 W AC PSUs, supporting 1+1 redundancy
Fan Modules	4 hot-swappable fan modules, supporting N+1 redundancy
Operating Temperature	5°C to 40°C
Dimensions (H x W x D)	86.1 mm x 447 mm x 790 mm

Atlas 800 Inference Server

Model: 3010

Powered by the Intel processors, the Atlas 800 inference server (model: 3010) supports up to 7 Atlas 300I inference cards for 560-channel real-time HD video analytics. It is widely used for AI inference in data centers.

Flexible configuration for various workloads

- Supports any combination of SAS/SATA/NVMe/M.2 SSD drives
- Supports LAN on motherboard (LOM) and FlexIO cards, providing rich network interface options

Smart video analysis

- Supports up to 7 Atlas 300I inference cards and 560-channel real-time HD video analytics (1080p 25 FPS)

Specifications

Form Factor	2U AI server
Processor	1 or 2 Intel® Xeon® Skylake or Cascade Lake Scalable processors, 205 W TDP
Processor Memory	24 DDR4 DIMM slots, up to 2933 MT/s
AI Accelerator Card	Up to 7 Atlas 300I inference cards
AI Computing Power	Up to 616 TOPS INT8
Local Storage	8 x 2.5" SAS/SATA drives 12 x 3.5" SAS/SATA drives 8 x 2.5" SAS/SATA + 12 x 2.5" NVMe 24 x 2.5" SAS/SATA drives 24 x 2.5" NVMe 25 x 2.5" SAS/SATA drives
RAID	RAID 0, 1, 5, 6, 10, 1E, 50, or 60
PCIe	10 PCIe Gen3.0 (including 1 RAID controller card and 1 FlexIO)
Power Supply	2 hot-swappable PSUs, with support for 1+1 redundancy. Supported options include: 550 W AC Platinum PSUs, 900 W AC Platinum/Titanium PSUs, and 1500 W AC Platinum PSUs 1500 W 380 V HVDC PSUs, 1200 W -48 V to -60 V DC PSUs
Fan Modules	4 hot-swappable fan modules, supporting N+1 redundancy
Operating Temperature	5°C to 45°C
Dimensions (H x W x D)	Chassis with 3.5" drives: 748 mm x 86.1 mm x 447 mm Chassis with 2.5" drives: 708 mm x 86.1 mm x 447 mm

Application Scenarios

Deployed in data centers to enable AI inference

Precision marketing | **Medical image analytics** | **Video analytics** | **OCR**

Smart retail | **Smart healthcare** | **Smart city** | **Smart finance**

Building a Fully Connected, Intelligent World

HUAWEI

Atlas 800 Training Server

Model: 9000

The Atlas 800 training server (model: 9000) is powered by the Kunpeng 920 and Ascend 910 processors. It features the industry's highest computing density, ultra-high energy efficiency, and high network bandwidth. The server is widely used in deep learning model development and training scenarios, and is an ideal option for computing-intensive industries, such as smart city, intelligent healthcare, astronomical exploration, and oil exploration.

The ultimate computing density

- 2.56 PFLOPS FP16 in a 4U space
- 1.7x the computing density of industry peers

Superior perf./watt

- Supports air cooling and liquid cooling
- 2.56 PFLOPS/5.6 kW[1] ultra-high energy efficiency, 1.3x that of its counterparts

High-speed network

- 8 100G RoCE v2 high-speed ports
- Slashes cross-server chip interconnect latency by 10–70%

Application Scenarios

Deployed in data centers to enable AI training

- **Model training**
- **HPC**
- **Smart city**
- **Smart healthcare**
- **Astronomical exploration**
- **Oil exploration**

Specifications

Form Factor	4U AI server
Processor	4 Kunpeng 920 processors
Processor Memory	• Up to 32 DDR4 DIMM slots, supporting RDIMMs • Up to 2933 MT/s • 32 GB or 64 GB per DIMM
AI Processor	8 Ascend 910 processors
HBM	8 * 32 GB
AI Computing Power	2.56 / 2.24 / 2 PFLOPS FP16
Local Storage	• 2 x 2.5" SAS/SATA + 3 x 2.5" NVMe • 2 x 2.5" SATA + 3 x 2.5" NVMe • 2 x 2.5" SAS/SATA + +6 x 2.5" NVMe • 2 x 2.5" SATA + +6 x 2.5" NVMe • 2 x 2.5" SATA + 8 x 2.5" SAS/SATA
RAID	RAID 0, 1, 10, 5, 50, 6, or 60
Network	8 100GE + 4 25GE/2 100GE
PCIe Expansion	Up to 2 PCIe 4.0 slots
PSUs	4 hot-swappable 2 kW or 3 kW AC PSUs, supporting 2+2 redundancy
Power Supply	• 200–240 V AC • 240 V DC
Power Consumption	Maximum: 5.6 kW[1]
Cooling Mode	Air or liquid cooling
Fan Modules	8 hot-swappable fan modules, supporting N + 1 redundancy
Operating Temperature	5℃ to 40℃ (Liquid Cooling) 5℃ to 35℃ (Air Cooling)
Dimensions (H x W x D)	790 mm x 175 mm x 447 mm

1. This specification item is in continuous optimization. The value is dynamically updated based on the optimization result.

Atlas 800 Training Server

Model: 9010

The Atlas 800 training server (model: 9010) is an AI training server based on the Intel processors and Huawei Ascend 910 processors. It features the industry's highest computing density and high network bandwidth. The server is widely used in deep learning model development and training scenarios, and is an ideal option for computing-intensive industries, such as smart city, intelligent healthcare, astronomical exploration, and oil exploration.

The ultimate computing density

- Up to 2.24 PFLOPS FP16 in a 4U space
- 1.5x the computing density of industry peers

High-speed network

- 8 100G RoCE v2 ports, slashing cross-server chip interconnect latency by 10–70%

Application Scenarios

Deployed in data centers to enable AI training

- **Model training**
- **HPC**
- **Smart city**
- **Smart healthcare**
- **Astronomical exploration**
- **Oil exploration**

Specifications

Form Factor	4U AI server
Processor	2 Intel V5 Cascade Lake processors
Processor Memory	Up to 24 DDR4 DIMM slots, supporting RDIMMs
AI Processor	8 Ascend 910 processors
HBM	8 * 32 GB
AI Computing Power	2.24 PFLOPS FP16 2 PFLOPS FP16
Local Storage	• 2 x 2.5" SATA + 8 x 2.5" SAS/SATA • 2 x 2.5" SAS/SATA + +6 x 2.5" NVMe
RAID	RAID 0, 1, 10, 5, 50, 6, or 60
Network	8 100GE 1 OCP NIC 3.0 standard card, supporting 2 25GE
PCIe Expansion	Up to 2 PCIe 3.0 x16 and 4 PCIe 3.0 x8 slots
Power Supply	4 hot-swappable 2 kW or 3 kW AC PSUs, supporting 2+2 redundancy
Power Consumption	Maximum: 5.6 kW[1]
Cooling Mode	Air cooling
Fan Modules	8 hot-swappable fan modules, supporting N + 1 redundancy
Operating Temperature	5℃ to 40℃ (Liquid Cooling) 5℃ to 35℃ (Air Cooling)
Dimensions (H x W x D)	790 mm x 175 mm x 447 mm

1. This specification item is in continuous optimization. The value is dynamically updated based on the optimization result.

Building a Fully Connected, Intelligent World

HUAWEI

Atlas 900 PoD

Model: 9000

The Atlas 900 PoD (model: 9000) is a basic unit of the AI training cluster based on Huawei Ascend 910 and Kunpeng 920 processors. It features powerful AI computing, optimal AI energy efficiency, and optimal AI scalability. The cluster basic unit is widely used in deep learning model development and training scenarios, and is an ideal option for computing-intensive industries, such as smart city, intelligent healthcare, astronomical exploration, and oil exploration.

Powerful AI computing

- Up to 20.48 PFLOPS FP16 in a 47U space

Superior AI energy efficiency

- 20.48 PFLOPS/43 kW ultra-high energy efficiency

Optimal AI scalability

- Supports scaling by basic units to an AI cluster of up to 4096 Ascend 910 processors, delivering up to 1 EFLOPS FP16

Application Scenarios

- **Model training**
- **HPC**
- **Smart city**
- **Smart healthcare**
- **Astronomical exploration**
- **Oil exploration**

Specifications

Form Factor	47U rack
Processor	32 Kunpeng 920 processors
Processor Memory	• Up to 256 DDR4 DIMM slots, supporting RDIMMs • 32 GB or 64 GB per DIMM
AI Processor	64 Ascend 910 processors
HBM	2048 GB
AI Computing Power	Up to 20.48 PFLOPS FP16
AI Computing Scalability	Up to 1 EFLOPS FP16
Local Storage	Up to 64 x 2.5" drives
RAID	RAID 0 or RAID 1
Power Supply	• AC: 6 PSUs in 3+3 redundancy mode: 380 V, 32 A • DC: 4 PSUs in 2+2 redundancy mode: 380 V, 32 A
Power Consumption	Maximum: 43 kW
Cooling Mode	Liquid cooling
Temperature	• Operating: 5℃ to 40℃ • Comply with ASHRAE Class A2/A3/A4
Dimensions (H x W x D)	• 2250mm×600mm×1200mm, half liquid-cooled, without air-to-liquid heat exchangers • 2250mm×600mm×1250mm, half liquid-cooled, with front and rear doors for liquid cooling • 2250mm×600mm×1350mm, fully liquid-cooled, without air-to-liquid heat exchangers • 2250mm×600mm×1375mm, fully liquid-cooled, with front and rear doors for liquid cooling

Building a Fully Connected, Intelligent World

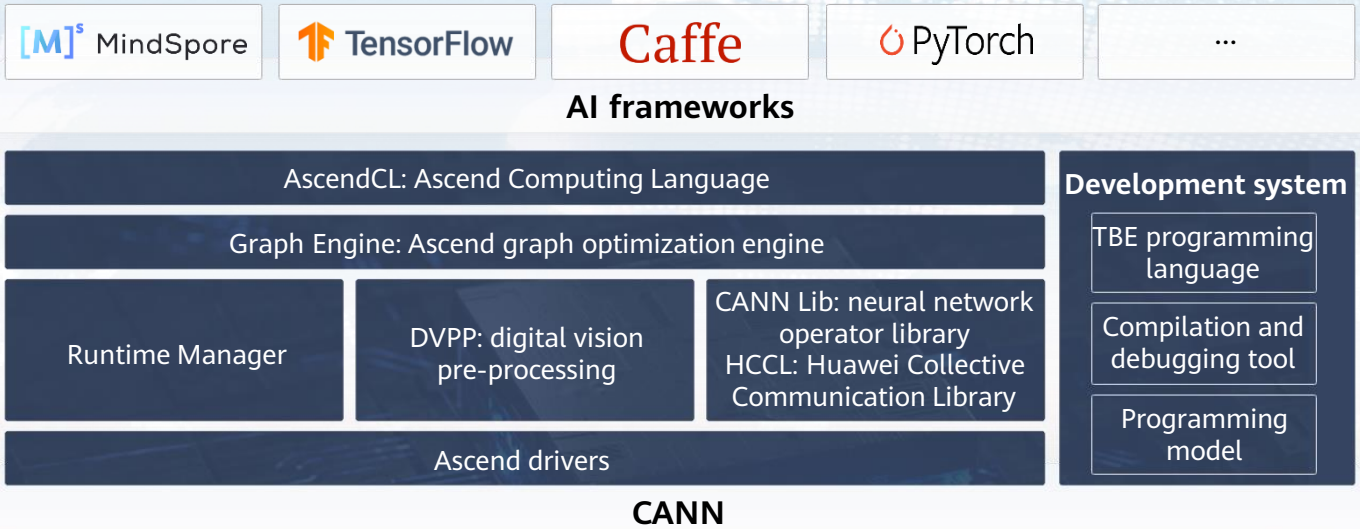
HUAWEI

CANN

Heterogeneous computing architecture

The Compute Architecture for Neural Networks (CANN) is a heterogeneous computing architecture designed for deep learning. Its core components exploit the computing power of the Ascend AI processor and allow users to quickly build AI applications and services on the Ascend platform. The core components include the AscendCL, DVPP, and HCCL.

Ascend Computing Language (AscendCL) provides a unified programming interface to implement software and hardware decoupling. Huawei Collective Communication Library (HCCL) facilitates efficient data transmission between different Ascend AI processors in distributed training. The digital vision pre-processing (DVPP) engine implements hardware acceleration to improve parallelism for image preprocessing.



CANN

Heterogeneous computing architecture

All-scenario enablement

14+ operating systems
10+ device/edge/cloud equipment form factors at the bottom layer
Adaptable to multiple AI frameworks

The ultimate performance

Graph compilation optimized for Ascend
A wide range of high-performance operators

Simplified development

Unified APIs adapted to the full range of hardware
Four open designs: plug-ins, graph fusion interfaces, Ascend-IR, and operator libraries

MindSpore

All-scenario AI Computing Framework

MindSpore is an AI computing framework tailored for use with Ascend AI processors. MindSpore supports flexible deployment in all scenarios, from devices and the edge to cloud. It creates a new AI programming paradigm, and reduces the threshold for AI development. MindSpore is built for user-friendly development, efficient runtime, and flexible deployment. It is instrumental in growing a vibrant AI software and hardware ecosystem.

AI framework

ModelArts	HiAI Service	Third-party platform

MindX	Third-party platform	MindX DL	ModelZoo	SDK

Application enablement

MindInsight Visualized tuning	Domain extension library
	ME: Unified APIs for all scenarios
MindArmour Enterprise-grade security	Computational graph compilation
	Distributed parallel architecture with on-demand collaboration across device-edge-cloud

MindSpore all-scenario AI computing framework

CANN

Atlas hardware

Public cloud	Private cloud	Edge computing	Consumer devices	Industry devices

All scenarios

Unrivaled E2E simplicity

- Model development kits for instant usability
- Model optimization kits for intuitive tuning
- One-click conversion to support third-party kits

Fully automatic parallelization

- Parallelization of serial algorithms takes just one line of code
- Automatic tensor splitting maximizes parallel efficiency

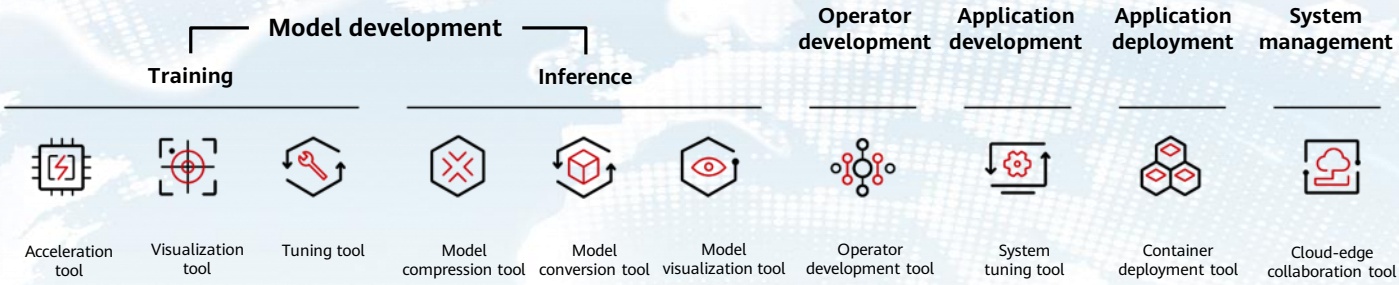
All-scenario collaboration

- Adaptive deployment and cross-heterogeneous hardware execution without model conversion
- Lightweight learning on the device side with customized models

MindStudio

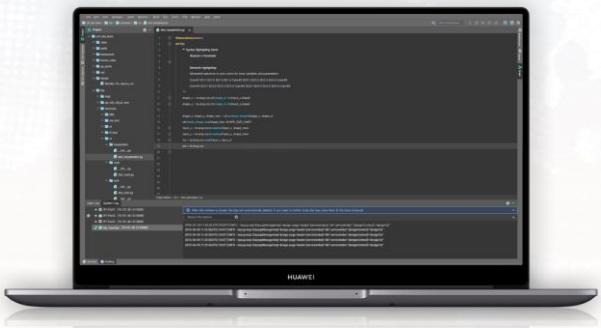
Full-Pipeline Development Toolchain

MindStudio is a one-stop environment for full-pipeline AI, operator, and application development based on Ascend. It provides model visualization, computing power test, and an integrated development environment (IDE) for local simulation debugging, making AI development easier and faster.



Model development

MindStudio provides a full series of tools for training and inference. In addition, it allows developers to invoke a large number of pre-trained AI models, model training scripts, and model development cases from ModelZoo, enabling more efficient model development.



Operator development

MindStudio delivers ease of use and flexibility for operator development and provides two operator development modes: Domain-Specific Language (DSL) and Tensor Iterator Kernel (TIK). During operator development, MindStudio also provides functions such as performance tuning and precision comparison.

TBE-DSL
Optimal development efficiency

- Automates data splitting and scheduling, so that developers can focus on computing representation
- Covers 70% of operators, and reduces the operator development time by 70% compared with industry counterparts

TBE-TIK
Optimal operator performance

- Provides instruction-level programming and tuning (with manual instruction set invocation, data splitting and orchestration)
- Covers all operators to fully unlock the chip performance

Application development

Supports AI application development such as system-level tuning and debugging transmission through AscendCL interfaces, and provides functions such as model/operator loading and execution and multiple C++ APIs

Application deployment

Manages and debugs devices in a unified manner through IP addresses, implementing remote management, debugging, and application push, and supporting seamless compatibility with different types of devices

System management

Ascend cloud-edge collaboration tool consists of FusionDirector and SmartKit. The tool allows developers to manage devices and deploy models in real time

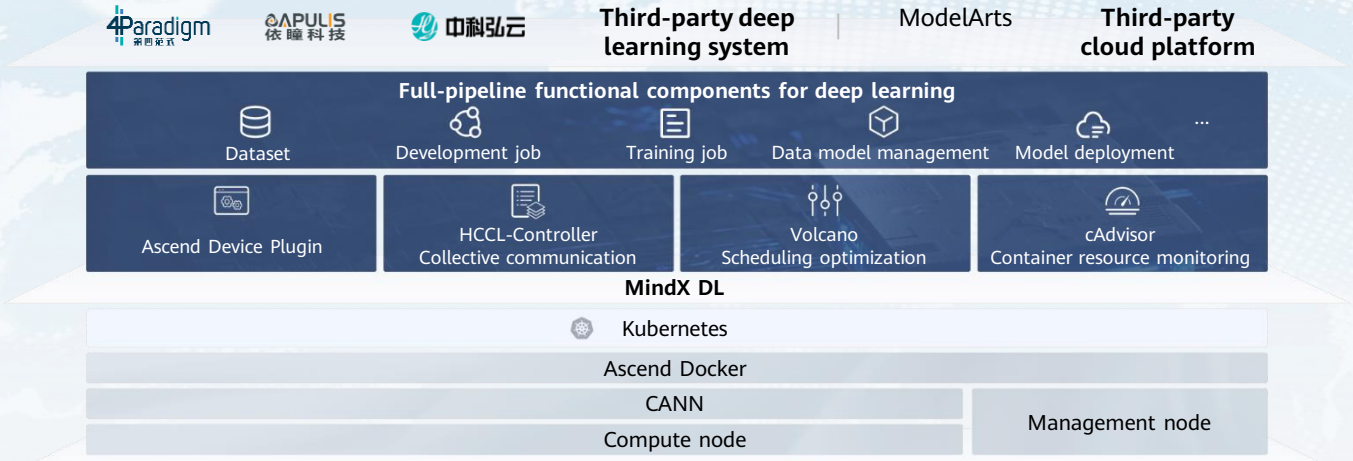
MindX

2+1+X for Ascend Application Enablement

MindX is designed to facilitate quick development of AI applications across different industries. MindX includes MindX DL for deep learning enablement, MindX Edge for intelligent edge enablement, the ModelZoo pre-trained model library, and multiple industry SDKs.

MindX DL for deep learning enablement

Unified management and scheduling of data center computing resources, enabling partners to quickly develop deep learning systems



Optimal computing resource scheduling

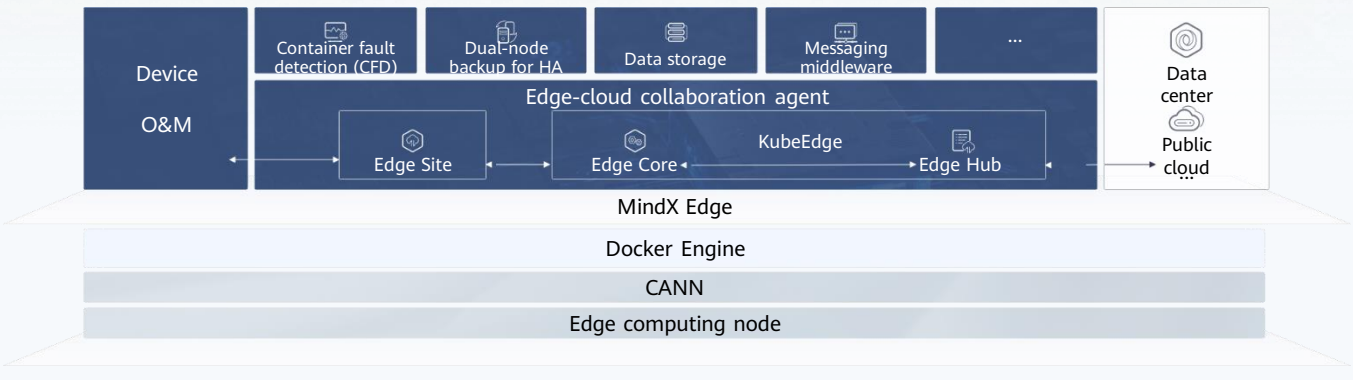
NPU device discovery, collective communication optimization, and group scheduling for massive amounts of data

Reference design for edge-cloud collaboration

Models at data centers can be released, updated, and pushed to the edge for inference

MindX Edge for intelligent edge enablement

Lightweight edge computing resource management and O&M, enabling industry customers to quickly build an edge-cloud collaborative inference platform



Edge-cloud collaborative design

Models pushed from the cloud to the edge for quick deployment
Edge data uploaded to the cloud for continuous training

Diversified hardware

Cameras, industrial computers, robots, drones, and edge inference servers

Lightweight deployment

Lightweight platform with only 256 MB memory overhead and 3% CPU usage

ModelZoo pre-trained model library

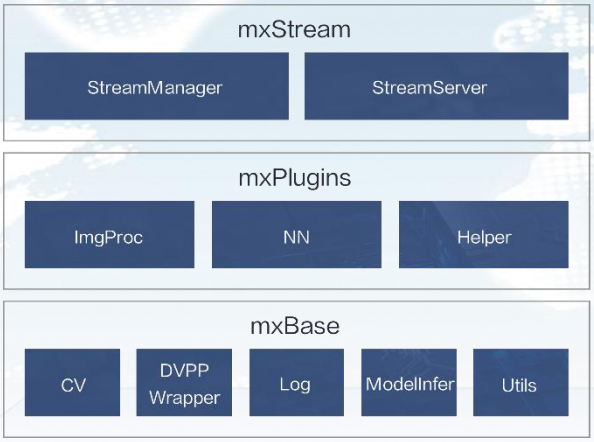
Provides developers with various scenario-specific pre-trained models, resolving the difficulties of model selection, training, and tuning

Easy access	ascend.huawei.com > ModelZoo	**Multiple AI frameworks**	MindSpore, TensorFlow, PyTorch, Caffe, etc.
Multiple scenarios	OCR, image detection, image classification, image segmentation, recommendation, NLP, machine translation, speech generation, enhanced learning, etc.	**High performance**	Pre-optimized models for dependable precision

MindX SDK

MindX SDK for Industry-specific Development

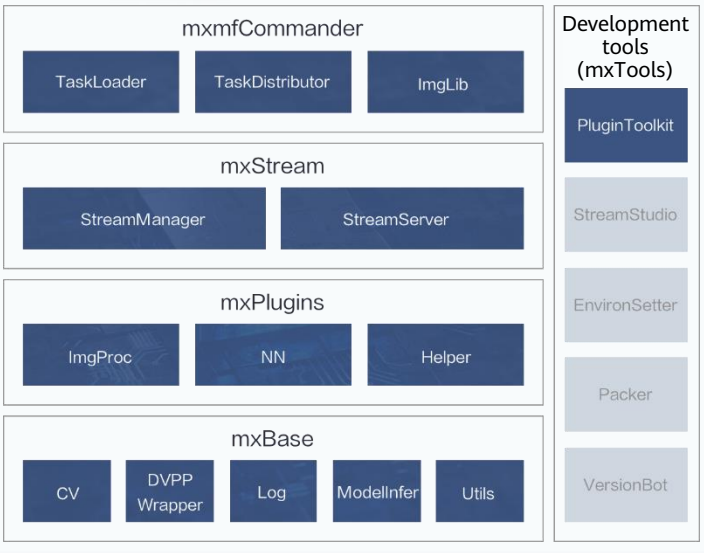
Comprehensive AI development kits for a range of industrial scenarios, providing easy-to-use APIs and graphical user interfaces to facilitate the development of AI applications with minimal code workload



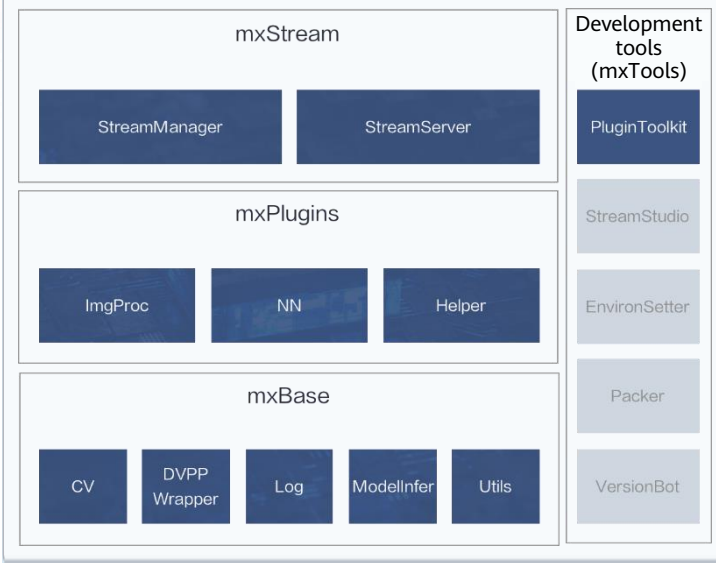
mxStream	Manages streams and generates inference services. StreamManager constructs and destroys streams, and provides interfaces for sending data and obtaining results. StreamServer constructs streams as an inference service and provides RESTful APIs to receive inference requests.
mxPlugins	Basic function plug-ins, such as model inference and image decoding plug-ins
mxBase	Basic function library of Ascend AI processors, including functions such as image encoding and decoding, cropping and resizing, and model inference
mxTools	Development tools provided by the SDK. PluginToolkit provides APIs for plug-in development.

Available in the **Ascend Community**

Manufacturing: mxManufacture



Computer vision: mxVision



Visit the Ascend Community to learn more.

ascend.huawei.com

Ascend Computing Industry Ecosystem

The Ascend computing industry ecosystem includes academic, technical, public, and business activities carried out using Ascend computing technology and product system, as well as various partners, including hardware partners, software algorithm partners, startups, universities, and industry developers. These elements constitute the partner ecosystem of the Ascend industry. Different roles interact with each other to bring AI to a wider range of industries.

Ascend Computing Industry

Open hardware, open source software, and partner enablement

1 AI computing industry: Enabling intelligent transformation across industries with Ascend

Healthcare | Carriers | Energy | Transportation | Internet | AI computing center | Manufacturing | Finance | Robotics | Smart city

2 business support plans: Enabling business partners to succeed

- Ascend ISV Program — **Enable business success of joint solutions**
- Ascend Startup Program — **Accelerate innovation and growth of startups**

3 talent cultivation initiatives: Cultivating talent for the Ascend ecosystem

- Ascend University Program
- MindSpore Paper + Model Incentive Program
- Ascend Developer Program

2 basic platforms for the Ascend ecosystem

- Ascend Ecosystem Innovation Center
- Ascend Developer Zone

Joint innovation with top ISVs

Manufacturing	Transportation	Energy	Finance	Internet	Healthcare
• Industrial quality inspection (semi-conductors, PCB, wires and tubes, and lens tubes)	• Free-flow tolling • Highway video cloud • Vehicle inspection	• Intelligent grid check • Smart substations • Smart customer service centers • Smart gas stations	• Smart branches • Financial OCR	• Precise recommendation • Content analysis	• COVID-19 diagnosis • Bone age evaluation

Ascend series tutorials

Promoting Ascend know-how in universities

Ascend AI Processor Architecture and Programming

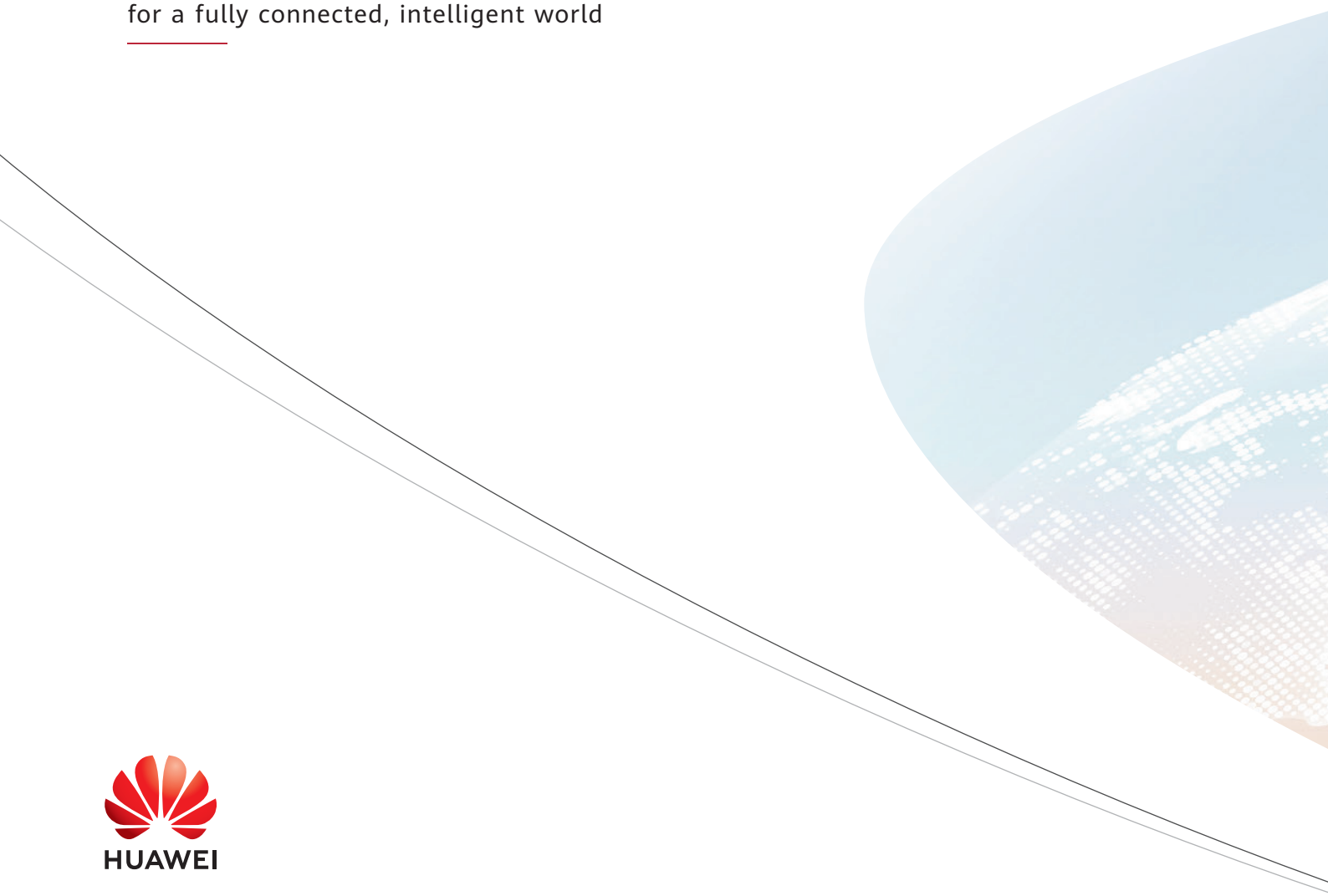
Deep Learning and Practice with MindSpore

Developing AI Applications with ModelArts

Collaboration with universities

Harnessing the synergy of academia and enterprise

Bring digital to every person, home and organization
for a fully connected, intelligent world



HUAWEI

HUAWEI TECHNOLOGIES CO., LTD.
Huawei Industrial Base, Bantian Longgang
Shenzhen 518129, P.R. China
Tel.: +86 755 28780808
ascend.huawei.com