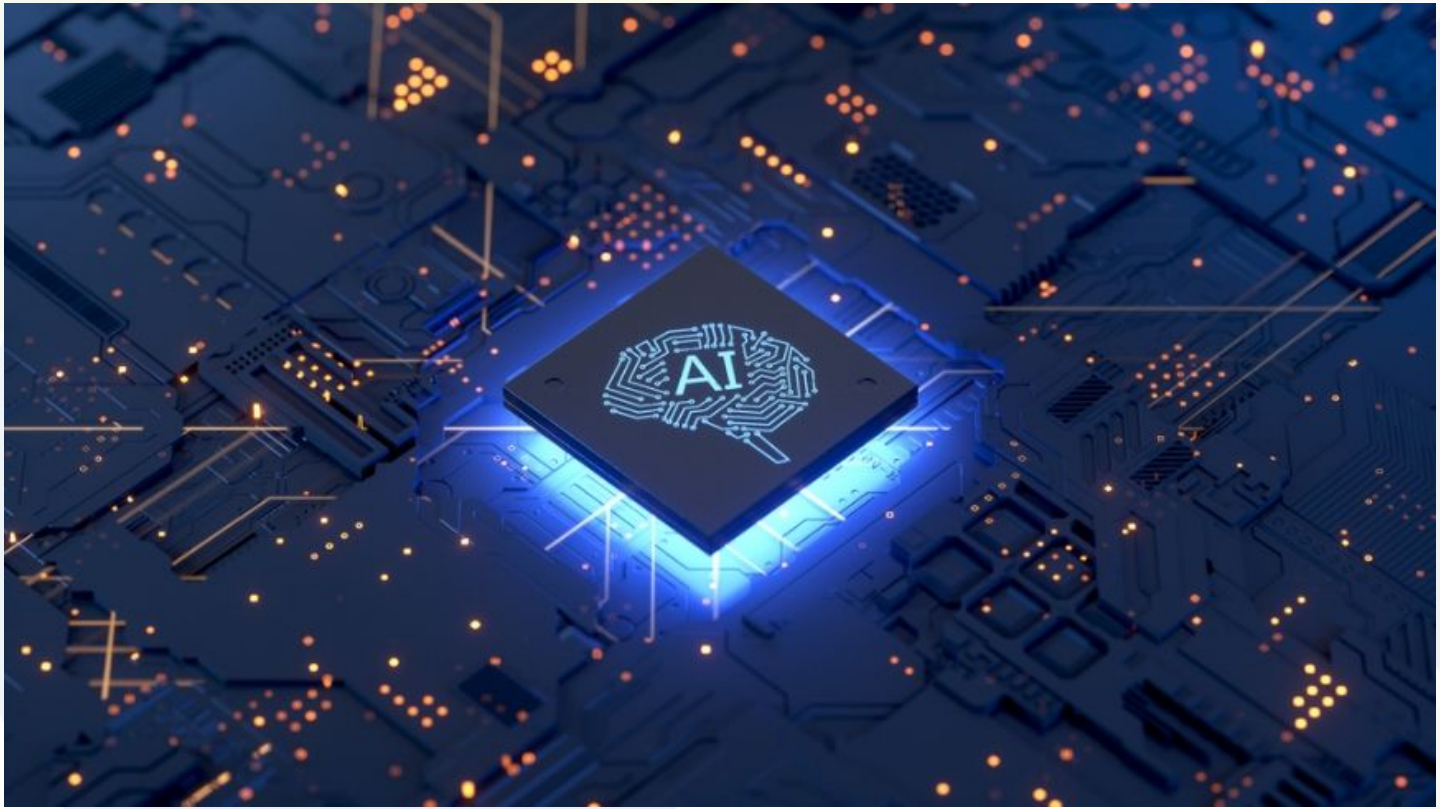
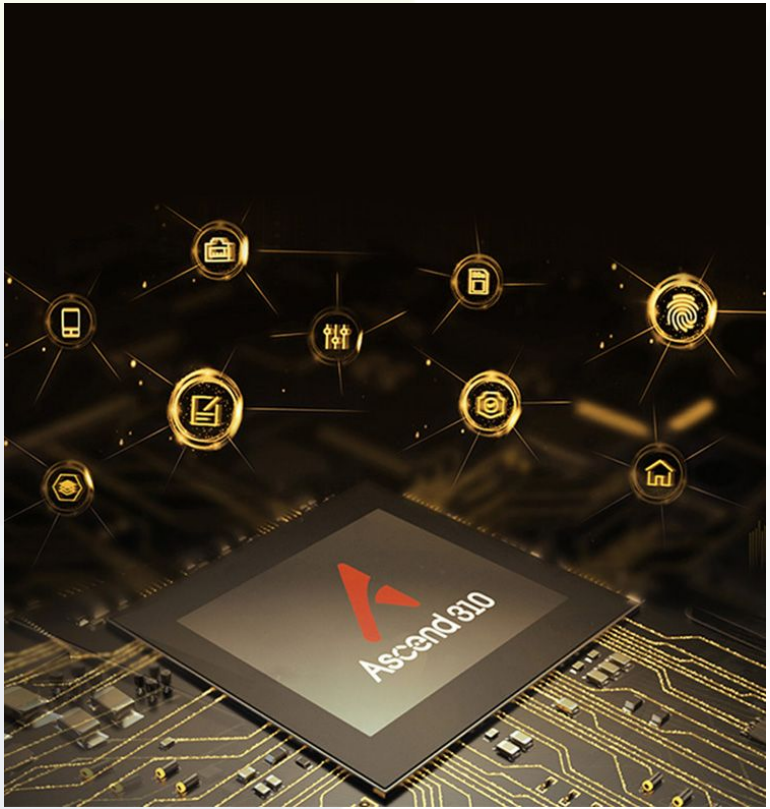
Huawei Ascend Chips

Prof. Carlos Moraes (Carlão)

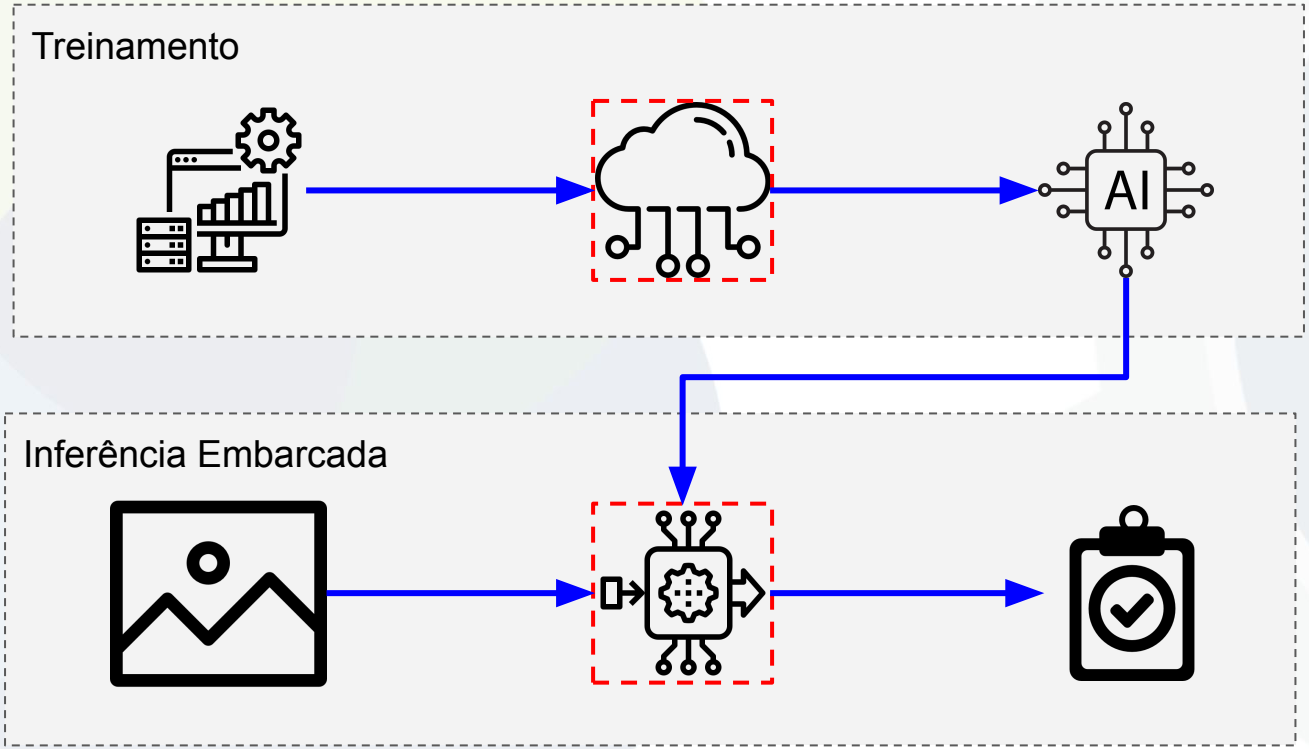
Inteligência Artificial para Sistemas Embarcados



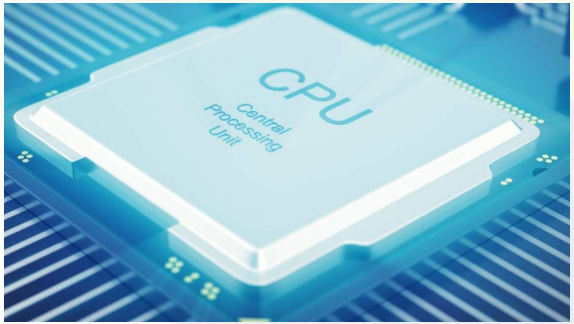
Microcontroladores (MCUs): Aceleração de I.A.

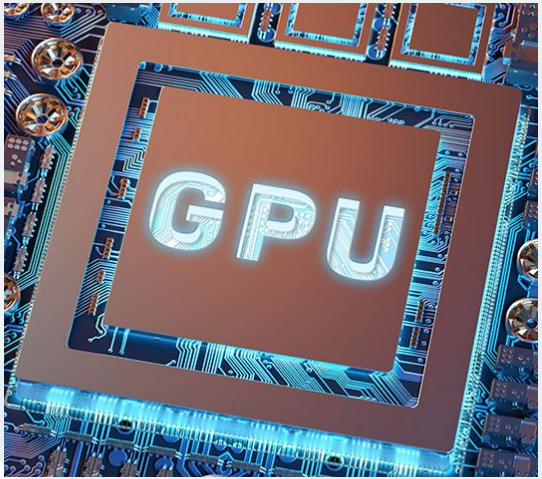


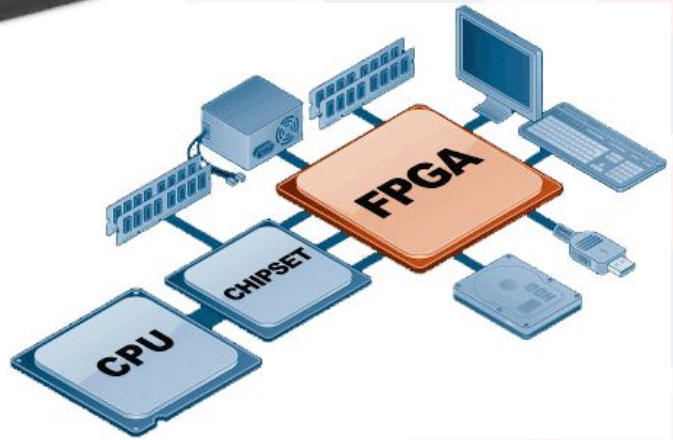
Treinamento e Inferência para Aplicações Reais



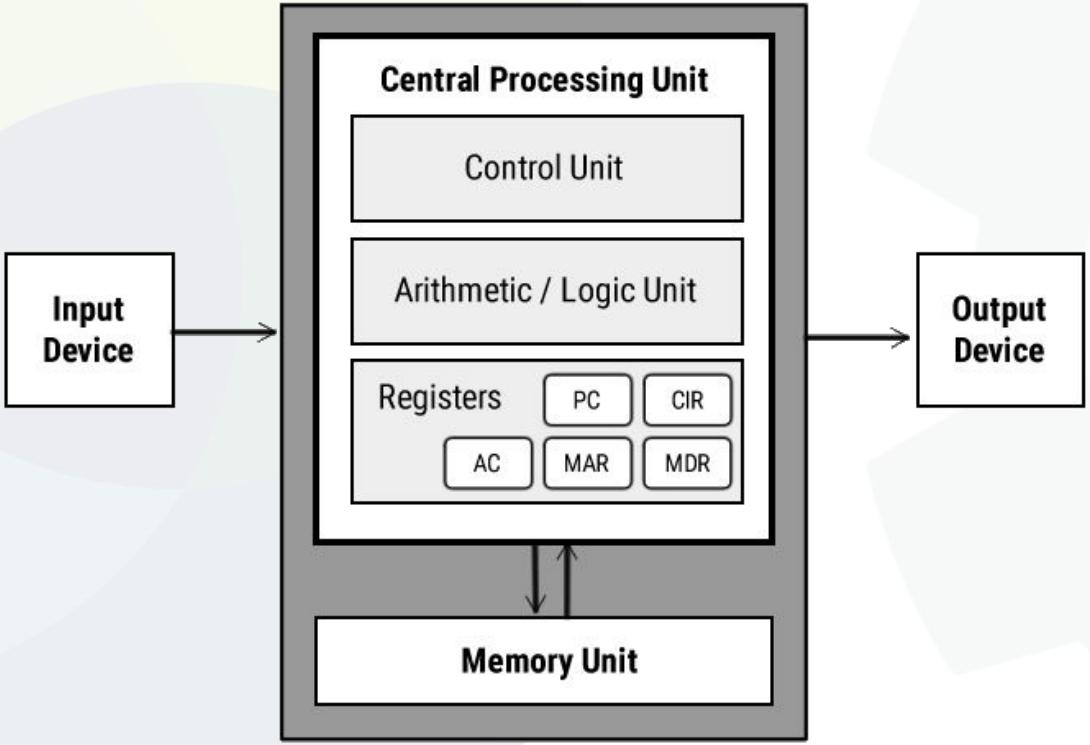
Classificação das Arquiteturas de Unidades com I.A.



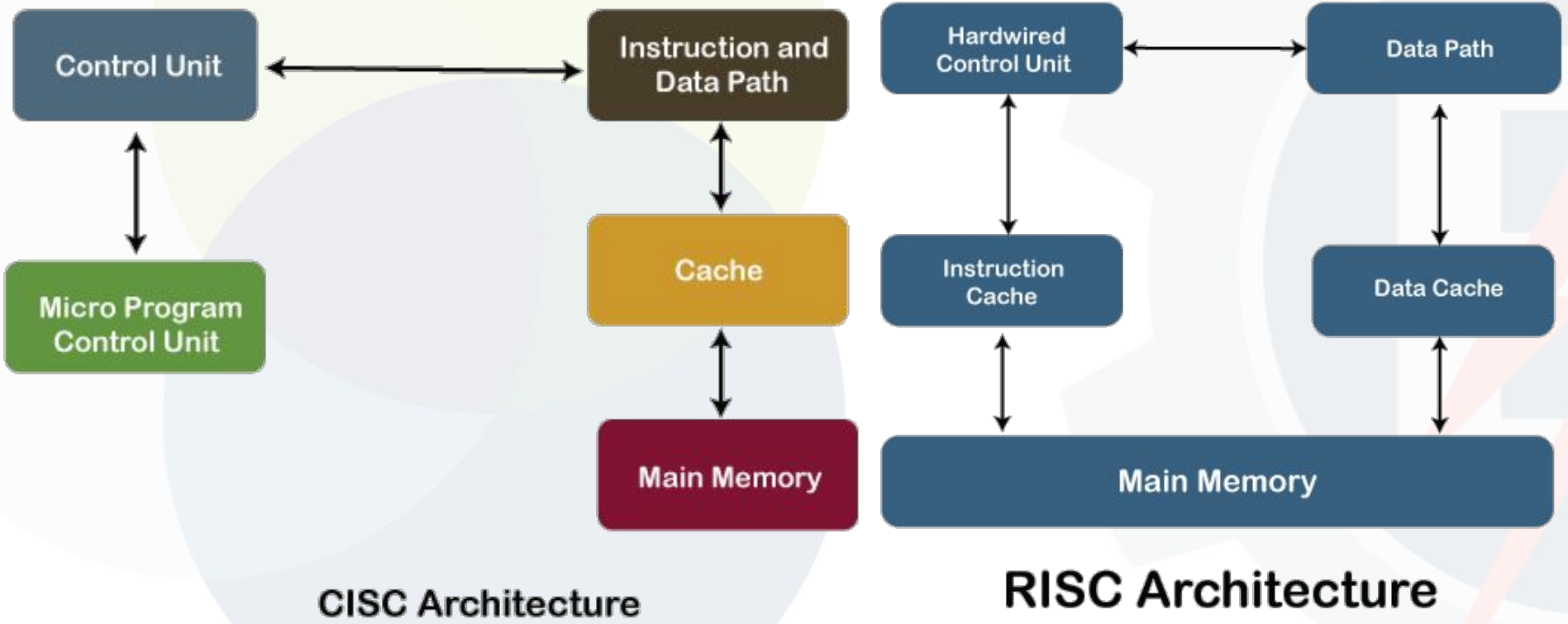




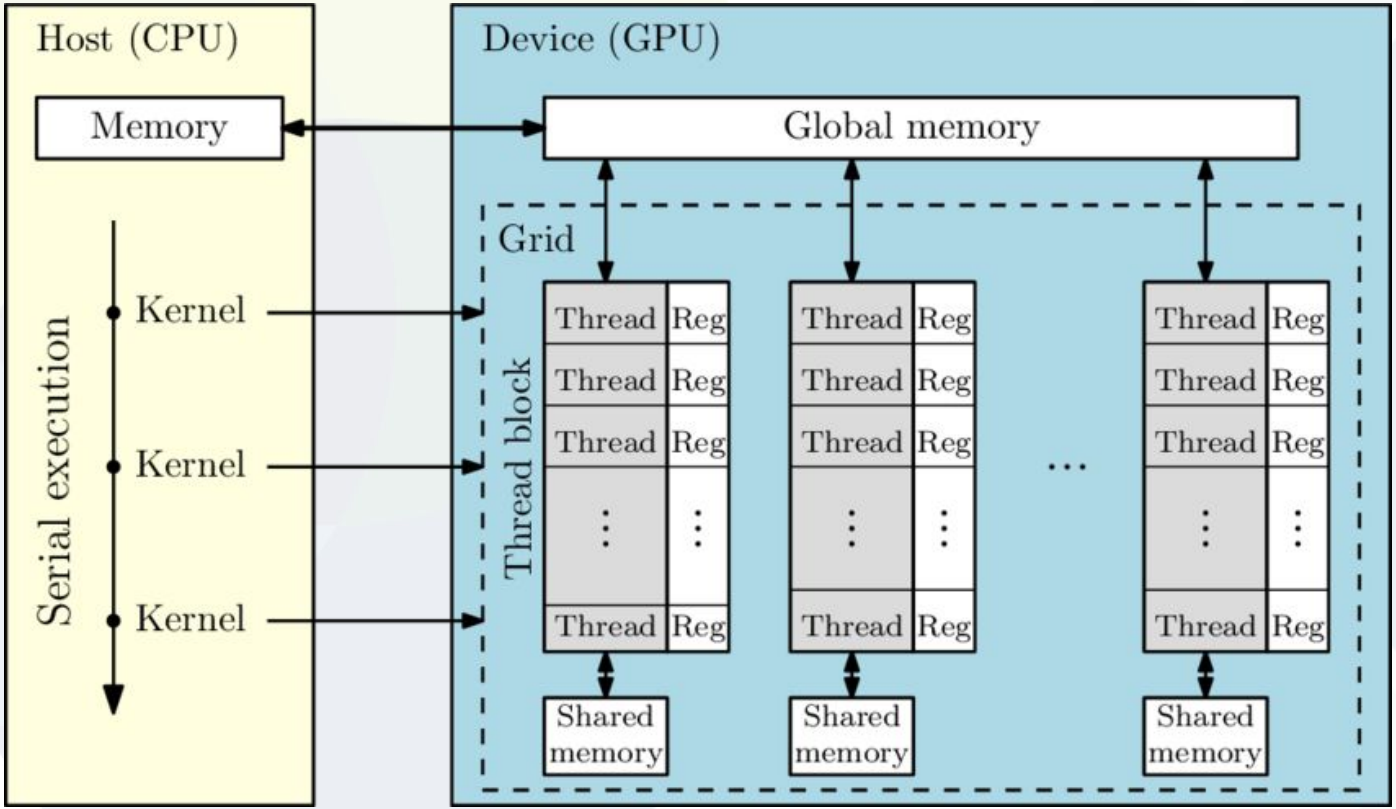
CPU


Central Processing Unit
Control Unit
Arithmetic / Logic Unit
Registers
PC
CIR
AC
MAR
MDR
Input Device
Output Device
Memory Unit

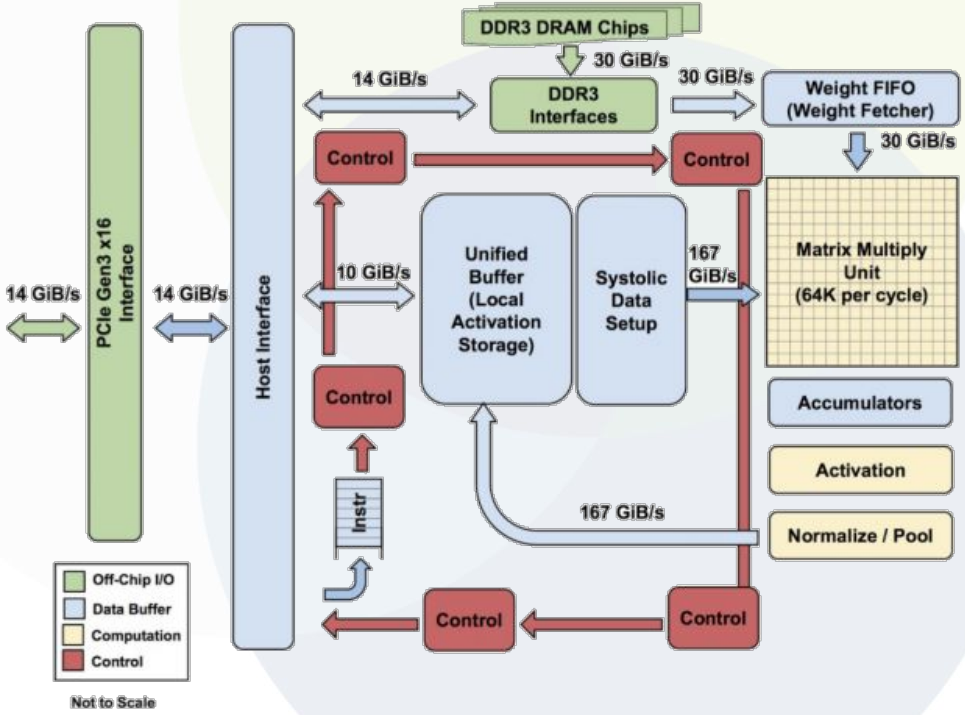
CPU: CISC (x86/x64) vs RISC (ARM)

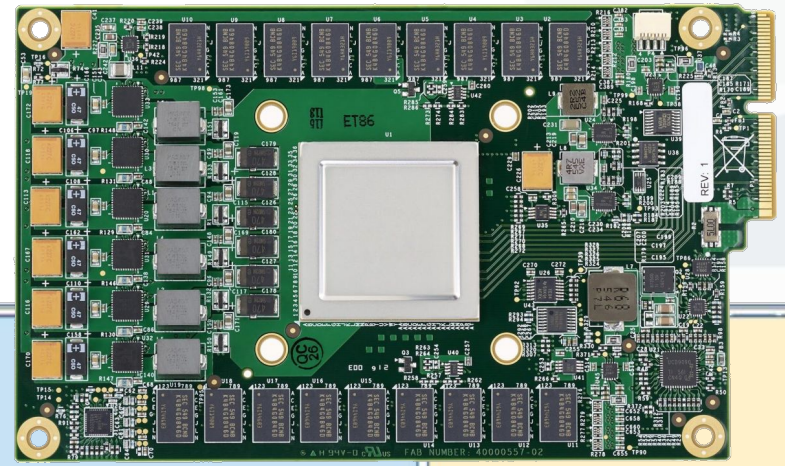


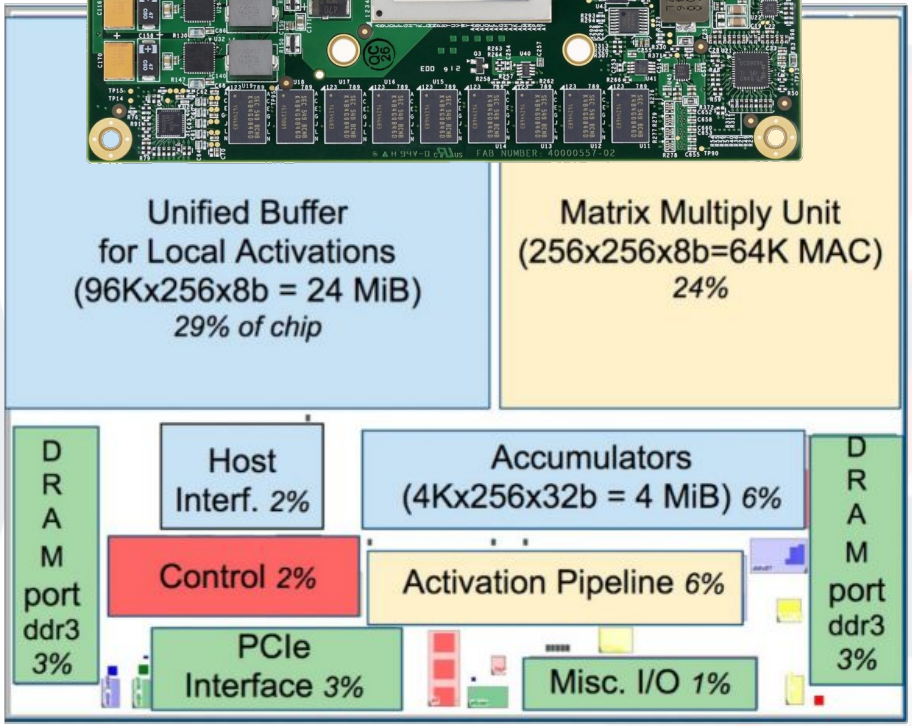
GPU


Host (CPU)
Memory
Serial execution
Kernel
Kernel
Kernel
Device (GPU)
Global memory
Grid
Thread block
Thread
Reg
Shared memory
...

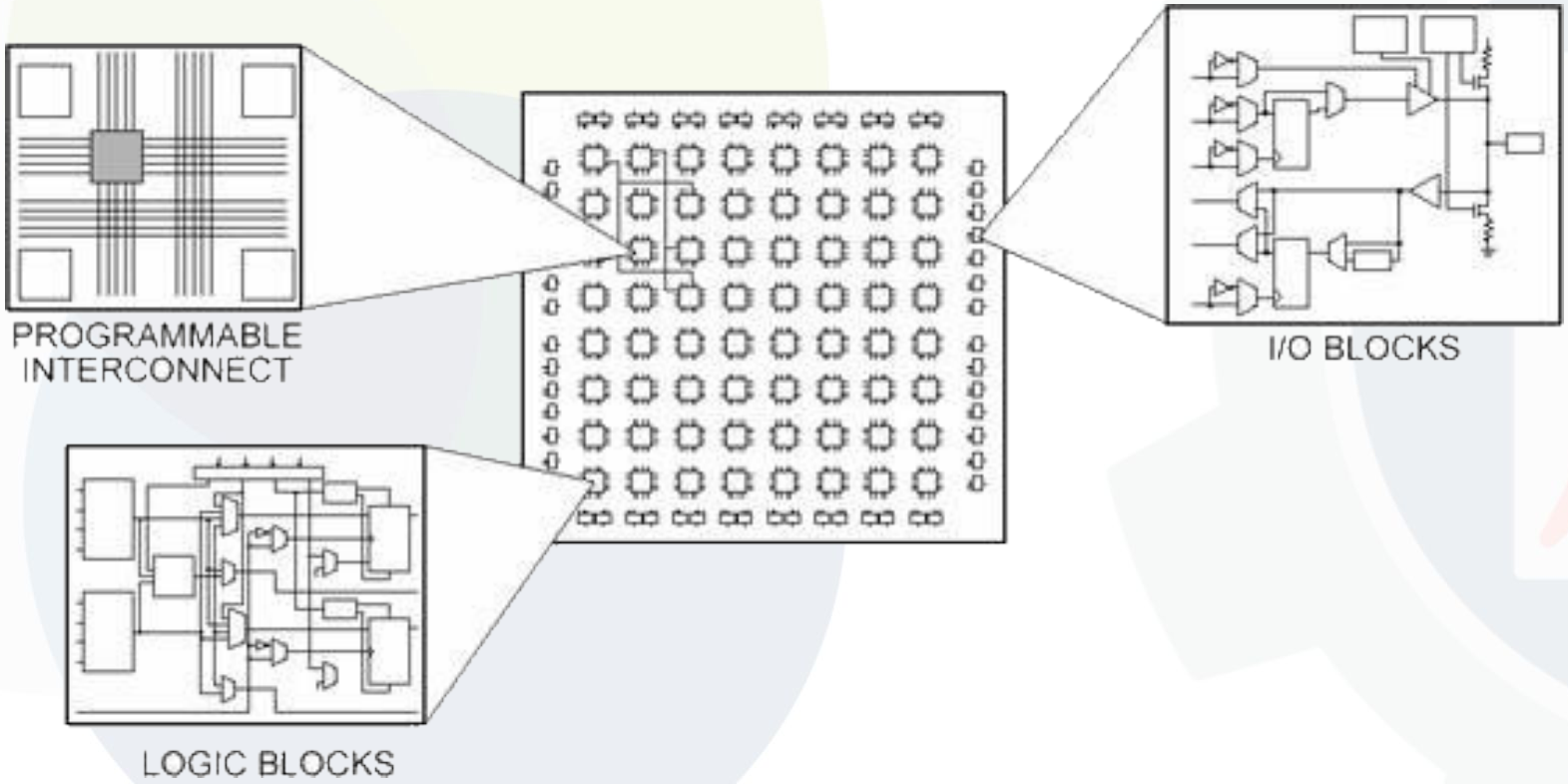
TPU


DDR3 DRAM Chips
30 GiB/s
14 GiB/s
DDR3 Interfaces
30 GiB/s
Weight FIFO (Weight Fetcher)
30 GiB/s
Control
Control
14 GiB/s
PCIe Gen3 x16 Interface
14 GiB/s
Host Interface
10 GiB/s
Unified Buffer (Local Activation Storage)
Systolic Data Setup
167 GiB/s
Matrix Multiply Unit (64K per cycle)
Control
Accumulators
Activation
Instr
167 GiB/s
Normalize / Pool
Control
Control
Off-Chip I/O
Data Buffer
Computation
Control
Not to Scale

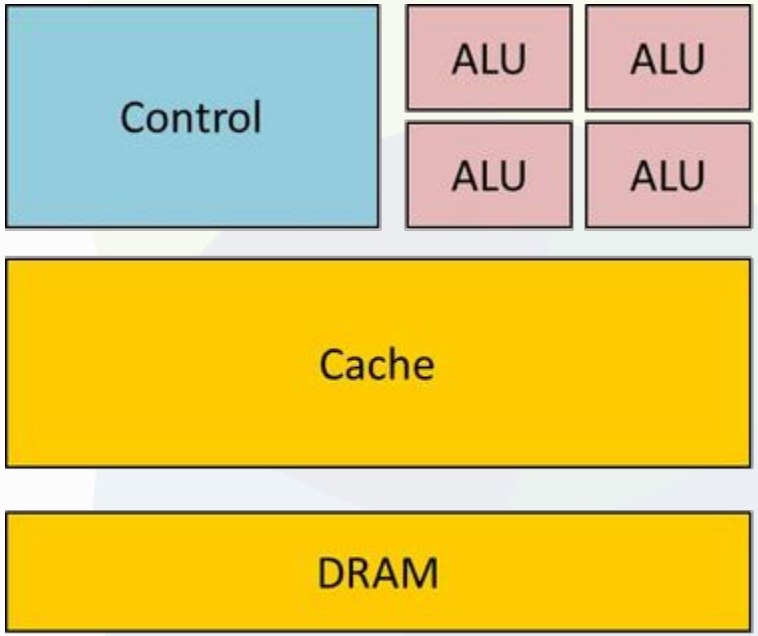



Unified Buffer
for Local Activations
(96Kx256x8b = 24 MiB)
29% of chip
Matrix Multiply Unit
(256x256x8b=64K MAC)
24%
DRAM port ddr3 3%
Host Interf. 2%
Accumulators
(4Kx256x32b = 4 MiB) 6%
Control 2%
Activation Pipeline 6%
PCIe Interface 3%
Misc. I/O 1%
DRAM port ddr3 3%

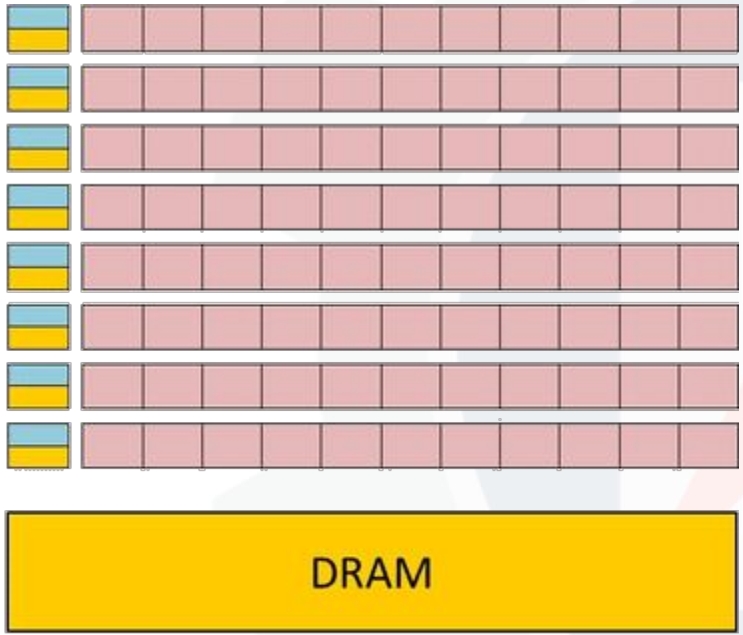
FPGA (Field-Programmable Gate Arrays)



CPU vs GPU/TPU



CPU



GPU

Processadores Huawei Ascend AI

Arquitetura Ascend AI

Ascend 310

High Power Efficiency

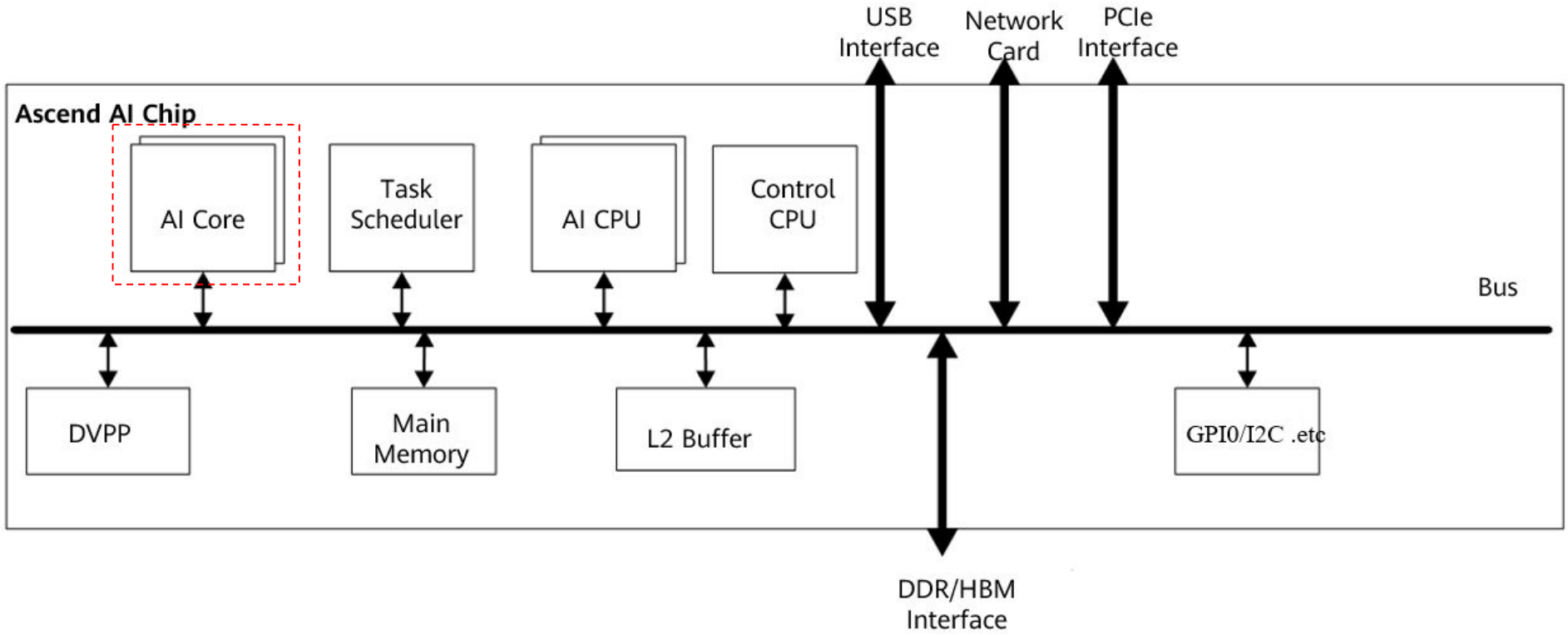
- Ascend-Mini
  Architecture: DaVinci
- FP16： 8 TeraFLOPS
- INT8 ： 16 TeraOPS
- 16 Channel Video Decode – H.264/265
- 1 Channel Video Encode – H.264/265
- Power： 8W
- Process: 12nm

Ascend 910

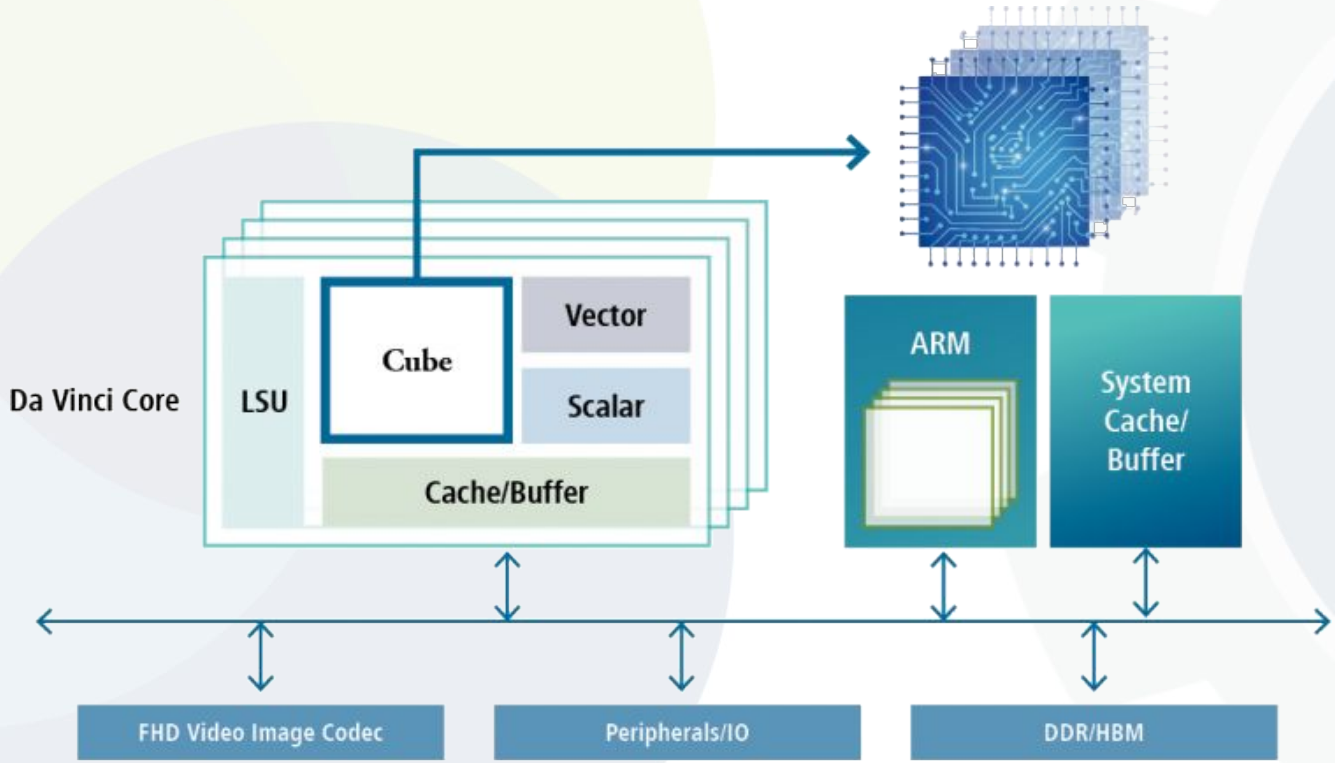
High Computing Density

- Ascend-Max
  Architecture: DaVinci
- FP16: 256 TeraFLOPS
- INT8: 512 TeraOPS
- 128 Channel Video Decode – H.264/265
- Power: 350W
- Process: 7+ nm EUV

Diagrama Ascend AI

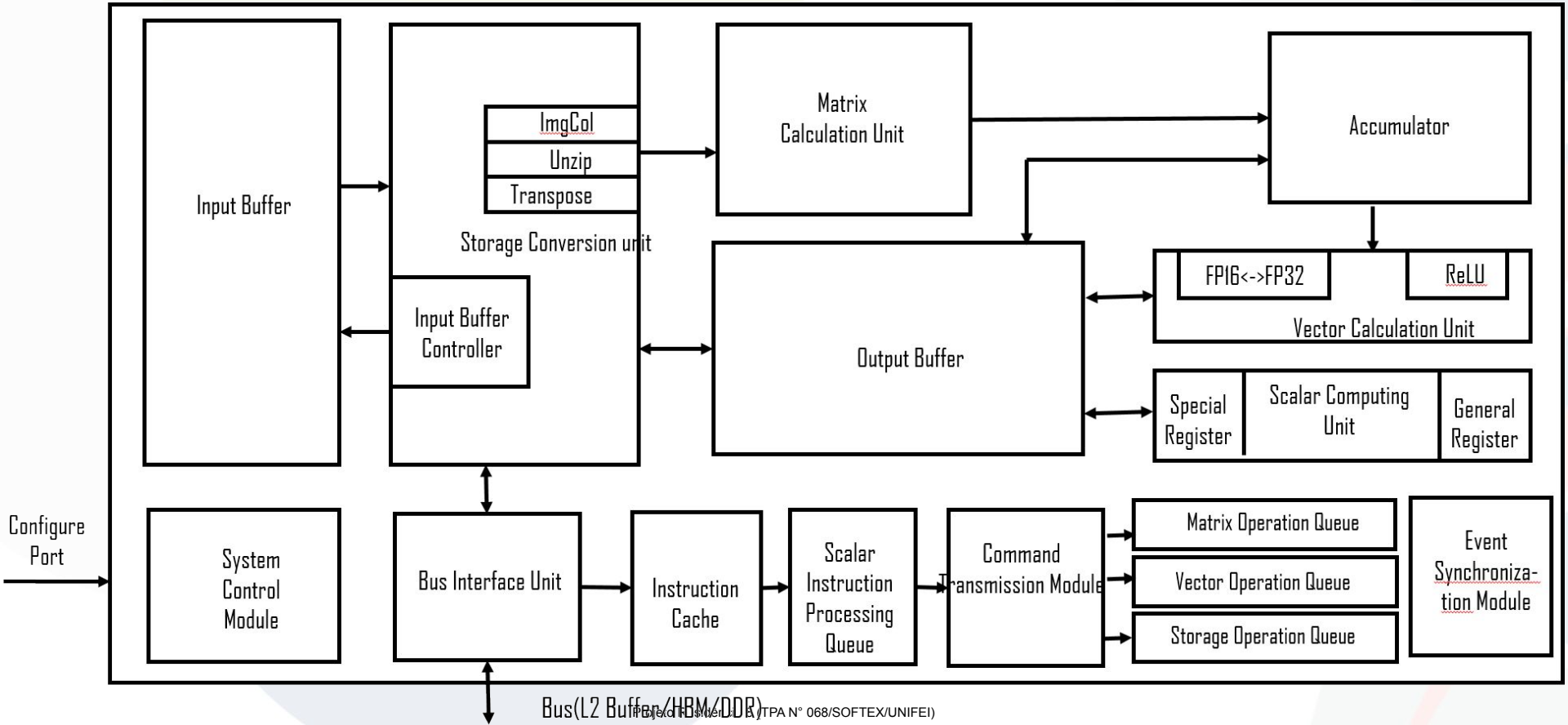


AI Core: Arquitetura Da Vinci

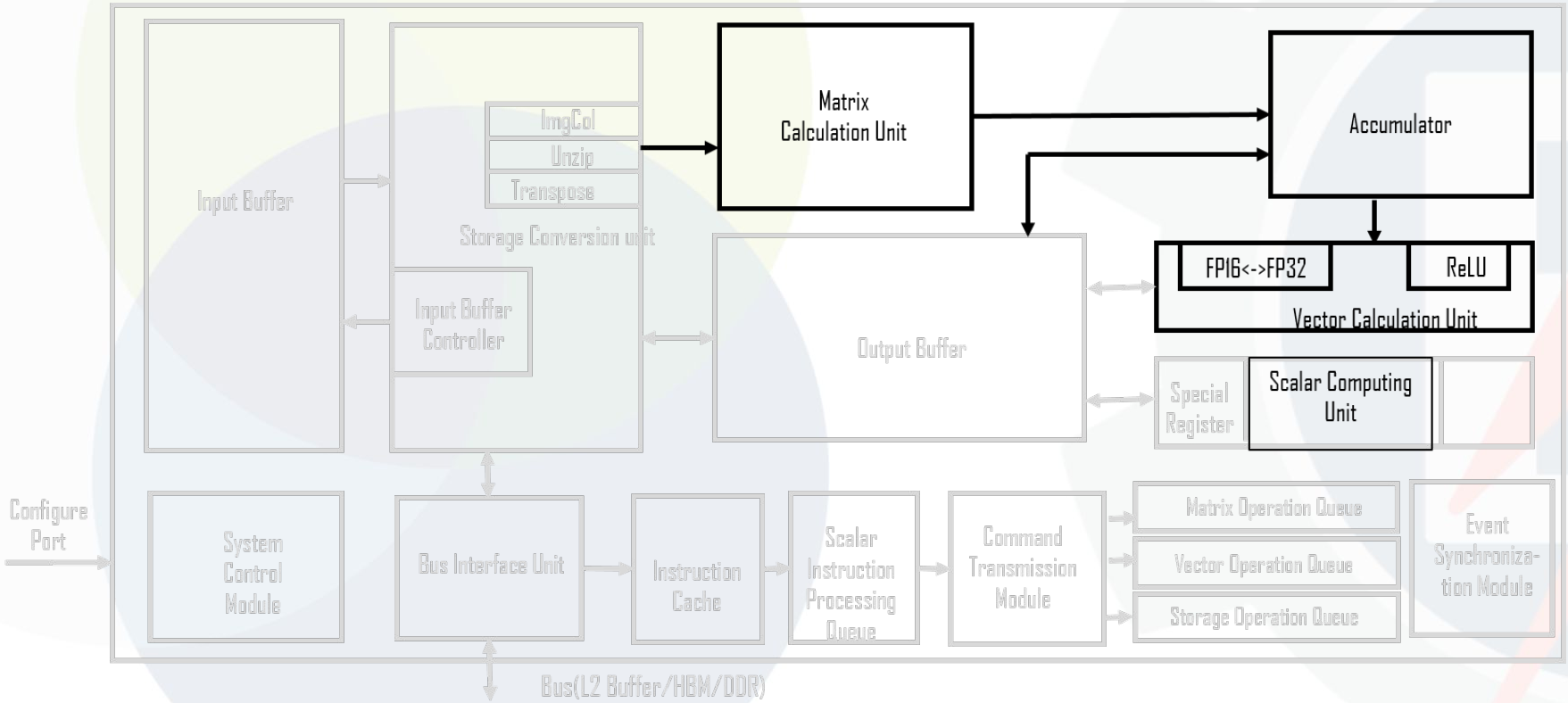


The unified Da Vinci architecture for Ascend chips

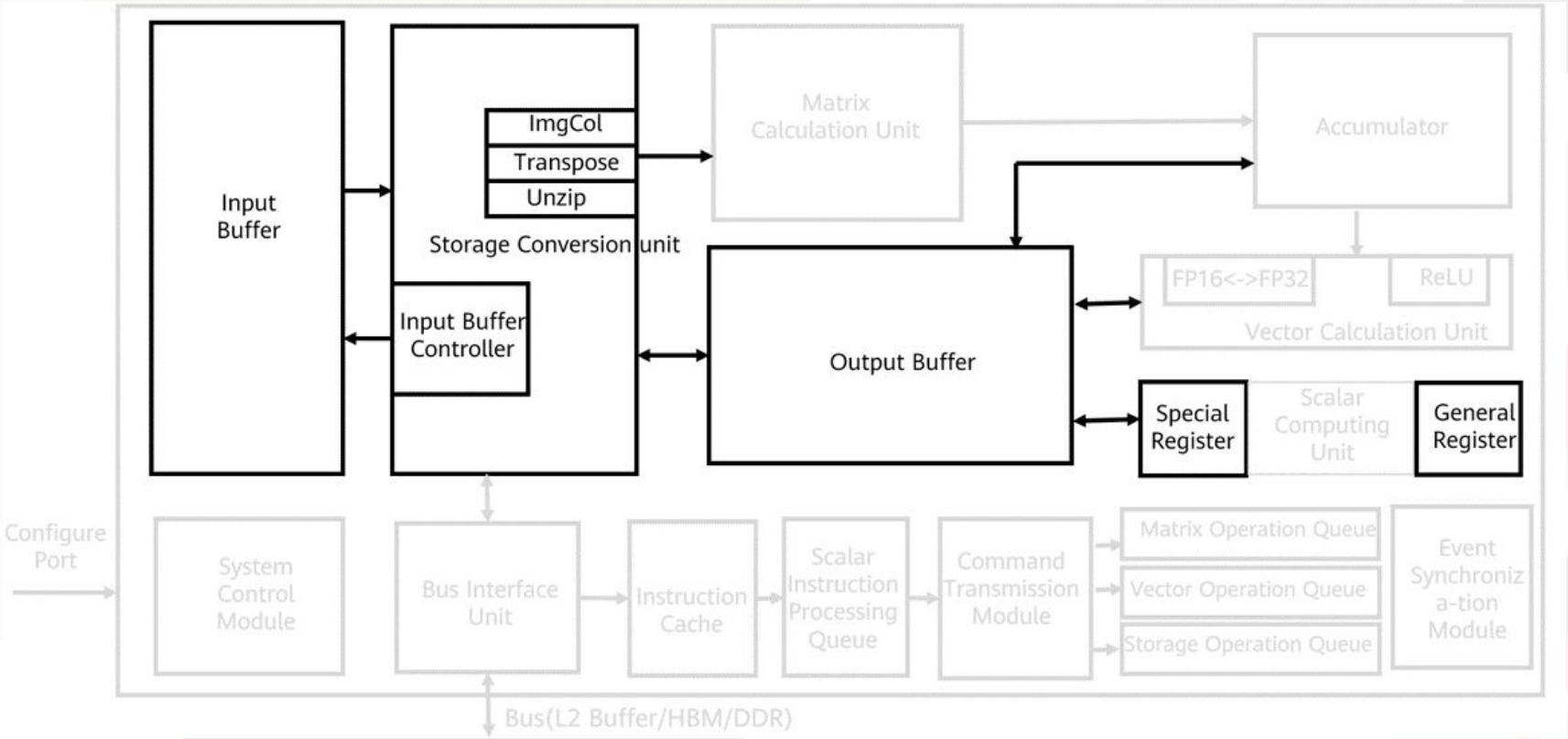
Componentes da Arquitetura Da Vinci



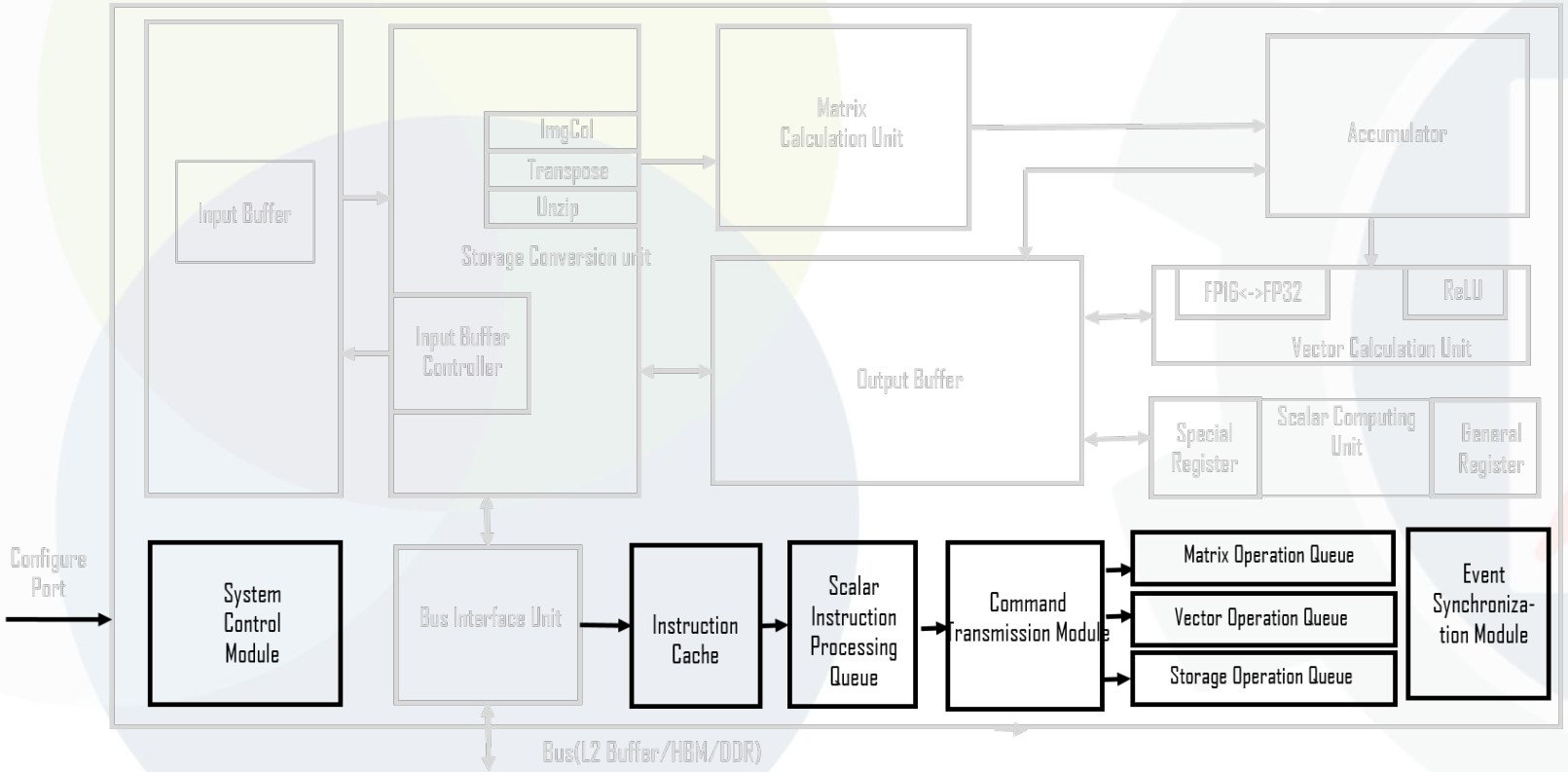
Da Vinci AI - Unidade de Computação



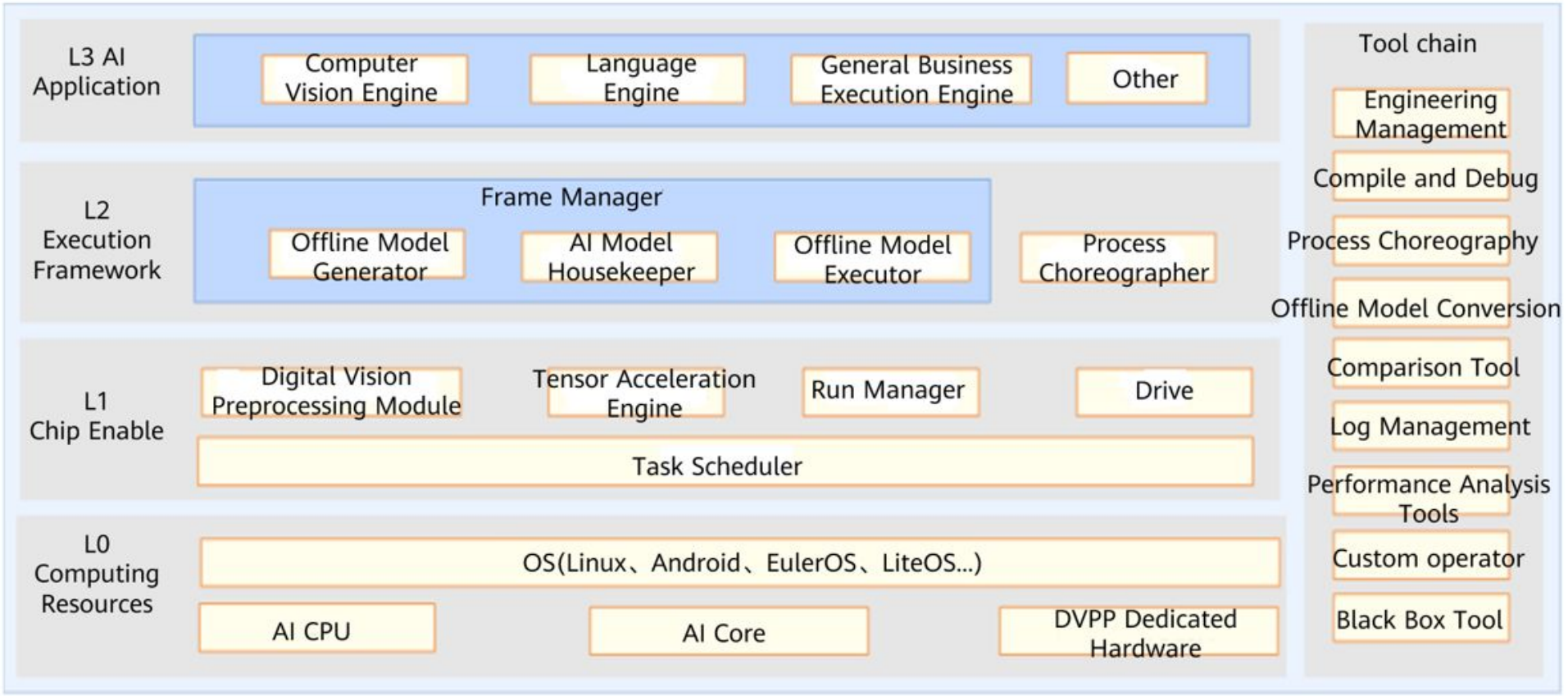
Da Vinci AI - Sistema de armazenamento



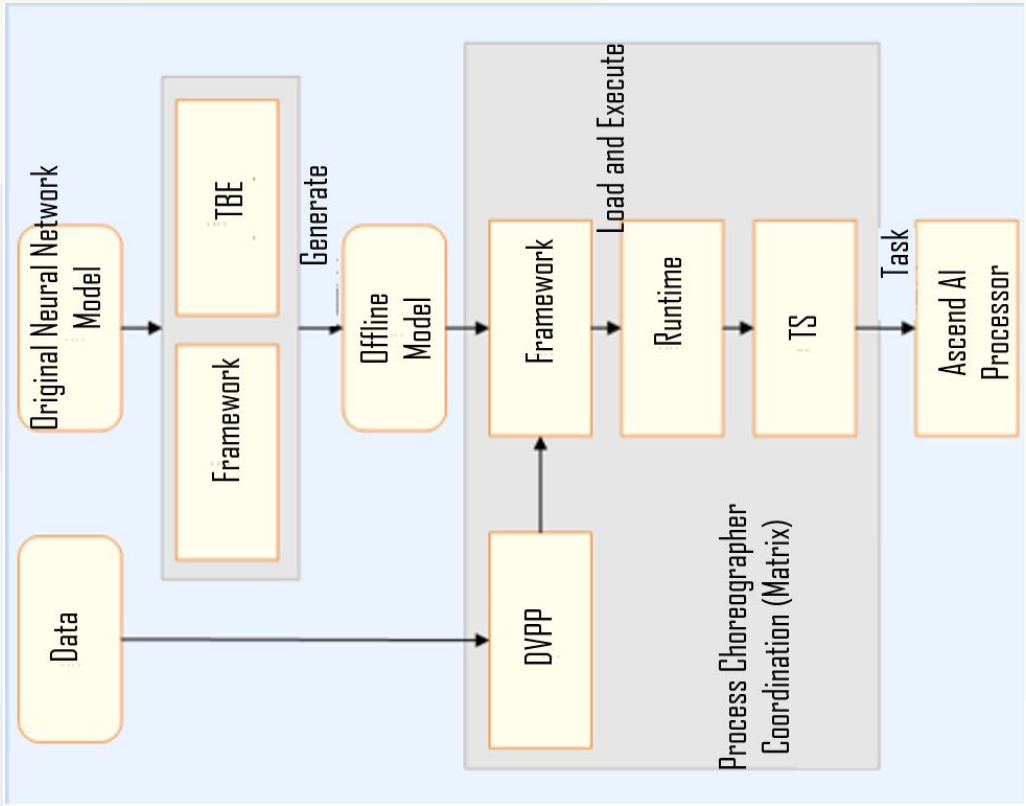
Da Vinci AI - Unidade de controle



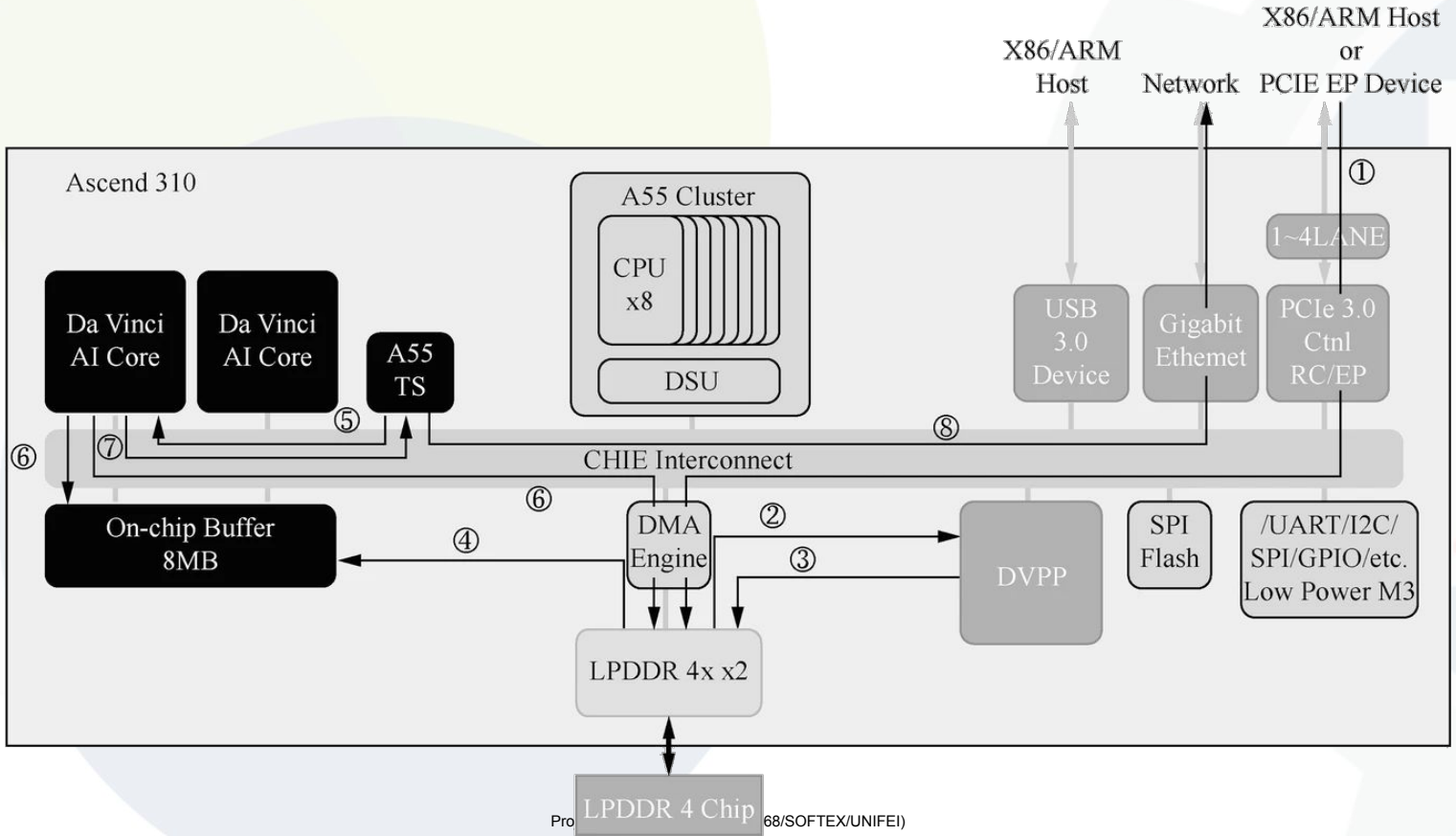
Arquitetura Lógica - Ascend AI



Fluxo de software de rede neural - Ascend AI

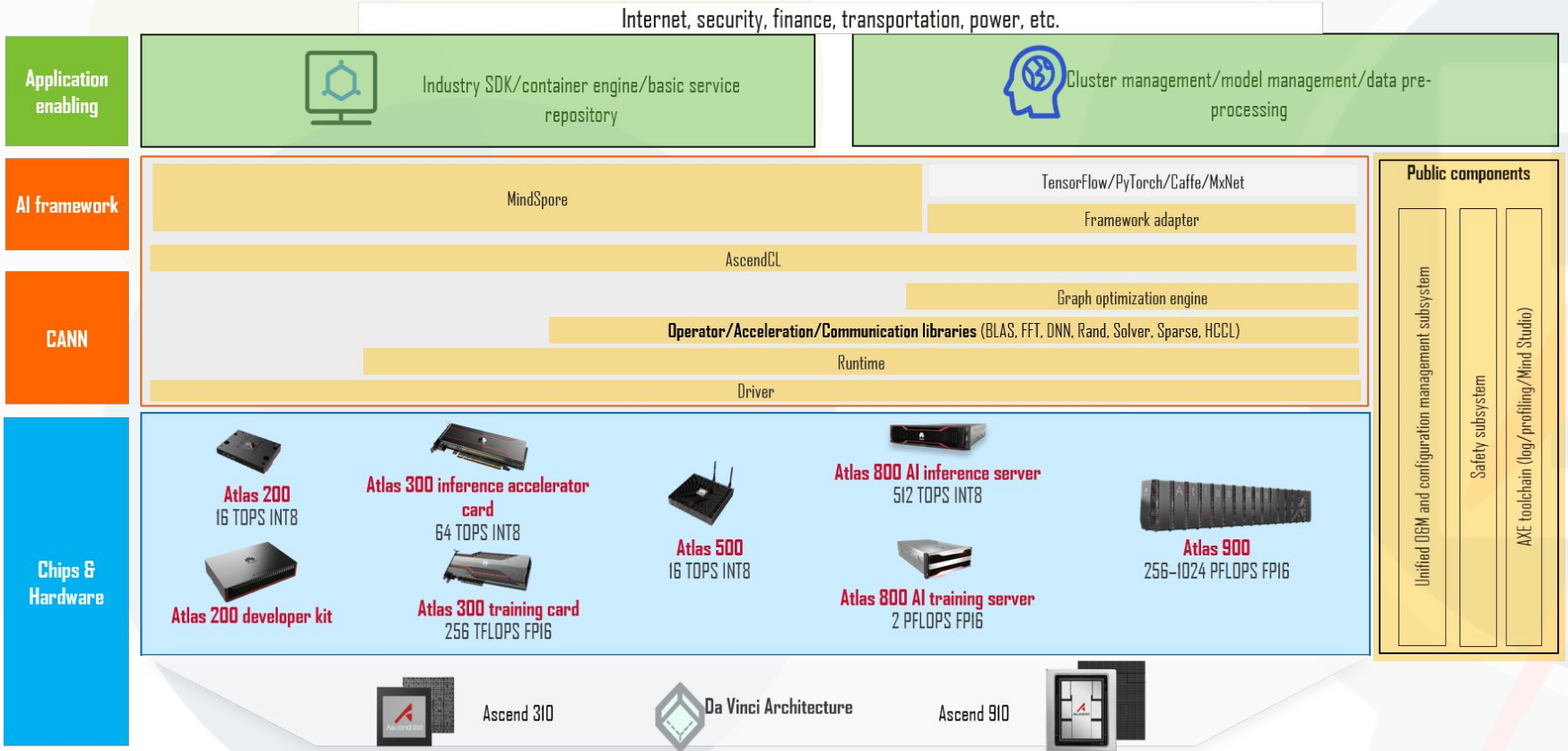


Fluxo de dados em Aplicação: Reconhecimento Facial



Huawei Atlas AI

Portfólio da Plataforma Atlas AI Computing



Inferência de IA com Atlas

Ascend 310
AI processor

Performance improved 7x for terminal devices

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

Atlas 200 AI accelerator module
Model: 3000

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

Atlas 300 AI accelerator card
Model: 3000

Edge intelligence and cloud-edge collaboration

Atlas 500 AI edge station
Model: 3000

Powerful computing platform for AI inference

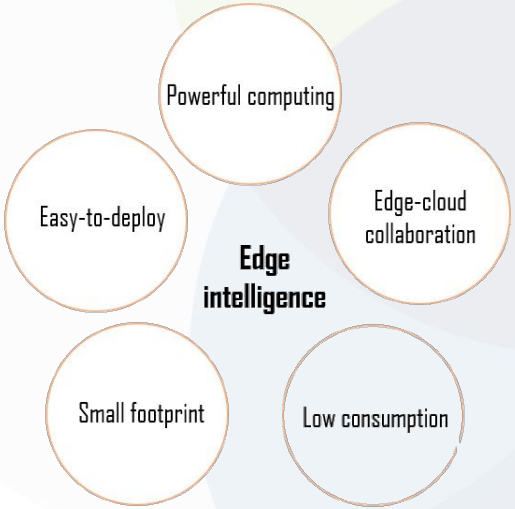
Atlas 800 AI server
Model: 3000/3010

Atlas 200DK: Poder de computação e facilidade de uso

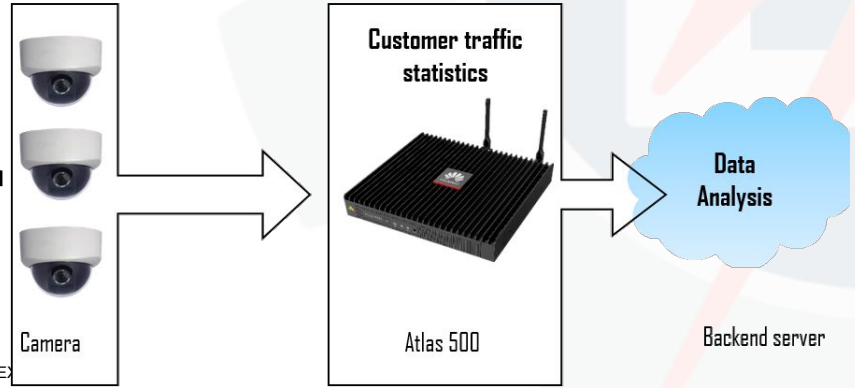
Full-Stack AI development on and off the cloud

- 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

Estação Edge Atlas 500 AI

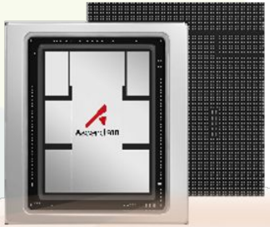


- 16 TOPS INT8
- 25 w to 40 w
- Wi-Fi & LTE
- 64-channel HD video analytics in real time
- Fanless design, stable from -40° C to +70° C



Servidor Atlas 800 AI

Treinamento de IA com Atlas



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

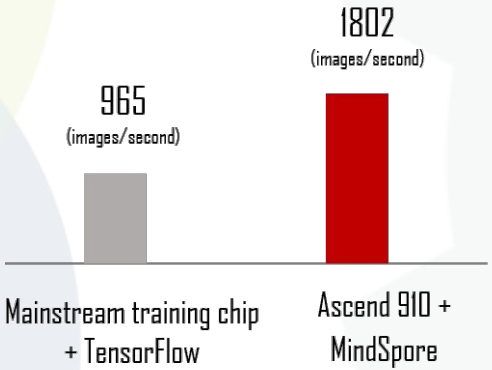
Placa Atlas 300 AI

Atlas 300

2x ↑

Computing power per card

256T FLOPS FP16



70% ↓

Gradient synchronization latency

Direct 100G RoCE

Servidor de Treinamento Atlas 800

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

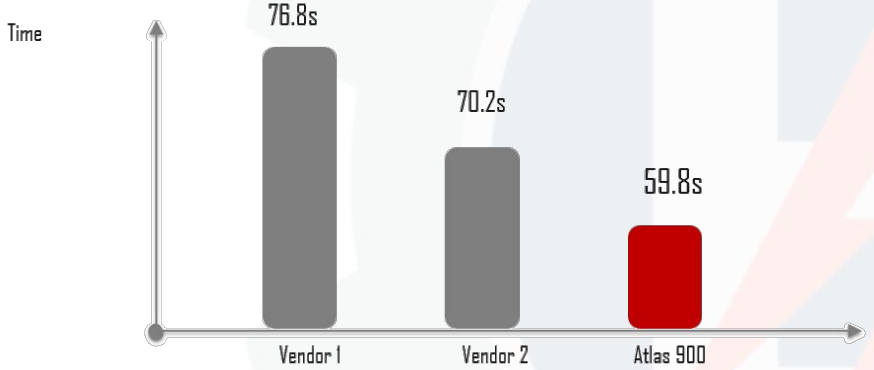
Atlas 900 AI Cluster

Atlas 900

256-1024 PFLOPS FP16

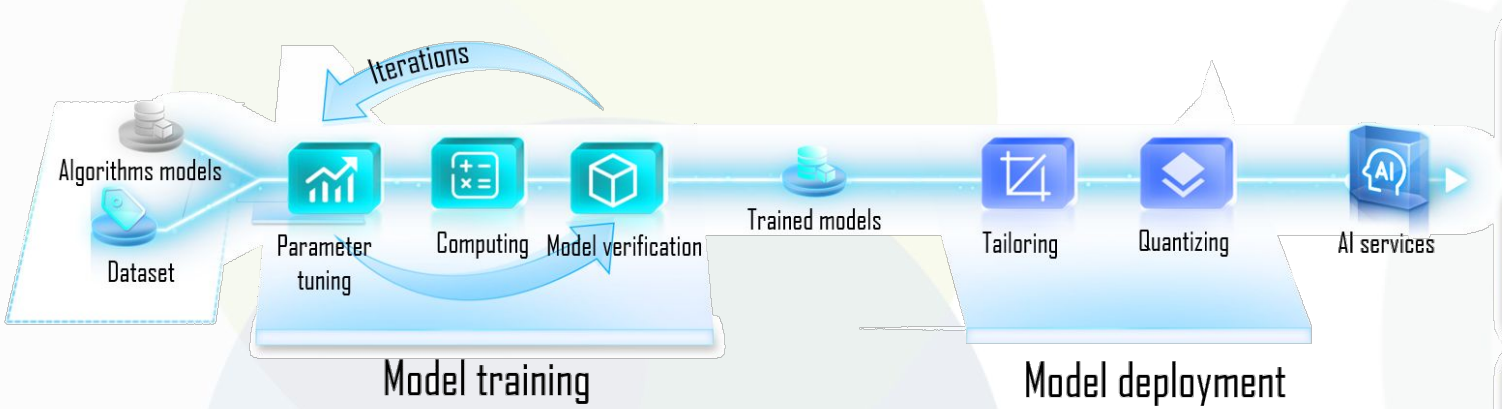
Industry-leading computing power | Best cluster network | Ultimate heat dissipation

Shortest time consumption: 59.8s

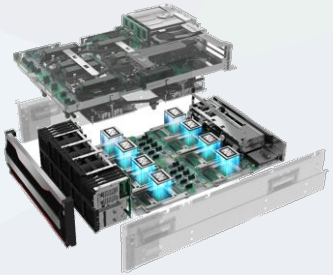


- **Test benchmark:**
  - Benchmark: ResNet-50 V1.5 model, ImageNet-1k dataset
  - Cluster: 1024 Ascend 910 AI processors
  - Accuracy: 75.9%

Aplicações Abrangentes com Atlas



Atlas 300 training acceleration card
Model: 9000



Atlas 800 AI training server
Model: 9000

Atlas 900 AI cluster

Video analysis

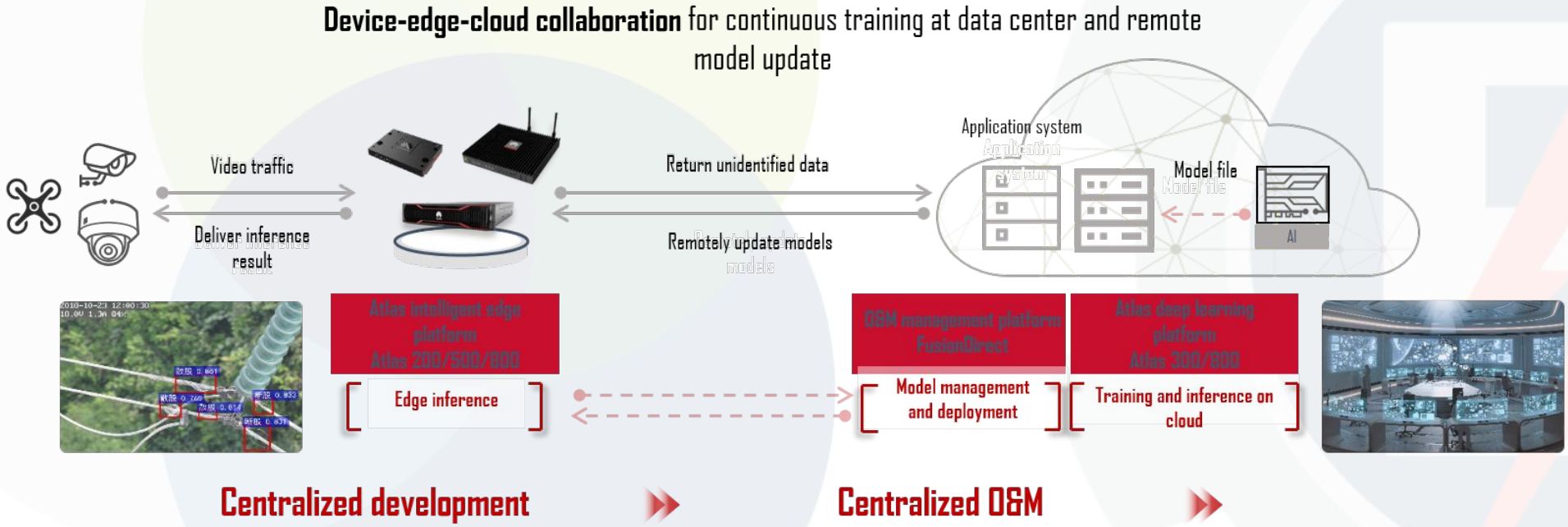
Gene research

Automated driving

Weather forecast

Oil exploration

Colaboração Dispositivo-Borda-Nuvem



Perguntas

Which of the following options are Huawei products?

- Atlas 200DK
- Ascend 310
- Ascend 910
- Atlas 400DK

Pergunta

The Tensor Boost Engine (TBE) is a processor enablement layer built by Huawei for deep neural networks and Ascend AI Processors. It provides a processor operators library and a highly automated operators development toolkit.

- True
- False

Pergunta

Which of the following are the major modules of the Ascend AI processor?

- Chip system control CPU
- AI computing engine
- Multi-level System-On-a-Chip cache or buffer
- Digital vision pre-processing module

Pergunta

Powered by Ascend AI Processors, the Huawei Atlas AI computing platform delivers products that come in various forms, such as modules, cards, edge stations, servers, and clusters, to build AI solutions for all scenarios across the device, edge, and processor.

- True
- False

Pergunta

Which of the following use the Da Vinci Architecture?

- Ascend 910
- Ascend 310
- RTX3080
- Kunpeng 920

Pergunta

Which of the following options is Huawei's artificial intelligence computing platform?

- Atlas
- ModelArts
- Huawei Cloud EI
- ECS

Apoio

Este projeto é apoiado pelo Ministério da Ciência, Tecnologia e Inovações, com recursos da Lei nº 8.248, de 23 de outubro de 1991, no âmbito do [PPI-Softex| PNM-Design], coordenado pela Softex.

Carlos Henrique Valério de Moraes
Universidade Federal de Itajubá
e-mail: valerio@unifei.edu.br

UNIFEI Softex MCTI FUTURO GOVERNO FEDERAL BRASIL UNIÃO E RECONSTRUÇÃO