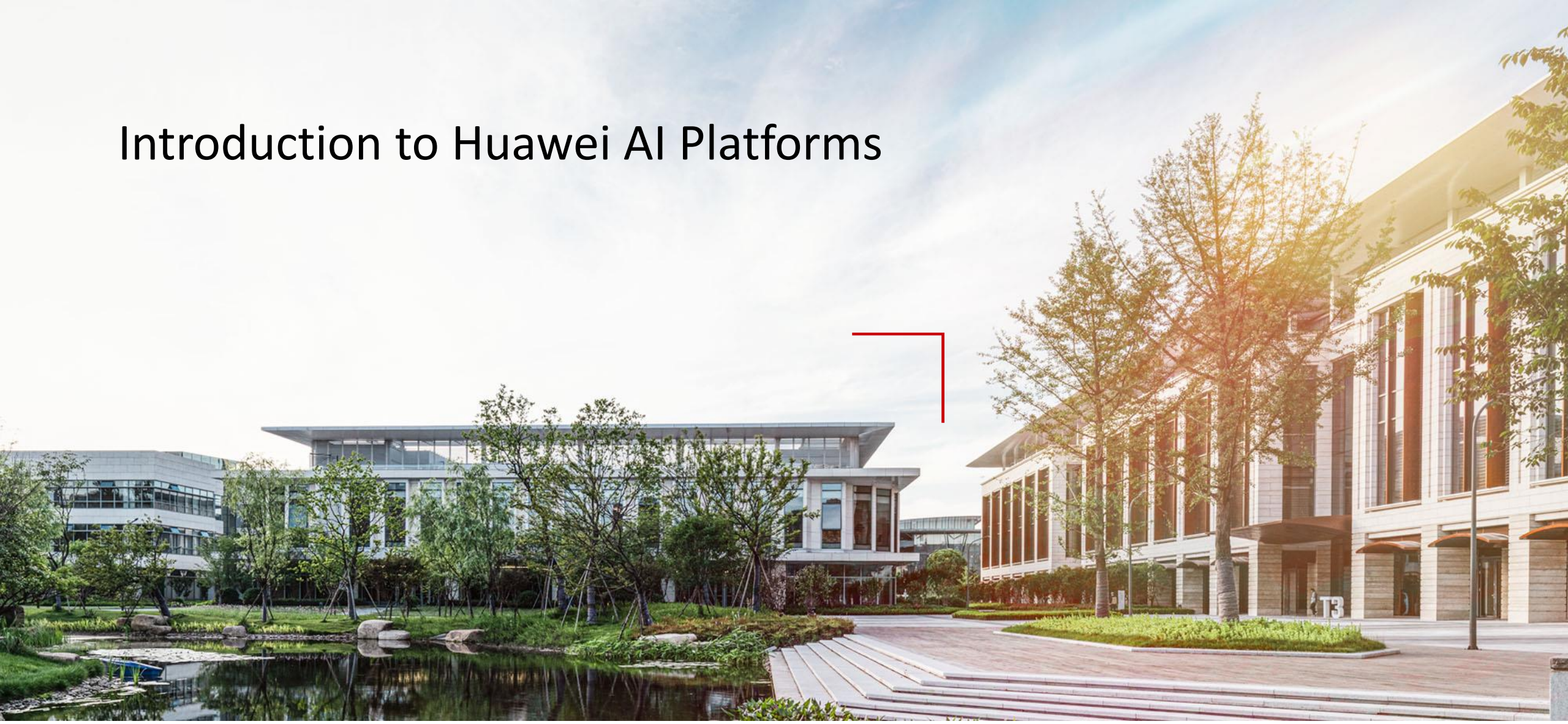
Introduction to Huawei AI Platforms



HUAWEI

Foreword

- This chapter describes Huawei's development platforms in the AI field, including the Ascend computing platform based on Da Vinci architecture, Huawei Cloud EI platform based on cloud services, and device AI platform that provides AI capabilities for devices.

Objectives

Upon completion of this course, you will understand:

- Software and hardware architectures of the Ascend processor.
- Features and application scenarios of the Atlas AI computing platform.
- AI services and application development process of Huawei Cloud EI.
- Features of HarmonyOS, HMS Core, ML Kit, HiAI, and MindSpore Lite.

Contents

Overview and Objectives

- This section describes the concept of AI processor, software and hardware architectures of the Ascend processor, and Atlas series overview and application scenarios.

AI Processor Definition

- More than 99% of operations in AI tasks are matrix operations.
- **AI processor**:
  - AI processor, in a broad sense, is a module dedicated to processing a large number of computing tasks in AI applications. In this sense, all processors oriented to the AI field are AI processors.
  - AI processor, in a narrow sense, is a type of processor specially designed for AI algorithm acceleration, which is also called an AI accelerator.
  - "An AI accelerator is a class of specialized hardware accelerator or computer system designed to accelerate artificial intelligence and machine learning applications, including artificial neural networks and machine vision. Typical applications include algorithms for robotics, internet of things, and other data-intensive or sensor-driven tasks. They are often manycore designs and generally focus on low-precision arithmetic, novel dataflow architectures or in-memory computing capability."

-- Wikipedia

AI Processor Types

- AI processors are classified into training processors and inference processors according to their application scenarios:
  - During training, a simple deep neural network (DNN) model is trained by inputting a large amount of data or using an unsupervised learning method like reinforcement learning. The training process involves massive training data, complex DNN structures, and large computing amounts, posing very high performance requirements on computing capability, accuracy, and scalability of a processor. AI processors used for training can be seen in Huawei Atlas 900 clusters and NVIDIA GPU clusters.
  - Inference is to use a trained model and new data to infer various conclusions. Facial authentication is an example of inference, where the device uses a DNN model to determine whether the face belongs to the device owner. Although the calculation workload of inference is much less than that of training, a large number of matrix operations are still required. CPUs, NPUs, GPUs, and FPGAs can be used during the inference process.

Application Fields of AI Processors

Cloud-based training

- Processor features: high power consumption, high throughput, high accuracy, distributed deployment, scalability, large memory, and high bandwidth
- Application: cloud, HPC, and data centers

Cloud-based inference

- Processor features: low power consumption, high throughput, high accuracy, distributed deployment, scalability, and low latency
- Application: cloud, HPC, and data centers

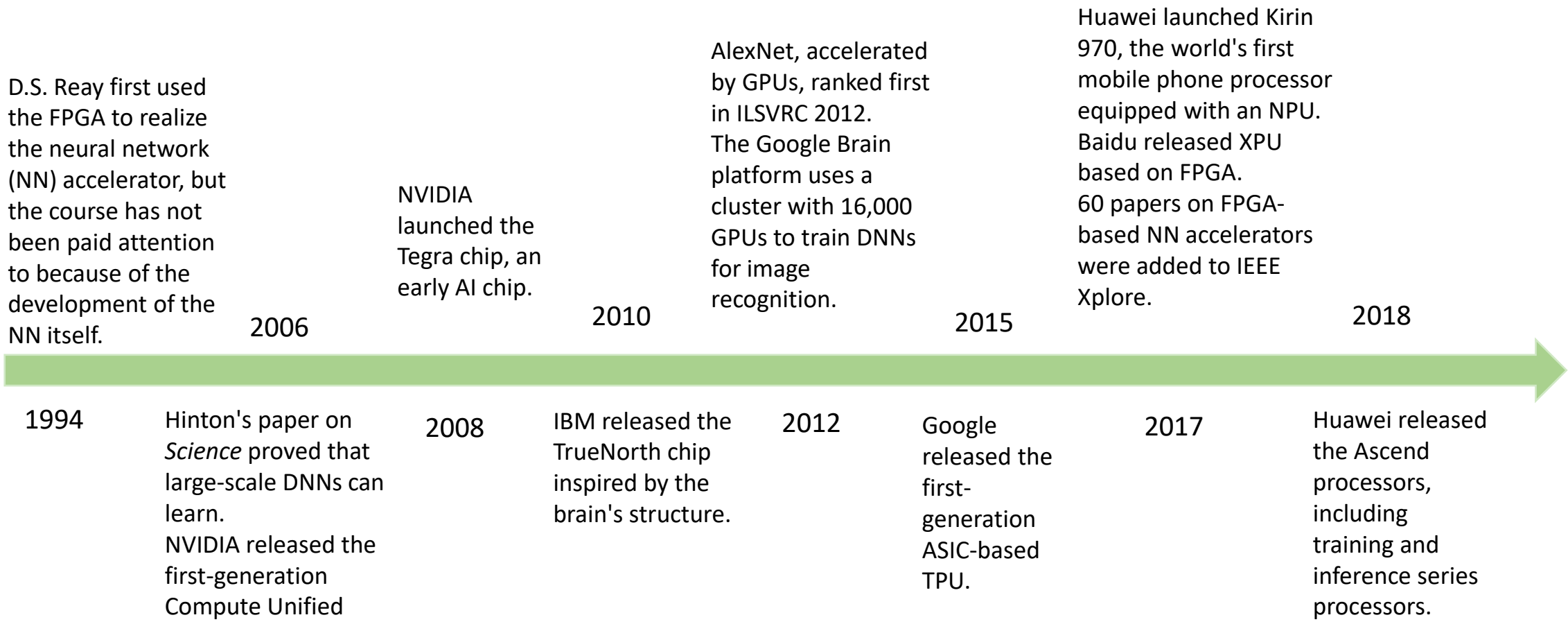
Edge computing

- Processor features: low power consumption, low latency, independent deployment or co-deployment with other devices, virtualization of multiple end users, and small rack space
- Application: smart manufacturing, smart home, smart transportation, and smart finance

End devices

- Processor features: ultra-low power consumption, high energy efficiency, inference orientation, low throughput, low latency, and cost-sensitivity
- Application: consumer electronics in diversified product forms, and IoT

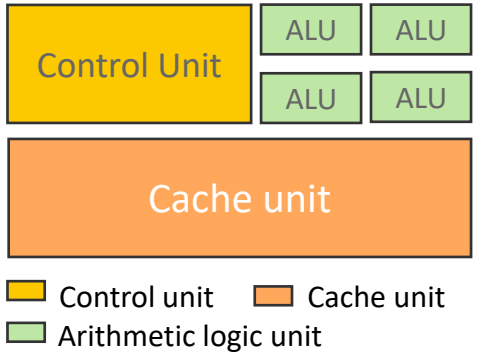
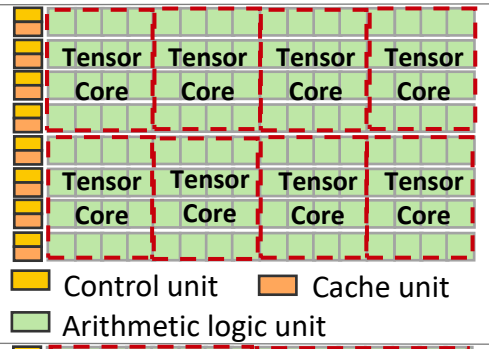
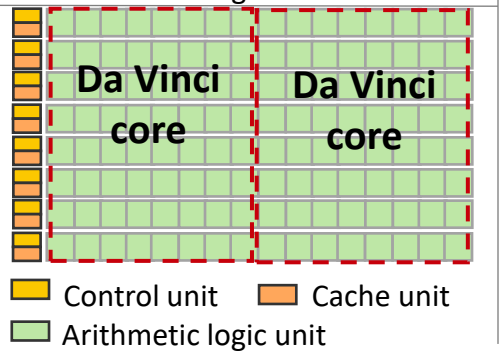
AI Processor Development Process



AI Processor Technology Directions

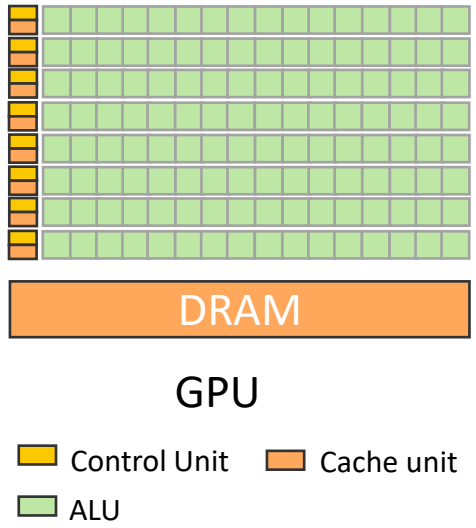
Technical Architecture	Customization Level	Program-mability	Computing Power	Price	Advantages	Disadvantages	Application Scenarios
GPU	General-purpose	Not program-mable	Medium	High	GPUs are universal, sophistically designed and manufactured, and suitable for large-scale parallel computing.	The parallel computing capability cannot be fully utilized at the inference end.	Advanced complex algorithms and universal AI platforms
FPGA	Semi-customized	High	High	Medium	FPGAs can be flexibly configured to adapt to algorithm iterations. The average performance is high, the power consumption is low, and the development period is short.	The unit price for mass production is high, the peak computing capability is low, and hardware programming is difficult.	Various specific industries
ASIC	Fully customized	Low	High	Low	Algorithms of ASICs are fixed to achieve optimal performance and energy efficiency. The average performance is high, the power consumption is low, the size is small. The cost for mass production is low.	High initial investment, long R&D time, and high technical risks.	Special scenarios with dedicated intelligent algorithm software
Brain-inspired processor	Human brain simulation	Not program-mable	High	-	Low power consumption, high communication efficiency, and strong cognitive ability.	It is still in the exploration phase.	Various specific industries

General Computing and AI Computing Build Diversified Computing Together

		Hardware Structure	Computing Features	Application Scenarios
General computing powered by **CPUs**	Kunpeng		**Suitable for complex logical operations, for example, most general-purpose software.** More than 70% transistors are used to build caches and control units. The number of computing cores ranges from several to dozens.	**General applications** Office, database, and numerical calculation (weather forecast, fluid simulation, and electromagnetic simulation).
AI computing powered by **GPUs**	NVIDIA AMD		**Suitable for compute-intensive and high-concurrency tasks with simple logic** More than 70% transistors are used to construct computing units, forming thousands or tens of thousands of computing cores.	**Specific applications** Image recognition: license plate recognition, object recognition, and object detection. Natural language processing (NLP): machine translation and text generation. Language processing: speech recognition and text to speech. Search recommendation, driving assistant, and trend prediction.
AI computing powered by **NPUs**	Ascend			

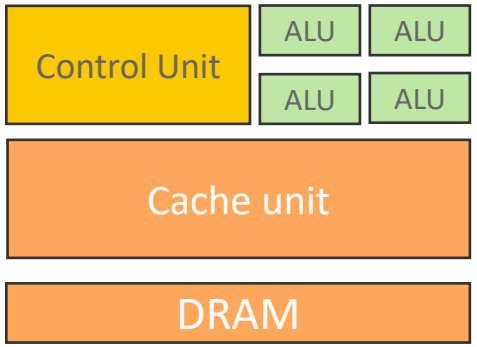
Comparison of CPU and GPU Designs (1)

- GPUs are mainly used for pure computing environments that have unified, large-scale independent data and are not interrupted.
  - Adopt a large-scale parallel computing architecture that consists of thousands of small and efficient cores designed for processing multiple tasks at the same time;
  - Designed for large throughput:
    - A GPU is configured with a large number of ALUs and few caches (serving the threads instead of CPU cores). DRAM accesses are coalesced in the caches, causing latency.
    - Control units coalesce the accesses.
    - The latency problem is masked by a large number of ALUs processing threads in parallel.
  - Good at programs that are compute-intensive and easy to run in parallel.



Comparison of CPU and GPU Designs (2)

- CPUs are used to process different data types in a highly universal manner, and perform logic judgment. In addition, CPUs need to handle a large number of branch directs and interrupts.
  - Consist of several cores optimized for serial processing;
  - Designed for low latency:
    - Powerful ALUs can complete the computation in few clock cycles.
    - A large number of caches reduce the delay.
    - High clock rate.
    - Complex ALUs can reduce the latency of multi-branch programs through branch prediction.
    - For some instructions that depend on previous instruction results, ALUs determine the positions of instructions in the pipeline to implement fast data forwarding.
  - Good at logic control and serial operations.



CPU

Evolution of CPU and GPU Designs

- CPU
  - General performance is improved by increasing the number and frequency of cores.
  - AI performance is improved by adding instructions (modifying the architecture):
    - Intel adds the AVX-512 FMA instruction sets to the CISC architecture.
    - ARM adds the Cortex A instruction set to the RISC architecture, which is planned to be continuously upgraded.
  - Intel plans to add vision processing units (VPUs) to CPUs.
- GPU
  - Dedicated AI computing units (Tensor Cores and Matrix Cores) are added.

Comparison of FPGA and ASIC

	FPGA	ASIC
Computing speed	Low due to inevitable redundancy caused by universality of the architecture and latency between different structures.	High because there is no special requirement on the architecture and specific modules can be placed close to each other to reduce latency.
Processor size	Large (when the functions are the same)	Small (when the functions are the same)
Power consumption	High (under the same process conditions)	Low (under the same process conditions)
Cost	The R&D risk and cost is low. The cost is mainly attributed to production.	The tool development and tape-out processes may incur significant cost because the hardware must be determined before production.
Running process	It takes time to load the configuration to the storage device.	Programs can run immediately.
Product positioning	Suitable for products that require rapid market occupation or flexible feature design.	Suitable for products with large-scale design or mature technologies, such as consumer electronics.
Development direction	Large capacity, low voltage, low power consumption, and SoC.	Larger scale, intellectual property reuse, and SoC.

FPGA and ASIC Processor Evolution

- FPGA
  - Xilinx adds the hardware/software programmable engine with many AI cores to the processor.
  - Intel upgrades the DSP module in the traditional FPGA.
- ASIC
  - Companies are deploying ASICs, including TPUs, NPUs, and VPUs.

Da Vinci Architecture: Born for Ultimate Efficiency of AI Computing

The proportion of Tensor Cores is low because the GPU needs to support both image rendering and AI computing.

First-generation Tensor Core	V100 – Volta architecture (2017) 4x4x4 64 operations in a clock cycle
Third-generation Tensor Core	A100 – Ampere architecture (2020) 8x8x4 256 operations in a clock cycle
Fourth-generation Tensor Core	H100 – Hopper architecture (2022) 16x8x4 512 operations in a clock cycle

Ascend NPU is specially designed for AI computing with a high proportion of Cube units.

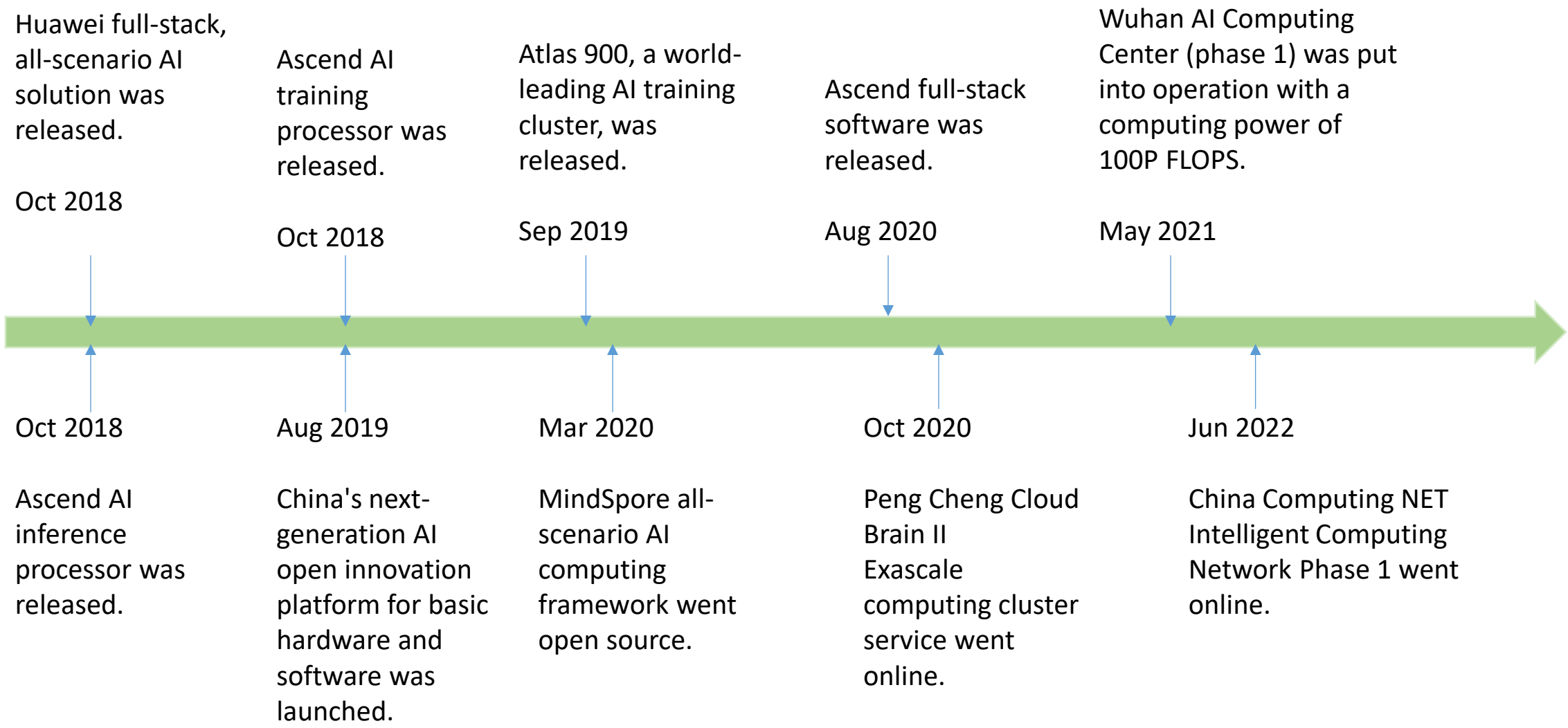
Ascend NPU – Da Vinci architecture (2018)

The proportion of Cube computing units is about 99.2%.

Cube=16x16xN
N=1/2/4/8/16

A maximum of 4096 operations in a clock cycle

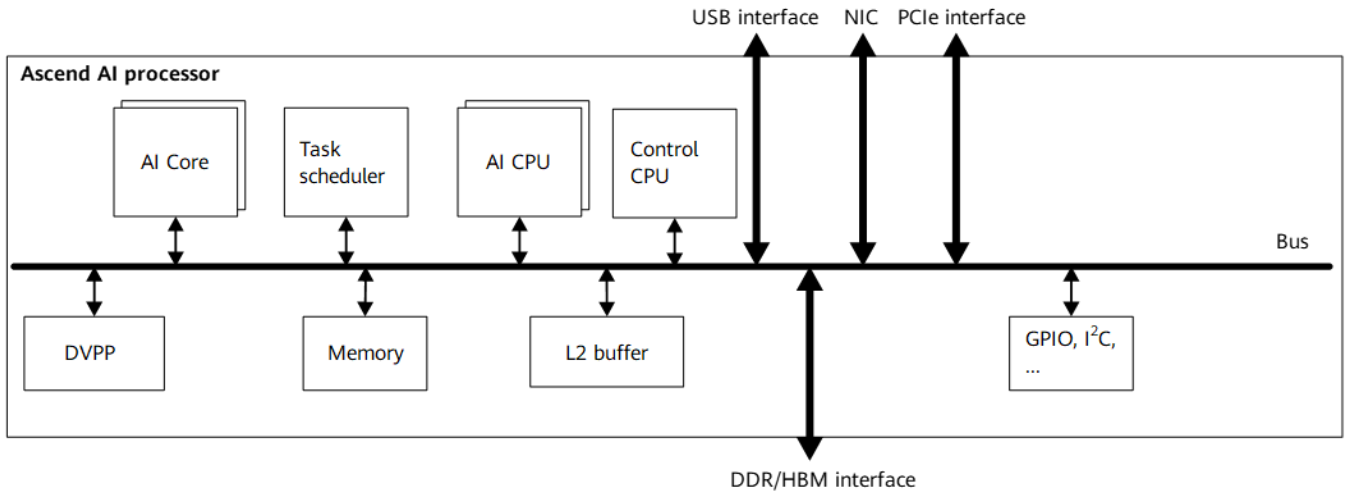
Huawei's Investment in the AI Industry



Contents

Logical Architecture of the Ascend AI Processor

- The Ascend AI processor consists of:
  - Processor system control CPU (control CPU)
  - AI computing engine (AI Cores and AI CPUs)
  - Multi-layer system-on-chip (SoC) cache or buffer
  - Digital vision pre-processing (DVPP) module

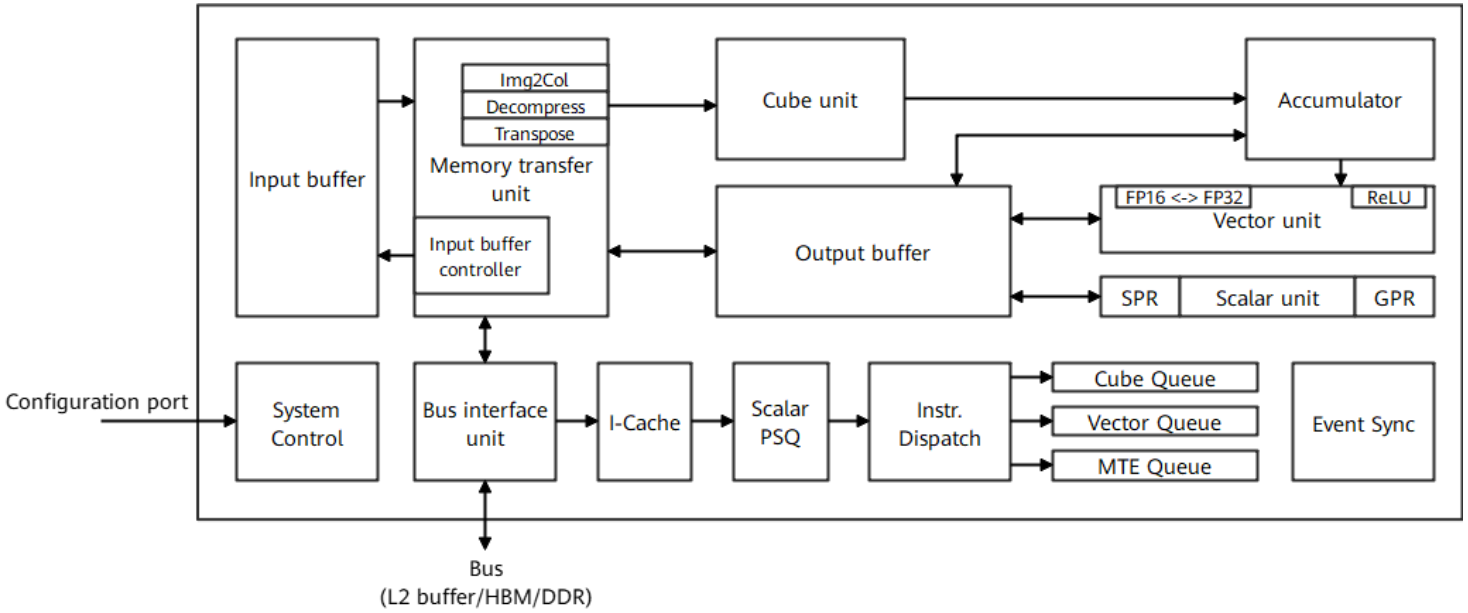


Ascend AI Computing Engine - Da Vinci Architecture

- One of the four major structures of the Ascend AI processor is the AI computing engine, which consists of AI Cores (Da Vinci architecture) and AI CPUs. Da Vinci architecture, an architecture dedicated to improving AI computing power, is the core of the Ascend AI computing engine and AI processor.

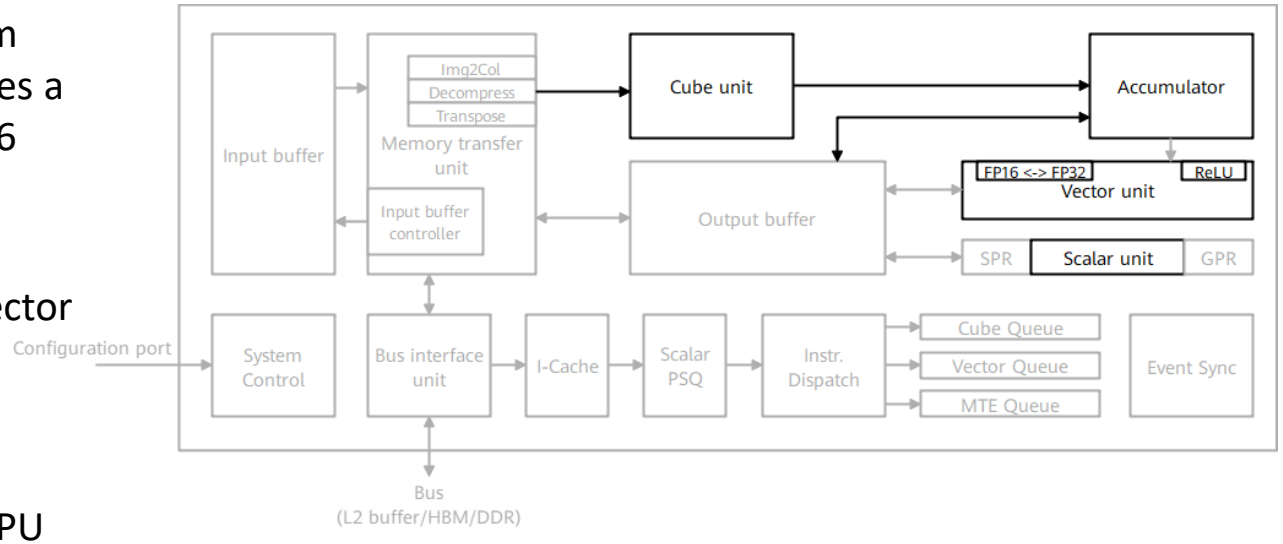
Da Vinci Architecture (AI Core)

- Main components of the Da Vinci architecture:
  - Compute units: Cube unit, Vector unit, and Scalar unit
  - Storage system: on-chip storage units of AI Core and the data paths
  - Control units: the brain of AI Core, responsible for AI Core runtime control with instructions.



Da Vinci Architecture (AI Core) — Computing Unit

- The three basic compute resources, Cube, Vector, and Scalar Units, perform computations related to matrices, vectors, and scalars, respectively.
  - The Cube unit works with the accumulator to perform matrix-related operations. Per clock cycle, it completes a 16 x 16 matrix and 16 x 16 matrix multiplication (4096 ops) at FP16 or a 16 x 32 matrix and 32 x 16 matrix multiplication (8192 ops) at INT8.
  - The Vector unit implements computing between a vector and a scalar or between two vectors, supporting precisions of FP16, FP32, INT32, INT8, and more customized types.
  - The Scalar unit controls the program flow as a mini CPU via iteration control, selection judgment, address and parameter computations of Cube/Vector instructions, and basic arithmetic operations.



Da Vinci Architecture (AI Core) — Storage System (1)

- Storage units and corresponding data paths form the storage system of AI Core.
- Storage units consist of the storage control unit, buffers, and registers:

▫ The storage control unit accesses lower level caches outside AI Core through the bus, and directly accesses DDR or HBM memory. A storage conversion unit is introduced as a transfer controller for the internal data paths in AI Core to implement read/write management of internal data of AI Core between different buffers, and for format conversion, such as padding, Img2Col, transpose, and decompress.

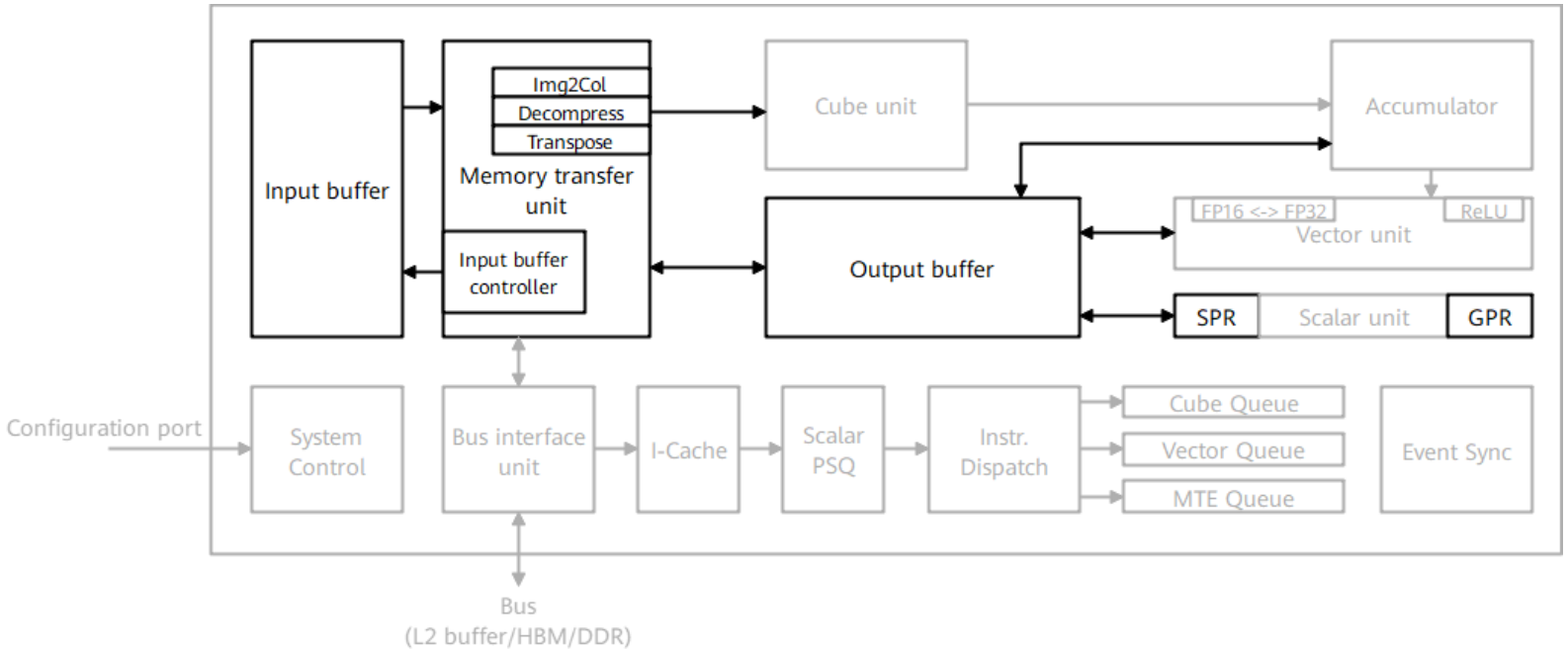
▫ The input buffer temporarily stores frequently-used data to reduce round-trip to memory outside AI Core, which decreases data accesses over the bus and avoids bus congestion, achieving improved performance with lower power consumption.

▫ The output buffer stores the intermediate results at each layer in NNs to facilitate data transfer to the next layer. Unlike the bus offering low bandwidth and severe latency, the output buffer greatly speeds up computation.

▫ Registers in AI Core are mainly used by the Scalar unit.

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

- Data paths: paths for data movement in AI Core during computations
  - Multiple-input single-output (parallel input) characterizes data paths in the Da Vinci architecture, in view of the diversity and large quantity of input data in NN computation. On the contrary, usually, only a feature matrix is output. A single output in data paths is enough, which frees up certain processor hardware resources.

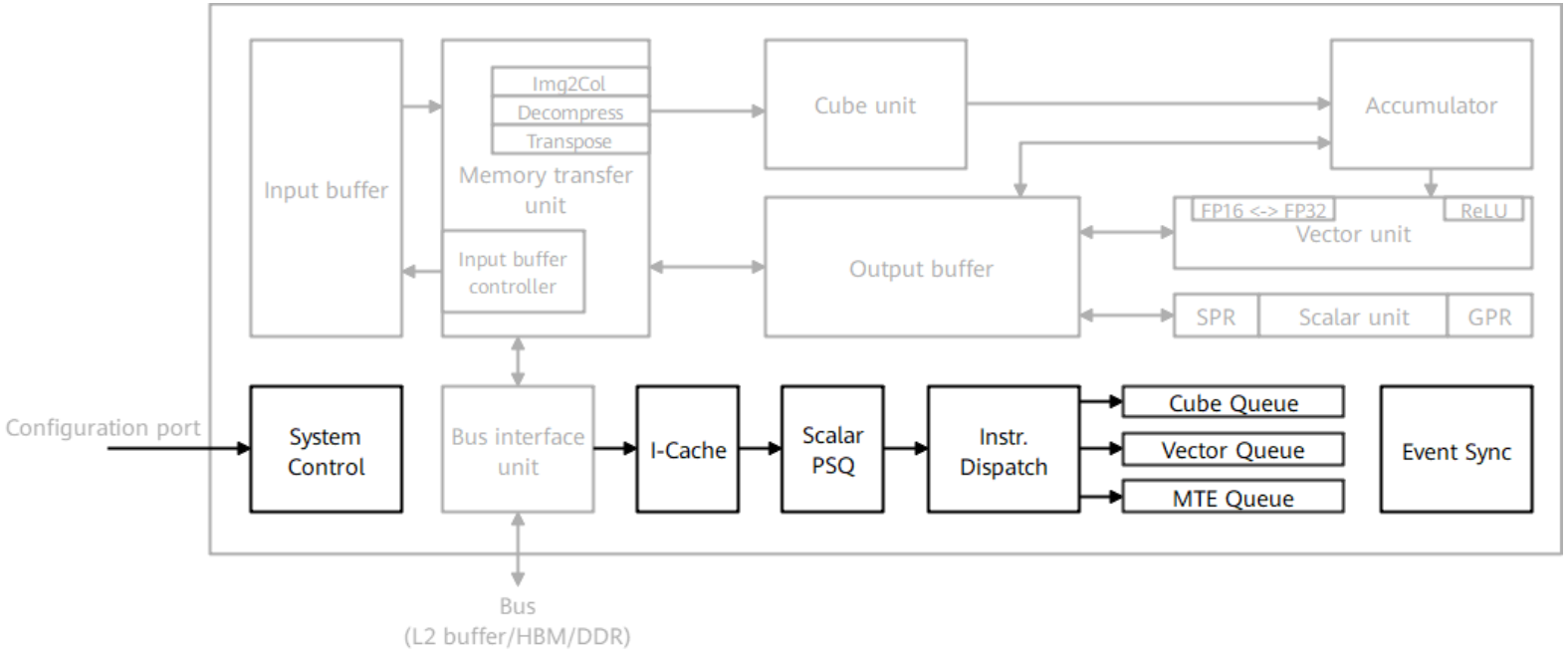


Da Vinci Architecture (AI Core) - Control Unit (1)

- Control units include System Control, I-Cache, Scalar PSQ, Instr. Dispatch, Cube Queue, Vector Queue, MTE Queue, and Event Sync.
  - System Control controls the execution process of the task block (the minimum computing task granularity of AI Core), and reports execution status through interrupts. If an execution error occurs, it reports the error to the task scheduler.
  - I-Cache is an instruction cache. During instruction execution, it prefetches and reads subsequent instructions at a time, accelerating instruction execution.
  - Scalar PSQ organizes decoded instructions, including those for matrix computation, vector computation, and storage conversion.
  - Instr. Dispatch reads the configured instruction addresses and decoded parameters in the Scalar PSQ, and sends instructions to corresponding execution queues by type. Scalar instructions reside in the Scalar PSQ.

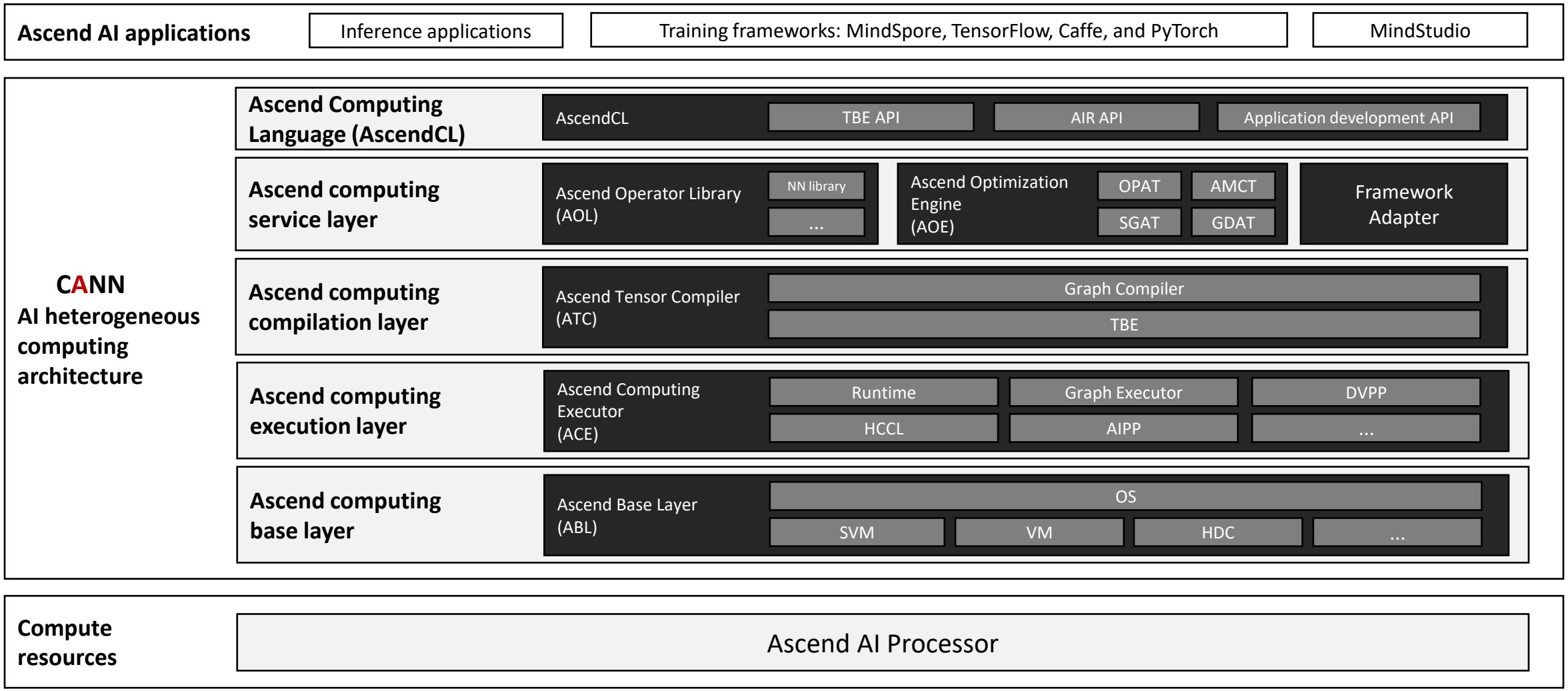
Da Vinci Architecture (AI Core) - Control Unit (2)

- Instruction execution queues include Cube Queue, Vector Queue, and MTE Queue. Different instructions go to corresponding queues and are executed in sequence.
- Event Sync controls the execution of each instruction pipeline in real time and analyzes the dependency between pipelines in consideration of data dependency and synchronization between instruction pipelines.



Contents

CANN AI Heterogeneous Computing Architecture (1)



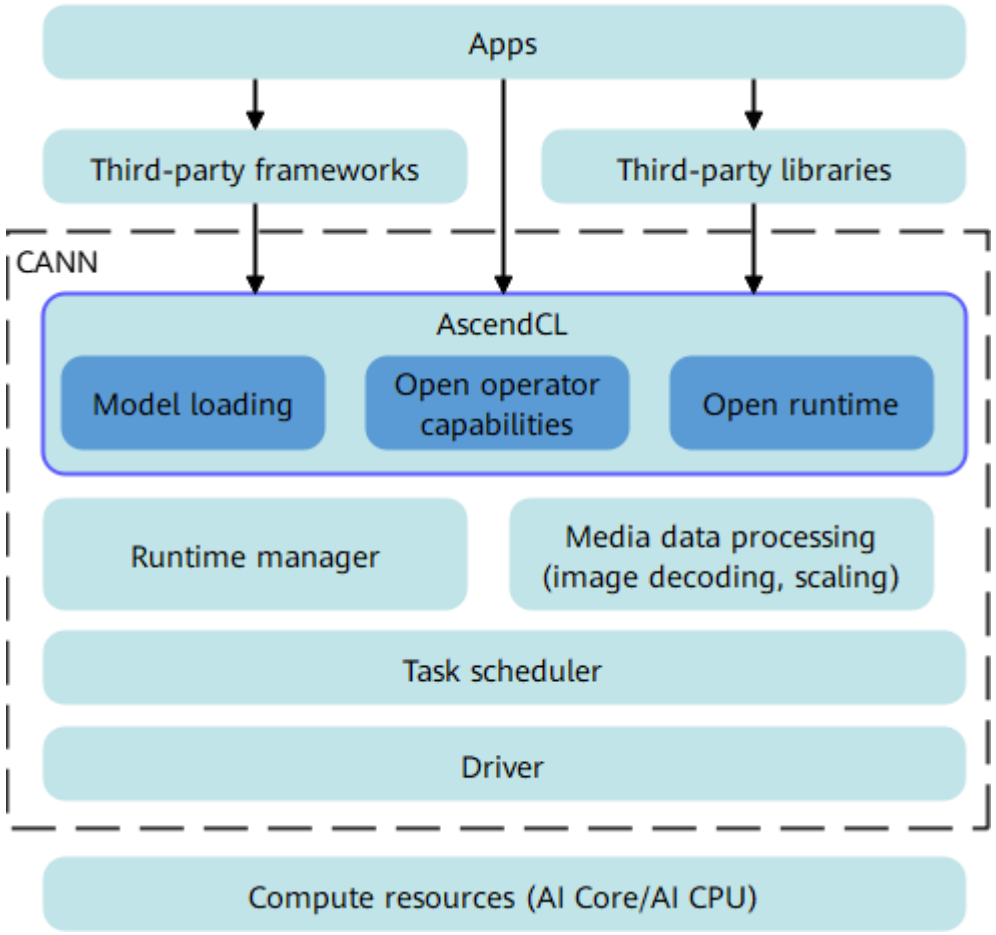
CANN AI Heterogeneous Computing Architecture (2)

- Ascend AI Applications
  - Ascend AI applications include inference applications, framework applications (for training), and MindStudio IDE.
- Ascend Computing Language
  - The Ascend computing language (AscendCL) API is an open programming framework for Ascend computing. It shields the differences between underlying processors and provides TBE operator development API, AIR standard graph development API, and application development API, allowing users to quickly build Ascend-based AI applications and services.
- Ascend Computing Service Layer
  - The Ascend computing service layer provides AOL and accelerates computing by using high-performance operators of the NN library and Basic Linear Algebra Subprograms (BLAS) library. The service layer also provides AOE to improve end-to-end model running speed through OPAT operator tuning, SGAT subgraph tuning, GDAT gradient tuning, and AMCT model compression. In addition, framework adapters are provided to adapt mainstream AI frameworks such as TensorFlow and PyTorch.

CANN AI Heterogeneous Computing Architecture (3)

- Ascend computing compilation layer
  - Ascend computing compilation layer uses Graph Compiler to compile the computational graph of the intermediate representation (IR) input by the user into a model executable by Ascend hardware. In addition, the automatic scheduling mechanism of tensor boost engine (TBE) is used to efficiently compile operators.
- Ascend computing execution layer
  - Ascend computing execution layer executes models and operators. The functional modules include the Runtime library, Graph Executor, DVPP, AI pre-processing (AIPP), and Huawei Collective Communication Library (HCCL).
- Ascend computing base layer
  - Ascend computing base layer provides base services for upper layers through shared virtual memory (SVM), virtual machines (VMs), and host-device communication (HDC).
- Compute resources
  - Compute resources, as the hardware computing power basis of the Ascend AI processor, execute specific computing tasks.

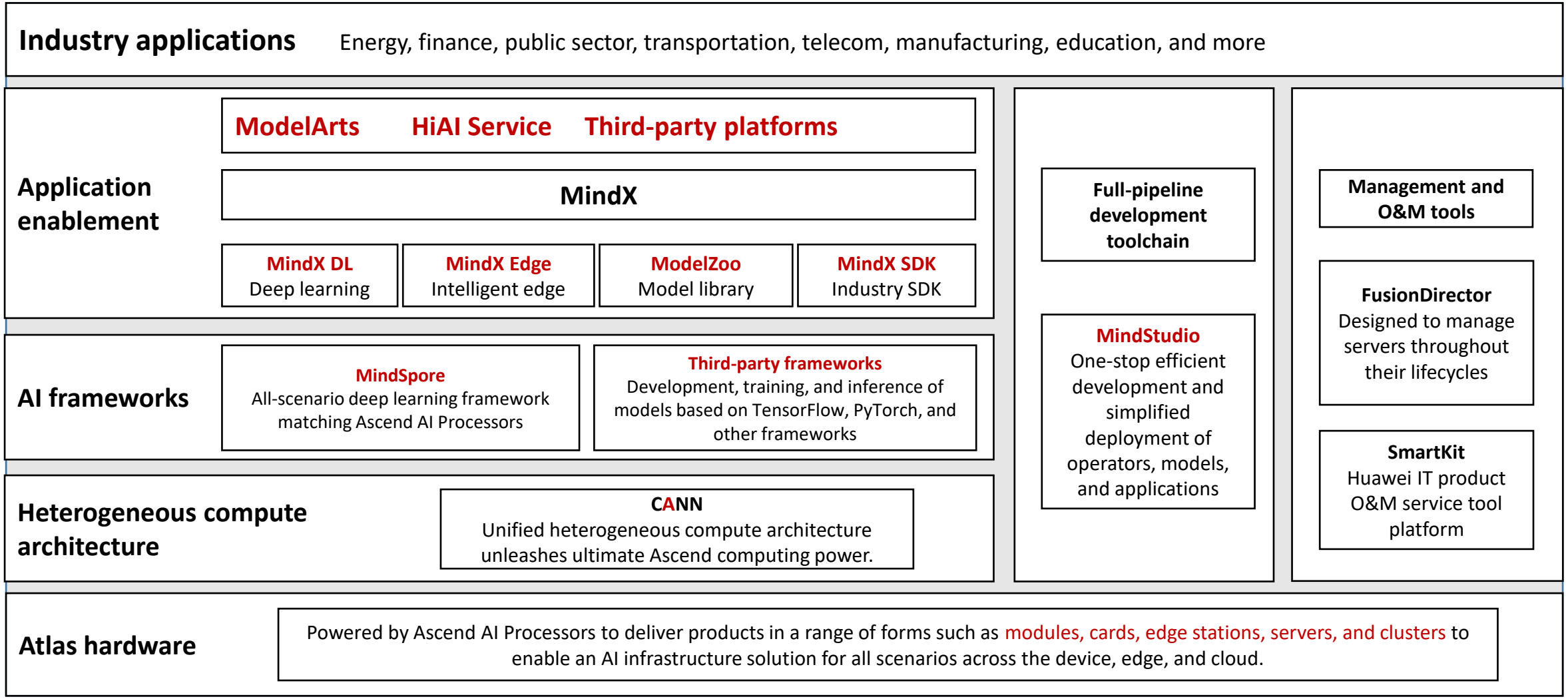
NN Software Flow of the Ascend AI Processor



- The NN software flow of the Ascend AI processor is a bridge between the deep learning framework and the Ascend AI processor. It implements and executes NN applications and integrates the following functional modules:
- DVPP: performs data processing and cleaning before input to meet format requirements for computing.
- Open operator capabilities (TBE): continuously provides powerful computing operators for NN models.
- Open runtime (runtime manager): provides various resource management paths for task delivery and allocation of the NNs.
- Task scheduler: provides specific target tasks for the Ascend AI processor as a task driver for hardware execution. The runtime manager and task scheduler work together to form a dam system for NN task flow to hardware resources, and monitor and distribute different types of tasks for execution in real time.

Contents

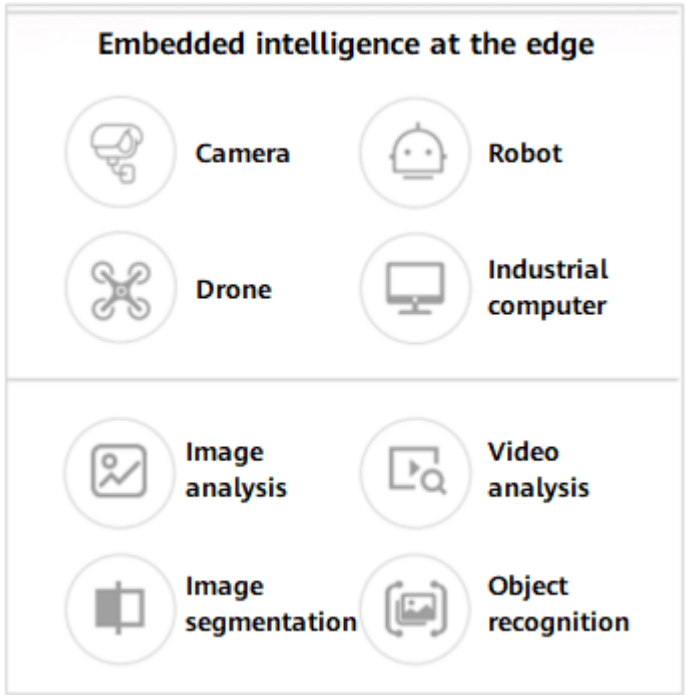
Ascend Full-Stack AI Software and Hardware Platform: Cornerstone of an Intelligent World



HUAWEI

Atlas 200 AI Accelerator Module

- The **Atlas 200 AI Accelerator Module 3000** is widely used in device-side AI scenarios such as smart cameras, robots, and drones to implement object recognition and image classification.



Optimal Performance

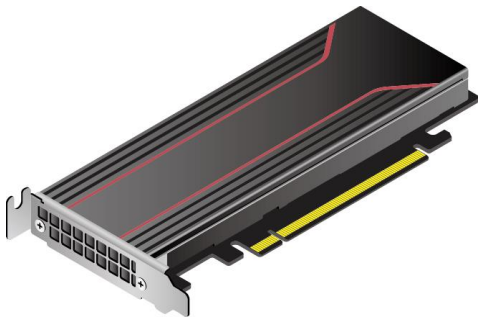
• Supporting real-time analysis of 20 HD video channels of 1080p at 25 fps

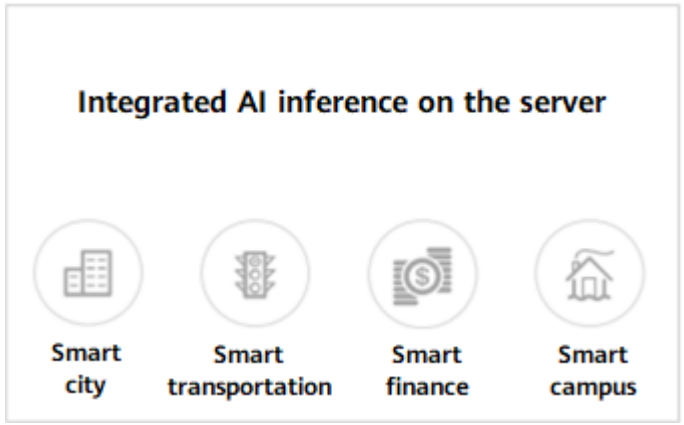
Ultra-low power consumption

• Milliwatt-level hibernation power consumption and millisecond-level wakeup enable edge AI applications with typical power consumption of 6.5 W.

Atlas 300V Pro Video Analysis Card

- The **Atlas 300V Pro video analysis card** integrates the general-purpose processor, AI Core, and codec to provide powerful AI inference and video and image encoding and decoding functions. With advantages such as numerous video analysis channels, high-performance feature retrieval, and secure boot, it supports real-time analysis of 128-channel HD videos and can be used in a wide array of AI application scenarios, such as smart city, smart transportation, and smart campus.





Analysis of numerous video channels

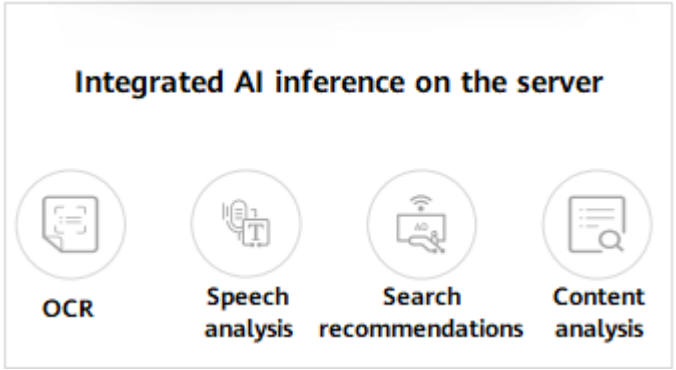
- Real-time analysis of 128 channels of HD videos
- JPEG and video hardware encoding/decoding, boosting performance of image and video applications

Secure boot

- Complete device boot chain and determined initial boot status of the device, preventing backdoor implantation

Atlas 300I Pro Inference Card

- The **Atlas 300I Pro inference card** integrates the general-purpose processor, AI Core, and codec, to provide powerful AI inference and object search functions. With advantages such as powerful computing power, ultra-high energy efficiency, high-performance feature retrieval, and secure boot, it can be widely used in a wide array of AI application scenarios, such as OCR, speech analysis, search recommendation, and content moderation.



Superior computing power

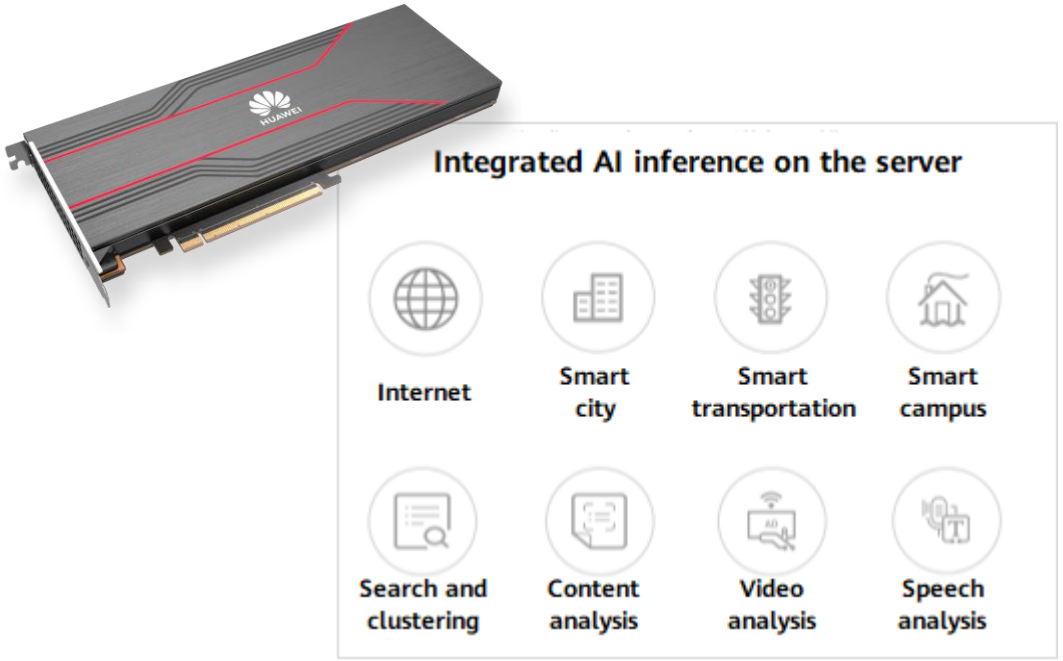
Ultra-high energy efficiency

Secure boot

• Complete device boot chain and determined initial boot status of the device, preventing backdoor implantation

Atlas 300I Duo Inference Card

- The **Atlas 300I Duo inference card** integrates the general-purpose processor, AI Core, and codec to provide powerful AI inference and video analysis functions. With advantages such as superior computing power, ultra-high energy efficiency, and high-performance video analysis, it is suitable for a wide array of scenarios such as Internet, smart city, and smart transportation, and supports multiple applications such as search clustering, content analysis, OCR, speech analysis, and video analysis.



Superior computing power

Ultra-high energy efficiency

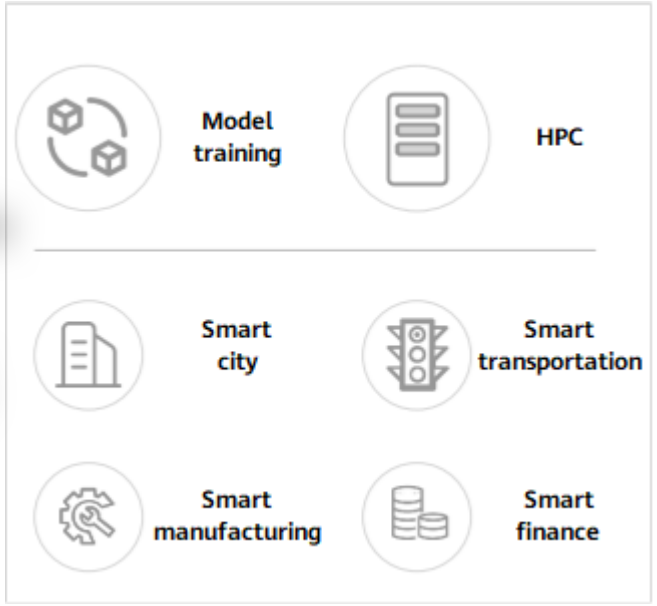
High-performance video analysis

- Real-time analysis of 256 channels of HD videos
- JPEG and video hardware encoding/decoding, boosting performance of image and video applications

HUAWEI

Atlas 300T Pro Training Card

- The **Atlas 300T training card** is an AI accelerator card that works with servers to provide powerful computing power for data centers. The Atlas 300T Pro features superior computing power, high integration, and high bandwidth to meet the requirements for AI training of Internet, carrier, and finance industries and computing power of high-performance computing.



Superior computing power

• 30 built-in Da Vinci AI Cores

High integration

• Three-in-one (AI computing power, general-purpose computing power, and I/O capabilities)

• Integration of 32 Huawei Da Vinci AI Cores,16 TaiShan cores, and one 100GE RoCE v2 NIC

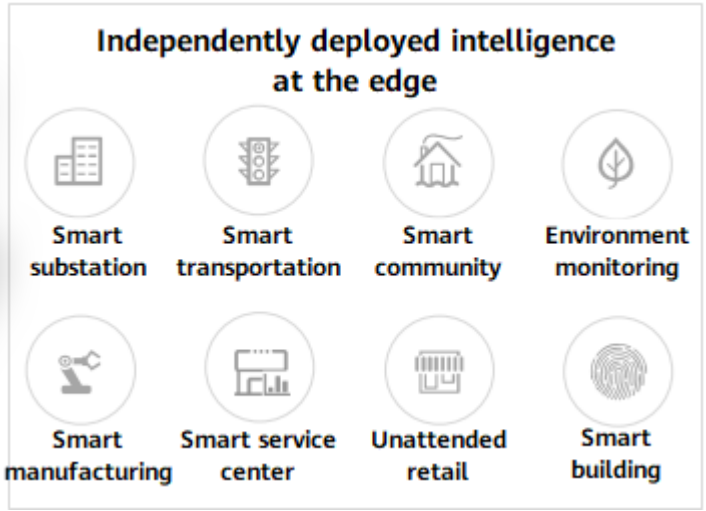
High bandwidth

• Supports PCIe 4.0 and 1 x 100G RoCE high-speed ports, achieving the total egress bandwidth of 56.5 Gbit/s

• Improves the training data and gradient synchronization efficiency by 10% to 70% without the need of external NICs

Atlas 500 AI Edge Station

- Oriented to edge applications, the **Atlas 500 AI edge station** (model 3000) features powerful computing performance, compact size, strong environment adaptability, easy maintenance, and support for cloud-edge collaboration. It can be widely deployed in edge scenarios to meet application requirements in complex environments such as security protection, transportation, communities, campuses, shopping malls, and supermarkets.



Intelligent edge

- Industry-leading edge product that integrates AI processing capabilities
- Operates from -40°C to 70°C outdoors with the fan-free design

Compact and powerful

- Processing of 20 channels of HD videos of 1080p at 25 fps

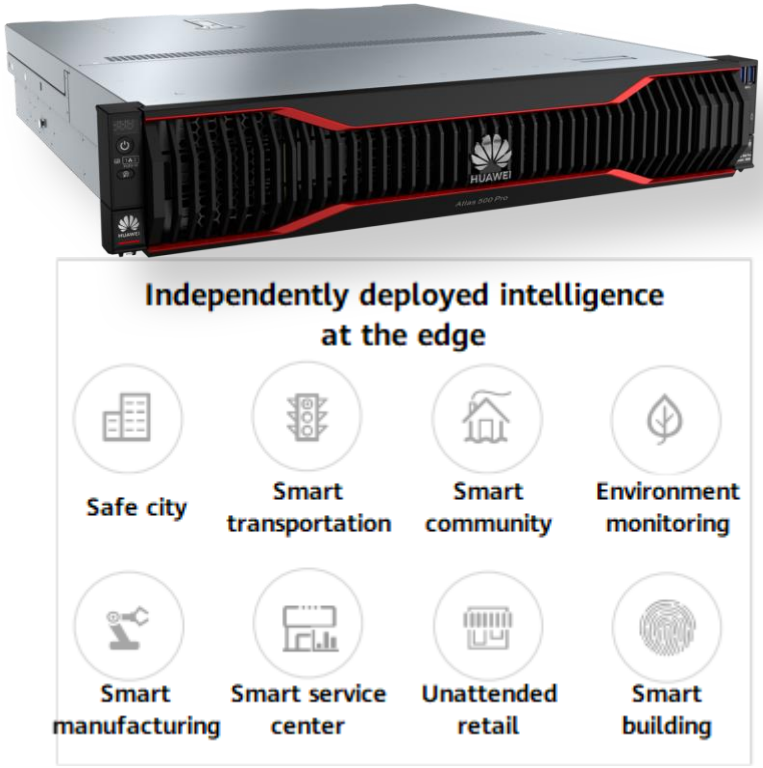
Edge-cloud synergy

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

HUAWEI

Atlas 500 Pro AI Edge Server

- The **Atlas 500 Pro AI edge server** (model 3000) is designed for edge applications. It features powerful computing performance, high environment adaptability, easy deployment and maintenance, and support for cloud-edge collaboration. It can be widely deployed in edge scenarios to meet application requirements in complex environments such as security protection, transportation, communities, campuses, shopping malls, and supermarkets.



Superior computing power

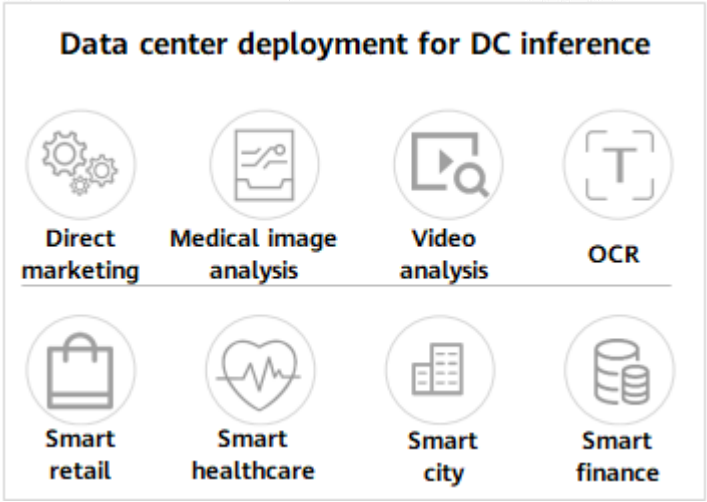
- Supports up to three Atlas 300 V/VI inference cards to meet inference requirements in multiple scenarios
- Supports real-time analysis of 384 channels of HD videos of 1080p at 30 fps
- Uses Kunpeng 920 processors to accelerate applications efficiently

Ultra-high energy efficiency

- Benefits from the multiple cores and low power consumption of the Kunpeng architecture to build an efficient and power-saving AI computing platform for inference scenarios
- Ultra-low power consumption of the Atlas 300I Pro (72 W per card), providing a better energy efficiency ratio while accelerating computing power of AI servers

Atlas 800 Inference Server (Model 3000)

- The **Atlas 800 inference server** (model 3000) supports up to 8 Atlas 300I/V Pro cards to provide powerful real-time inference and video analysis. It is perfectly suited for AI inference in data centers.



Superior computing power

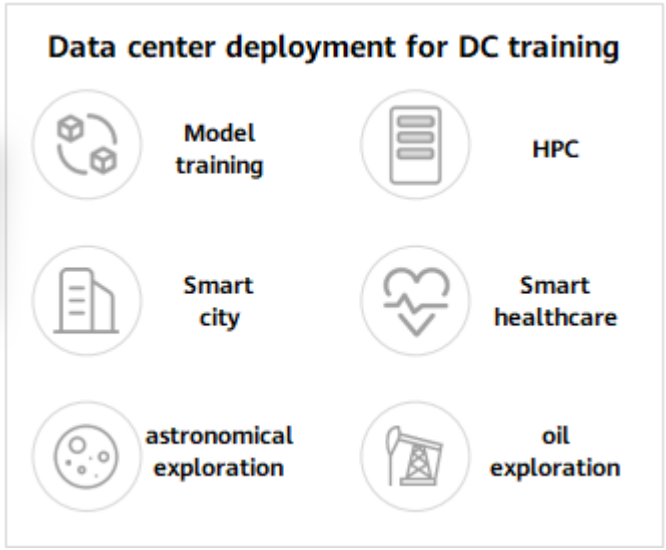
- Supports up to eight Atlas 300I/V Pro inference cards to meet inference requirements in multiple scenarios
- Supports real-time analysis of 1024 channels of HD videos of 1080p at 30 fps
- Uses Kunpeng 920 processors to accelerate applications efficiently

Ultra-high energy efficiency

- Benefits from the multiple cores and low power consumption of the Kunpeng architecture to build an efficient and power-saving AI computing platform for inference scenarios
- Ultra-low power consumption of the Atlas 300I/V Pro (72 W per card), providing a better energy efficiency ratio while accelerating computing power of AI servers

Atlas 800 Training Server (Model 9000)

- The **Atlas 800 training server** (model 9000) features high computing density, ultra-high energy efficiency, and high network bandwidth. With these competitive advantages, the server is perfect for deep learning model development and training scenarios. It is an ideal option for compute-intensive industries, such as smart city, smart healthcare, astronomical exploration, and oil exploration.



High computing density

Ultra-high energy efficiency

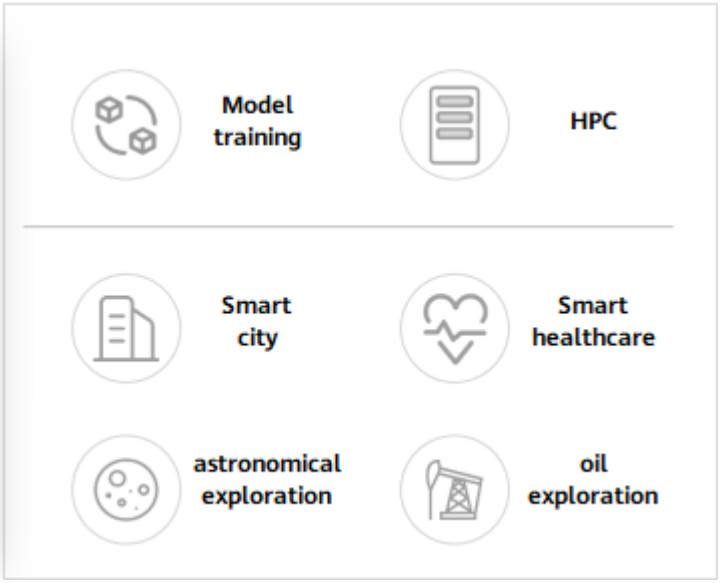
High network bandwidth

- 8 x 100GE RoCE v2 high-speed ports
- Cross-server processor interconnect latency slashed by 10-70%

HUAWEI

Atlas 900 PoD

- The **Atlas 900 PoD** (model 9000) is a basic AI training cluster unit that features powerful AI computing, optimal AI energy efficiency, and optimal AI expansion. With these competitive advantages, the basic unit is perfect for deep learning model development and training scenarios. It is an ideal option for AI compute-intensive fields, such as smart city, smart healthcare, astronomical exploration, and oil exploration.



Supreme AI computing

High AI energy efficiency

Optimal AI expansion

- Supports cabinet unit expansion to a maximum of 4096 Ascend AI training processors.

Contents

Atlas Facilitates Huawei's All Service Scenarios

Production services

Smart manufacturing

GTS intelligent O&M

AI-powered site survey, installation quality inspection, and device commissioning

IT systems

W3 NEXT

WeLink

VMALL.COM

Products & solutions

SoftCOM AI
Autonomous driving network

Autonomous driving
Noah's Ark Laboratory

NLP
Media Engineering Dept

Smartphone

Server

Huawei Cloud

Huawei Cloud EI

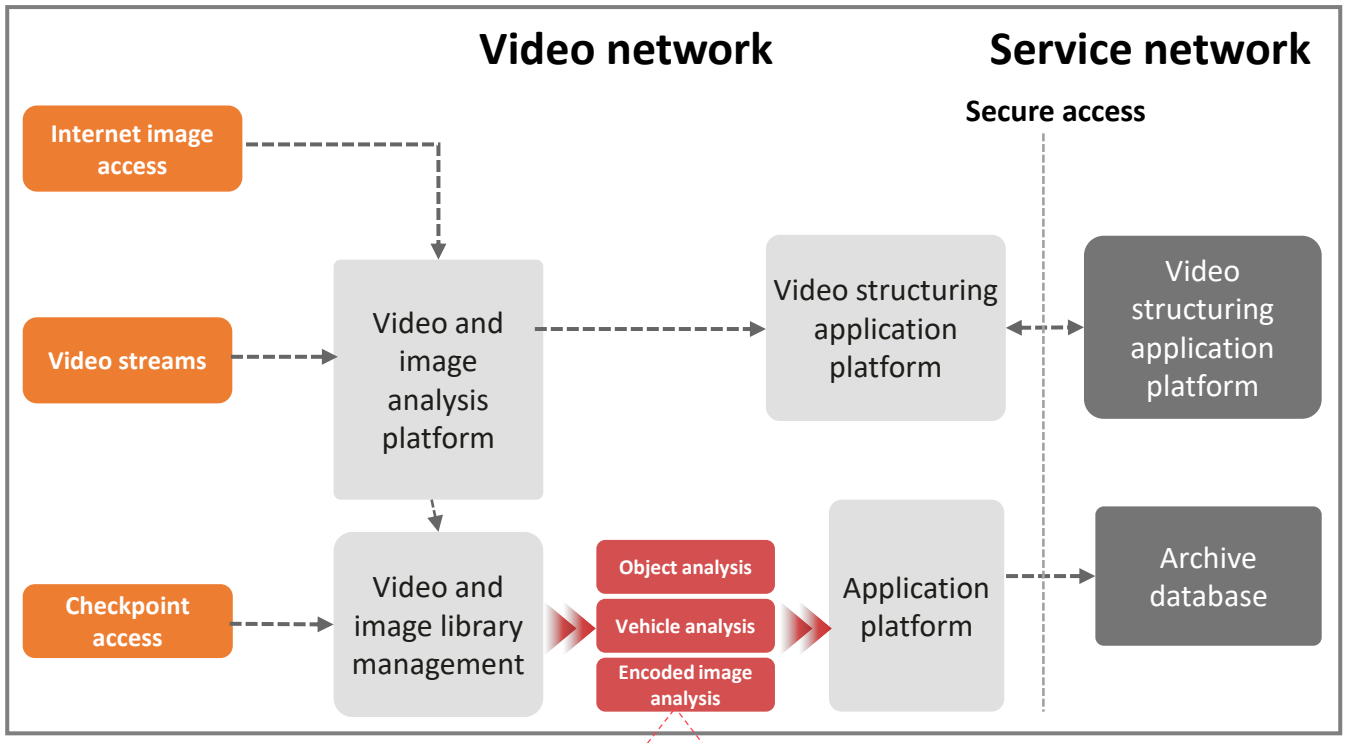
43 Ascend-powered cloud services

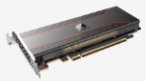
Digital foundation from Huawei with **Kunpeng** and **Ascend**

AI

Reliable operation of **100,000+ devices** over 3 years

Atlas-Based Video and Image Application Platform Improves Structuring Efficiency by 60%





Atlas 300 AI inference card

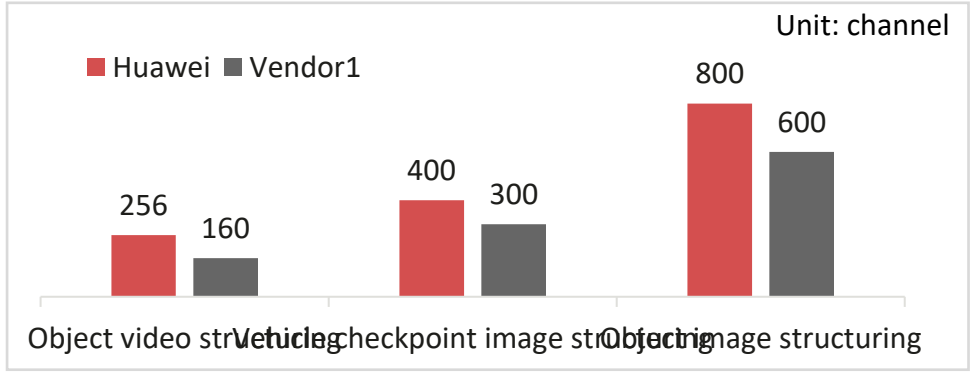
Atlas 800 AI server (model 3010)

1 **Object** Analysis of 15 million images from 5000 object checkpoint cameras per day

2 **Vehicle** Analysis of 10 million images from 2500 vehicle checkpoint cameras per day

3 **Video structuring** Analysis of data of motor vehicles, non-motorized vehicles, and pedestrians from 3000 cameras

Atlas 300 delivers more than twice the industry mainstream performance.



* Based on the configuration of 4 AI accelerator cards.

HUAWEI

Atlas Helps Peng Cheng Cloud Brain II Build a National AI Platform

Atlas 900

Building AI supercomputing based on Kunpeng + Ascend

6195 x86 cabinets = 208 GPU cabinets = 16 Atlas cabinets

40,268 kW 1,352 kW 736 kW

AI Industry Applications Facilitate the Development of the Greater Bay Area

Smart transportation	Smart healthcare	Smart finance

National Strategies	• International AI supercomputing platform • National open source platform for AI technologies • Open source AI and innovative ecosystem
Serving Shenzhen	• Supports major AI application requirements such as the intelligent computing system and robot system in Guangdong, Hong Kong, and Macao. • Improves the basic position and innovation of AI research on open source platforms and intelligent applications in the Guangdong-Hong Kong-Macao Greater Bay Area. • Attracts national AI resources, technologies, and talent.

Peng Cheng Cloud Brain II won three world No. 1s in AIPerf, IO500 full-node, and 10-node rankings. It participated in the industry-recognized MLPerf training v1.0 benchmark test and ranked second in the image classification track (with 1024 cards of the same scale).

Benchmarking Competitors: TCO Reduced by 9.3% with the Same Computing Power

1. Computing power improved by two times

1024+ Interconnected Ascend AI training processors, each delivering computing power twice the industry average

2. Network latency reduced by 70%

Integrates three high-speed interfaces: HCCS, PCIe 4.0, and 100G RoCE, reducing latency by 70%.

3. Cost on power cut by over 60%, space reduced by 80%

50 kW hybrid liquid cooling system for a single cabinet, PUE < 1.1

Ultra-high-density prefabricated modular equipment room, low power consumption, fast deployment, and exabyte-level cloud brain cluster rollout within six months

HUAWEI

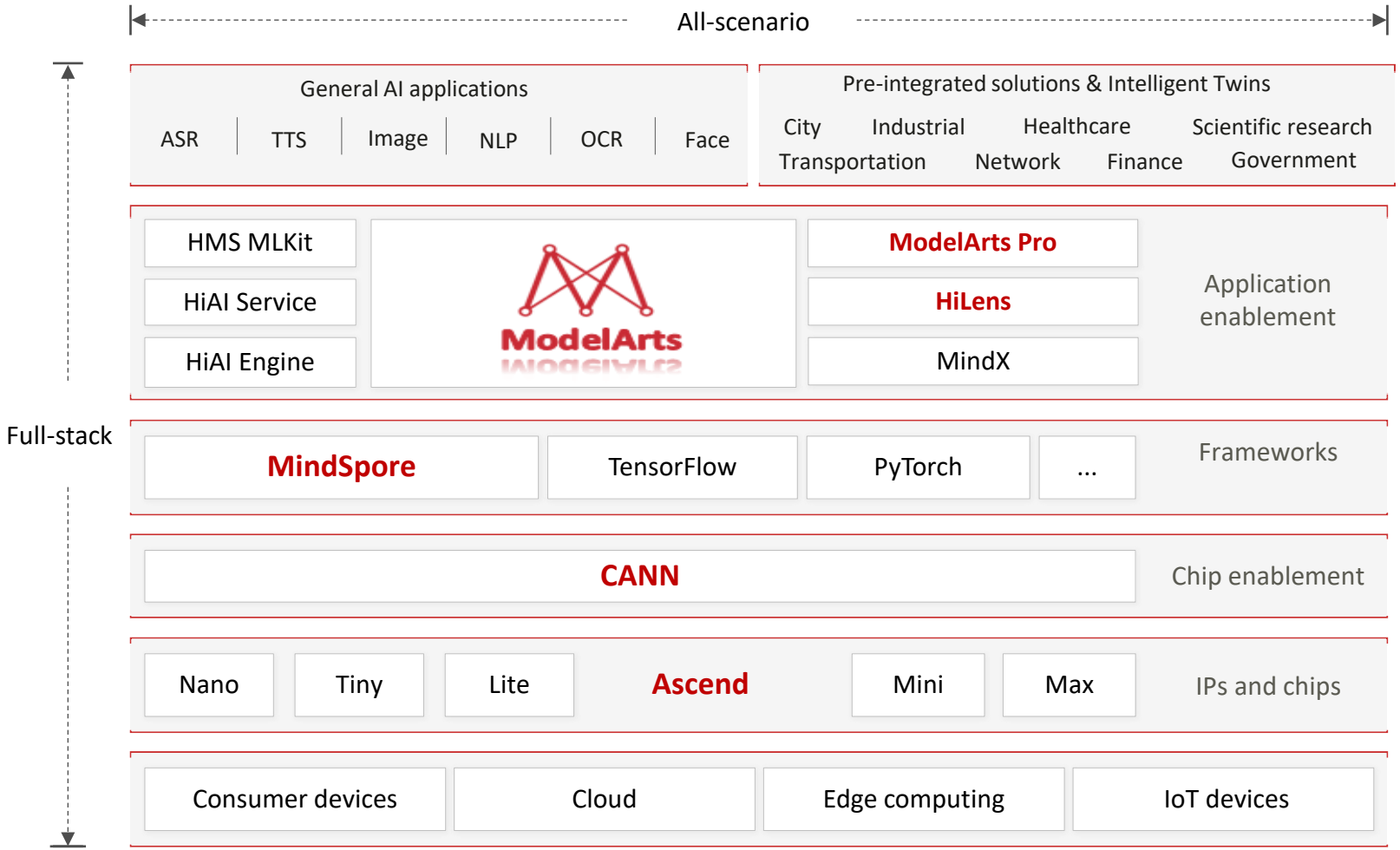
Huawei Helps Build a 'City-Wide Traffic Brain' for Predictive Travel

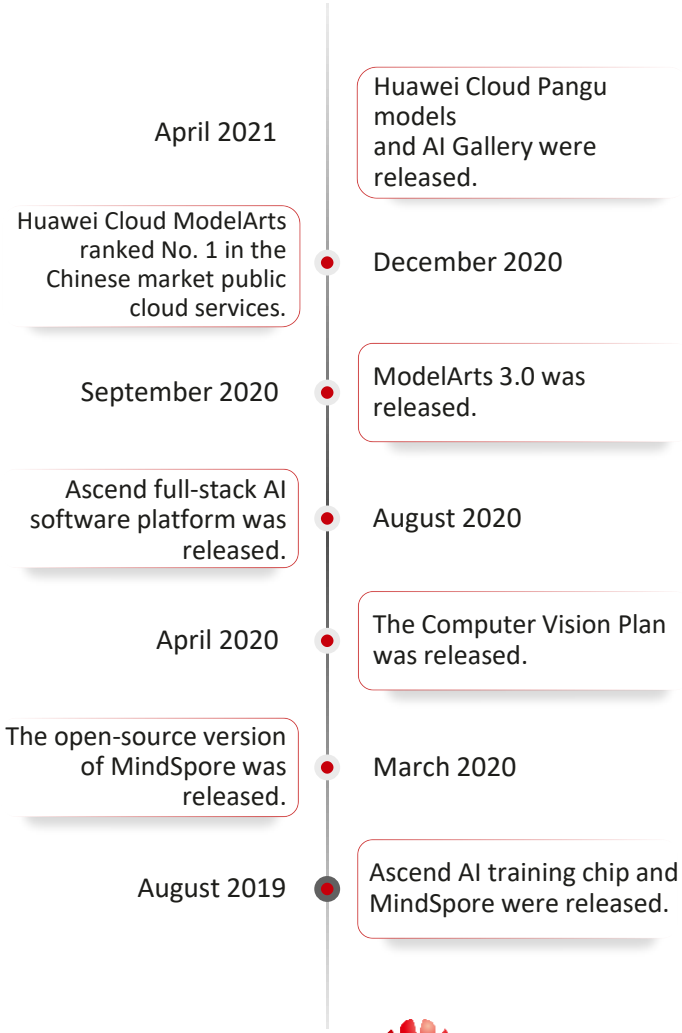
- With video cloud, big data, and AI being the core technologies, this 'Traffic Brain' presents a unified, open, and intelligent traffic control system. It focuses on the following five significant areas:

1	**Ultra-broadband traffic network**	Technologies such as a high-speed Optical Transport Network (OTN) are used to support transmission at 400G bandwidth, data storage of over 20 PB, and 10 billion-level data processing. The data bearing capability is approximately 40 times that of traditional public security networks.
2	**City-wide comprehensive traffic awareness**	A road monitoring system detects traffic conditions through license plate identification, video surveillance, and more, with a detection accuracy rate of up to 95%. The system collects about 700 million pieces of vehicle data every month, and integrates nearly 40 TB of data from 78 system databases, both internal and external. All these contribute to useful Big Data-enabled traffic congestion analysis and optimization.
3	**AI-assisted law enforcement**	The use of AI technology improves the efficiency of identifying traffic-violation images by 10 folds, ensuring the closed-loop processing of those images.
4	**Improved crime fighting efficiency**	The big data platform and a traffic analysis modeling engine are used to create multiple big data analytics models, including disqualified driving, drug driving, and multi-violation. Intelligence can now be generated and precisely pushed within 30 minutes, helping to pinpoint then clampdown on violations.
5	**Improved travel experience**	Scientific setting of intersection channelization, plus traffic organization innovation through big data management and control, is helping to improve road capacity by approximately 8%.

Contents

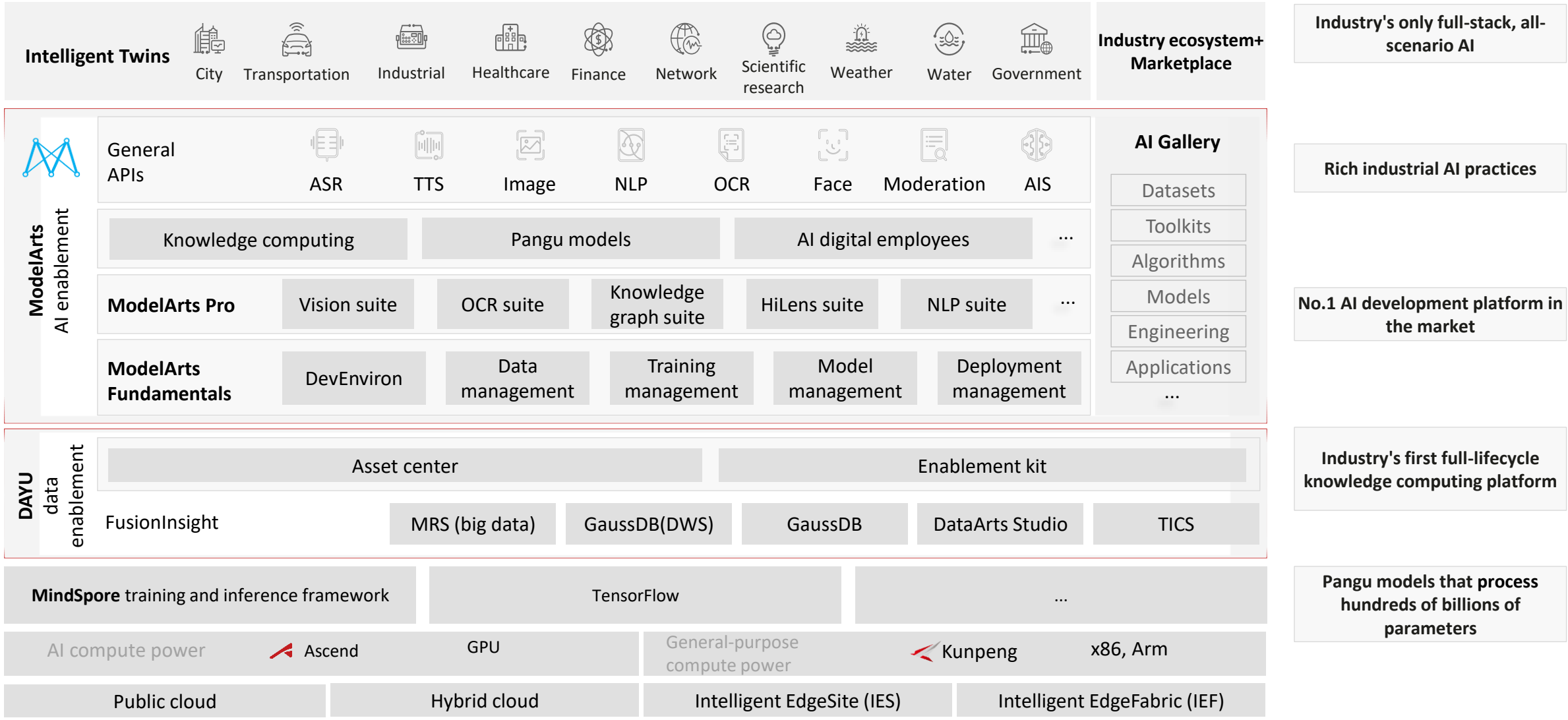
Huawei's Full-Stack, All-Scenario AI Promotes the Intelligent Upgrade of Thousands of Industries





HUAWEI

Huawei Cloud EI Open Capability Panorama



HUAWEI

Contents

From AI+ to +AI: AI Empowers Best Practices

AI+

Explore AI capabilities

EfficientNet model accuracy: 98.7% for **the top 5** labels, higher than manual accuracy (96%)

RNN-T model accuracy: 96.8%, higher than manual accuracy (94.17%)

BERT model accuracy: 90.9%, higher than manual accuracy (82%)

+AI

AI-empowered core production systems

Enormous Challenges in Deploying AI applications

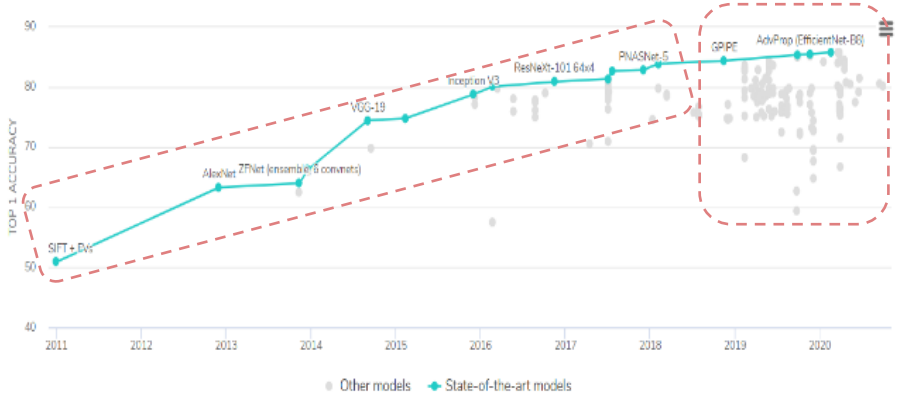
Algorithm vs. Precision iteration acceleration (image classification as an example)

AI development difficulties

Fast algorithm updates lead to difficult algorithm selection and tuning

50%

100+

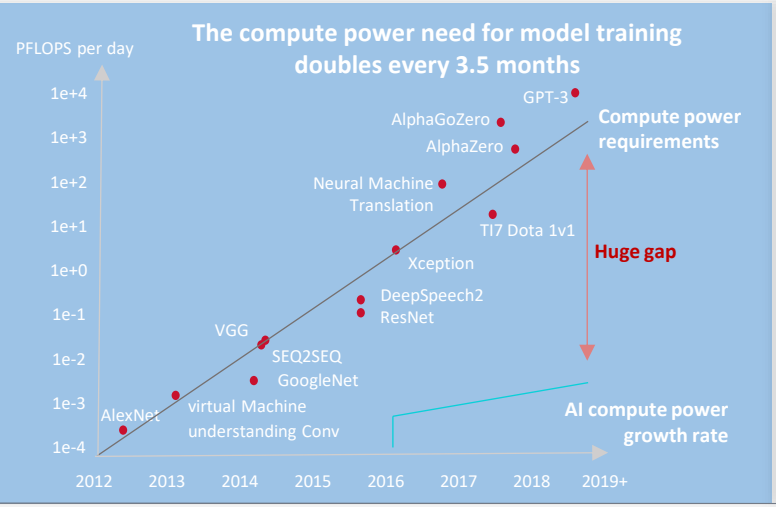


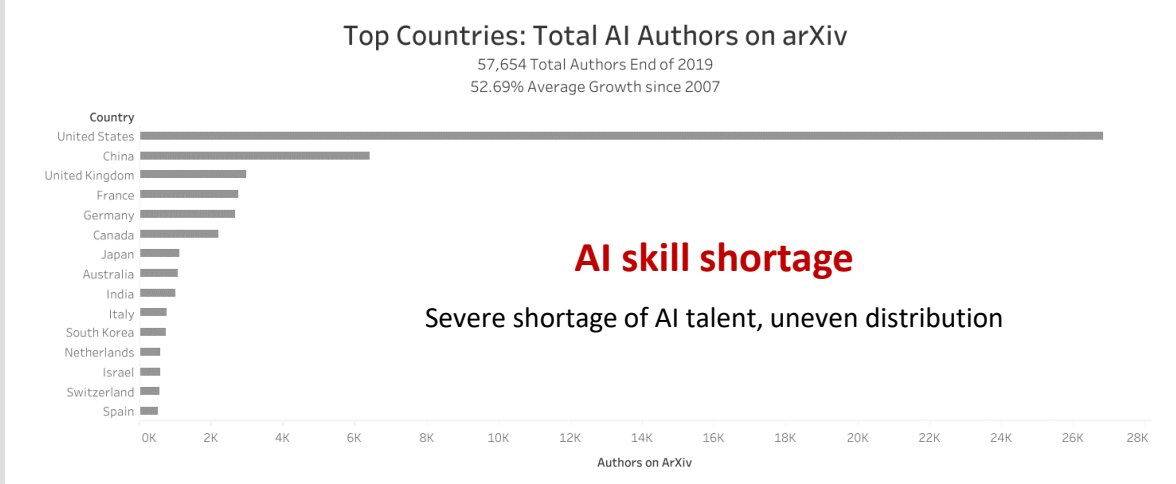
AI compute power scarcity

Expensive AI model training

GPT-3
US$4.6 million

BERT
US$15,000 on average





AI skill shortage

Severe shortage of AI talent, uneven distribution

HUAWEI

Rich Industry Practices Accelerate the Application of AI in Enterprise Production Systems

30%+

AI-powered core application systems

18%

Higher profitability

More than 10 years of experience and practices of 800+ Huawei projects accelerate the implementation of AI

Healthcare	Transportation	Industrial	Water management	Weather forecasts	Airport
Hours→Minutes AI-powered genome analysis	**17.7%** Reduced vehicle delay percentage	**CNY30 million** Costs saved per million tons of coal with intelligent coal blending	**81%** Water vision accuracy rate **vs.**	**2 hours** Short-term forecasts	**10%** O&M efficiency improvement
Months→Hours AI-assisted drug screening	**4.2%** Average vehicle speed	**80%** More precise steel quality inspection	**10%** Industry average	**10 minutes** Thunderstorm forecasting	**4 million person-times/year** Not taking shuttle buses

Huawei Has Been Developing a New AI Development Model Together With Developers

50,000+ jobs/month

840,000+ hours/month

50,000+ monthly active users

700,000+ AI developers

2500+ AI partners

30+ AI contests

70,000+ contestants

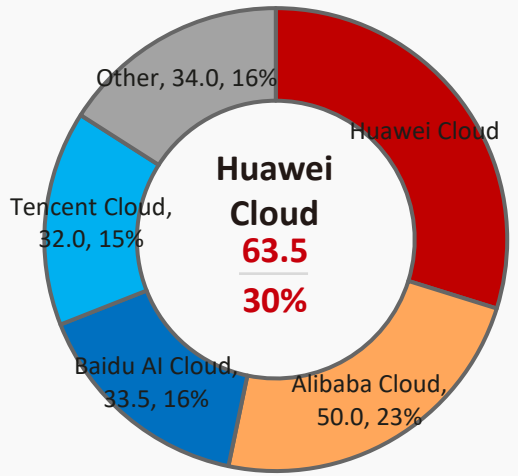
70+ universities

Statistics as of October 2021

ModelArts Continuously Sees Greater Influence in the Industry

Huawei Cloud ranks No. 1 in the Chinese machine learning market according to IDC's latest data on China's public cloud market.

China's public cloud service market share for machine learning
(Unit: CNY1 million)



IDC

The Forrester Wave: Huawei Cloud was named as a leader in PAML.

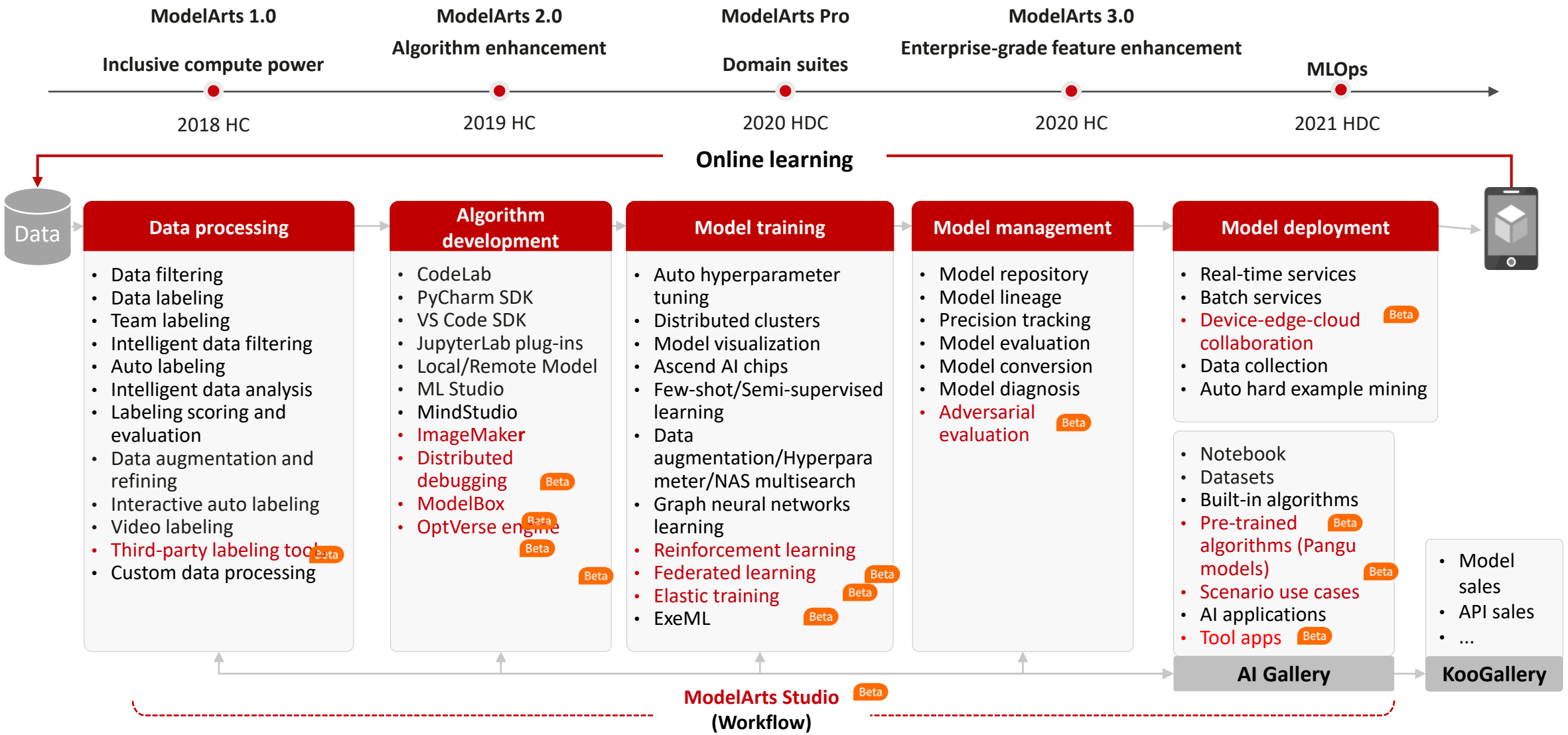
FORRESTER®

Huawei Cloud ModelArts has gained national recognition and its influence in the industry is continuously increasing.

At the World Artificial Intelligence Conference 2019, the Ministry of Science and Technology of China announced that Huawei was tasked with building China's next-generation AI open innovation platform for basic hardware and software. The platform combines cloud services and product software and hardware to provide full-process and inclusive basic platform services oriented to research and development of AI applications for various industries, startups, universities, and research institutions.

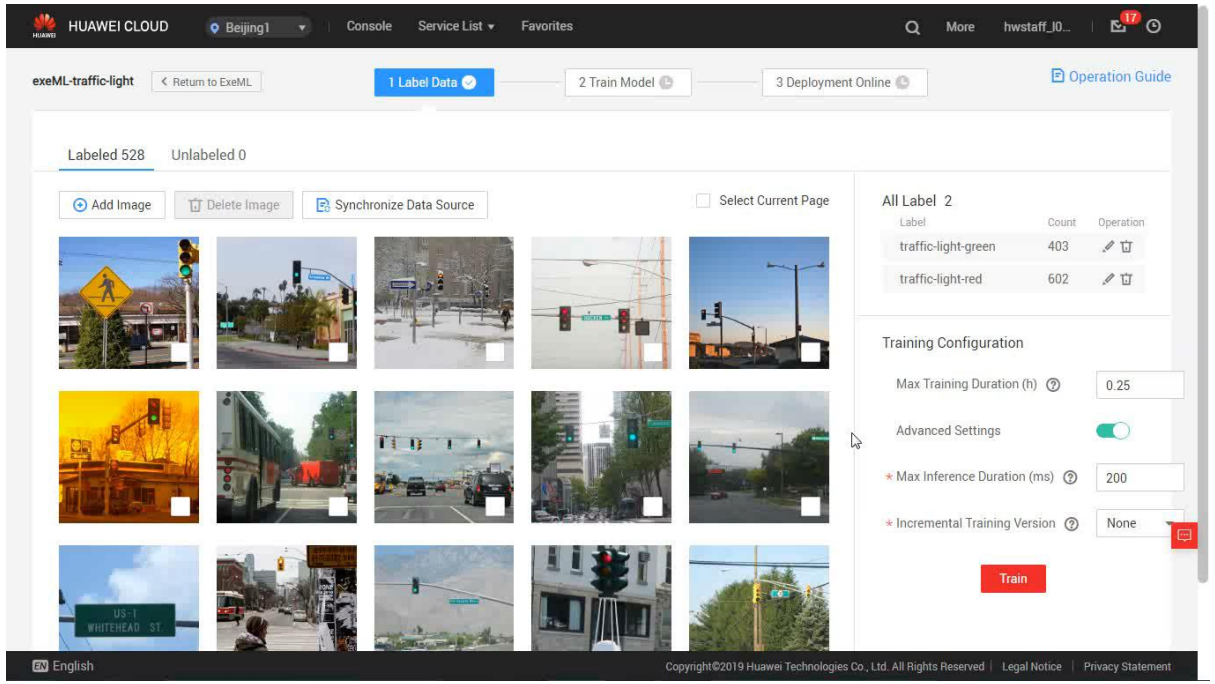
HUAWEI

ModelArts: Ideal Choice for Industry AI Implementation



HUAWEI

ExeML Engine Helps Beginners Easily Handle Common Application Scenarios



0

Code

0

AI experience

Step 1:

Upload data and label it.

Step 2:

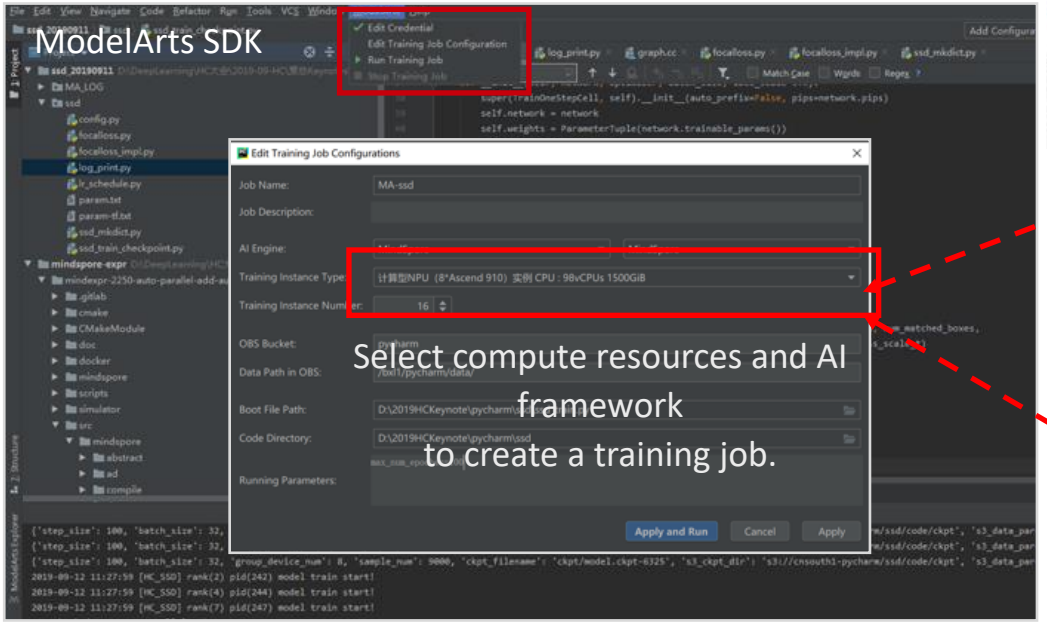
Train a model.

Step 3:

Check and publish the model.

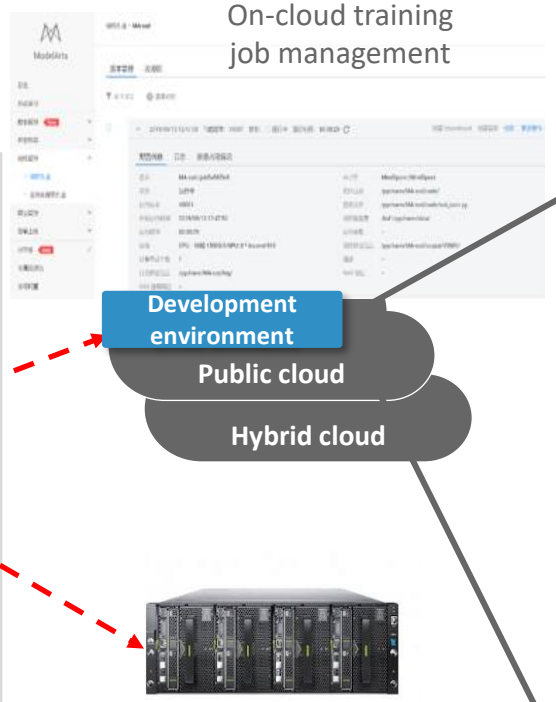
HUAWEI

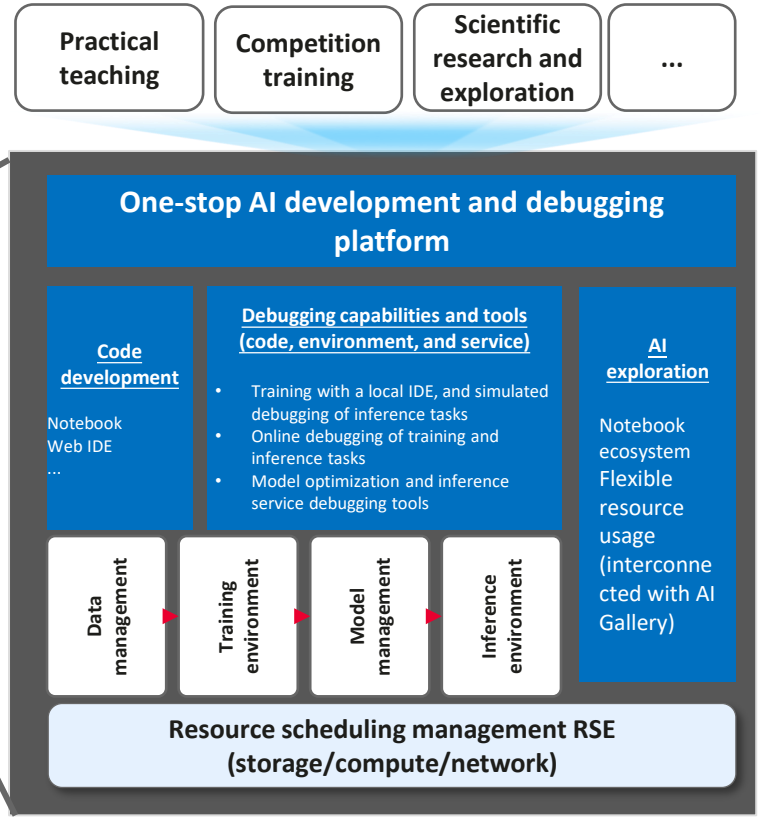
Development Environment 2.0, Providing the Ultimate Experience on Cloud and On-Premises



ModelArts > SDK Reference > Preparations > (Optional) Installing ModelArts SDKs Locally

https://support.huaweicloud.com/intl/en-us/sdkreference-modelarts/modelarts_04_0004.html





Experience: consistent experience on cloud and on-premises, out-of-the-box, and provisioning in seconds
Flexibility: flexible resource switching between dedicated and shared, 5 to 10 higher resource utilization
Experience: 300 cards for 1000 concurrent users (GPUs used within 25% of a 1-hour course)
Ecosystem: Seamless integration of AI Gallery (IPython Notebook) zone and GitHub

HUAWEI

Huawei Cloud OptVerse AI Solver Helps Customers Make Informed Decisions and Optimize Services

Schedule | Planning | Resource scheduling | Process optimization | Better user experience | Logistics | Inventory optimization | Path optimization

Global optimal

Resource utilization ↑ Operational efficiency ↑ Benefits ↑

OptVerse AI Solver

Mathematical programming | Heuristics | **Operations research** | **AI** | Machine learning | Deep learning

No.1 in the Hans Mittelmann's benchmark of simplex algorithm

- ↑ 30x Modeling speed: Manual modeling → NLP intelligent modeling
- ↑ 30% Solver efficiency: Manual parameter tuning → Adaptive tuning
- **100** million Variables ↑ 10-100x Computational speed: Single-node compute power → Distributed parallel acceleration

Empower Tianjin Port's intelligent port planning platform

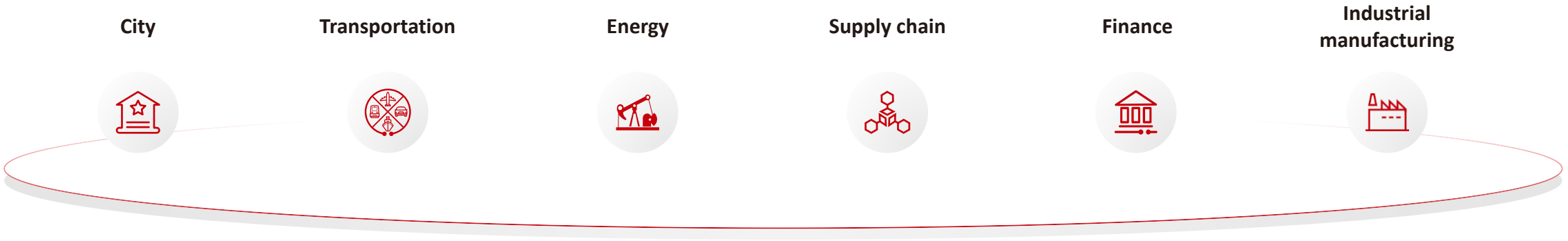
Tianjin Port: operation plan optimization

Time to develop a plan **24 hours** → **10 minutes**

- **Berth allocation plan**: 5% ↑ Berth utilization
- **Container crane operations plan**: 15% ↑ Equipment utilization
- **Smart container yard planning**: 20% ↓ Container yard reshuffles
- **Single ship stowage plan**: 10% ↓ Time in port

HUAWEI

Tailoring OptVerse AI Solver to the Specific Needs of Industry Scenarios for Easier Application



15% Risk-benefit ratio

Financial investment optimization

30x Efficiency improvement

Industrial batching system optimization

37% Inventory completeness

Production scheduling

2.6 million Fewer passengers ferried across the tarmac

Airport stand allocation optimization

15% Device utilization rate

Port operation plan optimization

10% Energy consumption

Smart heating optimization

33% Delivery timeliness

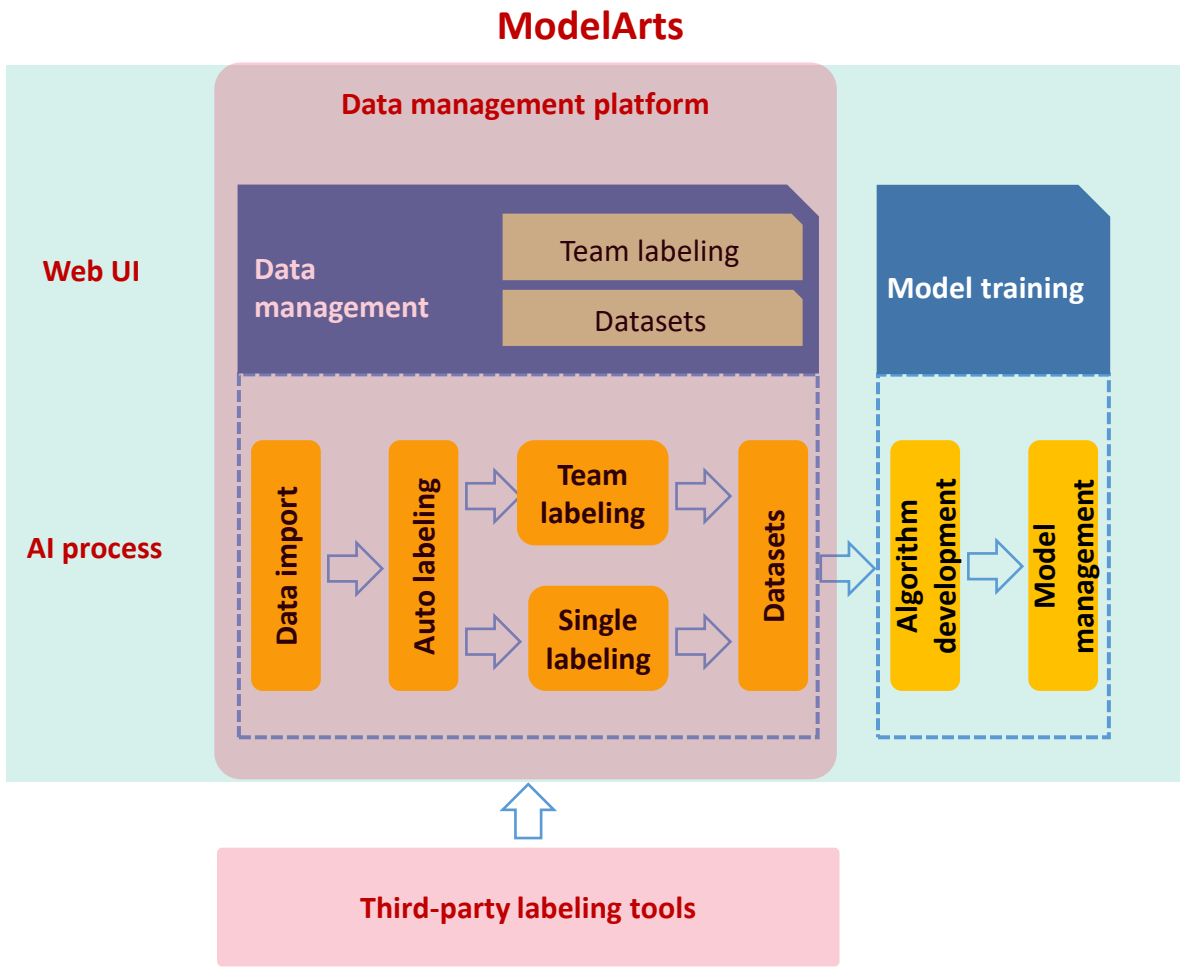
E-commerce order management optimization

3% Higher board utilization

Cutting

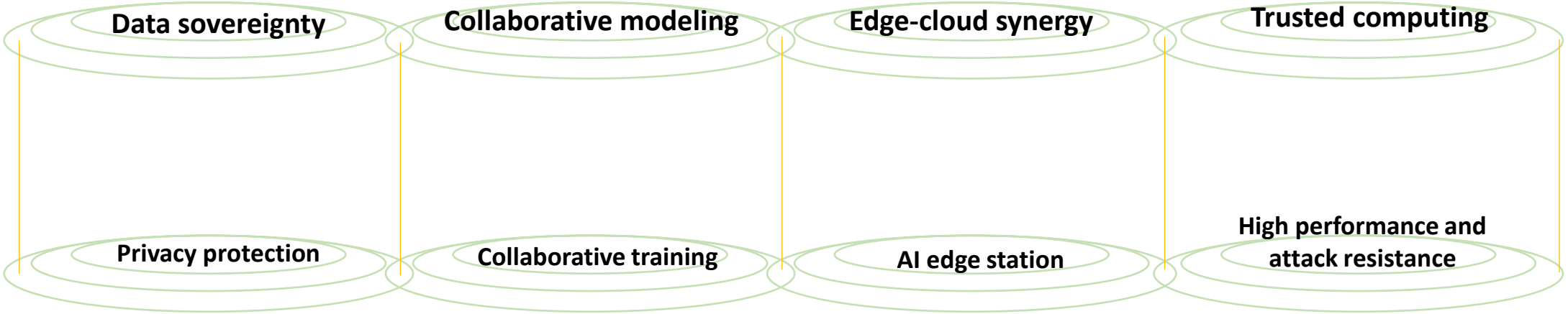
HUAWEI

Data Management Supports a Wide Range of Data Formats and Iterative Auto Labeling



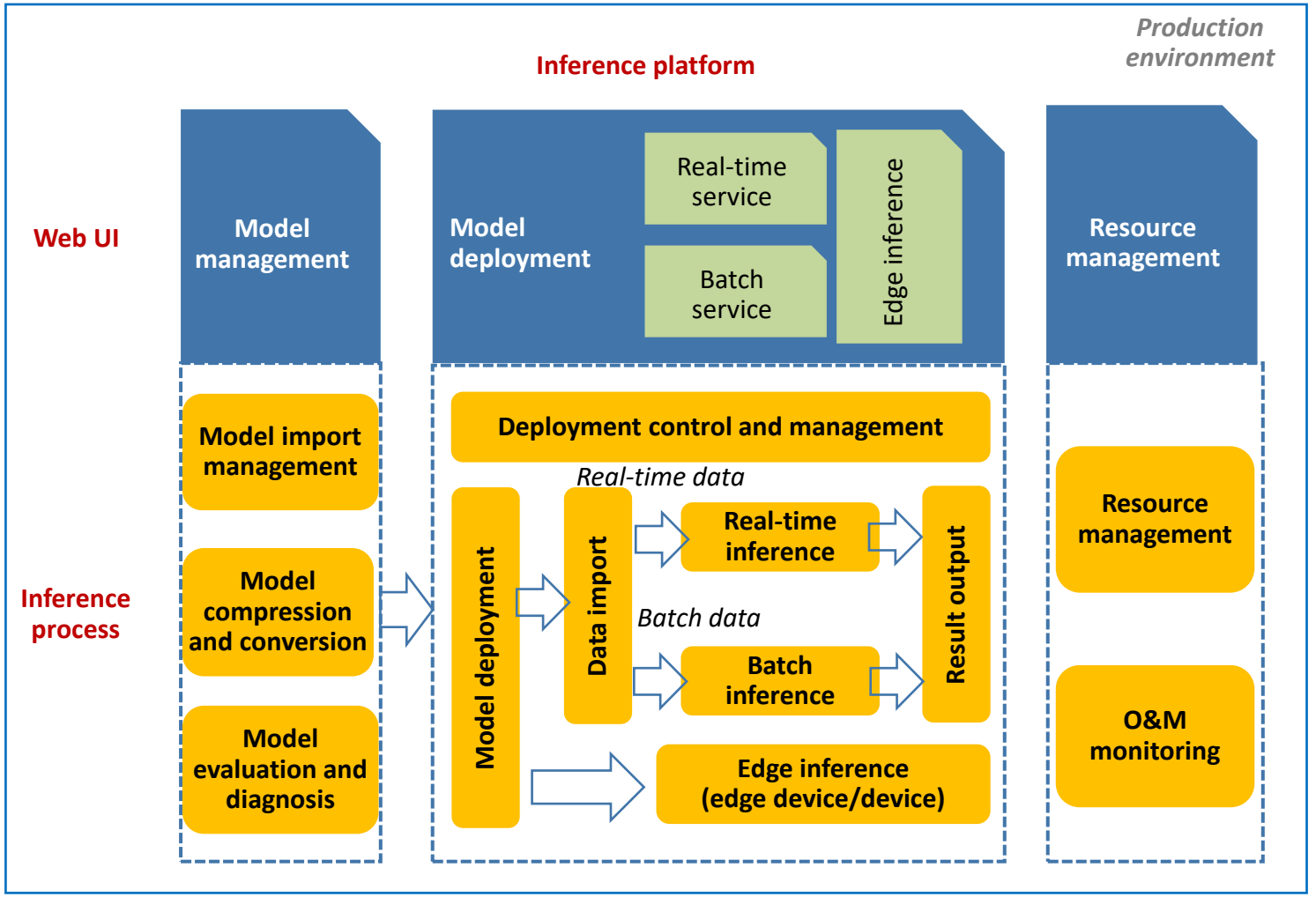
- A wide range of data formats
  - Five types of data
    - Image, audio, video, text, and table
  - Custom data formats
- Team labeling
  - Ideal for ultra-large-scale labeling
- Iterative intelligent labeling framework
  - Adaptive to data and algorithm changes
  - Intelligent data filtering and auto pre-labeling

Federated Learning: Eliminates Data Silos and Promotes Collaborative Modeling Across Industries



Horizontal federated	Vertical federated	Federated transfer
Homogeneous scenarios	Heterogeneous scenarios	Modeling across domains

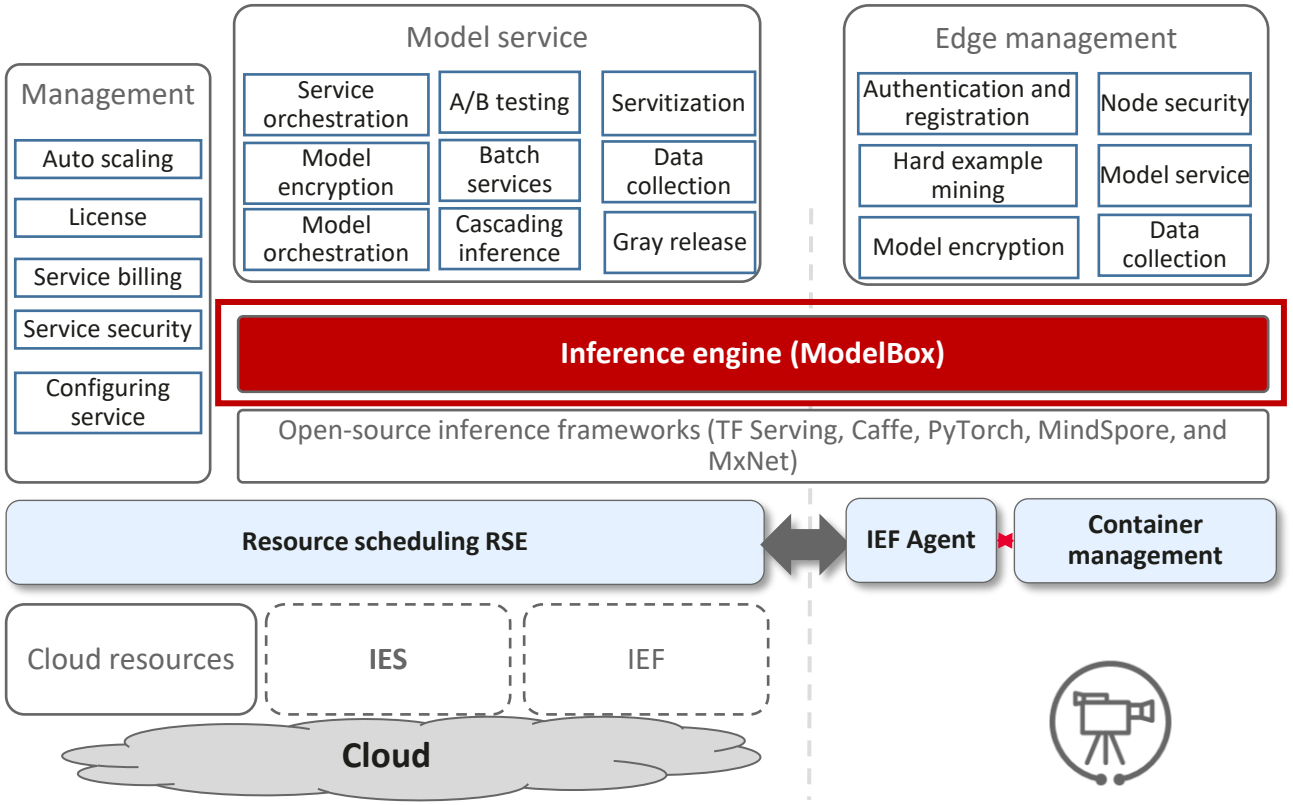
Inference Platform: Lifecycle Management of AI Application Running

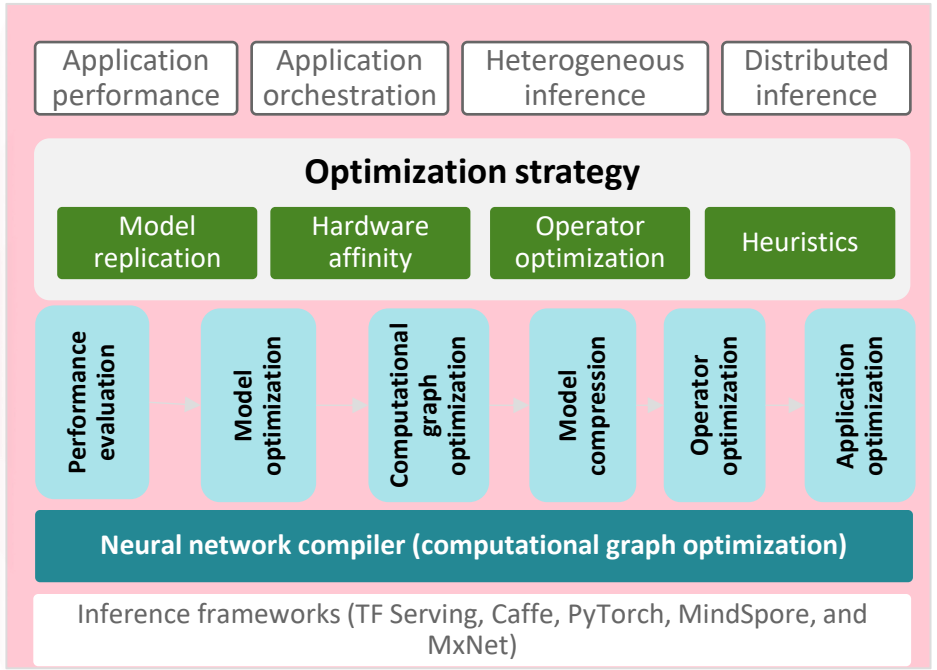


- Unified management
  - Unified management of models with different frameworks and functions from different vendors
  - High-concurrency model deployment, low-latency access, auto scaling, grayscale release, and rolling upgrade
- Flexible deployment
  - A rich array of deployment scenarios, including real-time inference services and batch inference jobs on the cloud, edge devices, and devices
  - Real-time inference of large models and model update in seconds, adapting to fast service iteration
- Higher performance
  - Huawei-developed inference frameworks hide underlying hardware and software differences from the upper layer software to improve performance.
- Iterative model update
  - Data collection and automatic hard example mining to quickly adapt to data changes

HUAWEI

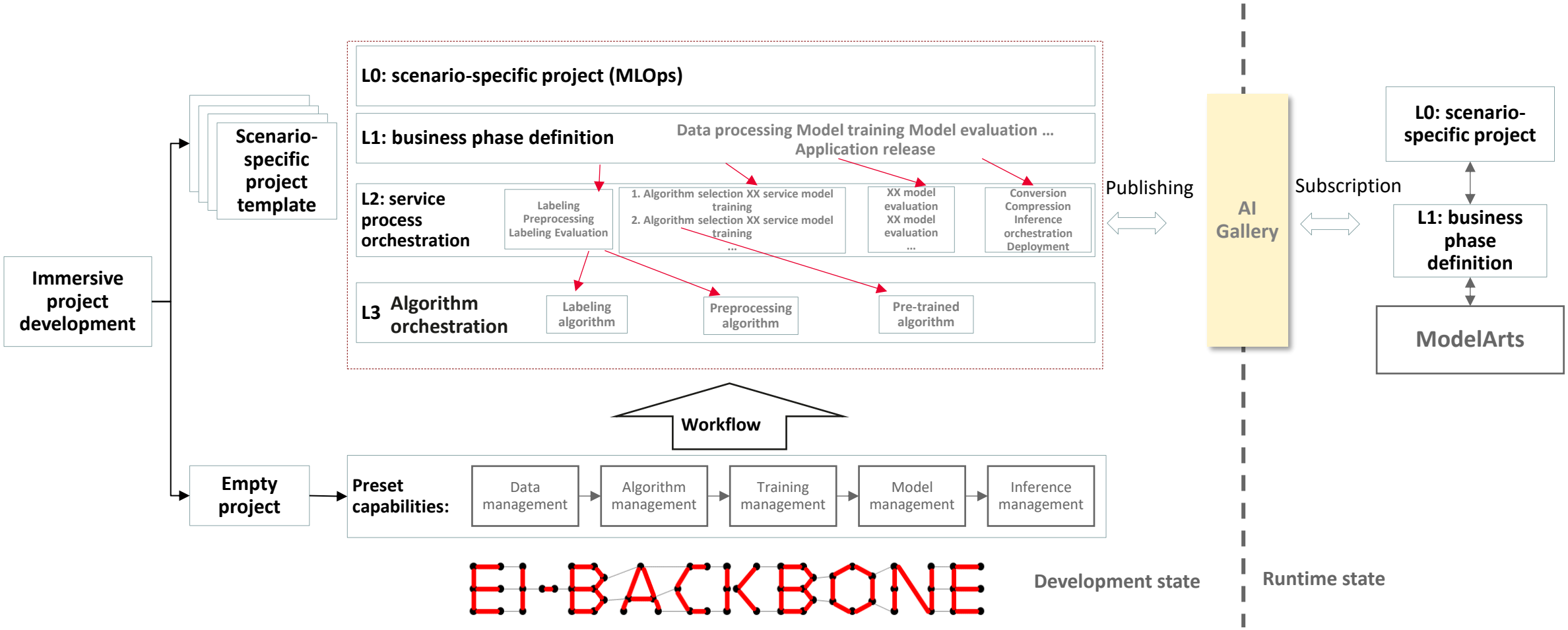
ModelBox: Device-Edge-Cloud Joint Development and Real-Time Model Update





Performance: The E2E inference performance is improved by 3 times, and the performance of image and gene services is improved by 3 and 22 times, respectively.
Efficiency: Code-free orchestration with more than 27 preset basic operators
Easy to use: 80% AI applications are developed and rolled out without coding.

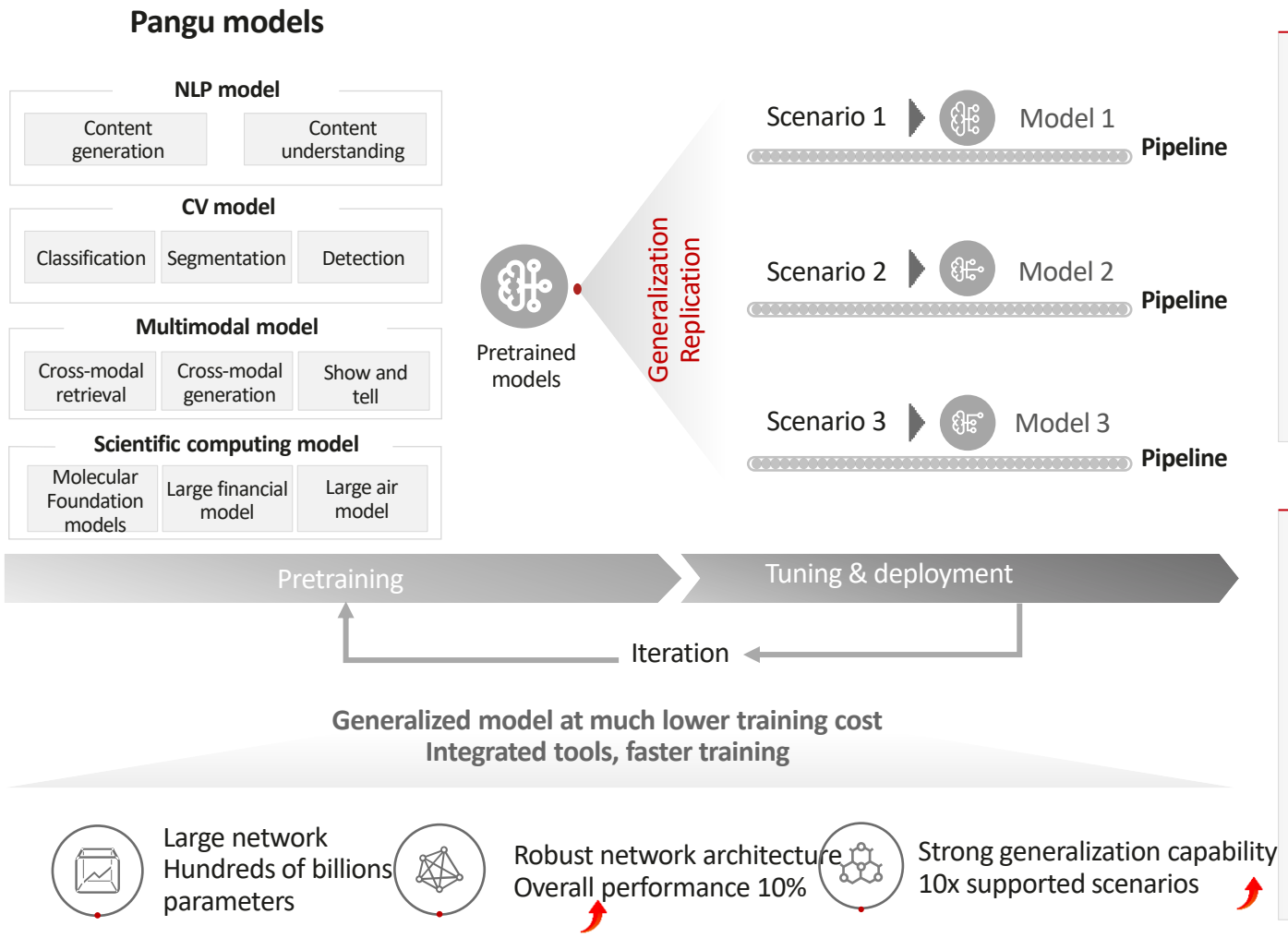
MLOps Immersive Development with Standardized Process

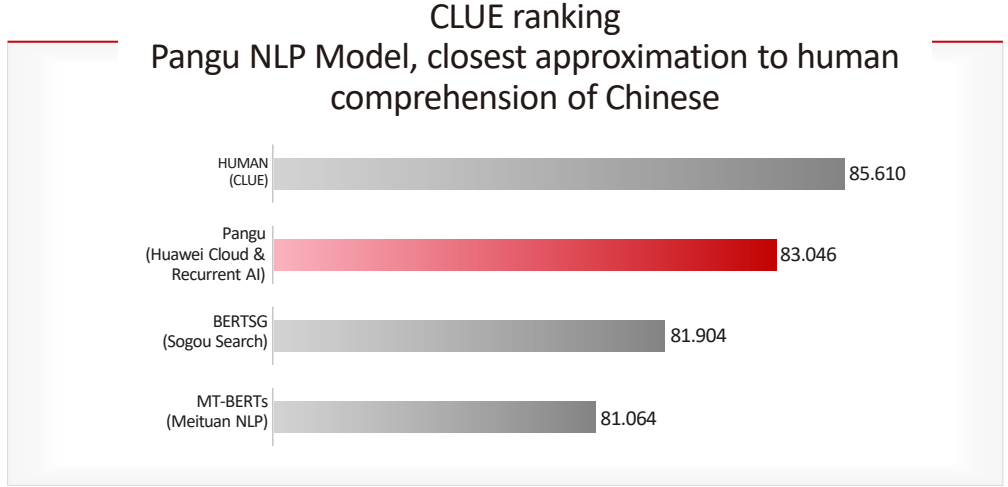


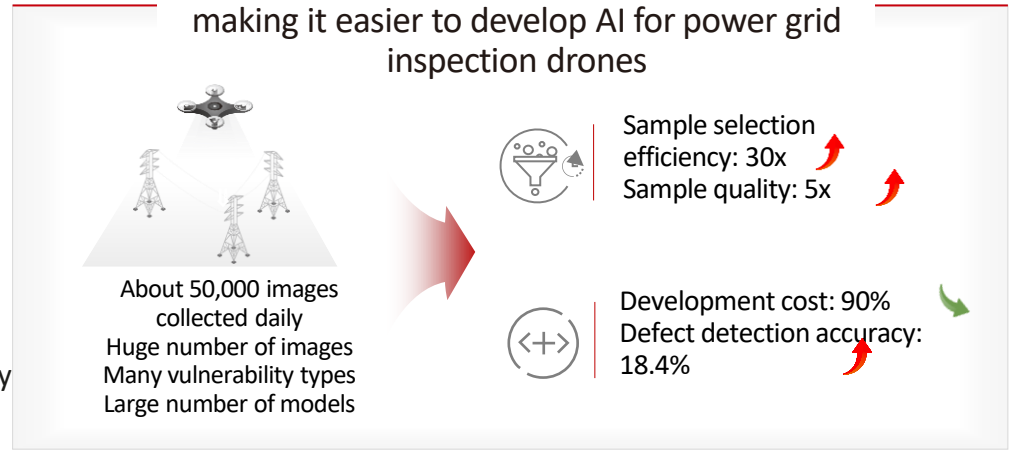
Non-professionals: Scenario-specific templates are provided for image recognition, object detection, natural language processing, video recognition, text recognition, and knowledge graph to better address industry problems using AI algorithms.

AI engineers: EI-Backbone is used for high-performance and efficient implementation. The project development workload is reduced by 50%, the application performance is improved by 5%, and the project processing capability of engineers is improved by 50%.

Huawei Cloud Pangu Models: Pioneering A New Paradigm for Industrial Scale AI Development

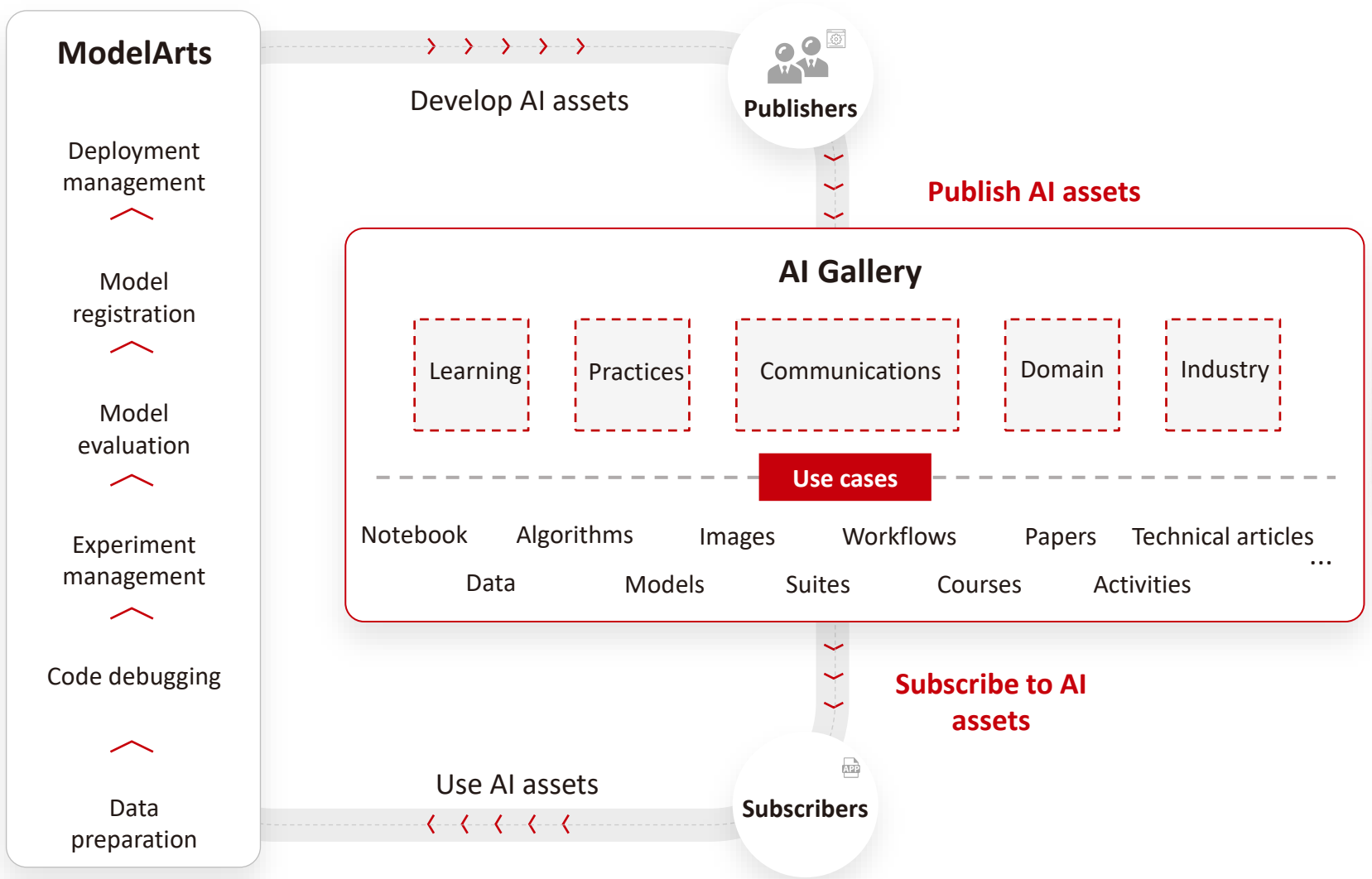






HUAWEI

AI Gallery: Bridging Supply and Demand in the AI Ecosystem



Pangu models

New paradigm for industrial AI development

OptVerse AI Solver

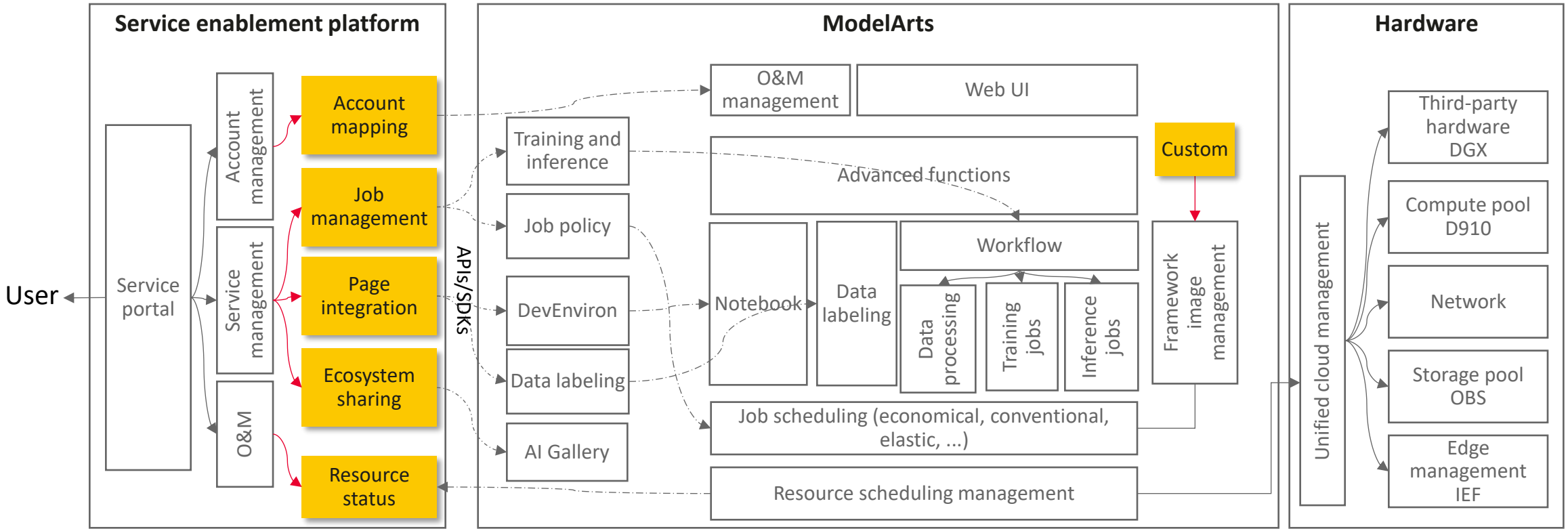
Setting a new bar for operations optimization

RL engines

New choice for intelligent decision making

HUAWEI

Third-Party Service Platform Integration Solution



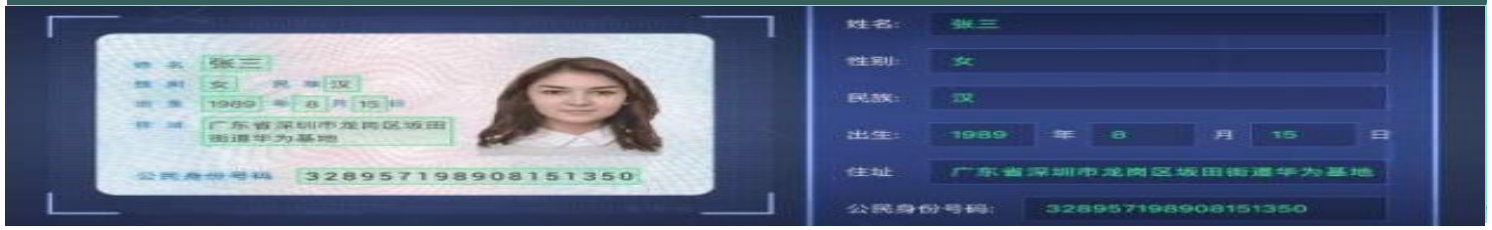
Account system: user authentication/...
Job management: training/development/inference/scheduling policy/custom image/...
Function integration: data labeling/federated learning/pipeline/Backbone/...
Resource status: resource management/job status/exception/...
Development ecosystem: AI Gallery/Huawei Cloud capability subscription/...

Contents

Vision Services: Empowering Service Systems with Awareness Capabilities

OCR

Recognition of general tables, ID cards, driving licenses, vehicle licenses, express waybills, contact cards, 1D code, QR code, bank cards, invoices used for reimbursement, network screenshots, and slogans, and quick customization...



100+ tables

Value recognition rate > 98%,
supporting recognition of letters and digits in sealed, twisted, tilted, or interlaced tables

Image recognition

Image tagging, image classification, content moderation, offering search, scenario identification, image recognition, image comparison, image interpretation, license plate recognition, and customization...

23 thousand+ labels

Semantic label, copyright search, reverse image search...

Facial recognition

Portrait recognition, facial recognition, facial attributes, face detection, human figure recognition, human body recognition, action recognition, human body analysis, face retrieval, face verification...

Recognizing a target from 100,000 records in seconds

95.50%+ search accuracy for a face library with 100,000 face records

HUAWEI

VAS Makes Video AI Available to All Industries

Huawei Cloud AI video analysis

The following information can be obtained after video or image analysis:

Smart City IOC Solution	Smart City Management Solution	Smart Emergency Response Solution
Garbage is piled up on *XX* street.	Unauthorized sidewalk sales occur on *XX* street.	Fires on *XX* street
Safe Construction Solution	**Smart Campus Solution**	**Food Safety Solution**
Whether workers wear safety helmets and whether workers operate in a standard manner	Determining crowd size and absence detection	Whether the kitchen staff are wearing health and safety gear

ModelArts	**Intelligent video analysis platform (IVA)**	**EI Intelligent Twins enablement foundation**
Kunpeng+Ascend chips	**Cloud-edge-device synergy architecture**	**Huawei Cloud Stack**

Natural Language Processing (NLP)

Named Entity Recognition (NER)

Extracts entities, such as person names, organization names, and place names, in text.

- ✓ 10+ industry-leading NER models
- ✓ 30+ supported categories (3 to 4 categories supported by other industry peers)
- ✓ 10% higher accuracy than that of major competitors

Word segmentation

Segments a text into sequences in unit of independent words and attaches a part-of-speech (POS) tag to each word.

- ✓ Supports multiple mainstream standards, such as PKU, CTB, and MSR.
- ✓ Supports multi-granularity word segmentation.
- ✓ 6% higher accuracy than that of major competitors

Sentiment analysis

Analyzes the sentiments involved in a text to determine whether the text is positive or negative.

- ✓ Supports sentiment analysis in multiple domains.
- ✓ Supports attribute-based sentiment analysis.
- ✓ 10% higher accuracy than that of major competitors

Text summarization

Automatically summarizes the main content of a text to form a summary.

- ✓ Automatic extraction and customizable content length
- ✓ 12% higher accuracy than that of major competitors

Full-stack NLP service

- Basic algorithm/Language understanding/Language generation/Machine translation
- Available for all mainstream service scenarios

Cost-effectiveness

- Industry-leading performance
- 5%+ ahead of the competition in accuracy

Stability and reliability

- Multi-active instance deployment and high service stability
- Fault recovery within seconds and multi-dimension fault isolation

Domain customization

- Named entity recognition supports multiple domains, such as business and entertainment. Sentiment analysis supports e-commerce and automobile.
- NLP supports customization in specific scenarios.

Won numerous international awards

World-leading precision, performance, and effect

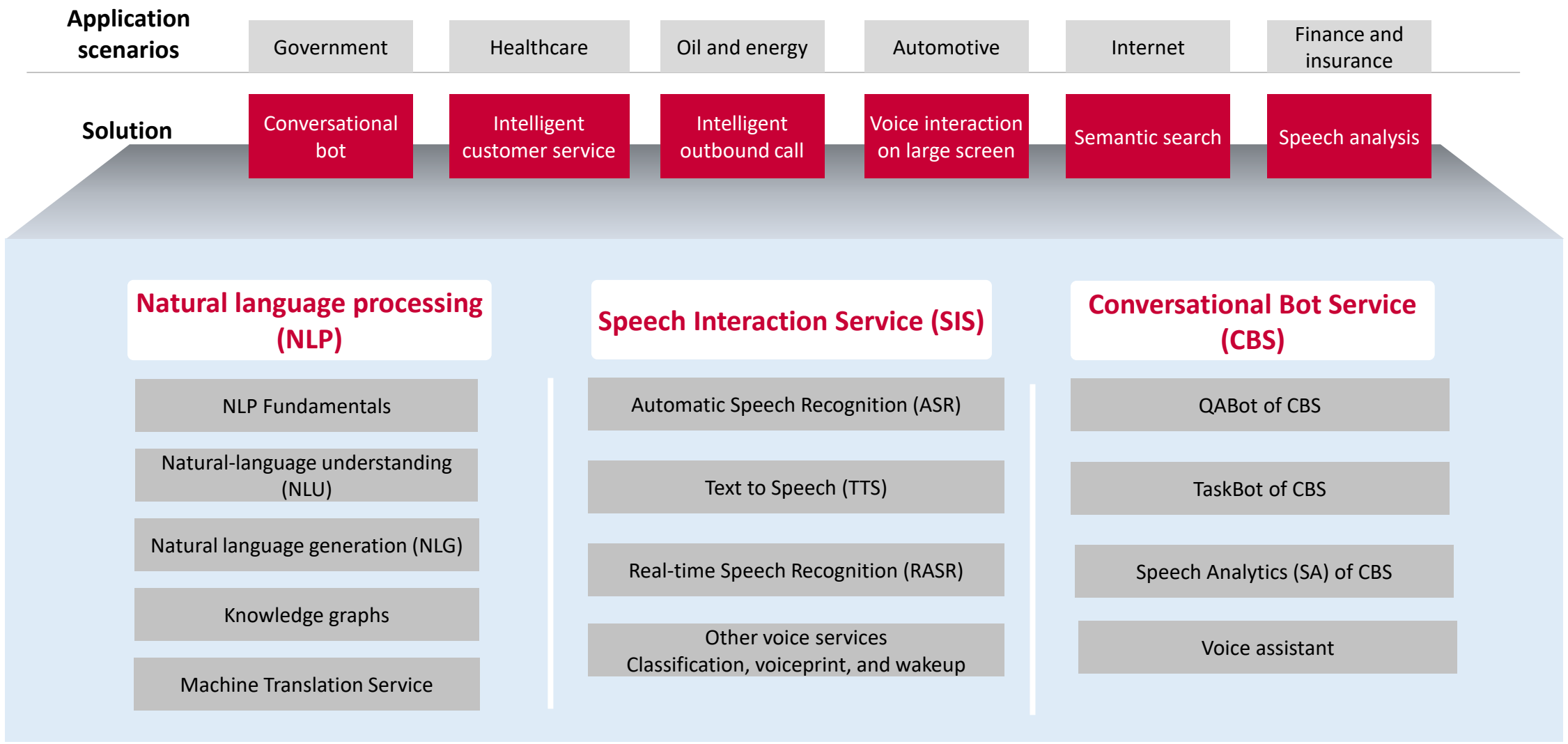
Big Data Expo 2019 Leading Technology Achievement Award	2019 CCKS Technical innovation award	2019 DiggScience Champion	2019 CCF BDCI Champion	2020 WSDM Champion

Knowledge Graph Service: Full-Stack, Full-Lifecycle, and First-Class in the Industry

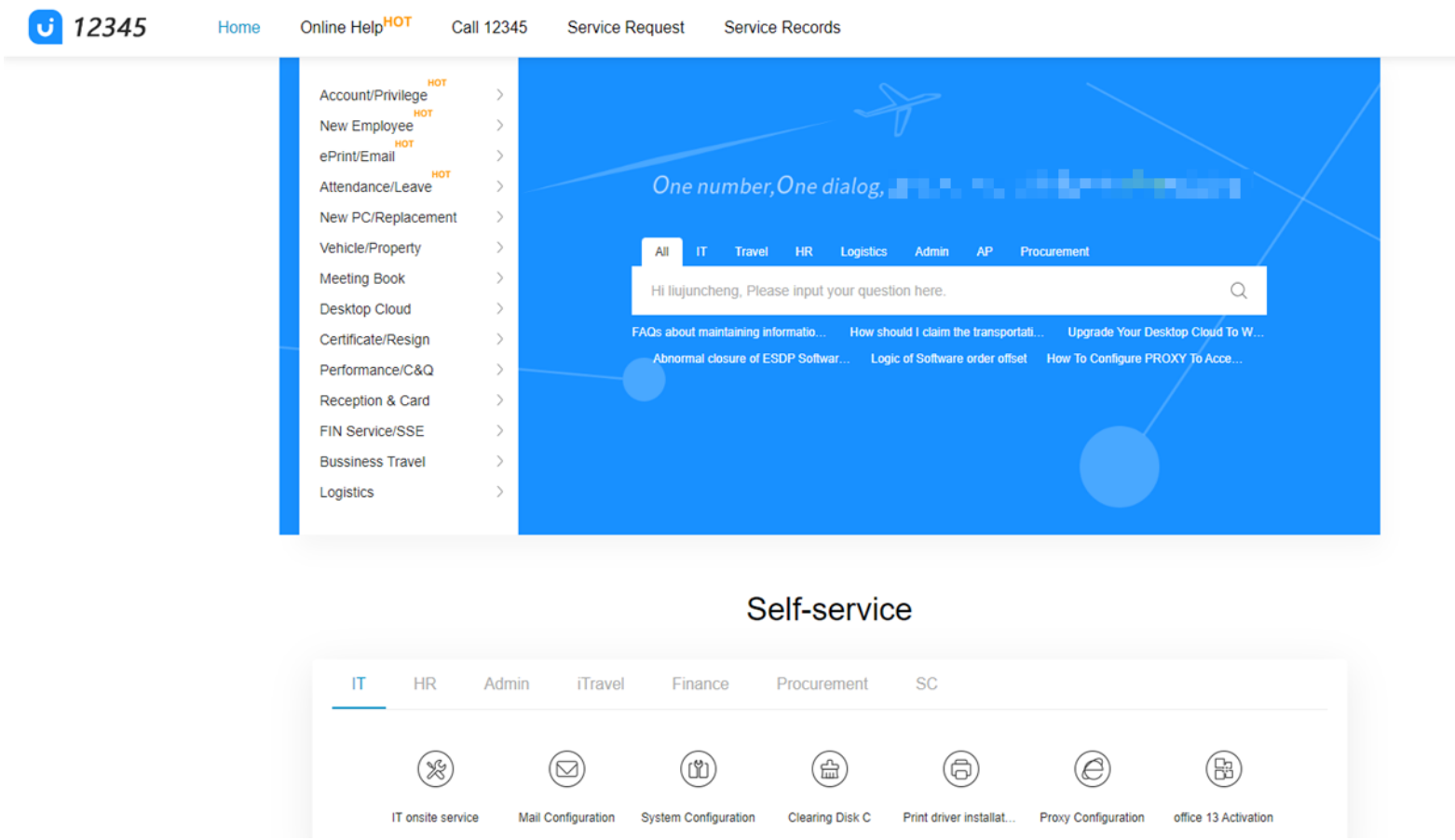
A knowledge graph is a structured semantic knowledge base, which is used to quickly depict concepts and relationships between concepts in the physical world.
The Knowledge Graph (KG) service provides one-stop knowledge graph lifecycle management, including graphical ontology designing, automatic graph pipeline building, and graph applications such as intelligent Q&A, intelligent search, and knowledge inference.

Knowledge acquisition		Knowledge generation		Inference and decision-making	
Knowledge Q&A	Intelligent customer service	Summary	Intelligent customer service	Association analysis	Fraud Crime groups
	Voice assistant		Voice assistant		Root cause analysis
Knowledge search	Search engine	Articles and reports	Blogs, reports, and bulletins	Classification and prediction	Public opinion analysis User profile
	Knowledge management system		Meeting minutes		Device fault warning
Knowledge recommendation	R&D process	Creations	Write work reports	Mechanism+AI	Oil and gas reservoir identification
	Search process		Compose and sing		Steelmaking ingredients

Rich Basic Speech and Semantics APIs+Conversational Bot



Customer Service+AI: Huawei Cloud Conversational Bot Service (CBS) Is Used to Quickly Build an Intelligent Q&A System for Customer Service



Service scenario

- Customer needs are increased, and thus more customer service personnel is required, increasing labor costs.
- Customer requirements are diverse, and some customer service personnel do not have comprehensive knowledge about the service. As a result, customer satisfaction decreases.
- It is difficult for manual customer service personnel to collect statistics and analyze data upon Q&A, and thus the recorded feedback is not enough to improve product quality.
- A large number of historical service tickets, logs, and cases are not used to extract service experience.

Solution

- Huawei Cloud Conversational Bot Service (CBS) is used to build an intelligent Q&A system for customer service, greatly improving customer service efficiency.
- Seamless transfer to manual processing provides better user experience.
- Closed-loop knowledge management quickly iterates and enriches the knowledge base. Active model learning makes robots smarter as long as it is used.

Benefits

- Over 350,000 Huawei staff are served.
- The time used for building the knowledge base is shortened by 48 times.
- At the early stage of commercial use, the problem hit rate exceeds 85%, and the manual substitution rate reaches 65%.
- CBS handles workloads of 179 human customer service personnel annually.

Contents

Huawei Cloud AI Explores Industry Best Practices with Customers and Partners

Intelligent Twins | City | Transportation | Industrial | Healthcare | Finance | Network | Scientific research | Weather | Water | Government

800+ Huawei projects

30%+
AI-powered core application systems

18%
Higher profitability

HUAWEI

Success Stories of City Intelligent Twins

Smart urban management

"Clean City" management in Longgang District

- AI-based investigation and handling of more than 40 littering incidents every day
- A showcase district for the "Clean City" campaign

Smart community

Smart community in Guangming District

- AI community governance, with dynamic event monitoring and rectification
- 24-hour e-government station, providing self-service handling of six types of services

City IOC

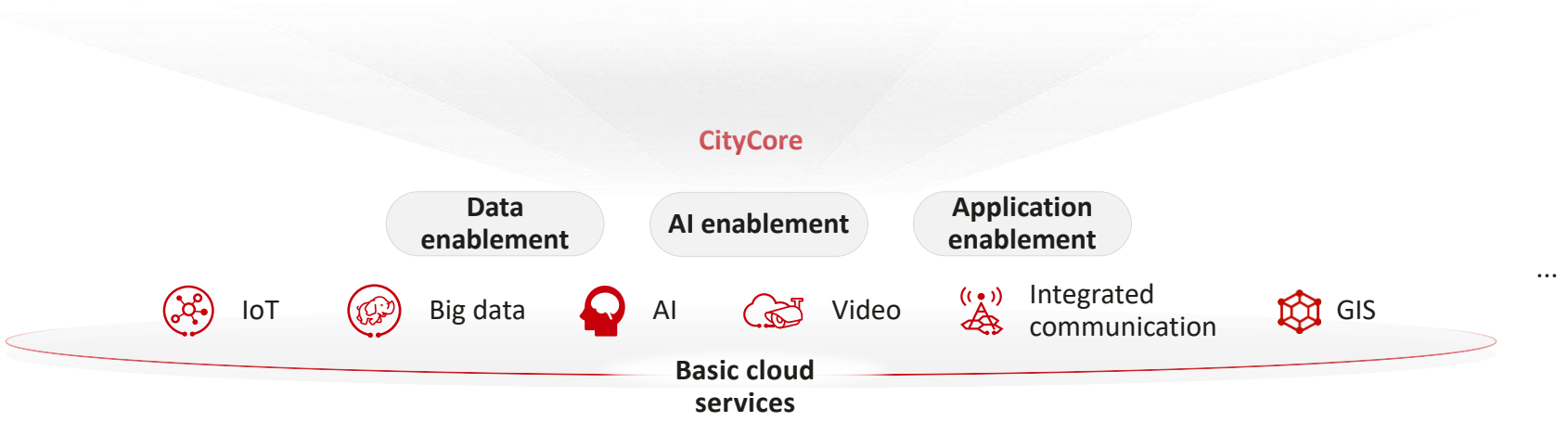
Voice assistant in Guangming District

- Precise response to voice commands
- Precise audio response to inquiries

Smart Q&A

Smart enterprise Q&A in Guangming District

- Policy consulting to help clarify doubts about frequent policy changes
- Enterprise-related affairs in the service hall (500+ items)
- Requirement issues (various types of requirements)



HUAWEI

Success Stories of TrafficGo

City-level traffic dashboards

- Infrastructure overview
- Real-time traffic
- Top congested roads
- Congestion alarms
- Incident detection

...

District-level traffic dashboards

- Best optimized roads
- Optimization result evaluation
- Intersection rotation display

...

Intersection traffic dashboards

- Intersection pedestrain density
- Intersection vehicle queue length
- Intersection video

...

Diagnostics dashboard

- Old solution vs New solution
- Entire diagnosis process displayed

...

Based on multi-source, converged data, build multi-dimensional **(macro-mid-micro)** traffic evaluation/metrics systems that are accurate and real-time.

Automatic evaluation of the results of each measure taken. Intelligent, iterative optimization in a "sense-understand-diagnose-optimize-evaluate" loop.

Success Stories of Industrial Intelligent Twins

Coking coal blending

Expert experience

Auto-learning of multidimensional parameters

E2E benefits prioritized

Quality prediction precision: > 95%

Cost saved per million tons on production line: over CNY20 million

Steelmaking ingredients

Constituent prediction accuracy

85%-90% → 95%

Alloying constituent costs saved

CNY20 million/year

Cutting

Higher board utilization

84%-86% → 86%-88%

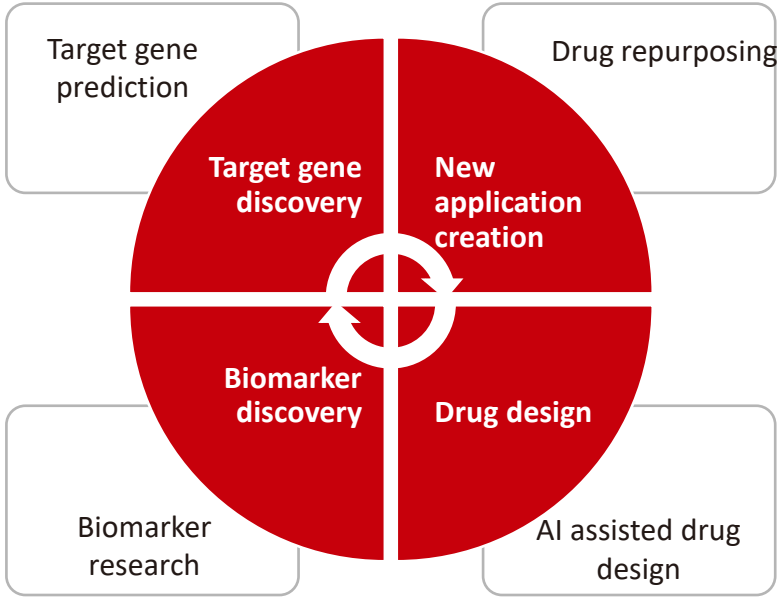
Plate cost reduction

40 million/year

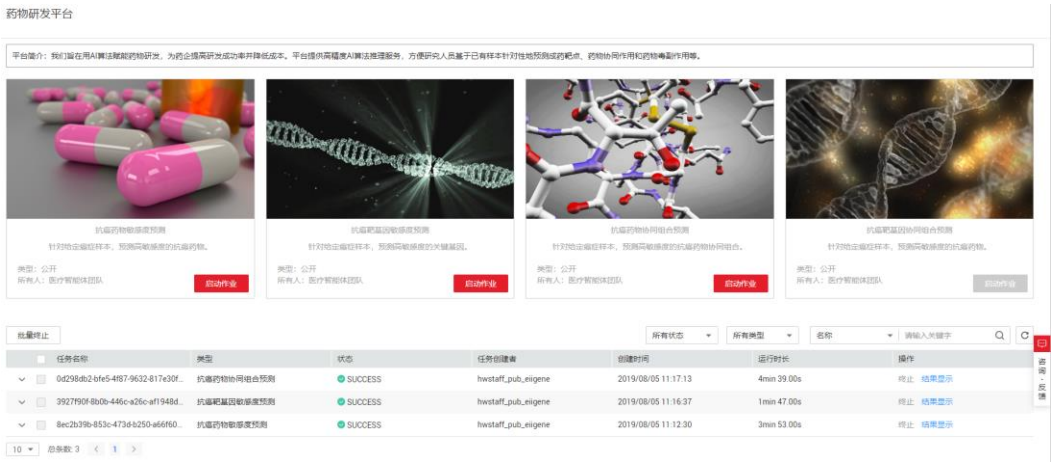
Success Stories of EIHealth

Scenarios

Hospital **Pharmaceutical** **Research institutes**



Real-time inference - AI prediction



DevEnviron - Notebook analysis

HUAWEI

Success Stories of Financial Intelligent Twins

Huawei Cloud

OCR

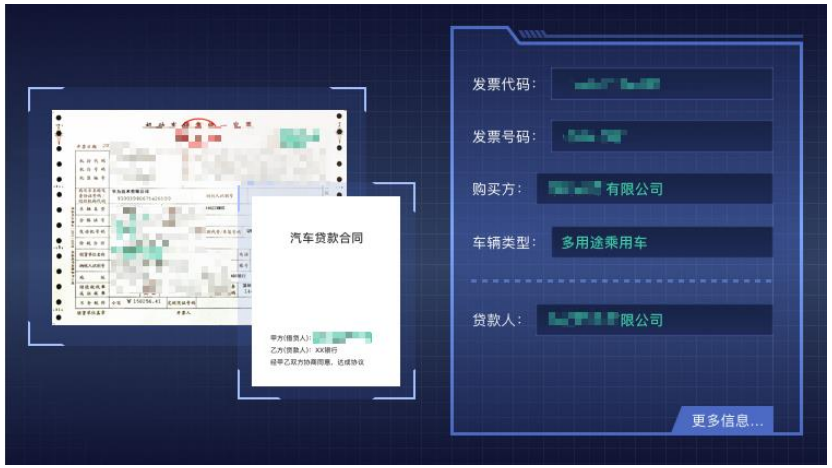
API call

Automatic identification of insurance policies

Automatic identification and recording of the text in ID cards, bank cards, and medical documents
Handling problems such as misaligned lines, overlapped words, and seal interference in medical records, and applying to various complex scenarios.

Claim information processing time
is reduced from 1*x* minutes to seconds.

Self-service appliance



>95%

Overall recognition accuracy > 95%; average recognition time < 1s

*Over **10-time** increase*

Over 10x efficiency in service processing, faster flow of letter of credit (LC) fund

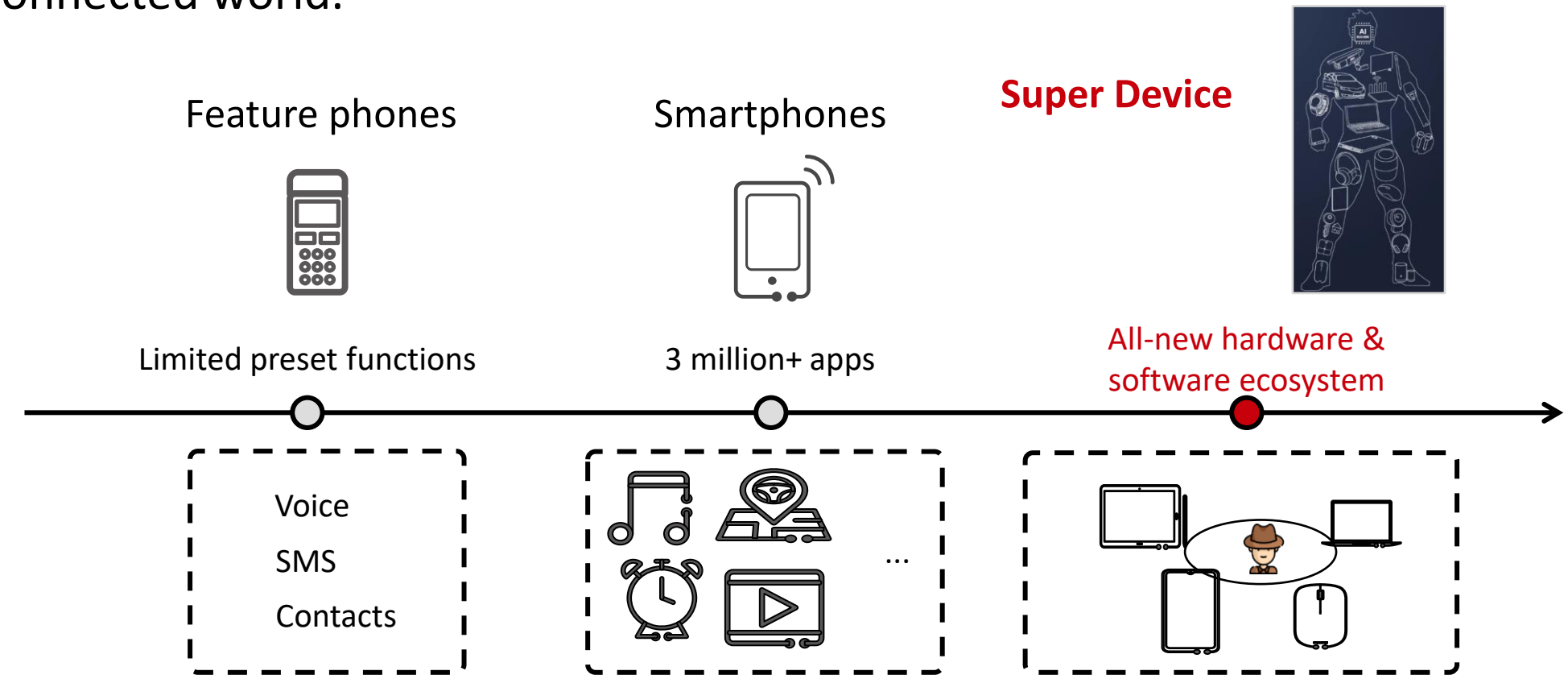
High-precision recognition of images with poor quality, such as rotation and wrinkles

HUAWEI

Contents

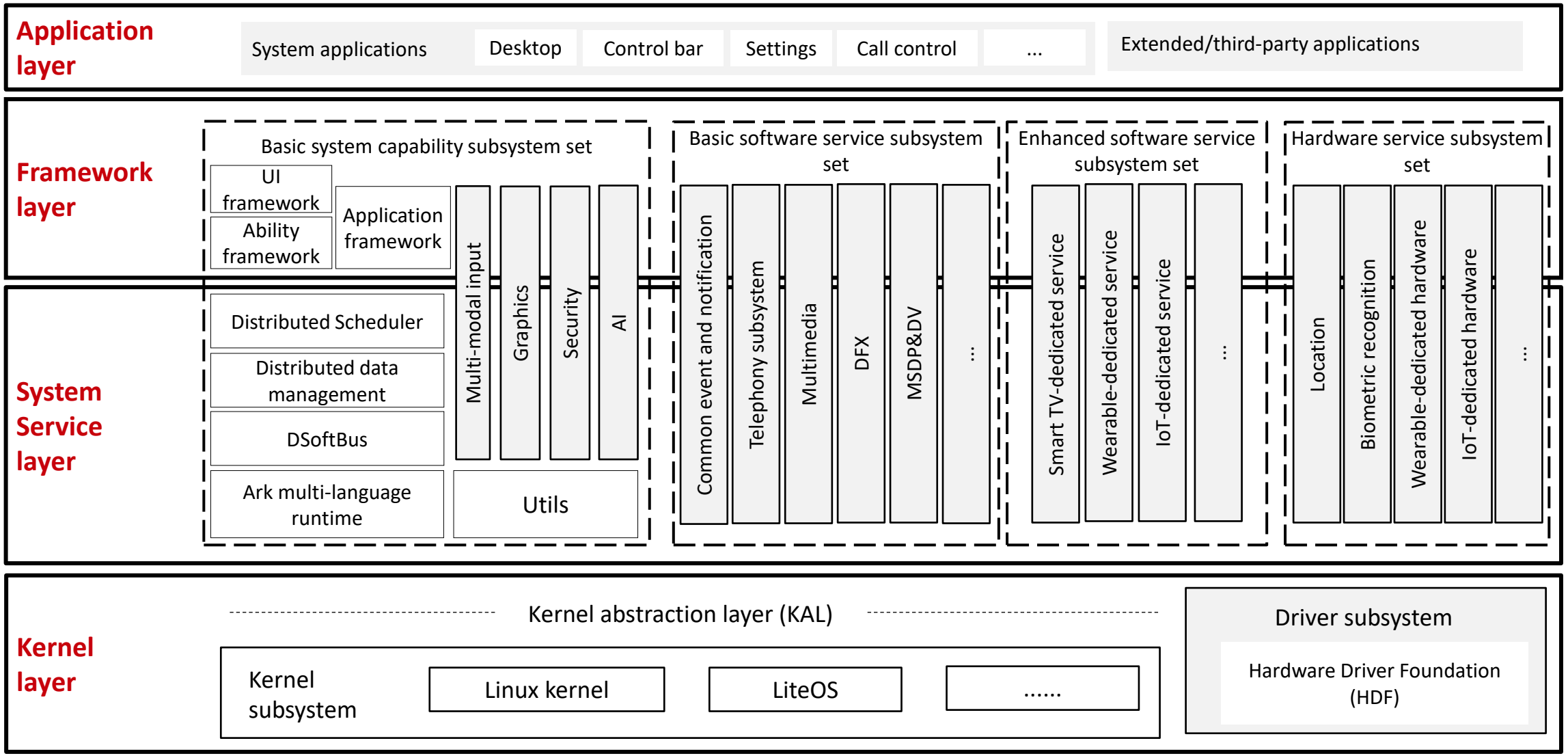
Introduction to HarmonyOS

- HarmonyOS is an innovative and distributed operating system designed for an interconnected world.



HUAWEI

HarmonyOS Architecture



HUAWEI

Key Capabilities of HarmonyOS

Service Widgets

Direct home screen access to application content.

Distributed Capabilities

Distributed soft bus, device virtualization, data management, and task scheduling.

IDE Tools

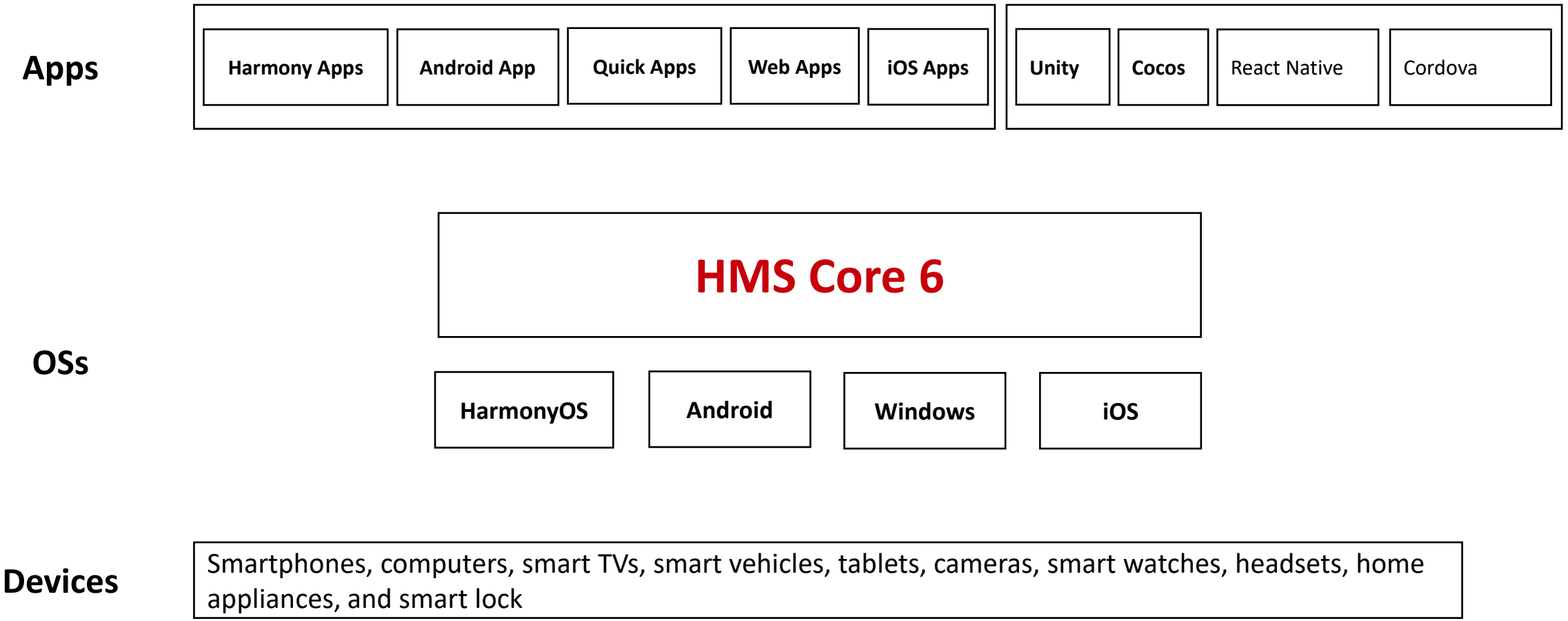
End-to-end HarmonyOS application development services, including project template creation, development, compilation, debugging, and release.

HMS Core

HMS Core provides open device and cloud capabilities to help developers achieve efficient app development and rapid growth.

Contents

HMS Core Introduction



Capabilities in Seven Domains Lay the Foundation for a Fully Connected Intelligent App Ecosystem

Efficiency

Quick application development for efficient connections

App Service

- Account Kit
- Ads Kit
- Analytics Kit
- Location Kit
- Awareness Kit
- Game Service
- App Linking
- Identity Kit
- In-App Purchases
- Business Touch Kit
- DCI Kit
- Dynamic Tag Manager
- SignPal Kit
- Map Kit
- Navi Kit
- Push Kit
- Quick apps
- Scan Kit
- Wallet Kit
- Search Kit
- Location Kit
- Weather Kit
- Health Kit
- UI Engine
- Drive Kit
- Membership Kit

Reliability

Secure applications for reliable connections

Graphics

- Accelerate Kit
- AR Engine
- VR Engine
- Computer Graphics Kit
- Scene Kit
- GameTurbo Engine
- 3D Modeling Kit
- 3D Engine

AI

- ML Kit
- HiAI Foundation
- HiAI Engine
- HiAI Service

Value

Continuously growing application value for business connections

Media

- Audio Kit
- Audio Engine
- Image Kit
- Video Kit
- Video Engine
- WisePlay DRM
- Camera Engine
- Panorama Kit
- Audio Editor Kit
- AV Pipeline Kit
- Video Editor Kit

System

- hQUIC Kit
- LinkTurbo Engine
- Nearby Service
- Network Kit
- MDM Engine
- HEM Kit
- Haptics Engine
- 5G Modem Kit

Intelligence

Intelligent applications for smart connections

Security

- FIDO
- Safety Detect
- Keyring
- LocalAuthentication Engine
- DataSecurity Engine
- iTrustee TEE

Smart Device

- CaaS Engine
- Cast Engine
- DeviceVirtualization Engine
- OneHop Engine
- Share Engine
- Wear Engine
- HUAWEI HiCar
- Pencil Engine

IDE/Tools — **Reality Studio** — **Theme Studio** — **Graphic Profiler** — **HMS Toolkit** — **Quick App IDE**

Contents

ML Kit Facilitates AI App Development With Brand New Experience

Text-related	Language/Voice-related	Image-related	Face/body-related	NLP	Custom models
Text recognition Document recognition Card recognition Form recognition	Translation Language detection Sound detection Text to speech (TTS) Automatic speech recognition (ASR) Audio file transcription Simultaneous interpretation (New)	Image classification Object detection Landmark recognition Image segmentation Product visual search Document skew correction Text-image super-resolution (TISR)	Face detection/verification Hand gesture recognition Skeleton detection Interactive/static biometric verification	Real-time text extraction Text embedding	Custom image classification Custom text classification Custom object detection

Core advantages

Wide device model coverage

All Android, iOS, and HarmonyOS devices are supported.

Global coverage

ML Kit is available around the globe.

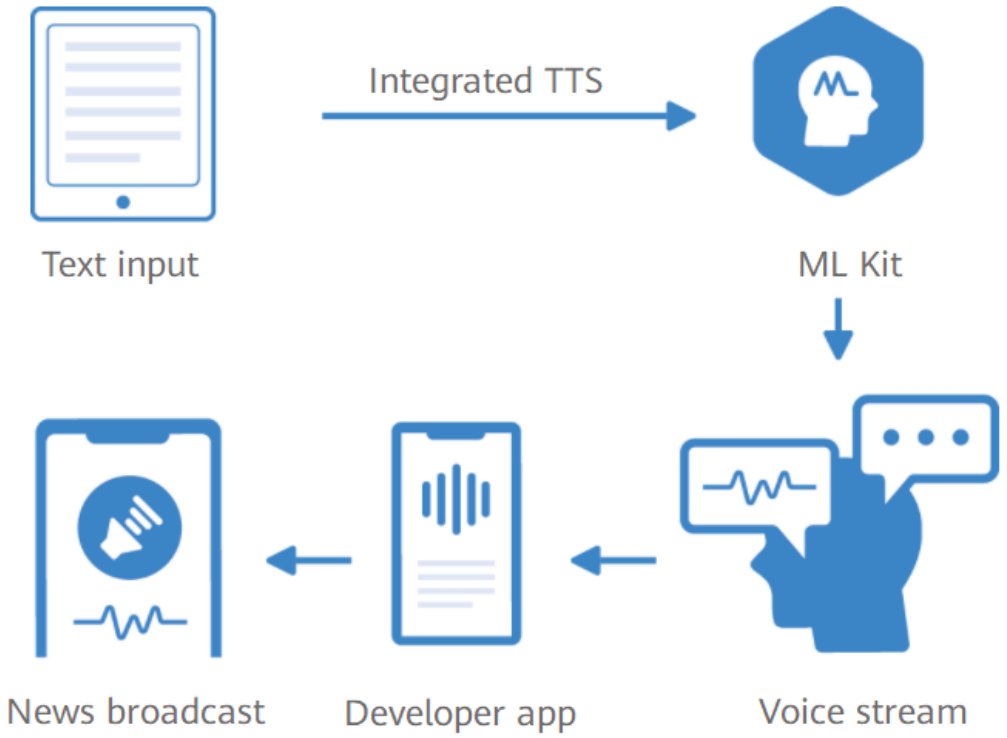
Data security

On-device data is not uploaded to the cloud. On-cloud data is stored and operated independently.

Customizable models

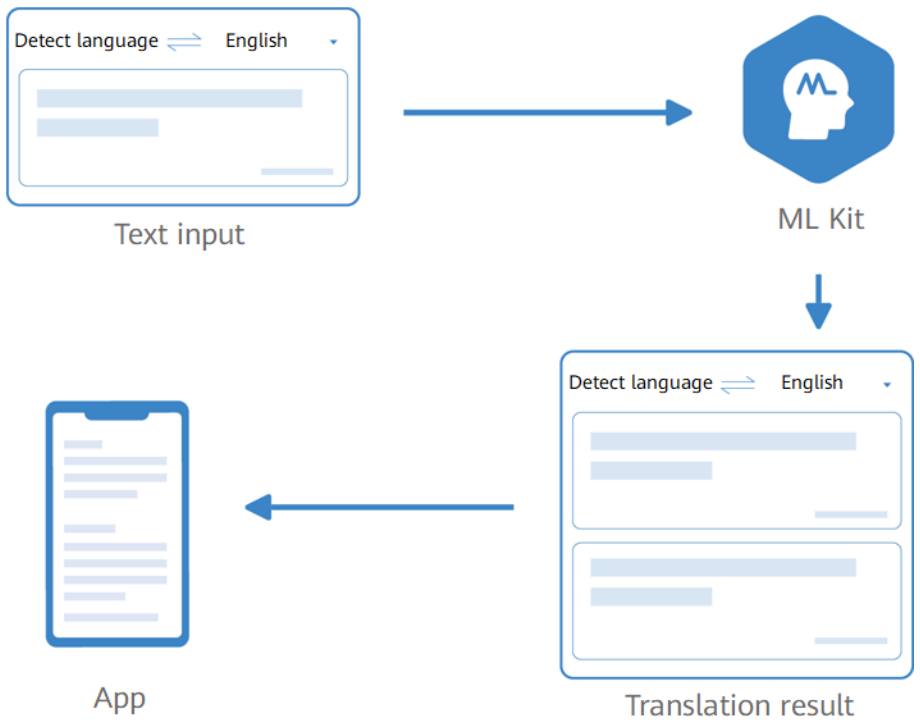
You can train and generate custom AI models to meet actual application requirements.

Text to Speech



- Supports Chinese, English, French, German, Italian, and Spanish
- Supports text containing both Chinese and English
- Free of charge for Huawei devices

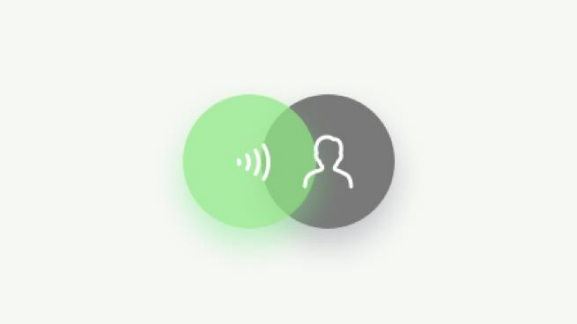
Translation



- Quickly detects 52 languages
- Online and offline translation
- Accurate and fast AI enablement
- Template-based integration, which is easy to use

Contents

HUAWEI HiAI - A Fully Open Intelligent Ecosystem to Drive Developer Business Success



Cloud

Provides intelligent digital services for an expanding range of application scenarios.

HUAWEI HiAI Service

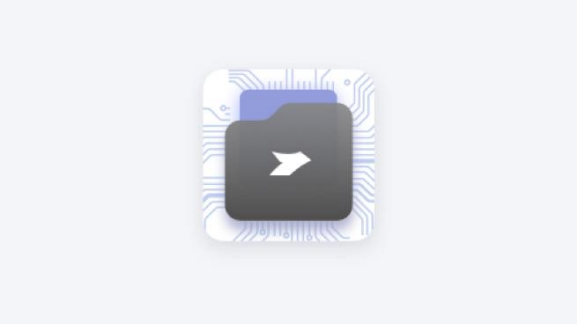
Open service capabilities better connect services and users.

Device

Creates an optimal user experience as a portal to the digital world.

HUAWEI HiAI Engine

Open application capabilities enable more intelligent and powerful apps.



Chip

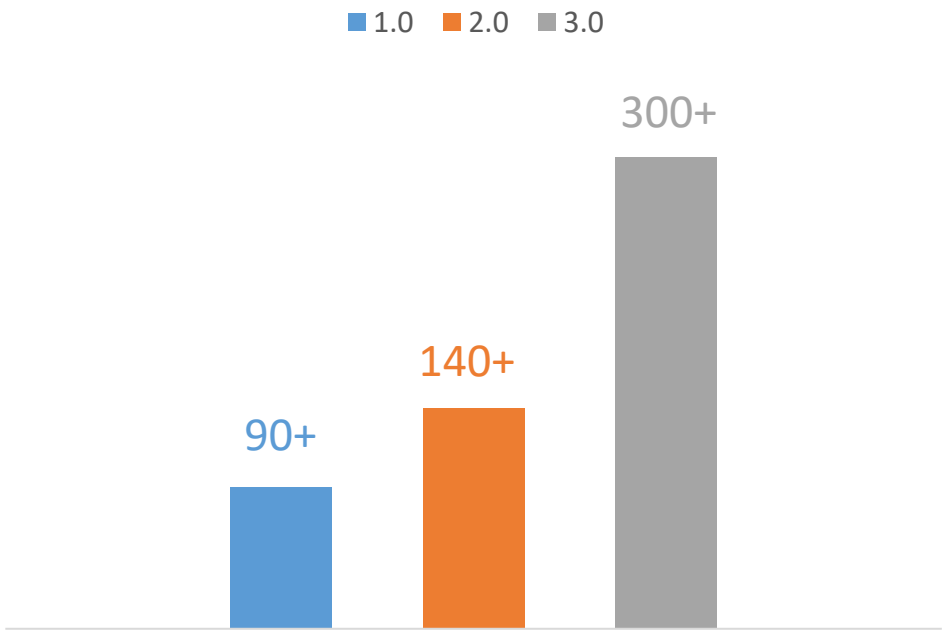
Achieves performance breakthroughs and better power efficiency.

HUAWEI HiAI Foundation

Open chip capabilities improve performance with NPU acceleration.

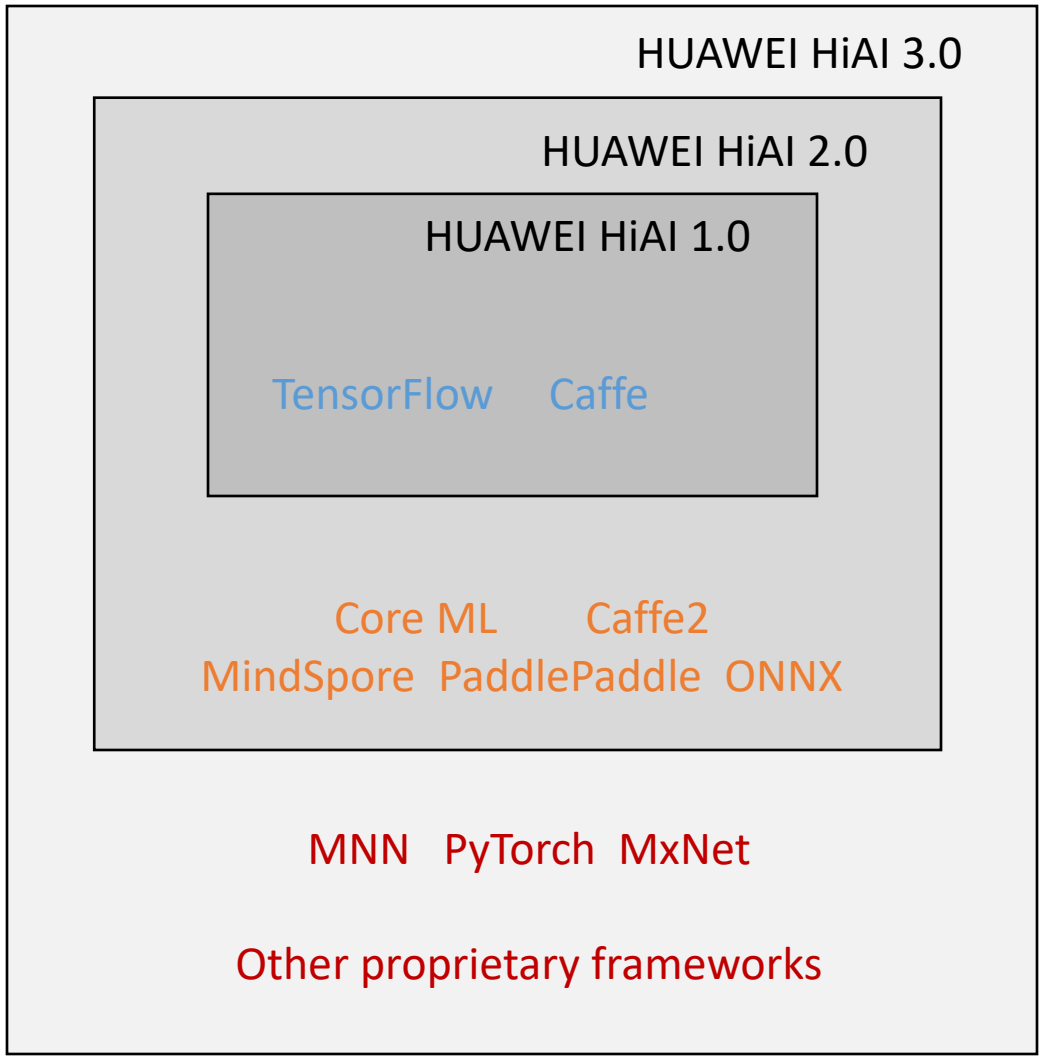
HUAWEI HiAI Foundation - Open Chip Capabilities Enable Powerful On-Device Computing

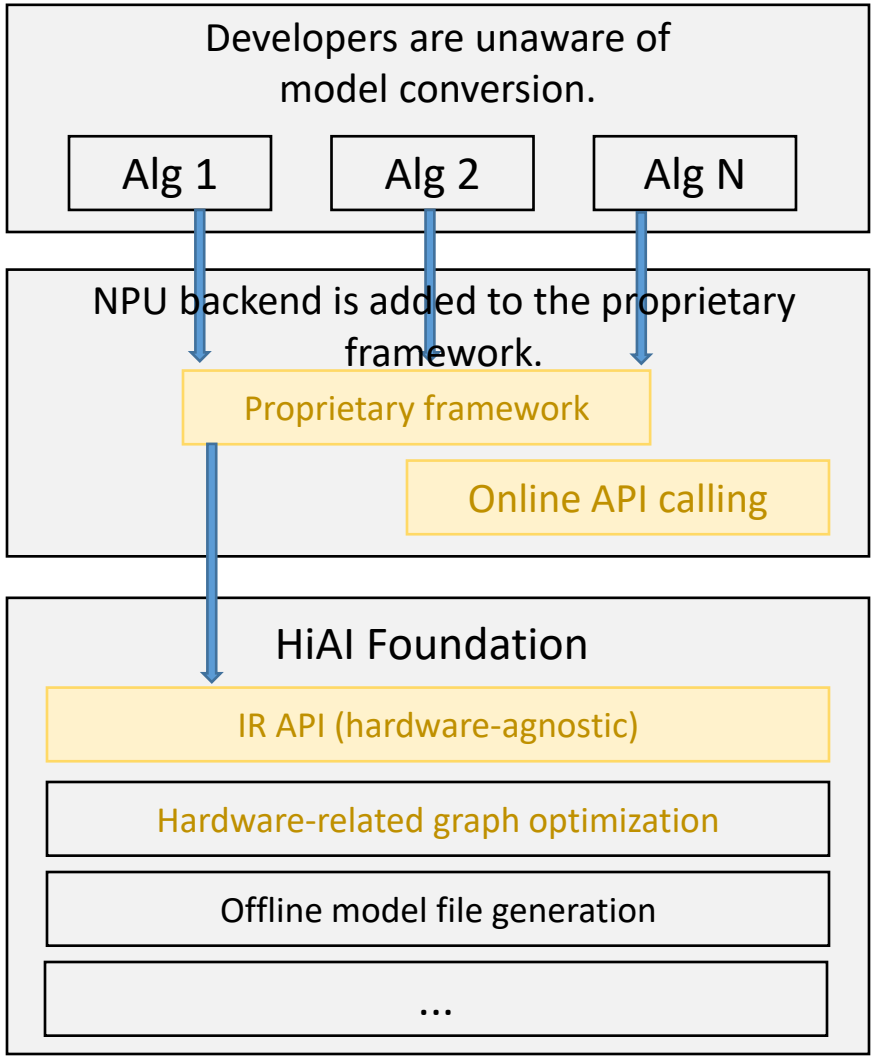
- Open chip capabilities are providing more and more operators for developers.



Operator Type	Number of Operators	Typical Operators
NN computing	60+	conv, deconv, pool
Mathematical operation	50+	sin, cos, add, mul, mean
Array operation	50+	concat, reverse, batch_to_space
Image operation	5	crop_and_resize, resize_bilinear
Logic control	30+	logicand, logicor, logicnot

Excellent Framework Compatibility Through IR APIs

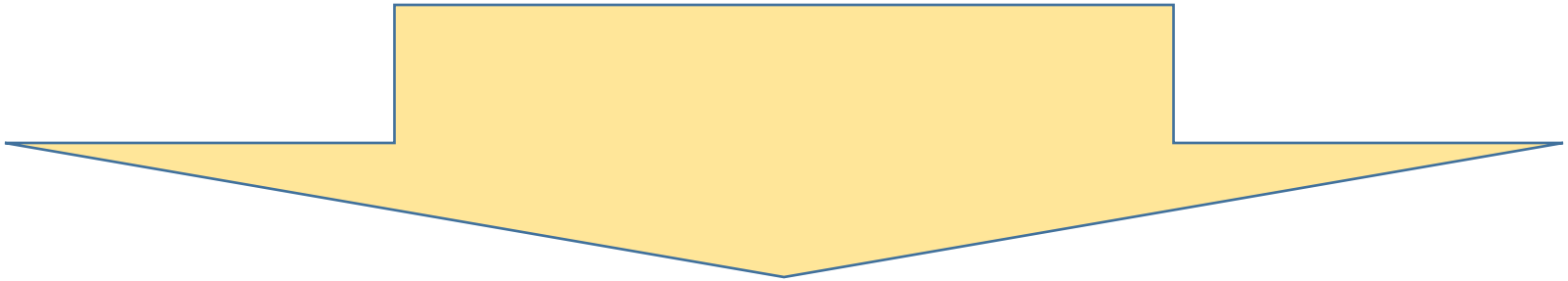




HUAWEI

AI Capabilities for Various Devices

Cars, computers, large screens, tablets, smartphones, earphones, VR glasses, speakers, and smart watches

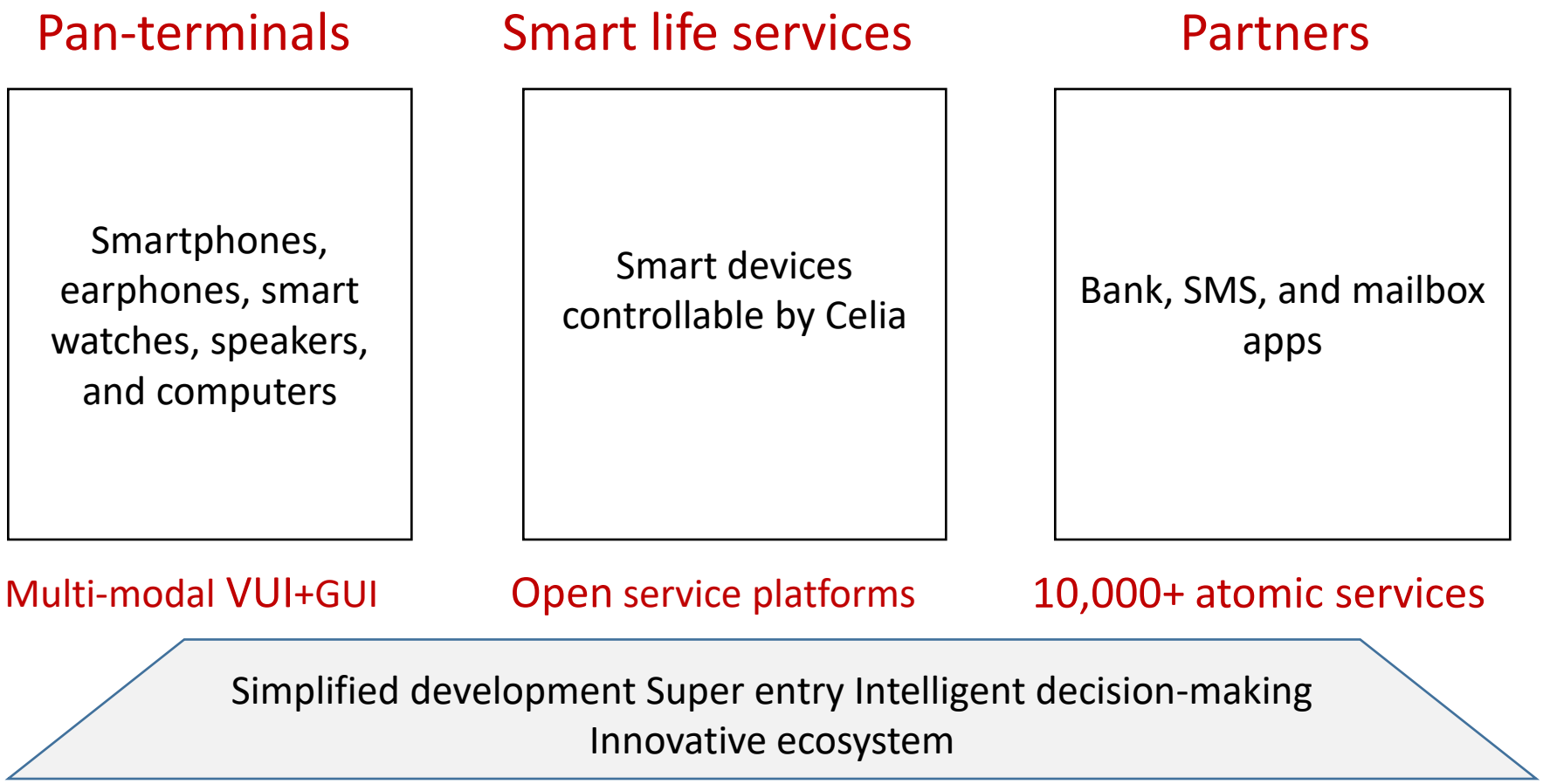


Honghu processors, Kirin processors, and AI camera processors

HUAWEi HiAI Engine – Rich AI Capabilities and Ultimate Experience Out of the Box

CV						NLU	ASR
Text recognition	**Image recognition**	**Facial recognition**	**Body recognition**	**Video technology**	**Code recognition**	**NLP**	**Speech recognition**
General text recognition Form recognition Passport recognition ID card recognition Driving license recognition Vehicle license recognition Document converter Bank card recognition	Aesthetics score Image category labeling Image super-resolution Scene detection Document skew correction Text-image super-resolution (TISR) Portrait segmentation Semantic segmentation	Facial comparison Face detection Face parsing Face attribute recognition Face orientation recognition Facial feature detection	Key skeletal feature recognition Video portrait segmentation	Video summarization Video thumbnail	Code recognition	Word segmentation Part-of-speech tagging Assistant-specific intention recognition IM-specific intention recognition Keyword extraction Entity recognition	Speech recognition

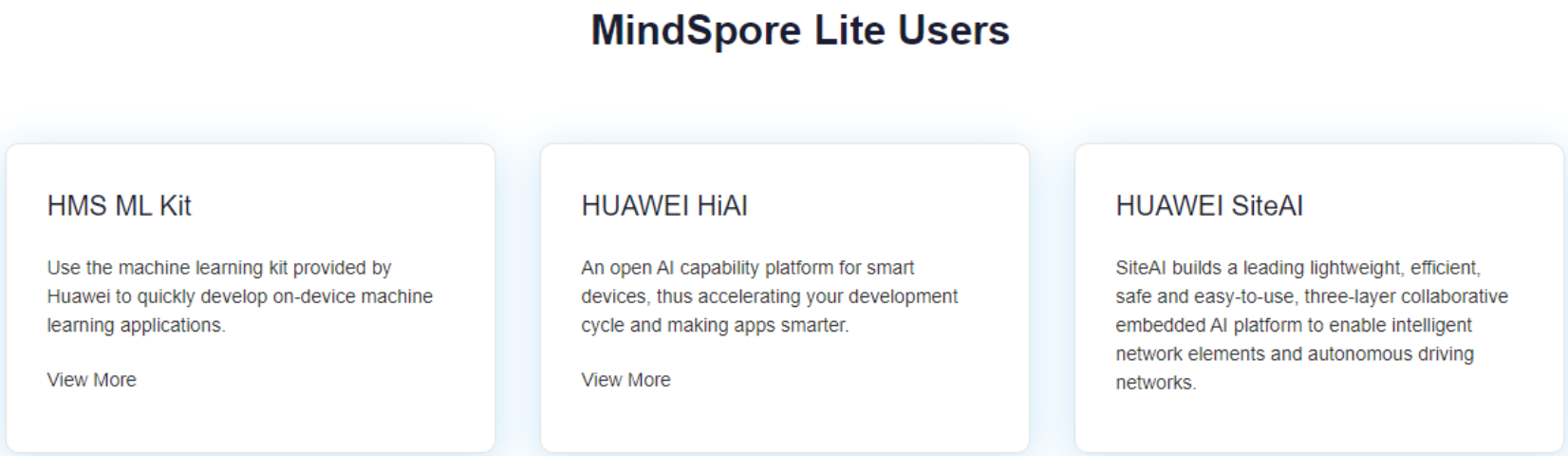
HUAWEI HiAI Service: One-Time Integration, Multi-Modal and Multi-Device Deployment



Contents

Introduction to MindSpore Lite

- MindSpore Lite is an ultra-fast, intelligent, and simplified AI engine that enables intelligent applications in all scenarios, provides end-to-end solutions for users, and helps users enable AI capabilities.



MindSpore Lite Features

Ultimate performance

- Provides efficient kernel algorithms and assembly-level optimization, and supports CPU, GPU, and NPU.
- Provides heterogeneous scheduling, maximizes hardware computing power, and minimizes inference latency and power consumption.

Lightweight

- Provides an ultra-lightweight solution to support model quantization and compression.
- Provides small AI models that run fast and can be quickly deployed in extreme environments.

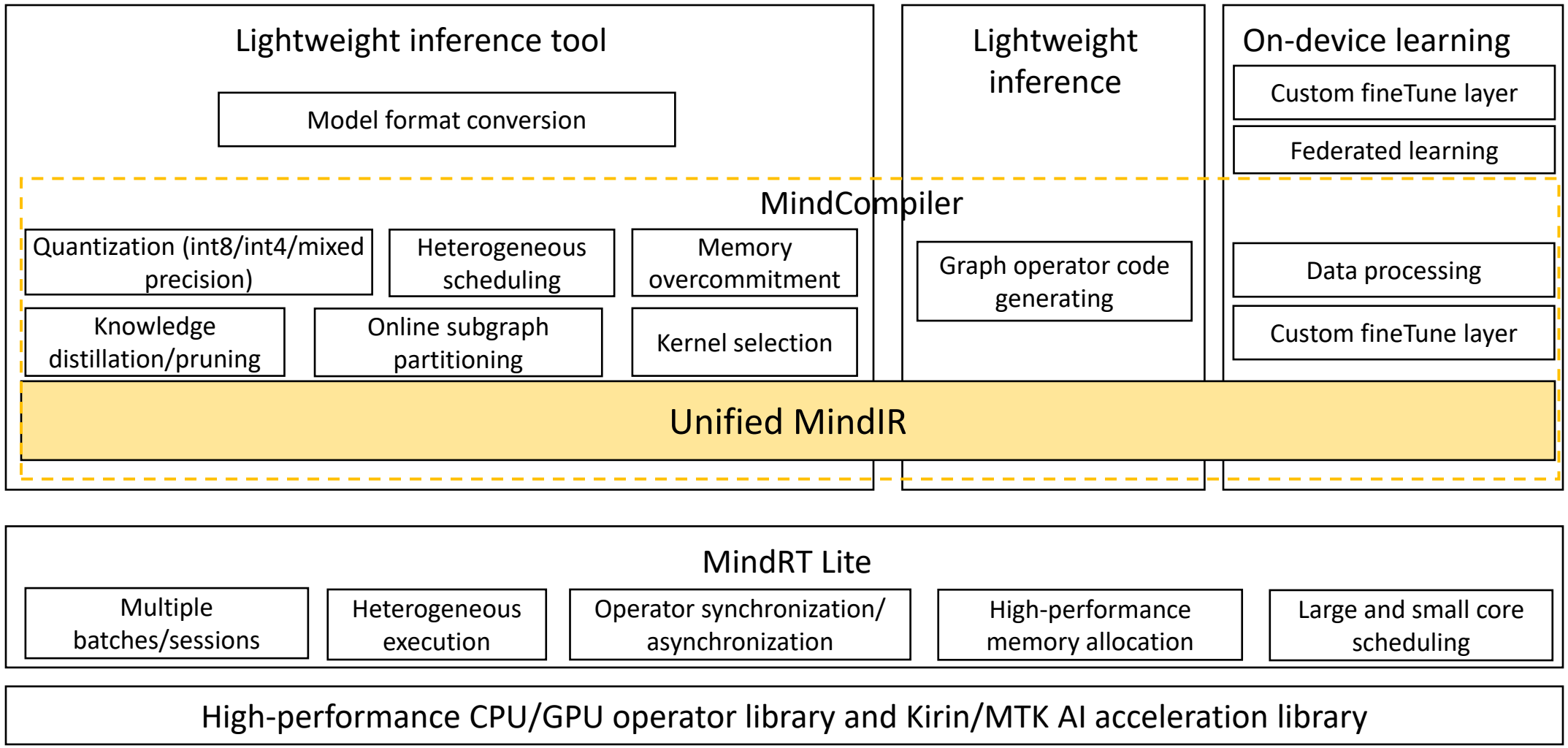
All-scenario support

- Supports mobile phone operating systems such as iOS and Android, LiteOS embedded operating system, and AI applications on various intelligent devices such as mobile phones, large screens, tablets, and IoT devices.

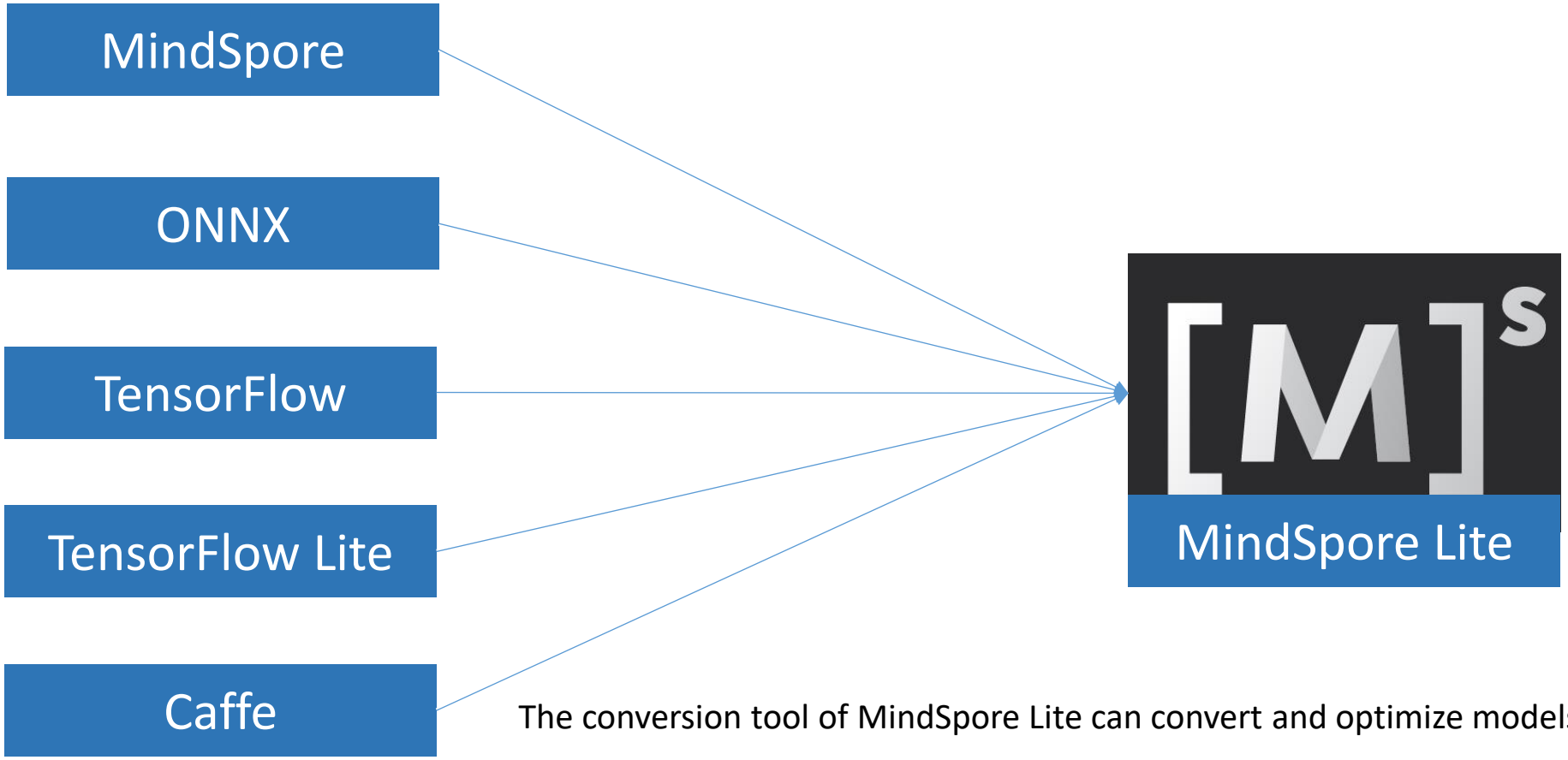
Efficient deployment

- Supports models such as MindSpore, TensorFlow Lite, Caffe, and ONNX, provides capabilities such as model compression and data processing, and supports unified training and inference IR, facilitating quick deployment.

MindSpore Lite Architecture



Compatibility With Mainstream AI Frameworks



The conversion tool of MindSpore Lite can convert and optimize models of mainstream AI frameworks to seamlessly support on-device learning and high-performance inference of models trained by MindSpore.

Quiz

1. What is the architecture used by the Ascend processor?
2. What is the functional unit responsible for image preprocessing in CANN?
3. Does using Huawei Cloud EI require programming basics?
4. How many AI capabilities does HMS provide?

Summary

- This chapter introduces Huawei's full-stack AI solution, including hardware products (Ascend processors and Atlas series devices), software (CANN), public cloud service (Huawei Cloud EI), and on-device products (ML Kit and HiAI).

Recommendations

- Ascend Developer Community
  - https://www.hiascend.com/
- Official MindSpore website
  - https://www.mindspore.cn/en
- HarmonyOS application development official website
  - https://developer.harmonyos.com/

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