Huawei's AI Development

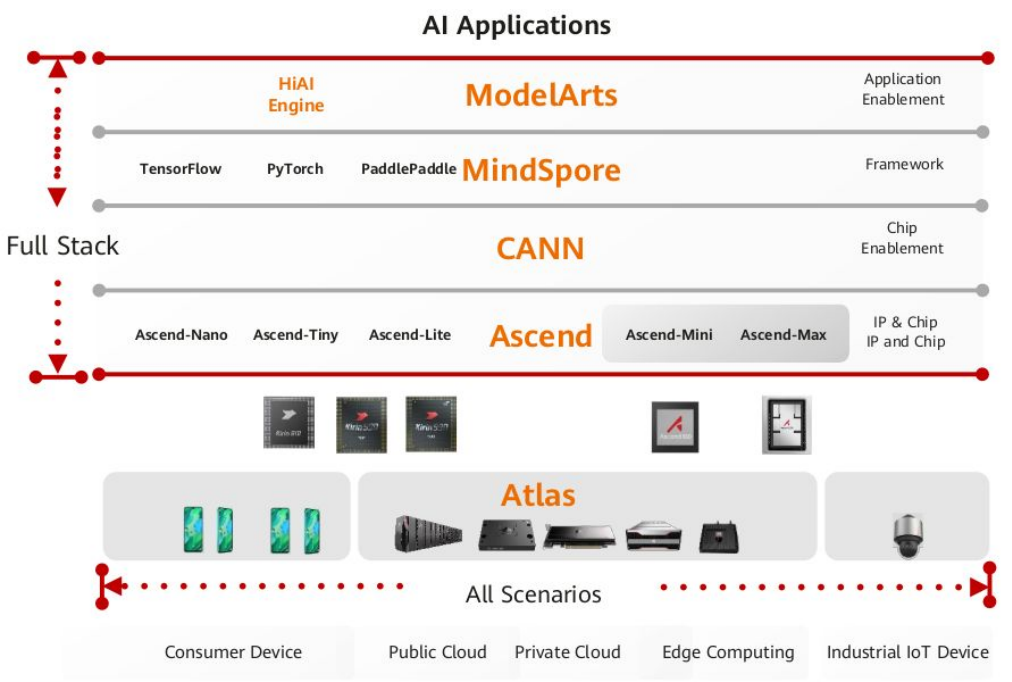
Thiago Alves Rocha
thiago.alves@ifce.edu.br

Full-Stack, All-Scenarios

Huawei's full-stack portfolio includes chips, chip enablement, a training and inference framework, and application enablement.

Huawei’s all scenarios means different deployment scenarios for AI, including public clouds, private clouds, edge computing, industrial IoT devices, and consumer devices.

Full-Stack, All-Scenarios



Application enablement: provides end-to-end services (ModelArts), layered APIs, and pre-integrated solutions.

MindSpore: supports the unified training and inference framework that is independent of the device, edge, and cloud.

CANN: a chip operator library and highly automated operator development tool.

Ascend: provides a series of NPU IPs and chips based on a unified, scalable architecture.

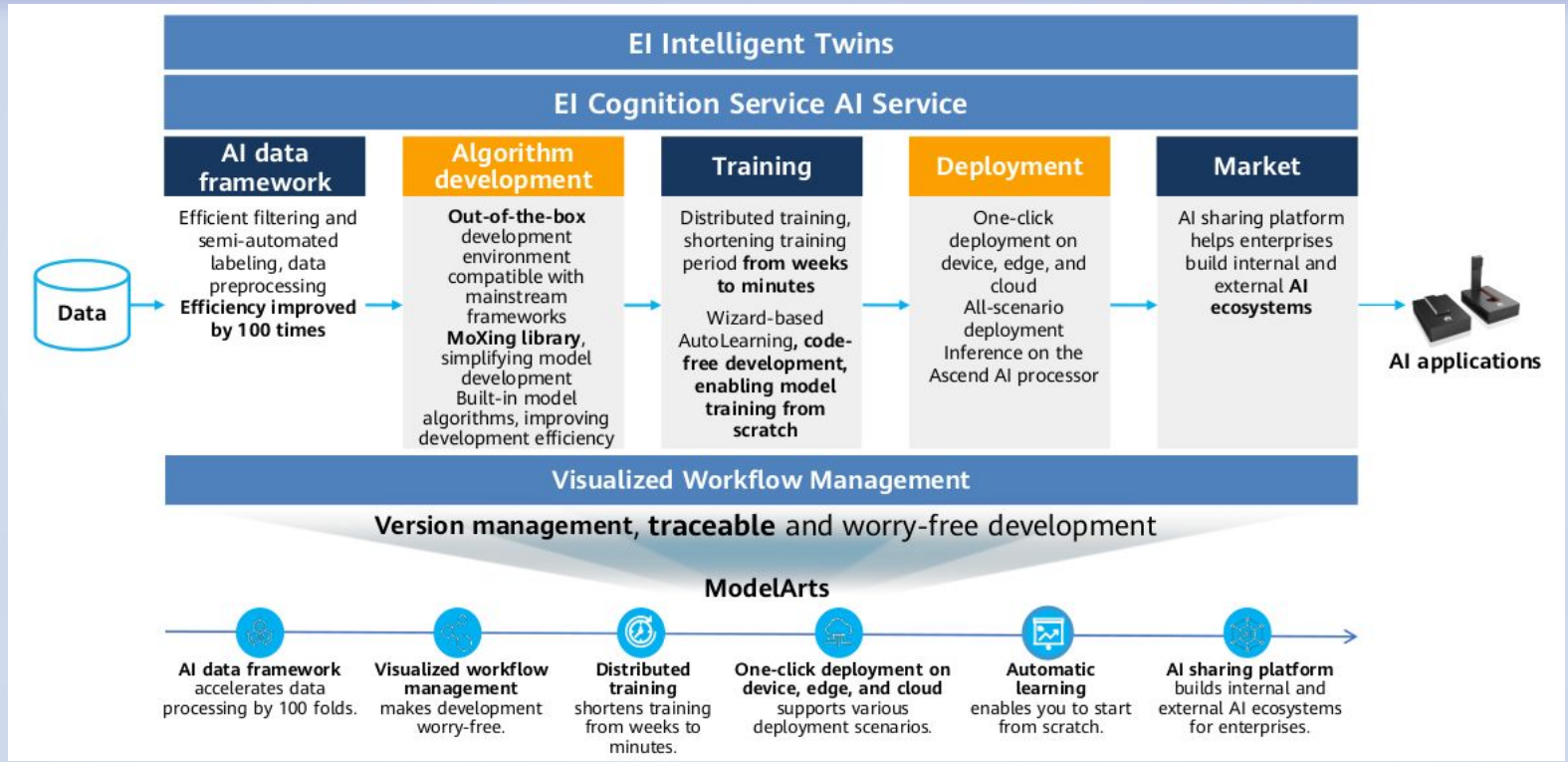
Atlas: enables an all-scenario AI infrastructure solution that is oriented to the device, edge, and cloud based on the Ascend series AI processors and various product forms.

ModelArts Full Cycle AI Workflow

ModelArts is a one-stop development platform for AI developers

ModelArts helps AI developers build models quickly and manage the lifecycle of AI development.

ModelArts Full Cycle AI Workflow

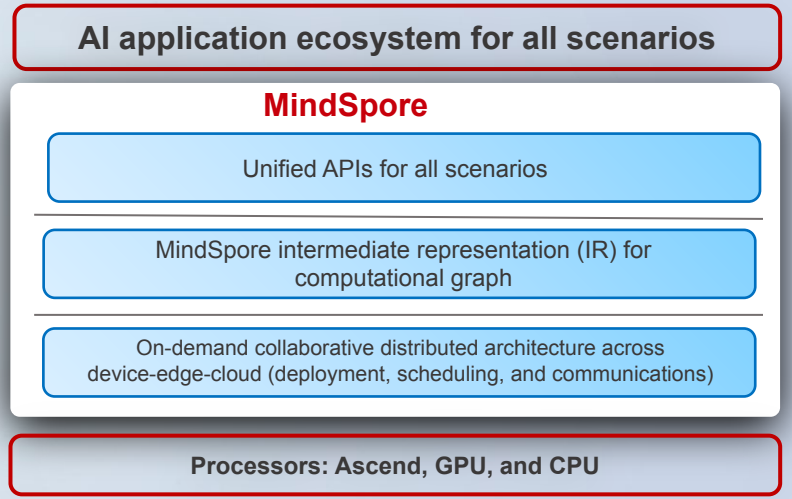


MindSpore

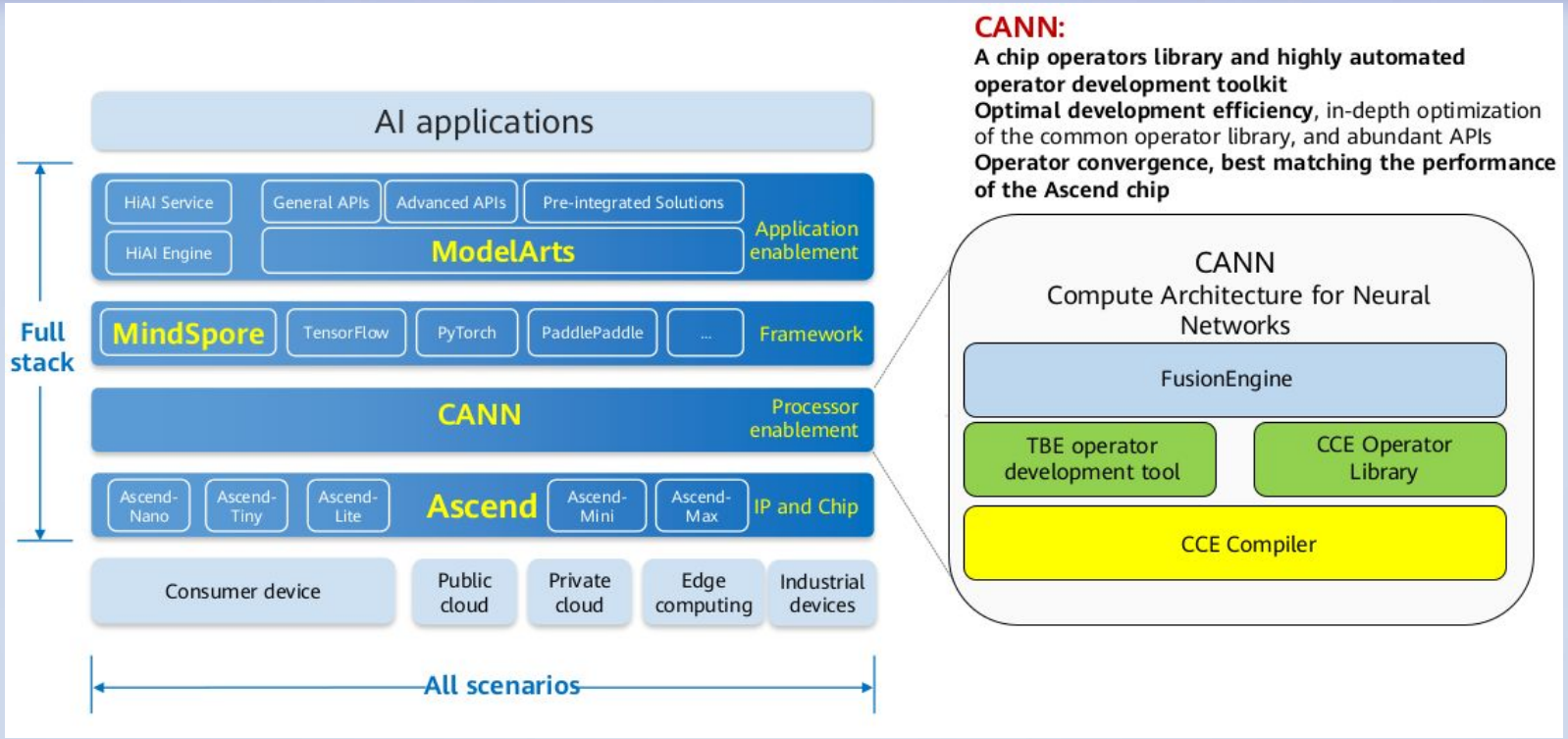
- Huawei's **open-source framework for AI development**.
- AI still faces huge challenges such as **technical barriers**, **high development** cost, and **long deployment period**.
- The all-scenario AI computing framework MindSpore is developed based on the principles of **friendly development**, **efficient operation**, and **flexible deployment**.

MindSpore

- MindSpore provides **automatic parallel** capabilities.
  - Run algorithms on dozens or even thousands of AI computing nodes with only a **few lines** of description.
- Supports large-scale and small-scale deployment.
- In addition to the Ascend AI processors, MindSpore also supports other processors such as GPUs and CPUs.



CANN - Computing Architecture



CANN

FusionEngine: Operator fusion, reducing memory transfer between operators, and improves the performance by 50%.

CCE operator library: The optimized general operator library provided by Huawei can meet the requirements of most mainstream vision and NLP neural networks.

Tensor Boost Engine: an efficient and high-performance custom operator development tool. It abstracts hardware resources as APIs, shortening the project duration.

CCE Compiler: bottom-layer compiler that optimizes performance and supports Ascend processors in all scenarios.

Ascend AI Processors

Demands for AI are soaring worldwide. However, with the market being dominated by only a few vendors, AI processors are sold at a very high price.

Ascend 310 processor for AI inference and Ascend 910 processor for AI training.

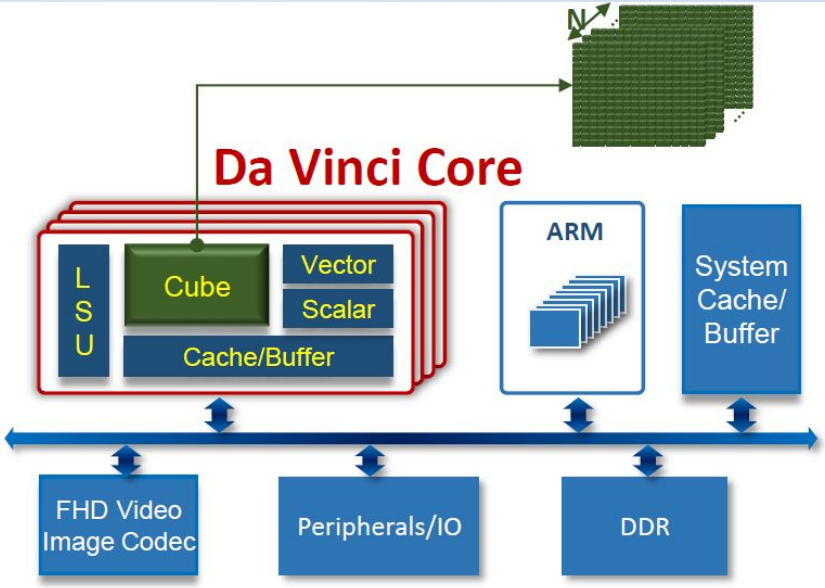
Built upon the unique Da Vinci 3D Cube architecture, Huawei's Ascend AI processors boast high computing power, energy efficiency, and scalability.

Ascend 310 provides up to 16 TOPS of computing power, with a power consumption of only 8 watts.

Ascend 910 AI processor delivers the industry's highest computing density on a single AI chip. It applies to AI training and delivers 512 TOPS of computing power, with a maximum power consumption of 310 watts.

Ascend 310 and Da Vinci Core

SPECIFICATIONS	Description
Architecture	AI co-processor
Performance	Up to 8T @FP16
	Up to 16T@INT8
Codec	16 Channel Decoder – H.264/265 1080P30 1 Channel Encoder
Memory Controller	LPDDR4X
Memory Bandwidth	2*64bit @3733MT/S
System Interface	PCIe3.0 /USB 3.0/GE
Package	15mm*15mm
Max Power	8Tops@4W, 16Tops@8W
Process	12nm FFC



Note: This is typical configuration, high performance and low power sku can be offered based on your requirement.

Ascend AI Processors

Ascend 310

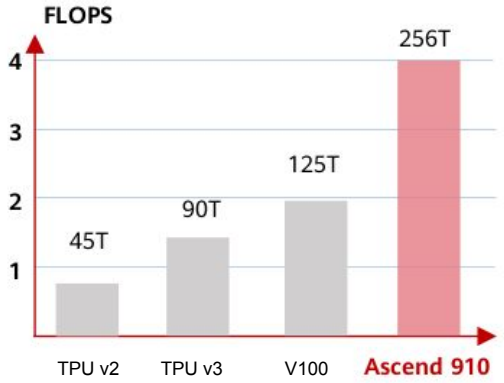
AI SoC with ultimate energy efficiency

- Ascend-Mini
  Architecture: Da Vinci
- **Half-precision (FP16): 8 TFLOPS**
- **Integer precision (INT8): 16 TOPS**
- **16-channel full-HD video decoder: H.264/265**
- 1-channel full-HD video encoder: H.264/265
- **Max. power: 8 W**

Ascend 910

Most powerful AI processor

- Ascend-Max
  Architecture: Da Vinci
- **Half-precision (FP16): 256 TFLOPS**
- **Integer precision (INT8): 512 TOPS**
- **128-channel full HD video decoder: H.264/265**
- **Max. power: 310 W**



Atlas AI Computing Platform Portfolio