Atlas AI Computing Platform

Foreword

- This chapter describes Huawei's Ascend AI chips, and hardware and software architectures of Ascend chips, and full-stack all-scenario solutions of Ascend AI chips.

Objectives

On completion of the course, you will be able to:

- Get an overview of AI chips.

- Understand hardware and software architectures of Huawei Ascend chips.

- Learn about Huawei Atlas AI computing platform.

- Understand industry applications of Atlas.

Contents

1. **Overview of AI Chips**

   ■ Summary

     ▫ Classification of AI Chips

     ▫ Current Status of AI Chips

     ▫ GPU, TPU, and Ascend 310 Design Comparison

     ▫ Ascend AI Processors
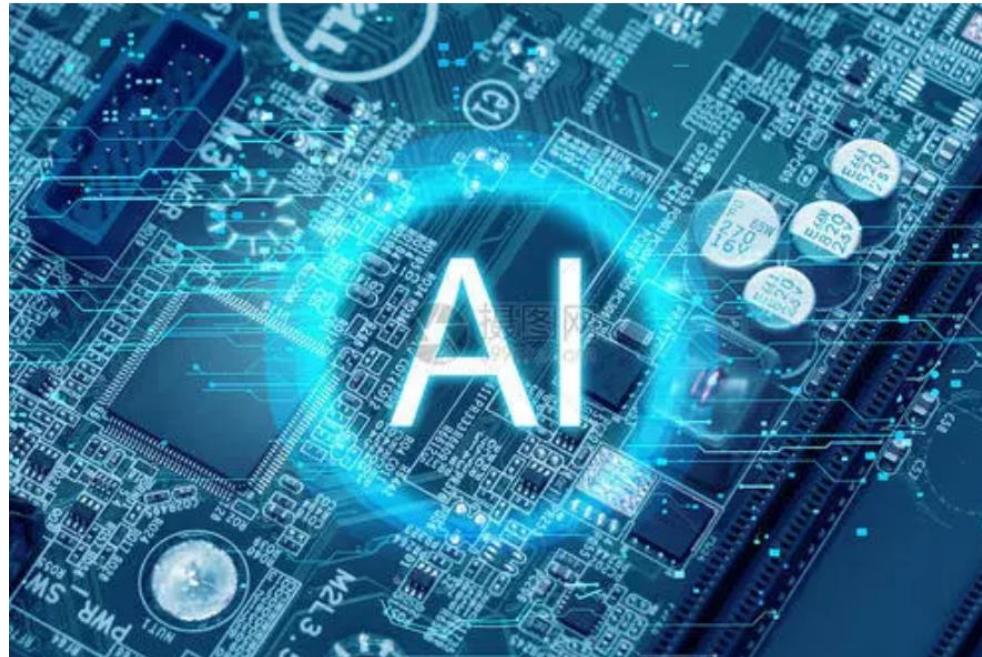
2. Hardware Architecture of Ascend Chips

3. ...

Overview and Objectives

- This section is an overview of AI chips, including the introduction, classification, and status of AI chips, comparison between GPUs and CPUs, and introduction of Ascend AI processors.

Definition

- Four elements of AI: data, algorithm, scenario, and computing power

- AI chips, also known as AI accelerators, are function modules that process massive computing tasks in AI applications.

Contents

1. Overview of AI Chips

▫ Summary

■ Classification of AI Chips

▫ Current Status of AI Chips

▫ Design Comparison of GPUs and CPUs

▫ Ascend AI Processors

2. Hardware Architecture of Ascend Chips

3. ...

Classification of AI Chips (1)

- AI Chips can be divided into four types by technical architecture:

  □ A central processing unit (CPU): a super-large-scale integrated circuit, which is the computing core and control unit of a computer. It can interpret computer instructions and process computer software data.

  □ A graphics processing unit (GPU): a display core, visual processor, and display chip. It is a microprocessor that processes images on personal computers, workstations, game consoles, and mobile devices, such as tablet computers and smart phones.

  □ An application specific integrated circuit (ASIC): an integrated circuit designed for a specific purpose.

  □ A field programmable gate array (FPGA): designed to implement functions of a semi-customized chip. The hardware structure can be flexibly configured and changed in real time based on requirements.

Classification of AI Chips (2)

- AI chips can be divided into training and inference by business application.

  - In the training phase, a complex deep neural network model needs to be trained through a large number of data inputs or an unsupervised learning method such as enhanced learning. The training process requires massive training data and a complex deep neural network structure. The huge computing amount requires ultra-high performance including computing power, precision, and scalability of processors. Nvidia GPU cluster and Google TPUs are commonly used in AI training.

  - Inferences are made using trained models and new data. For example, a video surveillance device uses the background deep neural network model to recognize a captured face. Although the calculation amount of the inference is much less than that of training, a large number of matrix operations are involved. GPU, FPGA and ASIC are also used in the inference process.

Contents

Current Status of AI Chips - CPU

- Central processing unit (CPU)

    - The computer performance has been steadily improved based on the Moore's Law.

    - The CPU cores added for performance enhancement also increase power consumption and cost.

    - Extra instructions have been introduced and the architecture has been modified to improve AI performance.

        - Instructions, such as AVX512, have been introduced into Intel processors (CISC architecture) and vector computing modules, such as FMA, into the ALU computing module.

        - Instruction sets including Cortex A have been introduced into ARM (RISC architecture), which will be upgraded continuously.

    - Despite that boosting the processor frequency can elevate the performance, the high frequency will cause huge power consumption and overheating of the chip as the frequency reaches the ceiling.

Current Status of AI Chips - GPU

- Graph processing unit (GPU)

  □ GPU performs remarkably in matrix computing and parallel computing and plays a key role in heterogeneous computing. It was first introduced to the AI field as an acceleration chip for deep learning. Currently, the GPU ecosystem has matured.

  □ Using the GPU architecture, NVIDIA focuses on the following two aspects of deep learning:

    ▪ Diversifying the ecosystem: It has launched the cuDNN optimization library for neural networks to improve usability and optimize the GPU underlying architecture.

    ▪ Improving customization: It supports various data types, including int8 in addition to float32; introduces modules dedicated for deep learning. For example, the optimized architecture of Tensor cores has been introduced, such as the TensorCore of V100.

  □ The existing problems include high costs and latency and low energy efficiency.

Current Status of AI Chips - TPU

- Tensor processing unit (TPU)

  - Since 2006, Google has sought to apply the design concept of ASICs to the neural network field and released TPU, a customized AI chip that supports TensorFlow, which is an open-source deep learning framework.

  - Massive systolic arrays and large-capacity on-chip storage are adopted to accelerate the most common convolution operations in deep neural networks.

    - Systolic arrays optimize matrix multiplication and convolution operations to elevate computing power and lower energy consumption.

HUAWEI

Current Status of AI Chips - FPGA

- Field programmable gate array (FPGA)

  - Using the HDL programmable mode, FPGAs are highly flexible, reconfigurable and re-programmable, and customizable.

  - Multiple FPGAs can be used to load the DNN model on the chips to lower computing latency. FPGAs outperform GPUs in terms of computing performance. However, the optimal performance cannot be achieved due to continuous erasing and programming. Besides, redundant transistors and cables, logic circuits with the same functions occupy a larger chip area.

  - The reconfigurable structure lowers supply and R&D risks. The cost is relatively flexible depending on the purchase quantity.

  - The design and tapeout processes are decoupled. The development period is long, generally half a year. The entry barrier is high.

HUAWEI

Contents

1. **Overview of AI Chips**

   □ Summary

   □ Classification of AI Chips

   □ Current Status of AI Chips

   ■ Design Comparison of GPUs and CPUs

   □ Ascend AI Processors

2. Hardware Architecture of Ascend Chips

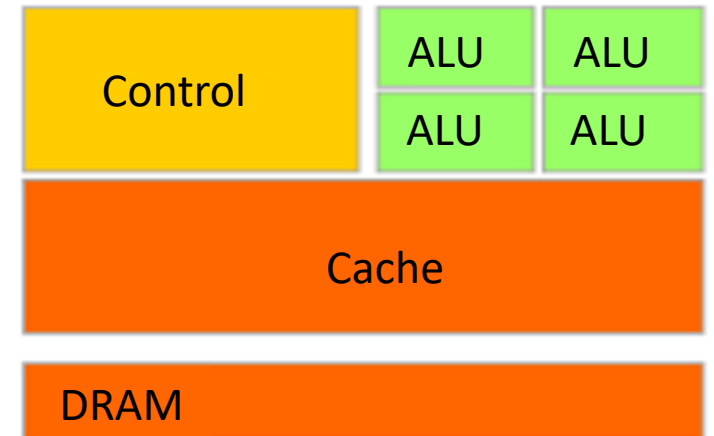3. Software Architecture of Ascend Chips

4. ...

Design Comparison of GPUs and CPUs

- GPUs are designed for massive data of the same type independent from each other and pure computing environments that do not need to be interrupted.

  - Each GPU comprises several large-sized parallel computing architectures with thousands of smaller cores designed to handle multiple tasks simultaneously.

  - Throughput-oriented design

    - With many ALUs and few caches, which improve services for threads, unlike those in CPU. The cache merges access to DRAM, causing latency.

    - The control unit performs combined access.

    - A large number of ALUs process numerous threads concurrently to cover up the latency.

  - Specialized in computing-intensive and easy-to-parallel programs

GPU
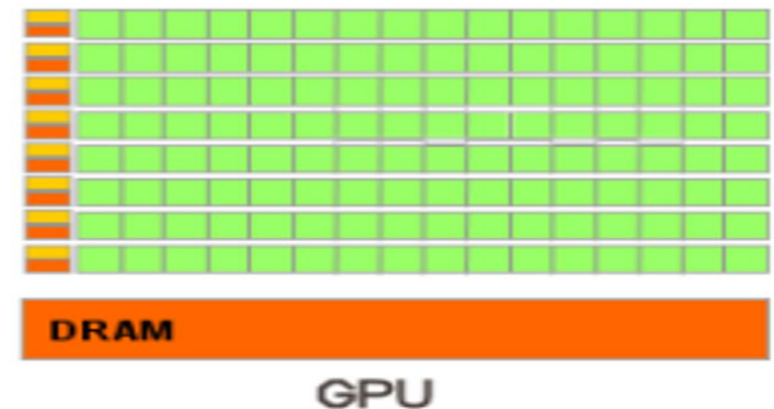
Design Comparison of GPUs and CPUs

- CPUs need to process different data types in a universal manner, perform logic judgment, and introduce massive branch jumps and interrupted processing.

  □ Composed of several cores optimized for sequential serial processing

  □ Low-latency design

    ▪ The powerful ALU unit can complete the calculation in a short clock cycle.

    ▪ The large cache lowers latency.

    ▪ High clock frequency

    ▪ Complex logic control unit, multi-branch programs can reduce latency through branch prediction.

    ▪ For instructions that depend on the previous instruction result, the logic unit determines the location of the instructions in the pipeline to speed up data forwarding.

  □ Specialized in logic control and serial operation

Control	ALU	ALU
	ALU	ALU
Cache		
DRAM		

Contents

1. **Overview of AI Chips**

   □ Summary

   □ Classification of AI Chips

   □ Current Status of AI Chips

   □ Design Comparison of GPUs and CPUs

   ■ Ascend AI Processors

2. Hardware Architecture of Ascend Chips

3. Software Architecture of Ascend Chips

4. …

Ascend AI Processors

- Neural-network processing unit (NPU): uses a deep learning instruction set to process a large number of human neurons and synapses simulated at the circuit layer. One instruction is used to process a group of neurons.

- Typical NPUs: Huawei Ascend AI chips, Cambricon chips, and IBM TrueNorth

- Ascend-Mini
- Architecture: Da Vinci
- Half precision (FP16): 8 Tera-FLOPS
- Integer precision (INT8): 16 Tera-OPS
- 16-channel full-HD video decoder: H.264/H.265
- 1-channel full-HD video decoder: H.264/H.265
- **Max. power: 8W**
- 12nm FFC

- Ascend-Max
- Architecture: Da Vinci
- Half precision (FP16): 256 Tera-FLOPS
- Integer precision (INT8): 512 Tera-OPS
- 128-channel full-HD video decoder: H.264/H.265
- Max. power: 350W
- 7nm

HUAWEI

Section Summary

- This section describes AI chips, including classification of AI chips by technologies and functions, AI chip ecosystem, and comparison between GPUs and CPUs.

HUAWEI

Contents

Overview and Objectives

- This section describes the hardware architecture of Ascend chips, including the logic architecture of the Ascend AI processors and Da Vinci architecture.
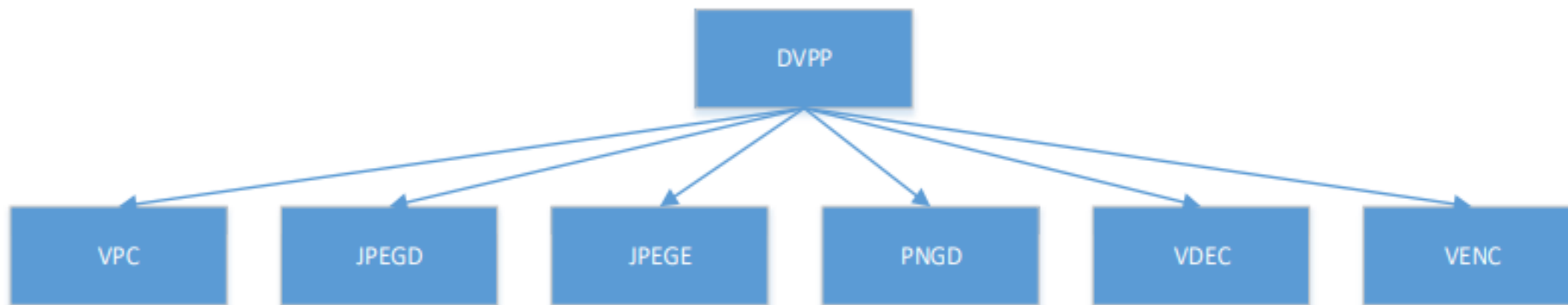
Huawei Confidential

Logic Architecture of Ascend AI Processors

- Ascend AI processor consist of:

  ▫ Control CPU (synchronization)

  ▫ AI computing engine, including AI core and AI CPU (昇腾AI芯片的计算核心，主要负责执行矩阵、向量、标量计算密集的算子任务)

  ▫ Multi-layer system-on-chip (SoC) caches or buffers (temp storage)(专用于AI Core、AI CPU，提供高带宽、低延迟的memory访问)

  ▫ Digital vision pre-processing (DVPP) module (数字视觉预处理子系统，完成图像视频的编解码)
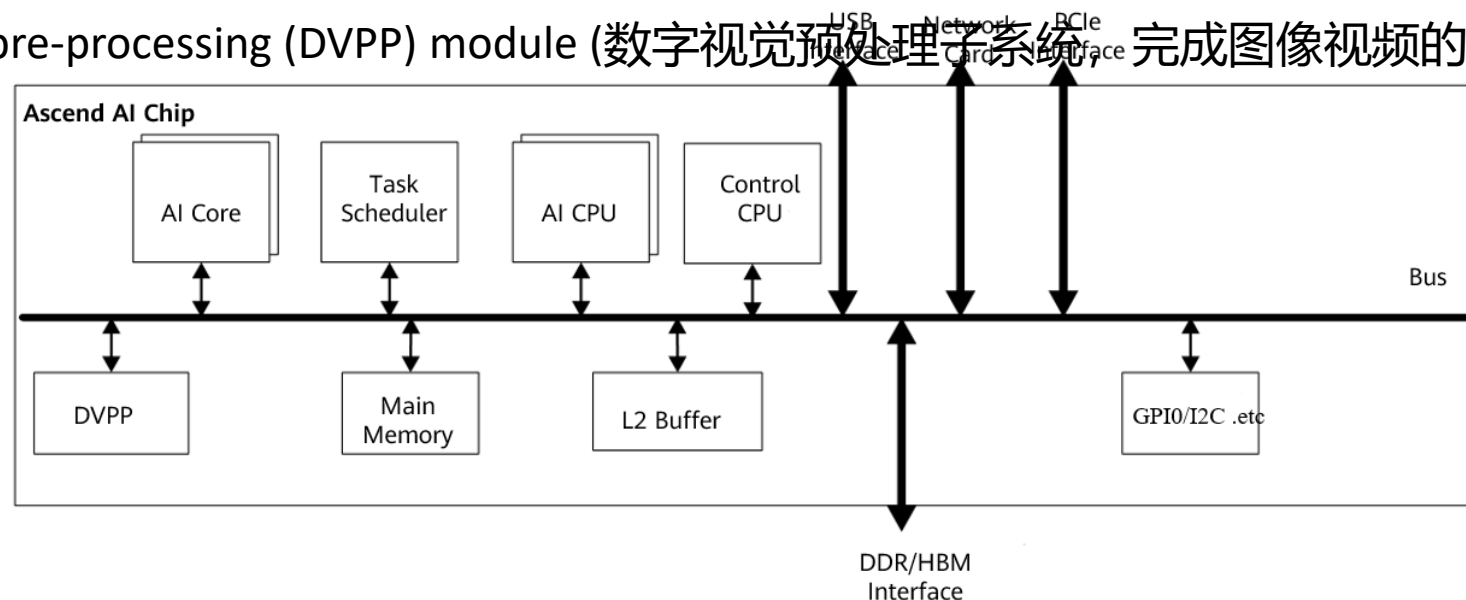
DVPP

DVPP(Digital Video Pre-Processor) : image/video encoding, decoding, resizing, cropping, format conversion.

```
                              ┌──────────┐
                              │   DVPP   │
                              └────┬─────┘
        ┌────────┬────────┬────────┼────────┬────────┐
   ┌────┴───┐ ┌──┴────┐ ┌─┴─────┐ ┌┴─────┐ ┌┴─────┐ ┌┴─────┐
   │  VPC   │ │ JPEGD │ │ JPEGE │ │ PNGD │ │ VDEC │ │ VENC │
   └────────┘ └───────┘ └───────┘ └──────┘ └──────┘ └──────┘
```

| DVPP	Digital Vision Pre-Process
VPC	Vision Preprocess Core
JPEGD	JPEG Decode
JPEGE	JPEG Encoder
PNGD	Portable Network Graphics Decoder
VDEC	Video Decoder
VENC	Video Encoder

Contents

1. Overview of AI Chips

2. **Hardware Architecture of Ascend Chips**

   ▫ Logic Architecture of Ascend AI Processors

   ■ Da Vinci Architecture

3. Software Architecture of Ascend Chips

4. Huawei Atlas AI Computing Platform

5. Industry Applications of Atlas

Ascend AI Computing Engine - Da Vinci Architecture

- One of the four major architectures of Ascend AI processors is the AI computing engine, which consists of the AI core (Da Vinci architecture) and AI CPU. The Da Vinci architecture developed to improve the AI computing power serves as the core of the Ascend AI computing engine and AI processor.

Da Vinci Architecture (AI Core)

- Main components of the Da Vinci architecture:

  □ Computing unit: It consists of the cube unit, vector unit, and scalar unit.

  □ Storage system: It consists of the on-chip storage unit of the AI core and data channels.

  □ Control unit provides instruction control for the entire computing process. It is equivalent to the command center of the AI core and is responsible for the running of the entire AI core.
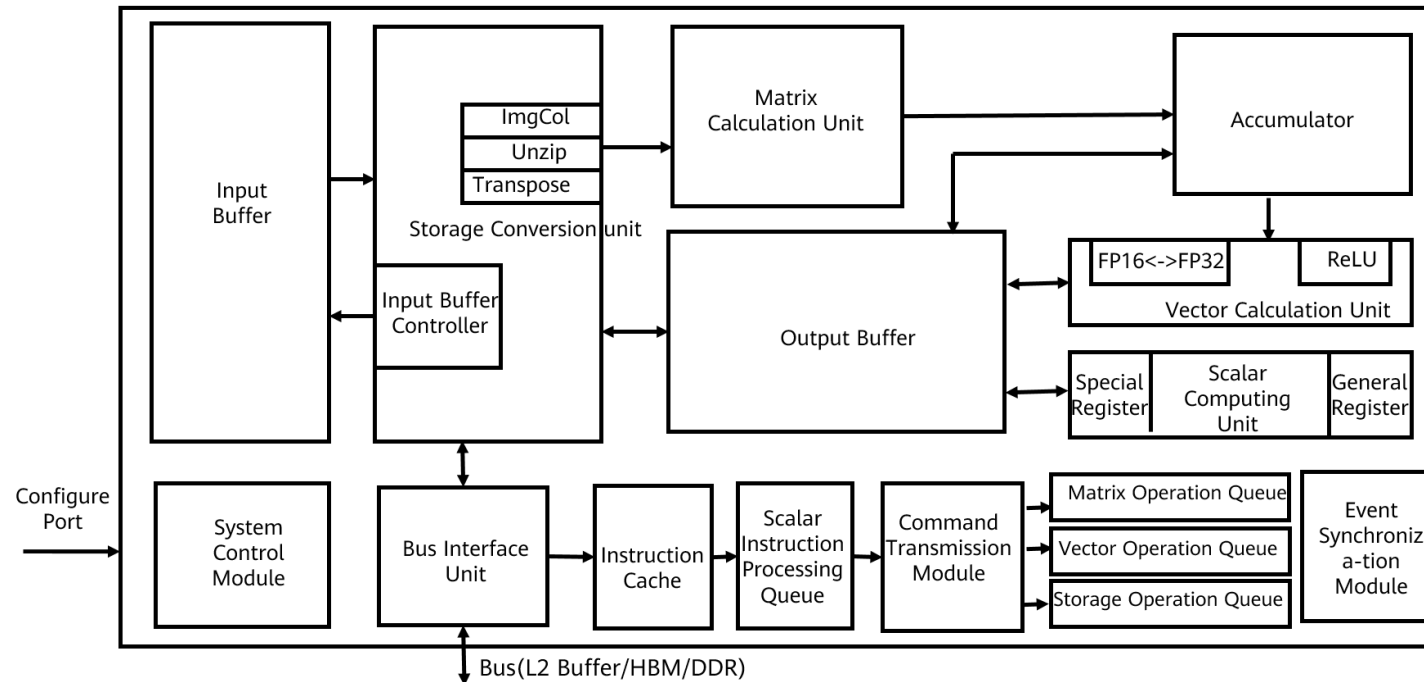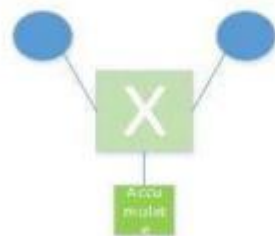
Da Vinci Architecture (AI Core) - Computing Unit

- Three types of basic computing units: cube, vector, and scalar units, which correspond to matrix, vector and scalar computing modes respectively.

- Cube computing unit: The matrix computing unit and accumulator are used to perform matrix-related operations. Completes a matrix (4096) of 16x16 multiplied by 16x16 for FP16, or a matrix (8192) of 16x32 multiplied by 32x16 for the INT8 input in a shot.

- Vector computing unit: Implements computing between vectors and scalars or between vectors. This function covers various basic computing types and many customized computing types, including computing of data types such as FP16, FP32, INT32, and INT8.

- Scalar computing unit: Equivalent to a micro CPU, the scalar unit controls the running of the entire AI core. It implements loop control and branch judgment for the entire program, and provides the computing of data addresses and related parameters for cubes or vectors as well as basic arithmetic operations.

HUAWEI

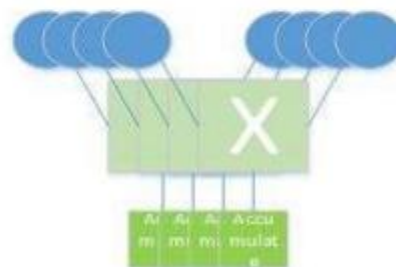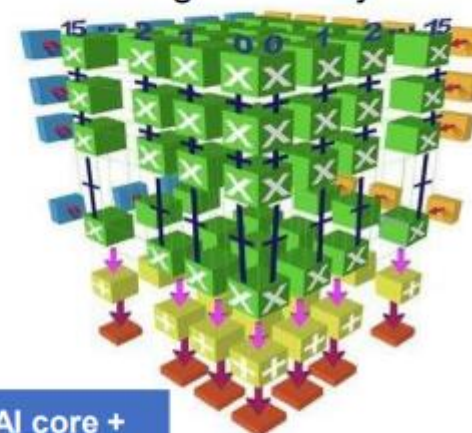# Building Blocks and their Computation Intensity

1D Scalar Unit + **2D Vector Unit** + **3D Matrix Unit**

Full flexibility | Rich & efficient operations | High intensity

N	N²	N³
1	1	1
2	4	8
4	16	128
8	64	512
16	**256**	**4096**
32	1024	32768
64	4096	262144

	GPU + Tensor core	AI core + SRAM
Area (normalized to 12 nm)	5.2mm^2	13.2mm^2
Compute power	1.7Tops fp16	8Tops fp16

HUAWEI

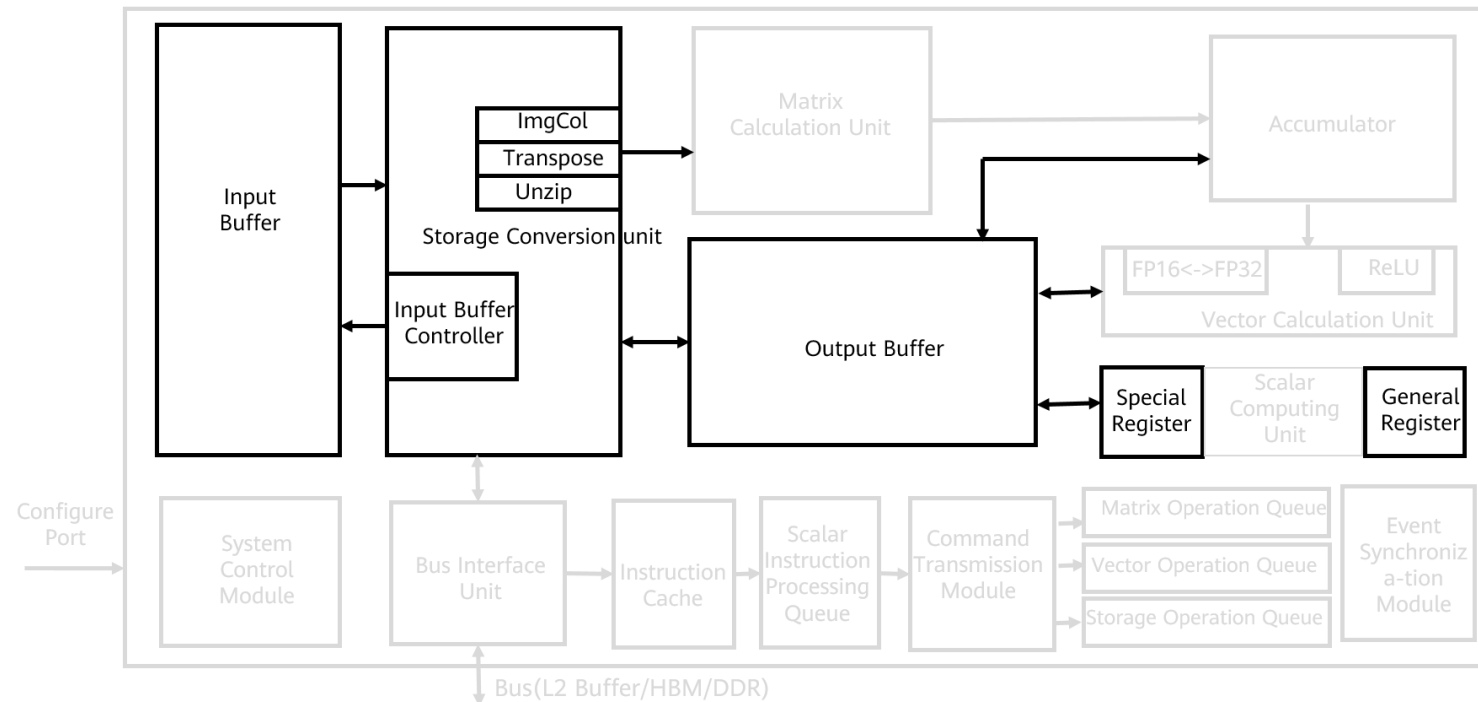# Da Vinci Architecture (AI Core) - Computing Unit
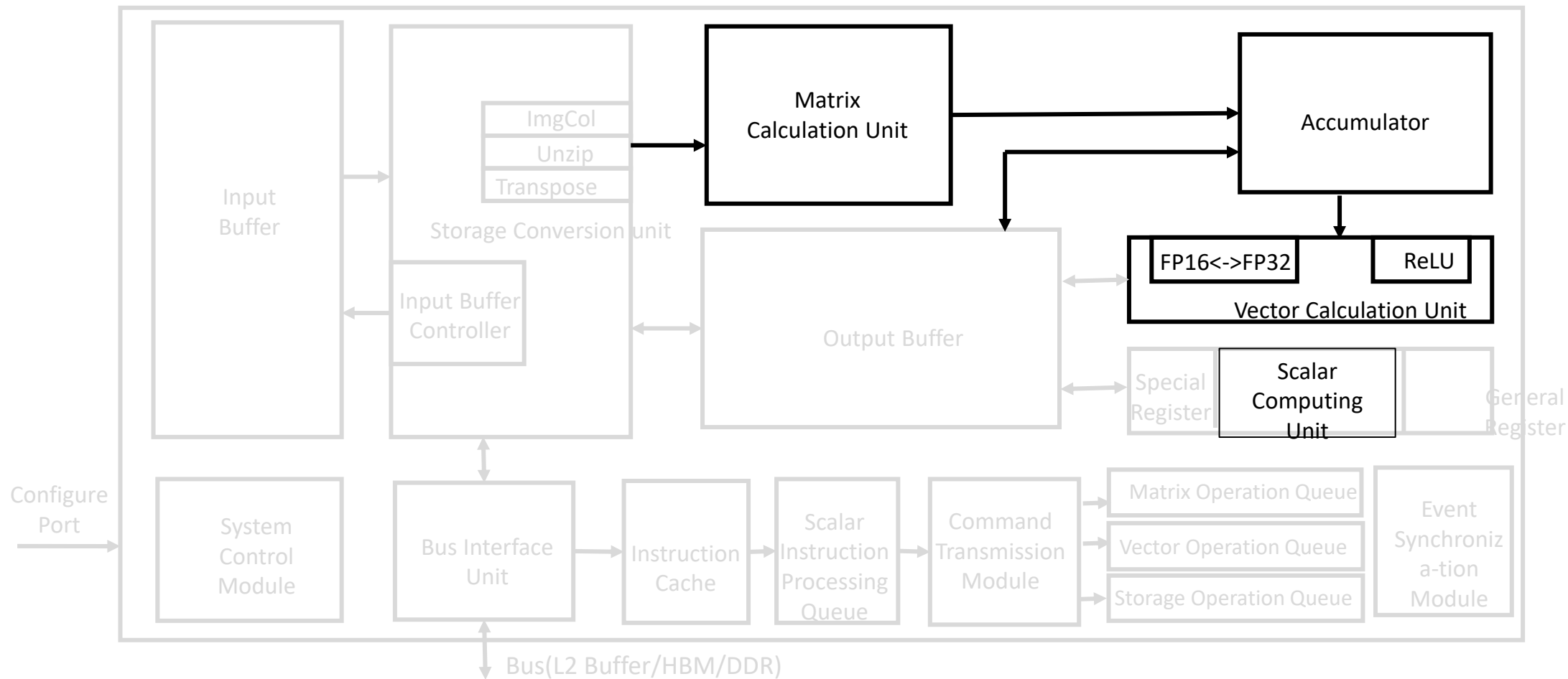
Da Vinci Architecture (AI Core) - Storage System (1)

- The storage system of the AI core is composed of the storage unit and corresponding data channel.

- The storage unit consists of the storage control unit, buffer, and registers:

- Storage control unit: The cache at a lower level than the AI core can be directly accessed through the bus interface. The memory can also be directly accessed through the DDR or HBM. A storage conversion unit is set as a transmission controller of the internal data channel of the AI core to implement read/write management of internal data of the AI core between different buffers. It also completes a series of format conversion operations, such as zero padding, Img2Col, transposing, and decompression.

- Input buffer: The buffer temporarily stores the data that needs to be frequently used so the data does not need to be read from the AI core through the bus interface each time. **This mode reduces the frequency of data access on the bus and the risk of bus congestion, thereby reducing power consumption and improving performance.**

- Output buffer: **The buffer stores the intermediate results of computing at each layer in the neural network, so that the data can be easily obtained for next-layer computing**. Reading data through the bus involves low bandwidth and long latency, whereas using the output buffer greatly improves the computing efficiency.

- Register: Various registers in the AI core are mainly used by the scalar unit.

Da Vinci Architecture (AI Core) - Storage System (2)

- Data channel: path for data flowing in the AI core during execution of computing tasks

  - A data channel of the Da Vinci architecture is characterized by **multiple-input single-output**. Considering various types and a large quantity of input data in the computing process on the neural network, parallel inputs can improve data inflow efficiency. On the contrary, only an output feature matrix is generated after multiple types of input data are processed**. The data channel with a single output of data reduces the use of chip hardware resources**.
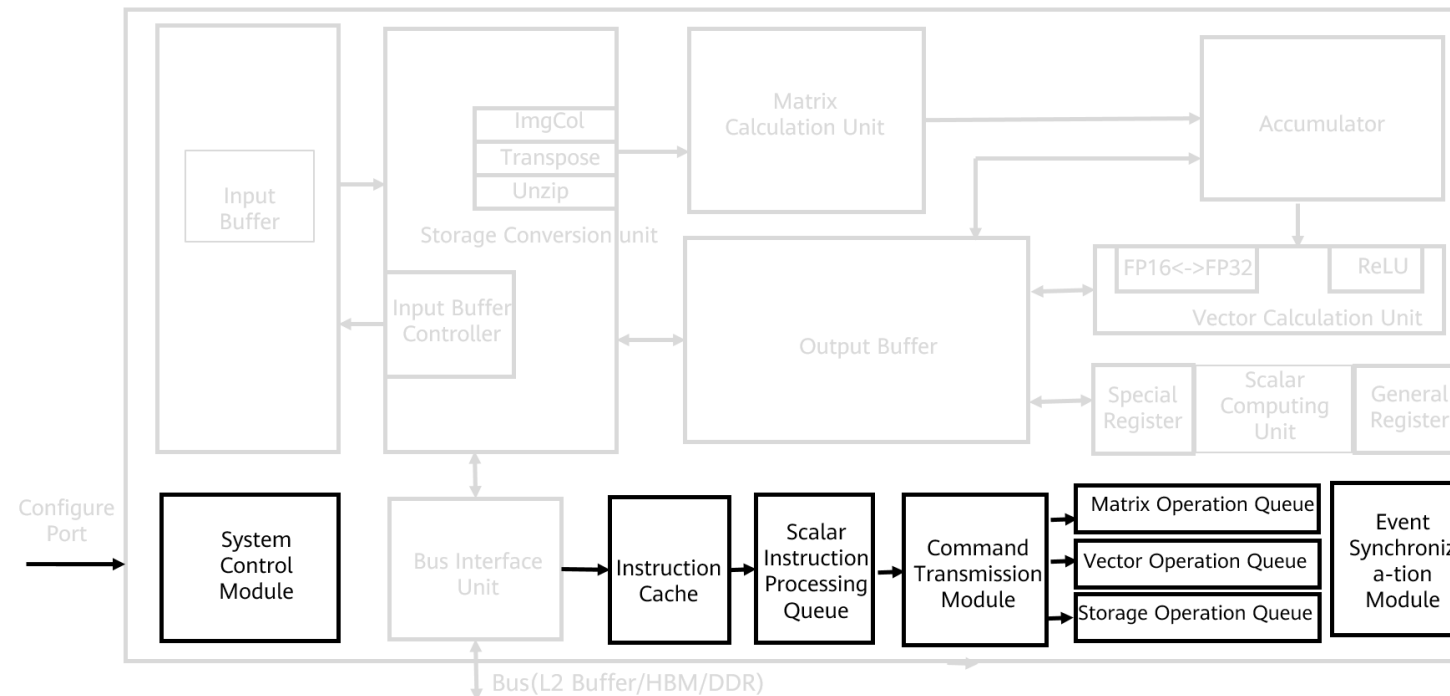
Da Vinci Architecture (AI Core) - Control Unit (1)

- The control unit consists of the system control module, instruction cache, scalar instruction processing queue, instruction transmitting module, matrix operation queue, vector operation queue, storage conversion queue, and **event synchronization** module.

  - System control module: Controls the execution process of a task block (minimum task computing granularity for the AI core). After the task block is executed, the system control module processes the interruption and reports the status. If an error occurs during the execution, the error status is reported to the task scheduler.

  - Instruction cache: Prefetches subsequent instructions in advance during instruction execution and reads multiple instructions into the cache at a time, improving instruction execution efficiency.

  - Scalar instruction procession queue: After being decoded, the instructions are imported into a scalar queue to implement address decoding and operation control. The instructions include matrix computing instructions, vector calculation instructions, and storage conversion instructions.

  - Instruction transmitting module: Reads the configured instruction addresses and decoded parameters in the scalar instruction queue, and sends them to the corresponding instruction execution queue according to the instruction type. The scalar instructions reside in the scalar instruction processing queue for subsequent execution.

Da Vinci Architecture (AI Core) - Control Unit (2)

- Instruction execution queue: Includes a matrix operation queue, vector operation queue, and storage conversion queue. Different instructions enter corresponding operation queues, and instructions in the queues are executed according to the entry sequence.

- Event synchronization module: Controls the execution status of each instruction pipeline in real time, and analyzes dependence relationships between different pipelines to resolve problems of data dependence and synchronization between instruction pipelines.

Section Summary

- This section describes the hardware architecture of Ascend chips, including the computing unit, storage unit, and control unit of the core Da Vinci architecture.

HUAWEI

# Contents

Overview and Objectives

- This section describes the software architecture of Ascend chips, including the logic architecture and neural network software flow of Ascend AI processors.

Logic Architecture of Ascend AI Processor Software Stack (1)

- L3 application enabling layer: It is an application-level encapsulation layer that provides different processing algorithms for specific application fields. L3 provides various fields with computing and processing engines. It can directly use the framework scheduling capability provided by L2 to generate corresponding NNs and implement specific engine functions.

  - Generic engine: provides the generic neural network inference capability.

  - Computer vision engine: encapsulates video or image processing algorithms.

  - Language and text engine: encapsulates basic processing algorithms for voice and text data.

Huawei Confidential
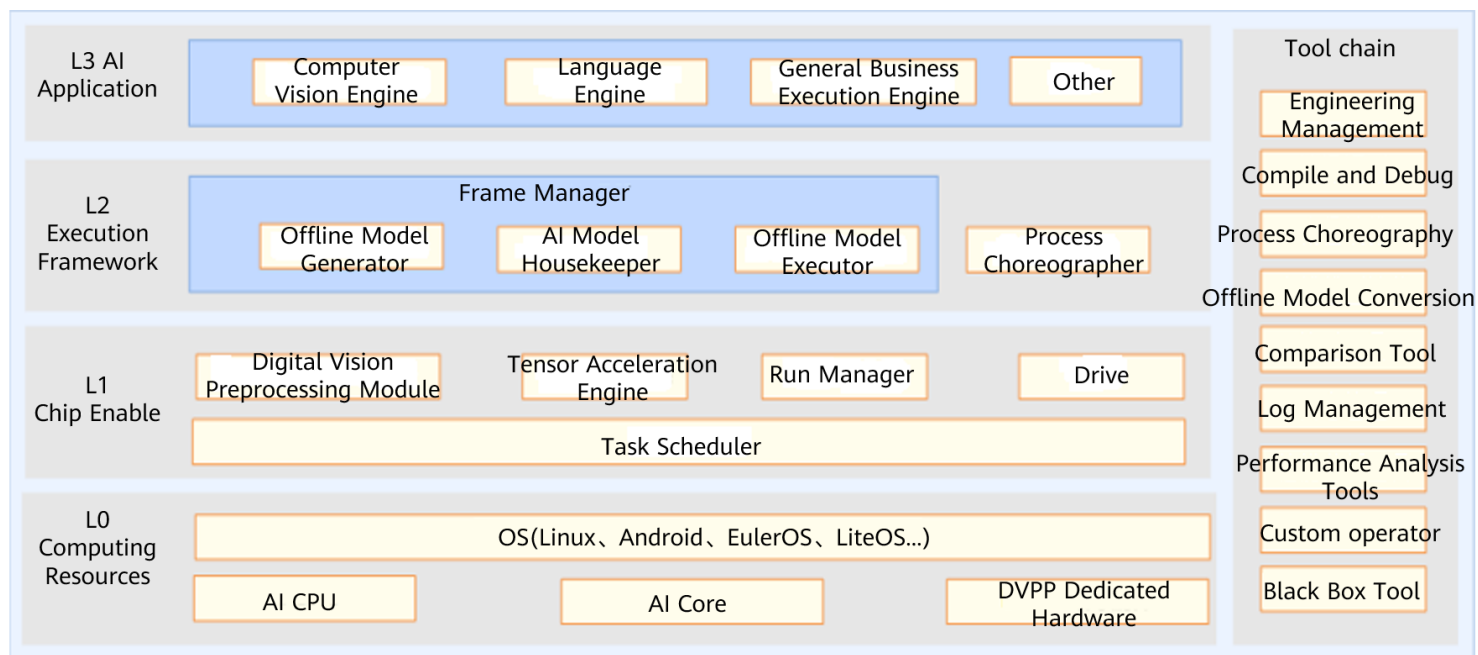
Logic Architecture of Ascend AI Processor Software Stack (2)

- L2 execution framework layer: encapsulates the framework calling capability and offline model generation capability. After the application algorithm is developed and encapsulated into an engine at L3, L2 calls the appropriate deep learning framework, such as Caffe or TensorFlow, based on the features of the algorithm to obtain the neural network of the corresponding function, and generates an offline model through the framework manager. After L2 converts the original neural network model into an offline model that can be executed on Ascend AI chips, the offline model executor (OME) transfers the offline model to Layer 1 for task allocation.

- L1 chip enabling layer: bridges the offline model to Ascend AI chips. L1 accelerates the offline model for different computing tasks via libraries. Nearest to the bottom-layer computing resources, L1 outputs operator-layer tasks to the hardware.

- L0 computing resource layer: provides computing resources and executes specific computing tasks. It is the hardware computing basis of the Ascend AI chip.
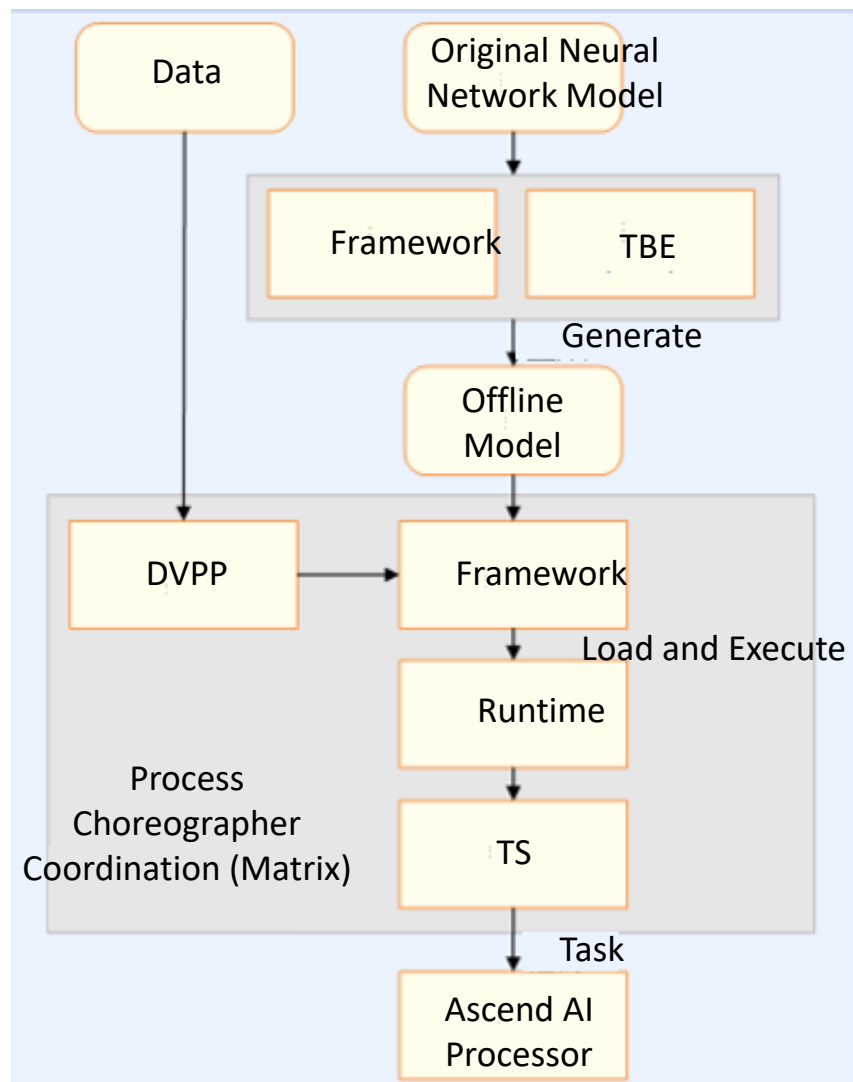
Contents

Neural Network Software Flow of Ascend AI Processors

- The neural network software flow of Ascend AI processors is a bridge between the deep learning framework and Ascend AI chips. It realizes and executes a neural network application and integrates the following functional modules.

- Process orchestrator: implements the neural network on Ascend AI chips, coordinates the whole process of effecting the neural network, and controls the loading and execution of offline models.

- Digital vision pre-processing (DVPP) module: performs data processing and cleaning before input to meet format requirements for computing.

- Tensor boosting engine (TBE): functions as a neural network operator factory that provides powerful computing operators for neural network models.

- Framework manager: builds an original neural network model into a form supported by Ascend AI chips, and integrates the new model into Ascend AI chips to ensure efficient running of the neural network.

- Runtime manager: provides various resource management paths for task delivery and allocation of the neural network.

- Task scheduler: As a task driver for hardware execution, it provides specific target tasks for Ascend AI chips. The operation manager and task scheduler work together to form a dam system for neural network task flow to hardware resources, and monitor and distribute different types of execution tasks in real time.
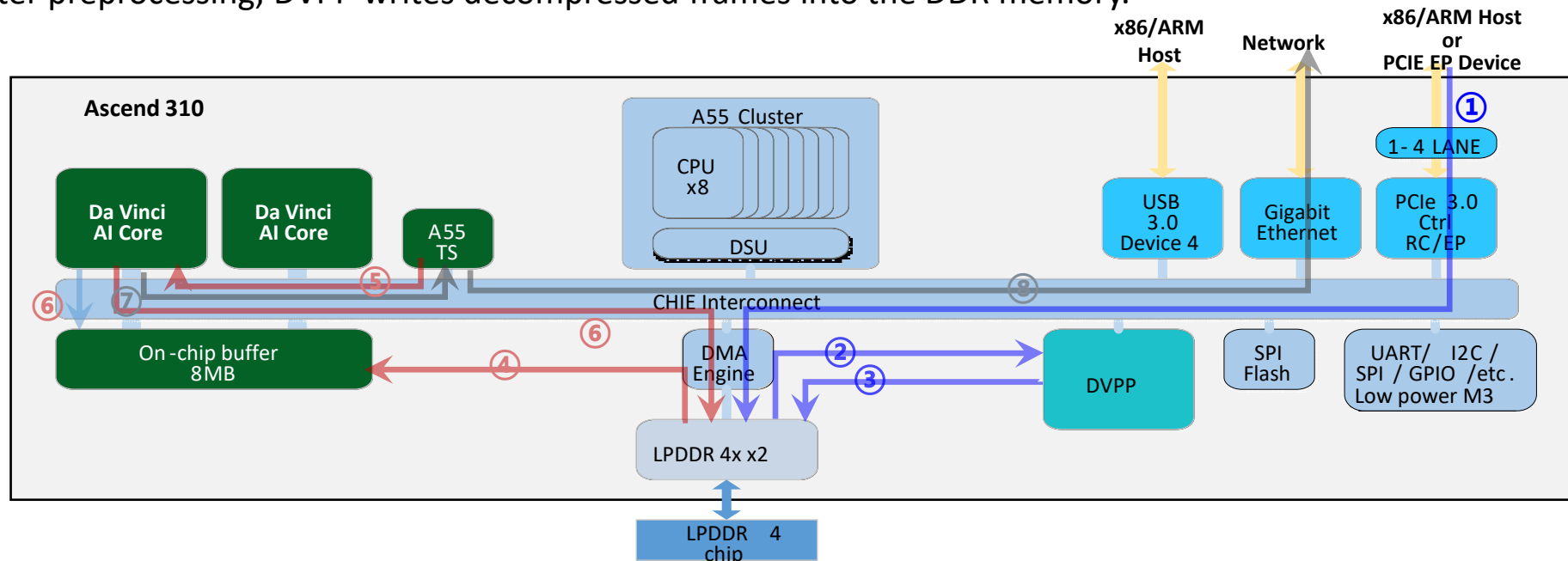
Neural Network Software Flow of Ascend AI Processors

Contents

Data Flowchart of the Ascend AI Processor - Facial Recognition Inference Application (1)

- Camera data collection and processing

  □ Compressed video streams are transmitted from the camera to the DDR memory through PCIe.

  □ DVPP reads the compressed video streams into the cache.

  □ After preprocessing, DVPP writes decompressed frames into the DDR memory.

Data Flowchart of the Ascend AI Processor - Facial Recognition Inference Application (2)

- Data inference

  - The task scheduler (TS) sends an instruction to the DMA engine to pre-load the AI resources from the DDR to the on-chip buffer.

  - The TS configures the AI core to execute tasks.

  - The AI core reads the feature map and weight, and writes the result to the DDR or on-chip buffer.

- Facial recognition result output

  - After processing, the AI core sends the signals to the TS, which checks the result. If another task needs to be allocated, the operation in step ④ is performed.

  - When the last AI task is complete, the TS reports the result to the host.

Section Summary

- This section describes the software architecture of Ascend chips, including the L0-L3 software architectures and the neural network software flow of the Ascend AI processor.
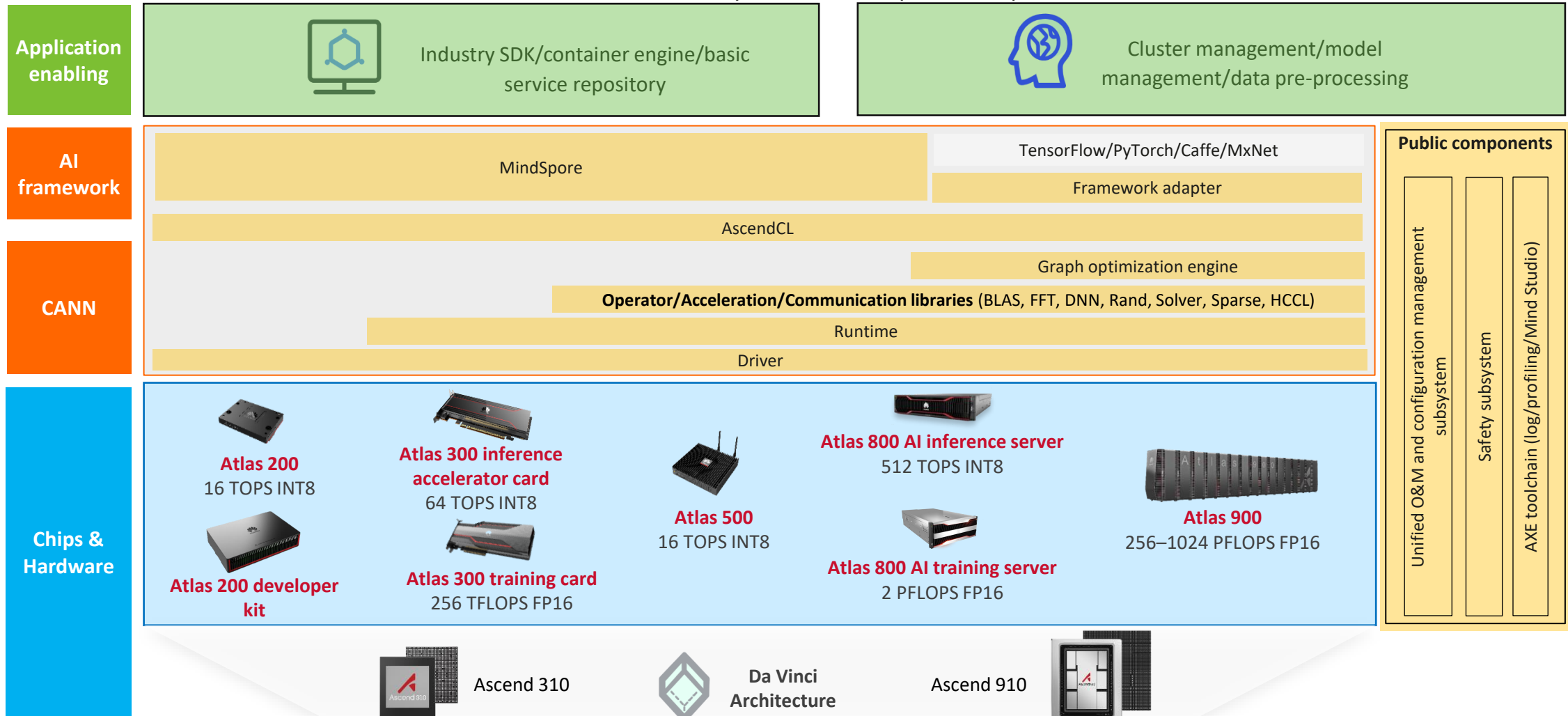
Contents

Overview and Objectives

- This section describes the main products of Huawei Atlas AI computing platform, including inference and training.

Atlas AI Computing Platform Portfolio

Internet, security, finance, transportation, power, etc.

Application enabling

Industry SDK/container engine/basic service repository

Cluster management/model management/data pre-processing

AI framework

MindSpore

TensorFlow/PyTorch/Caffe/MxNet

Framework adapter

AscendCL

CANN

Graph optimization engine

Operator/Acceleration/Communication libraries (BLAS, FFT, DNN, Rand, Solver, Sparse, HCCL)

Runtime

Driver

Public components

Unified O&M and configuration management subsystem

Safety subsystem

AXE toolchain (log/profiling/Mind Studio)

Chips & Hardware

Atlas 200
16 TOPS INT8

Atlas 300 inference accelerator card
64 TOPS INT8

Atlas 500
16 TOPS INT8

Atlas 800 AI inference server
512 TOPS INT8

Atlas 900
256–1024 PFLOPS FP16

Atlas 200 developer kit

Atlas 300 training card
256 TFLOPS FP16

Atlas 800 AI training server
2 PFLOPS FP16

Ascend 310

Da Vinci Architecture

Ascend 910

HUAWEI

Atlas Accelerates AI Inference

Ascend 310
AI processor

Performance improved 7x for terminal devices

Highest density in the industry (64-channel) for video inference

Edge intelligence and cloud-edge collaboration

Powerful computing platform for AI inference

Atlas 200 Developer Kit (DK)
AI developer kit
Model: 3000

Atlas 200 AI accelerator module
Model: 3000

Atlas 300 AI accelerator card
Model: 3000

Atlas 500 AI edge station
Model: 3000

Atlas 800 AI server
Model: 3000/3010

Atlas 200DK: Strong Computing Power and Ease-of-Use

Full-Stack AI development on and off the cloud

- **16TOPS INT8 24W**

- **1 USB type-C, 2 camera ports, 1 GE port, 1 SD card slot**

- **8 GB memory**

- **Operating temperature: 0°C to 45°C**

- **Dimensions (H x W x D): 24 mm x 125 mm x 80 mm**

Developers

Set up a dev environment with one laptop
Ultra low cost for local independent environment, with multiple functions and interfaces to meet basic requirements

Researchers

Local dev + cloud training collaboration
Same protocol stack for Huawei Cloud and the developer kit; training on the cloud and deployment at local; no modification required

Startups

Code-level demo
Implementing the algorithm function by modifying 10% code based on the reference architecture; interaction with the Developer Community; seamless migration of commercial products

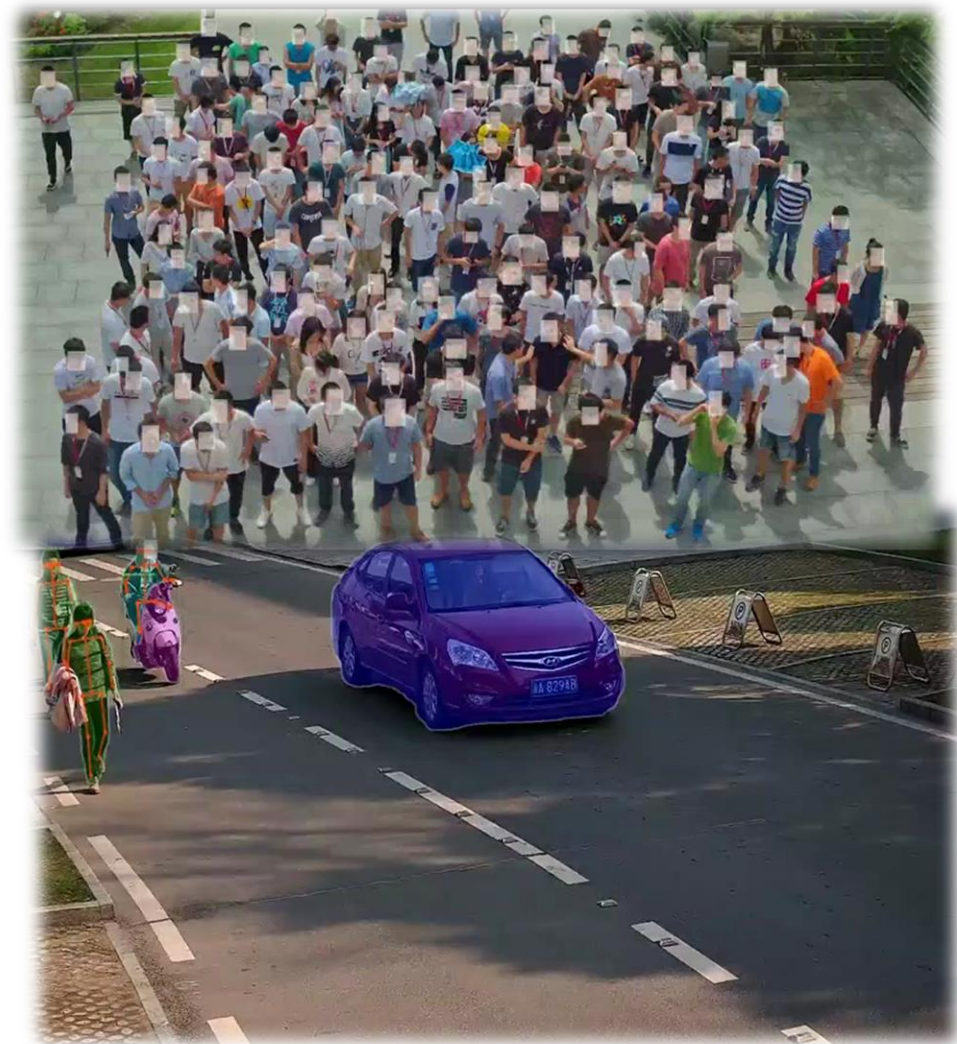# Atlas 200DK: Strong Computing Power and Ease-of-Use

30 → **200**

Number of faces recognized per frame

- **16 TOPS INT8, 9.5 w**
- **16-channel HD video real-time analytics, JPEG decoding**
- **4 GB /8 GB memory, four PCIe 3.0 interfaces**
- **Operating temperature: –25°C to +80°C**

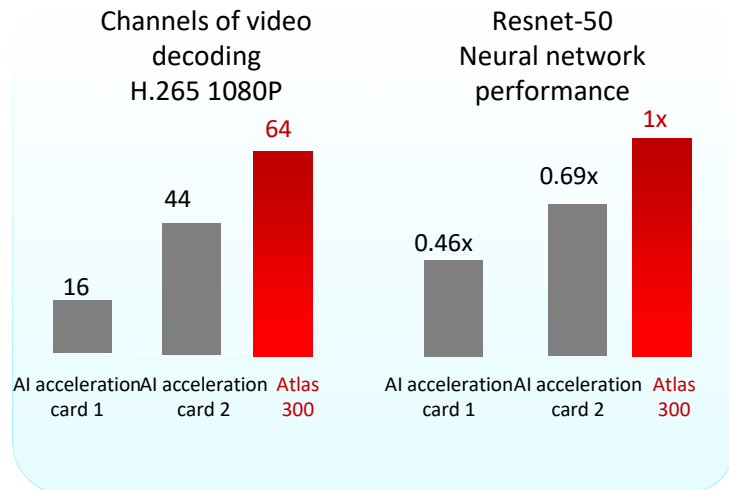- **Concurrent running of multiple algorithms on AI camera**

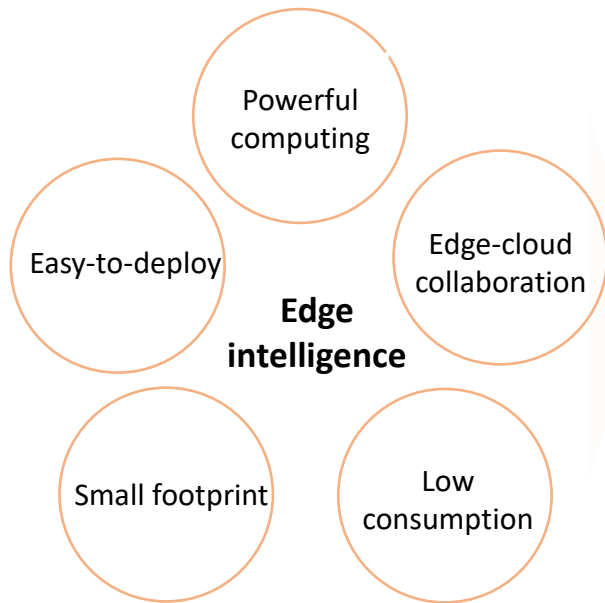# Atlas 300: Highest-Density, 64-Channel Video Inference Accelerator Card

- 64 TOPS INT8, 67 w
- 32 GB memory
- 64-channel HD video analytics in real time

Model: 3000

Channels of video decoding H.265 1080P
- AI acceleration card 1: 16
- AI acceleration card 2: 44
- Atlas 300: 64

Resnet-50 Neural network performance
- AI acceleration card 1: 0.46x
- AI acceleration card 2: 0.69x
- Atlas 300: 1x

Video Analytics | OCR | Speech Recognition | Precision Marketing | Medical Image Analysis

HUAWEI

Atlas 500 AI Edge Station

Edge intelligence:
- Powerful computing
- Edge-cloud collaboration
- Easy-to-deploy
- Small footprint
- Low consumption

- **16 TOPS INT8**
- **25 w to 40 w**
- **Wi-Fi & LTE**
- **64-channel HD video analytics in real time**
- **Fanless design, stable from 40°C to +70°C**

Camera → Customer traffic statistics / Atlas 500 → Data Analysis / Backend server

HUAWEI
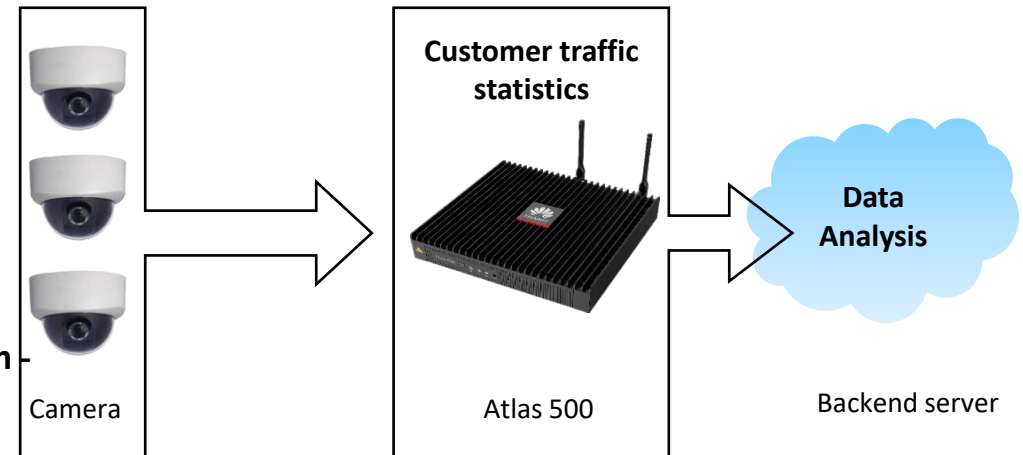
Atlas 800 AI Server

Model: 3000

Model: 3010

- **An efficient inference platform powered by Kunpeng**

- **Key functions:**
  - ◻ 2 Kunpeng 920 processors in a 2U space
  - ◻ 8 PCIe slots, supporting up to 8 Atlas 300 AI accelerator cards
  - ◻ Up to 512-channel HD video real-time analytics
  - ◻ Air-cooled, stable at 5°C to 40°C

- **A flexible inference platform powered by Intel**

- **Key functions:**
  - ◻ 2 Intel® Xeon® SP Skylake or Cascade Lake processors in a 2U space
  - ◻ 8 PCIe slots, supporting up to 7 Atlas 300/NVIDIA T4 AI accelerator cards
  - ◻ Up to 448-channel HD video real-time analytics
  - ◻ Air-cooled, stable at 5°C to 35°C

Atlas Accelerates AI Training
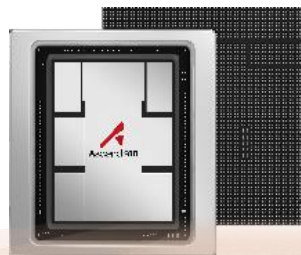
Ascend 910
AI processor

Training card with ultimate computing power	World's most powerful training server	World's fastest AI training cluster

Atlas 300 AI accelerator card
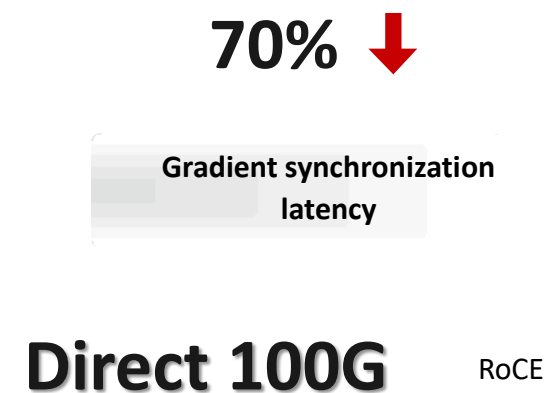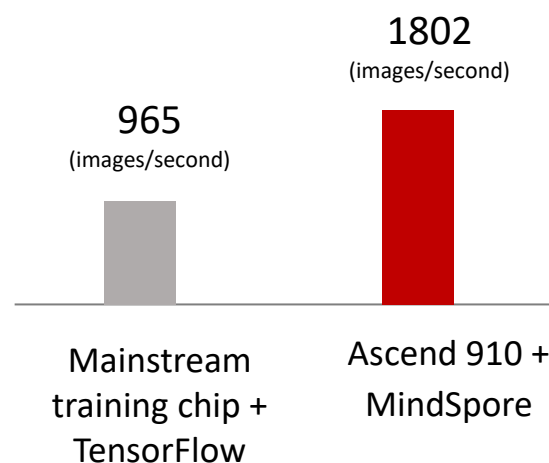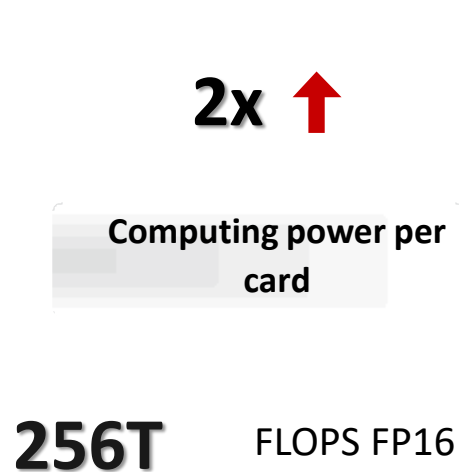Model: 9000

Atlas 800 AI server
Model: 9000/9010

Atlas 900 AI cluster

HUAWEI

Atlas 300 AI Accelerator Card: Highest-Performing Accelerator Card for AI Training

Atlas 300

Model: 9000

2x ⬆

Computing power per card

256T FLOPS FP16

1802
(images/second)

965
(images/second)

Mainstream training chip + TensorFlow

Ascend 910 + MindSpore
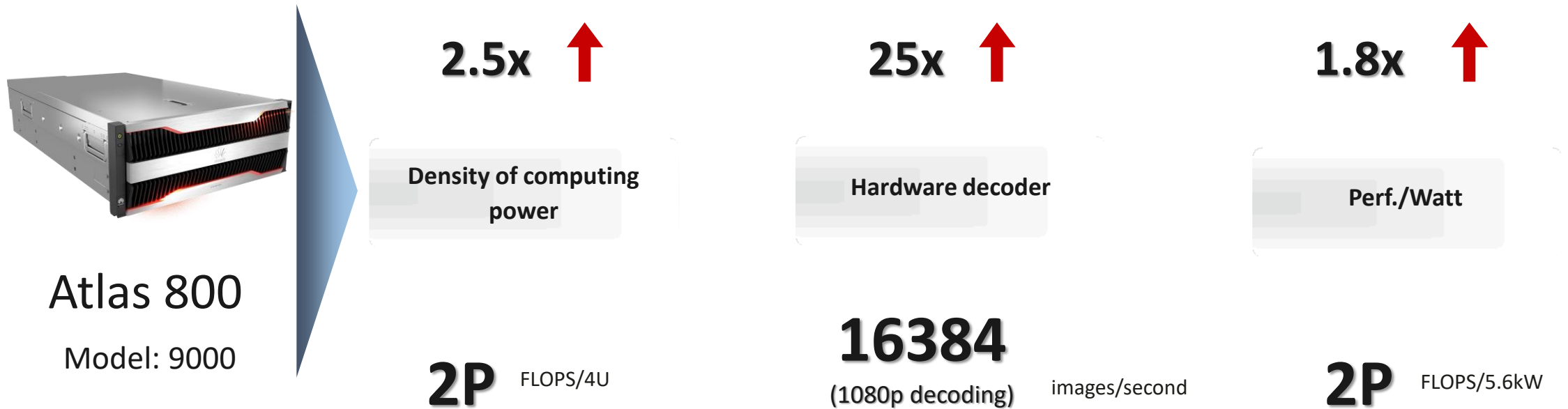
70% ⬇

Gradient synchronization latency

Direct 100G RoCE

- **Test benchmark:**
  - ResNet 50 V1.5
  - ImageNet 2012
  - Optimal batch size respectively

Atlas 800 Training Server: Industry's Most Powerful Server for AI Training

Atlas 800

Model: 9000

2.5x ↑

Density of computing power

2P FLOPS/4U

25x ↑

Hardware decoder

16384

(1080p decoding)   images/second

1.8x ↑

Perf./Watt

2P FLOPS/5.6kW

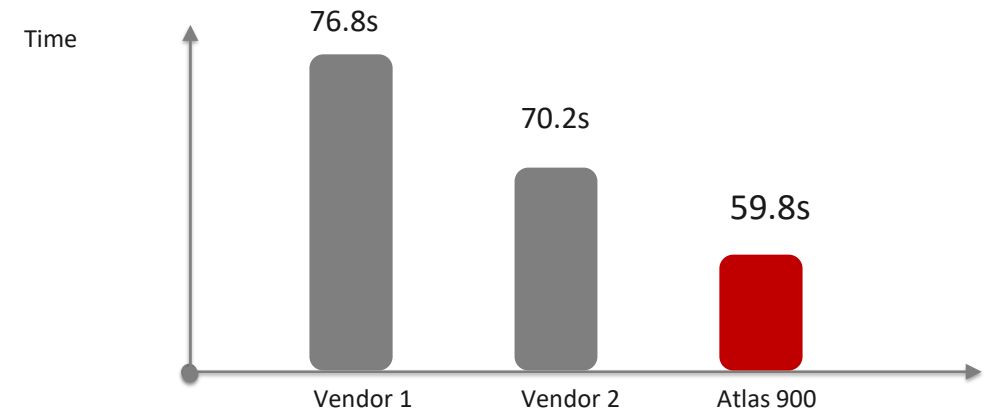# Atlas 900 AI Cluster: Fastest Cluster for AI Training
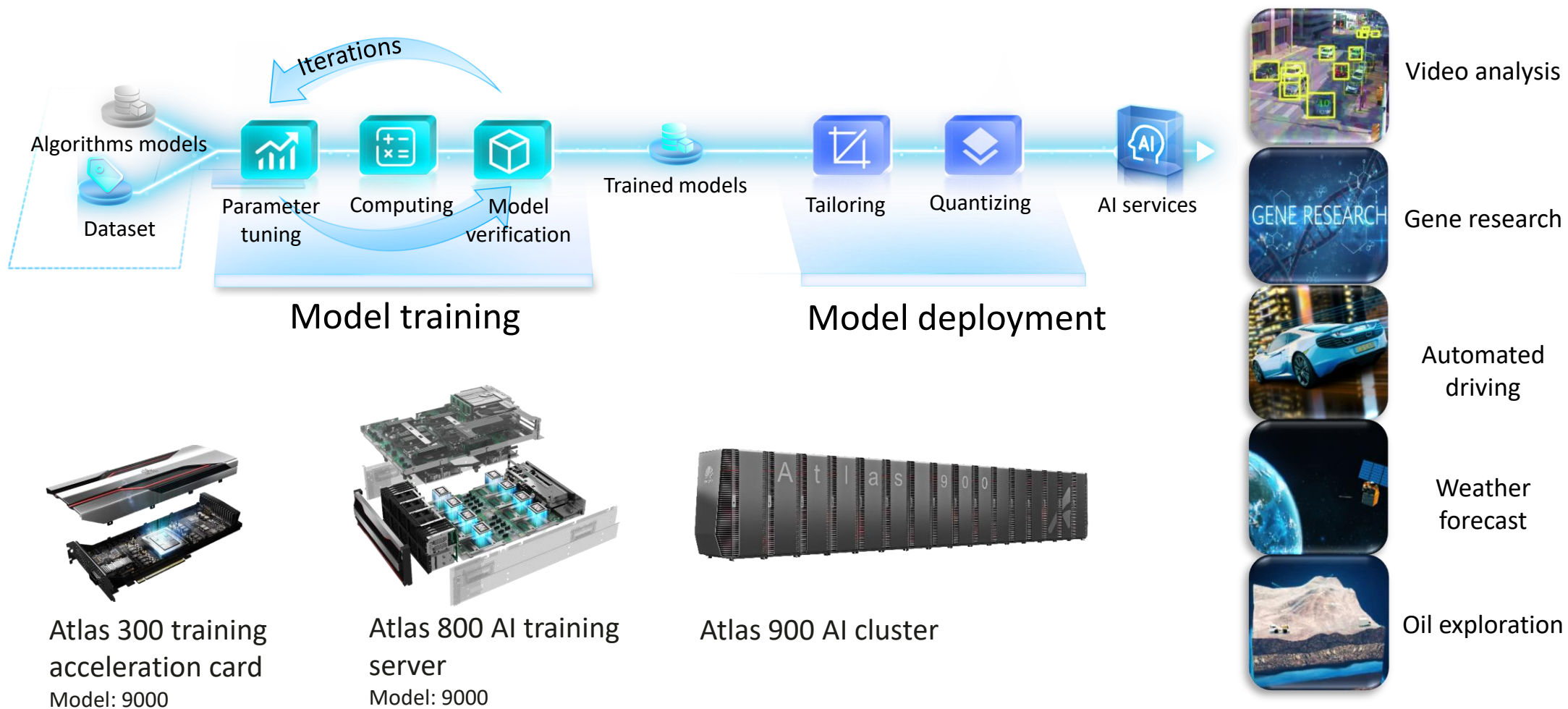

Atlas 900

256-1024 PFLOPS FP16

Industry-leading computing power | Best cluster network | Ultimate heat dissipation
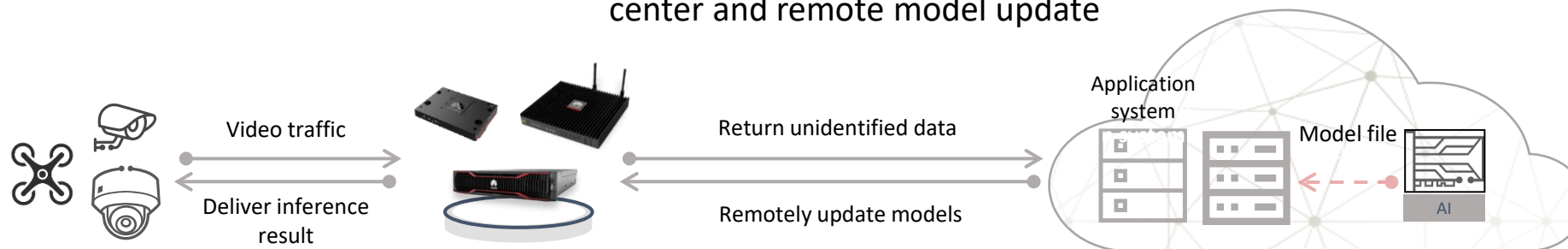
Shortest time consumption: 59.8s

- **Test benchmark:**
  - Benchmark: ResNet-50 V1.5 model, ImageNet-1k dataset
  - Cluster: 1024 Ascend 910 AI processors
  - Accuracy: 75.9%

Atlas Deep Learning System Accelerates AI Model Training and Builds Extensive Applications

Iterations

Algorithms models

Dataset

Parameter tuning

Computing

Model verification

Model training

Trained models

Tailoring

Quantizing

AI services

Model deployment

Atlas 300 training acceleration card
Model: 9000

Atlas 800 AI training server
Model: 9000

Atlas 900 AI cluster

Video analysis

Gene research

Automated driving

Weather forecast

Oil exploration

Device-Edge-Cloud Collaboration Enables the Ultimate Development and User Experience

Device-edge-cloud collaboration for continuous training at data center and remote model update

Video traffic

Deliver inference result

Return unidentified data

Remotely update models

Application system

Model file

AI

Atlas intelligent edge platform
Atlas 200/500/800

O&M management platform
FusionDirect

Atlas deep learning platform
Atlas 300/800

Edge inference

Model management and deployment

Training and inference on cloud

Centralized development ➤➤ Centralized O&M ➤➤ Enhanced security

	Centralized development	Centralized O&M	Enhanced security
Atlas	• Centralized development architecture based on Da Vinci and CANN, **develop once, deploy everywhere**	• FusionDirector manages up to **50,000 nodes**, **manages central and edge devices**, and remotely pushes models and upgrades devices	• Transmission channel encryption • **Model encryption, double assurance**
Industry solution	• Edge and data centers use different development architectures. **Models cannot transfer freely, requiring secondary development.**	• **No O&M management tools**; provides only APIs, so customers need to develop APIs by themselves.	• **No encryption/decryption engine**; models are not encrypted.

HUAWEI

Section Summary

- This section introduces products of Huawei Atlas AI computing platform, mainly inference products including Atlas 200 DK, Atlas 200 AI accelerator module, Atlas 300 AI accelerator card, Atlas 500 AI edge station, and Atlas 800 AI server. The computing platforms used for training include Atlas 300 AI accelerator card, Atlas 800 AI server, and Atlas 900 AI cluster.
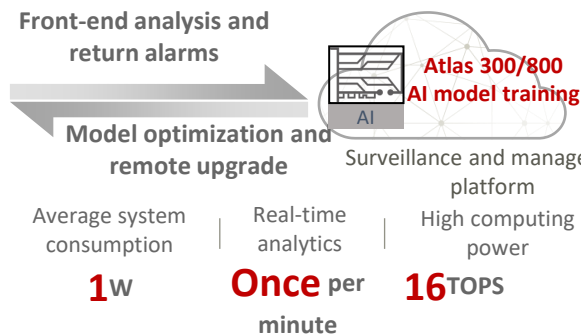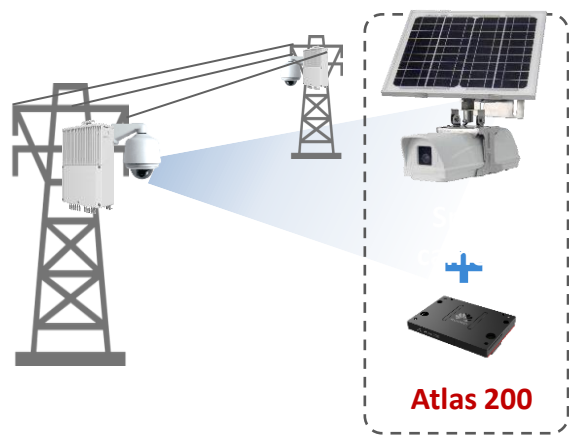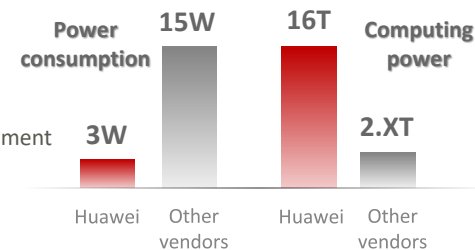
Contents

Overview and Objectives

- This section introduces industry application scenarios of the Atlas AI computing platform.

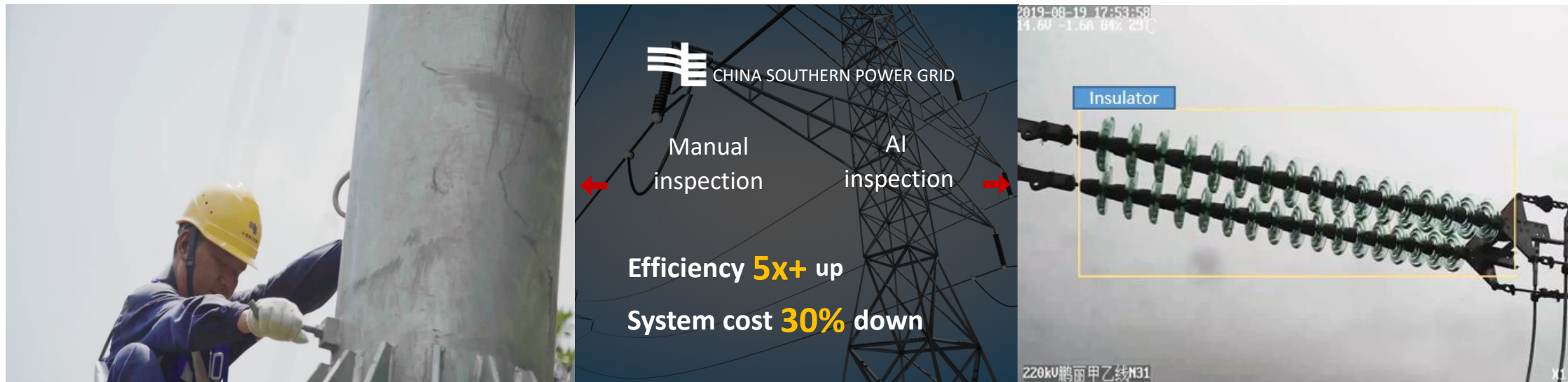# Electric Power: Industry's First Intelligent Unattended Inspection Solution, with 5x Efficiency

Front-end analysis and return alarms

Model optimization and remote upgrade

Atlas 200

+

Atlas 300/800 AI model training

Surveillance and management platform

High computing power

Average system consumption	Real-time analytics	High computing power
1W	**Once** per minute	**16**TOPS

Intelligent analysis at the edge with ultra-low consumption

Power consumption: 15W (Other vendors), 3W (Huawei)

Computing power: 16T (Huawei), 2.XT (Other vendors)

Power consumption **80%↓**

Ideal for edge computing

CHINA SOUTHERN POWER GRID

Manual inspection

AI inspection

Efficiency **5x+** up

System cost **30%** down

Insulator

2019-08-19 17:53:58
14.6V -1.6A 84% 29°C

220kV 鹅丽甲乙线N31

HUAWEI

Finance: AI Enables Smart Branches for Banks

Past: Customer experience needs improvement

Now: cloud-edge collaboration, smart finance

Poor experience
Long queuing time and low efficiency

Poor results
Potential VIP customers missed out

Branch outlet

Atlas 500 AI edge station

HQ

**Atlas 300/800 Training platform**

Precise VIP identification
Potential customer conversion rate 60%

Facial recognition-based authentication
Service processing time 70%

Customer queuing duration analysis
Customer complaints 50%

HUAWEI

Manufacturing: AI Enables Intelligent Production Line with Machine Vision

Cloud-edge collaboration enables intelligent EL detection for PV modules

Atlas 500 AI edge station

HUAWEI CLOUD

AI

Atlas 300/800 Training platform
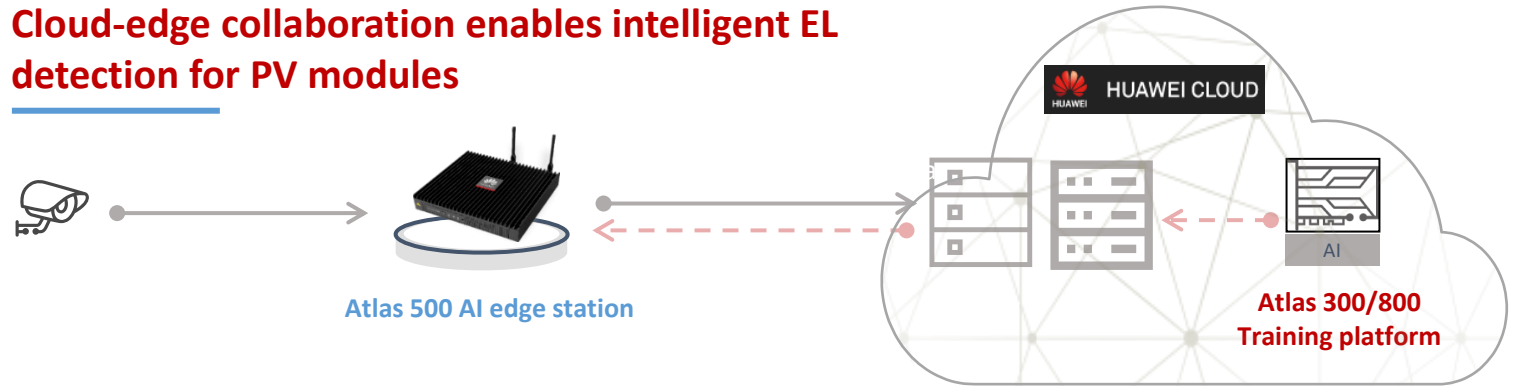
Manual inspection

Unstable results	Low production efficiency	Discontinuous process	High labor costs

Number of defective battery chips	*2*
Detection duration	*about 5s*
Accuracy	*33.33%*

Smart inspection

Zero omission	High production efficiency	Cloud-edge synergy	Reduced labor costs

Defective battery chip	*6*
Detection duration	*1.36s*
Accuracy	*100%*

HUAWEI

Transportation: AI Smooths Highways with 5x Efficiency Boost

Networking center

Service aggregation

Model training and AI algorithm upgrade

Charging & credit Big data platform

Algorithm and application

Rate | Transaction | Audit | ...

Atlas 500 Lane controller

Intelligent License Plate Recognition (LPR)

Camera

RSU antenna

Smart site

ETC gantry system

s3mmo

Deployed 15,000+ units in China

Manual charging
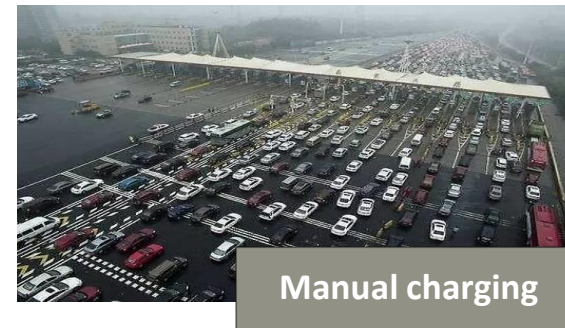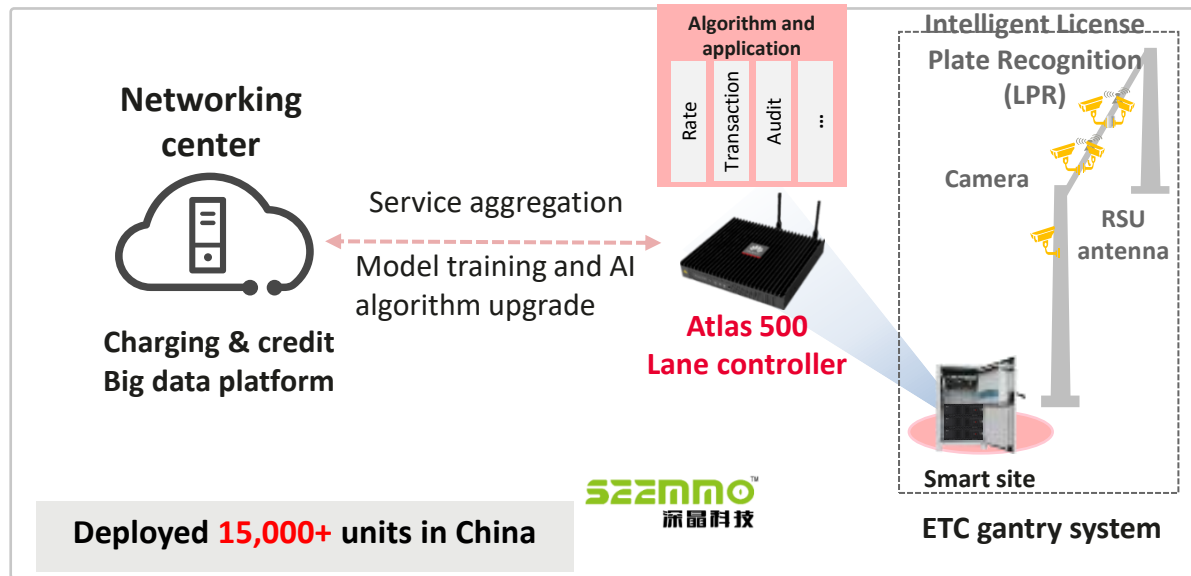
- Low efficiency
- Long queuing time

Free-flow charging

- Passing efficiency **5X** Saving energy and reducing emission

Vehicle-route collaboration
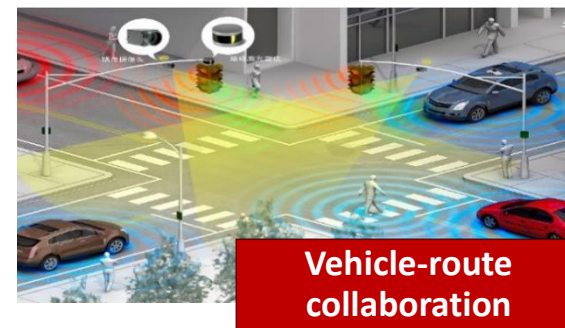
- Proactive security control
- Road cooperation management
- Autonomous vehicle driving

Reliable
Automatic active/standby switchover

Wide temperature range
-40°C to +70°C

Ultimate computing power
16 TOPS INT8

Easy O&M
Unified cloud management

AI-enabled
Lightweight AI inference, real-time computing, and vehicle feature library extraction

AI evolution
Remote algorithm upgrade, continuous evolution for toll audit and vehicle-road cooperation

HUAWEI

Supercomputing: Atlas Helps PCL Build CloudBrain Phase II

Peng Cheng Laboratory (PCL)
Innovative basic platform for national missions

Automated driving   City brain   Smart healthcare   Speech recognition   Natural language processing

Application layer

Basic platform layer   Critical infrastructure for AI

Peng Cheng CloudBrain Phase I   Peng Cheng CloudBrain Phase II

Physical resource layer

Peng Cheng CloudBrain Phase II mainly built
**Atlas 900,
the world's fastest training cluster**

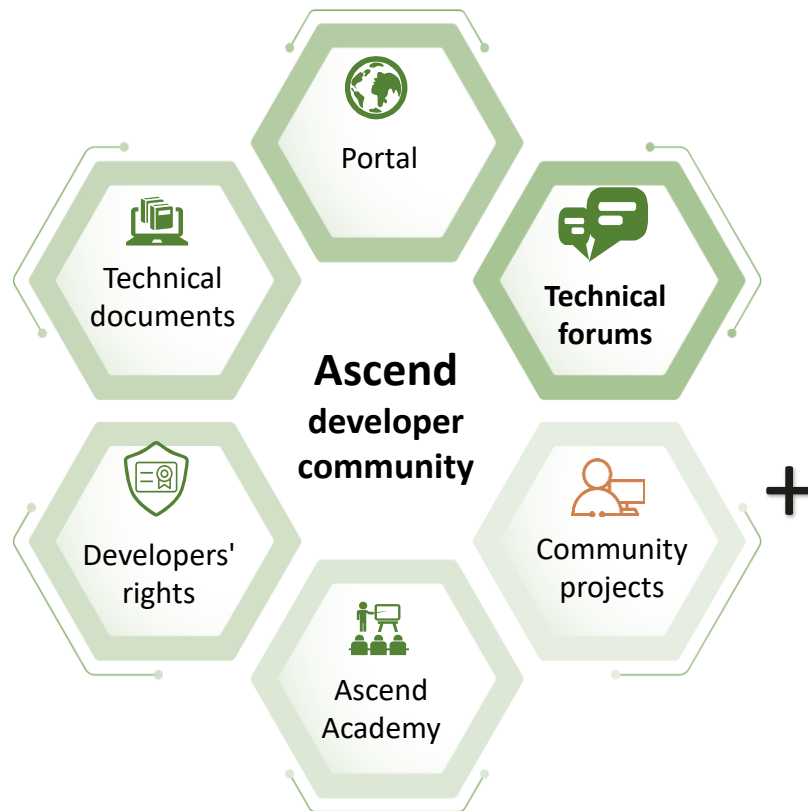**Ultimate computing power**

Level-E AI computing power

Top cluster network
HCCL communication supports 100 TB/s non-blocking parameter plane networking

Ultimate energy efficiency

AI cluster PUE < 1.1

HUAWEI

Attract More Developers Based on the Ascend Developer Community

Ascend developer community

- Portal
- Technical documents
- Technical forums
- Developers' rights
- Ascend Academy
- Community projects

- **Developer-centric enabling platform**
- **https://ascend.huawei.com/home**

+

Annual technical salon
- Held in **10+** cities
- 20+ senior trainers
- Dozens of speeches

Developer contest
- **1,500+** teams
- Annual prize of over RMB1 million
- Equal opportunities for enterprises and universities

Developer support
- Public cloud vouchers
- Free certification course tickets
- Free Atlas developer kits

Summary

- This chapter describes the products of Huawei Atlas computing platform and helps you to understand the working principles of Huawei Ascend chips. It focuses on the hardware and software architectures of Ascend chips and application scenarios of the Atlas AI computing platform.

Quiz

1. What are the main applications of Ascend 310? (   )

   A. Model inference

   B. Model training

   C. Model building

Recommendations

- Ascend community:

    - https://ascend.huawei.com

Huawei Confidential

Thank you.

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。

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

HUAWEI