



ZOMI

Ascend 芯片

Ascend



# About

- **昇腾 SOC 架构:** 昇腾 310 芯片 - 昇腾 910 芯片
- **AICore 的灵魂:** 达芬奇架构内部细节
- **AICore 计算模式:** Vector 和 Cube 计算方法
- **服务器爆炸图:** 从芯片到服务器



# 训练推理系列硬件产品

边缘端

云端推理

云端训练

< 20 TFLOPS

1000 TFLOPS

> 2000 TFLOPS



Atlas 200I A2  
(20 TOPS)



Atlas 300V Pro 视频  
解析卡



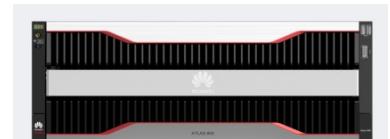
Atlas 300V Pro 视  
频解析卡



Atlas 800 推理服务  
器 (3000)



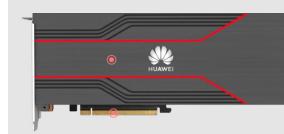
Atlas 500 Pro 智能边缘服  
务器



Atlas 800T A2训练服务  
器



Atlas 300I Pro 推  
理卡



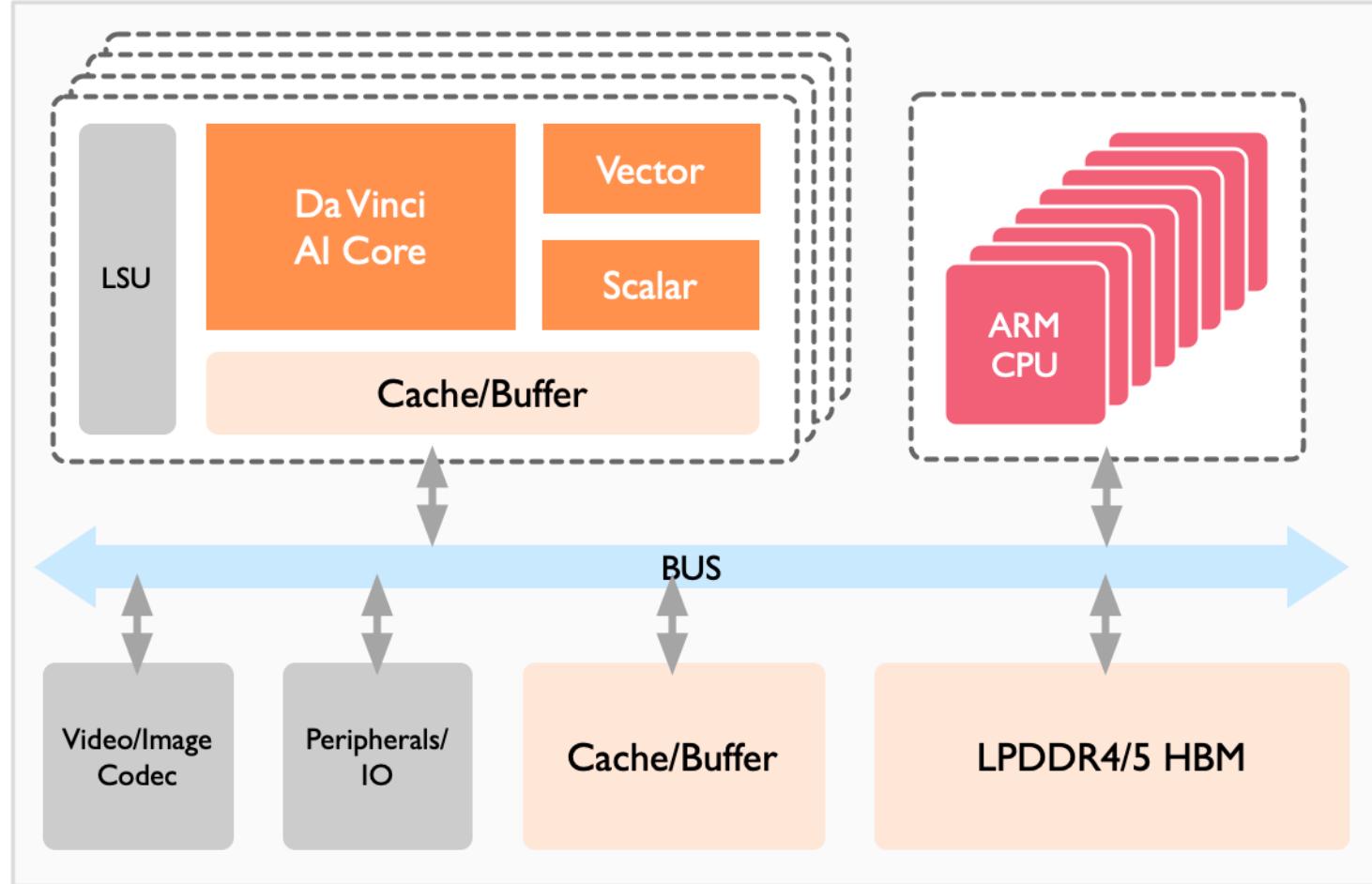
Atlas 300I Duo 推  
理卡



Atlas 800 推理服务  
器 (3010)

# 1. 昇腾 SOC 架构

# 基于 DaVinci AI 技术架构



- 3D Cube:  $16^3$  三维弹性立方：
  1. **高算力**: 可在一个时钟周期内完成4096个FPI6 MAC 运算
  2. **高效**: 支持几十毫瓦IP到几百瓦芯片，适应端、边和云的平滑架构扩展

# 昇腾 AI 处理器：达芬奇架构



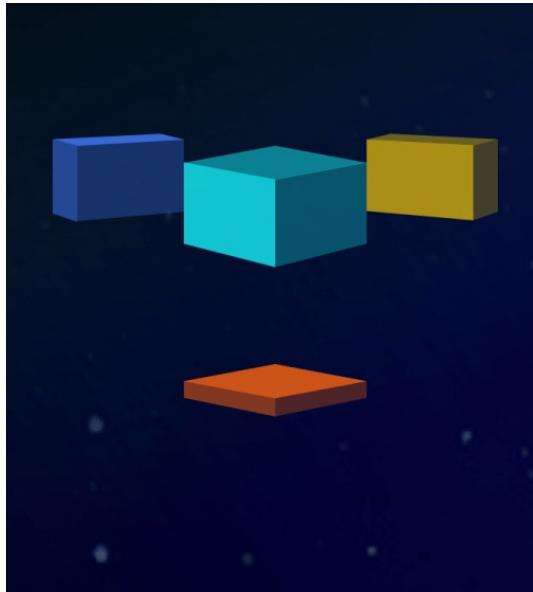
- 半精度 (FP16): XX Tera-FLOPS
- 整数精度 (INT8) : XX Tera-OPS
- 16 通道 全高清 视频解码器 – H.264/265;
- 1 通道 全高清 视频编码器 – H.264/265;
- 最大功耗: 8W
- XXnm FFC



- 半精度 (FP16): XXX Tera-FLOPS
- 整数精度 (INT8) : XXX Tera-OPS
- 128 通道 全高清 视频解码器 – H.264/265;
- 最大功耗: 310W
- XXnm FFC

# 芯片层：基于DaVinci AI技术架构

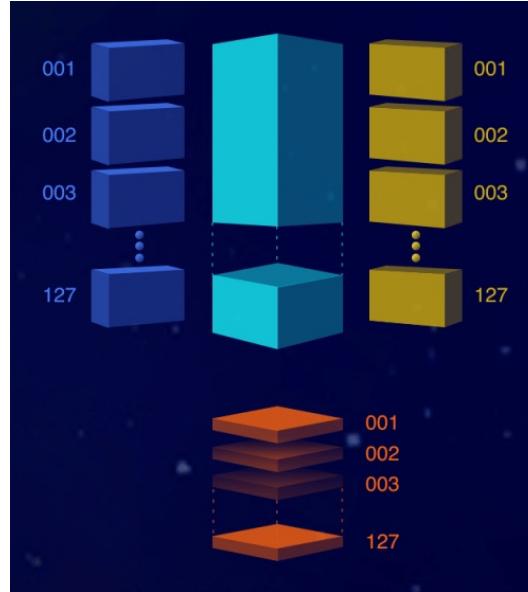
Scalar Compute



0.00X TOPS / W

**ID:**  $N^2$  Cycle  
 $N \uparrow$  ID MAC

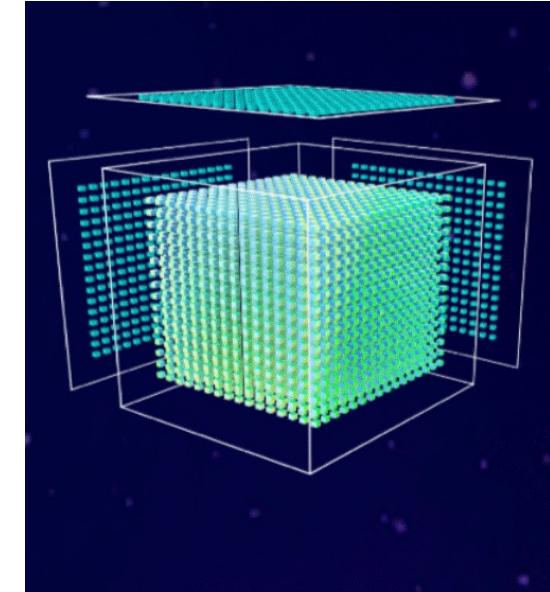
Vector Compute



0.X TOPS / W

**2D:** N Cycle  
1个 $N^2$  2D MAC

Da Vinci  
Tensor Compute



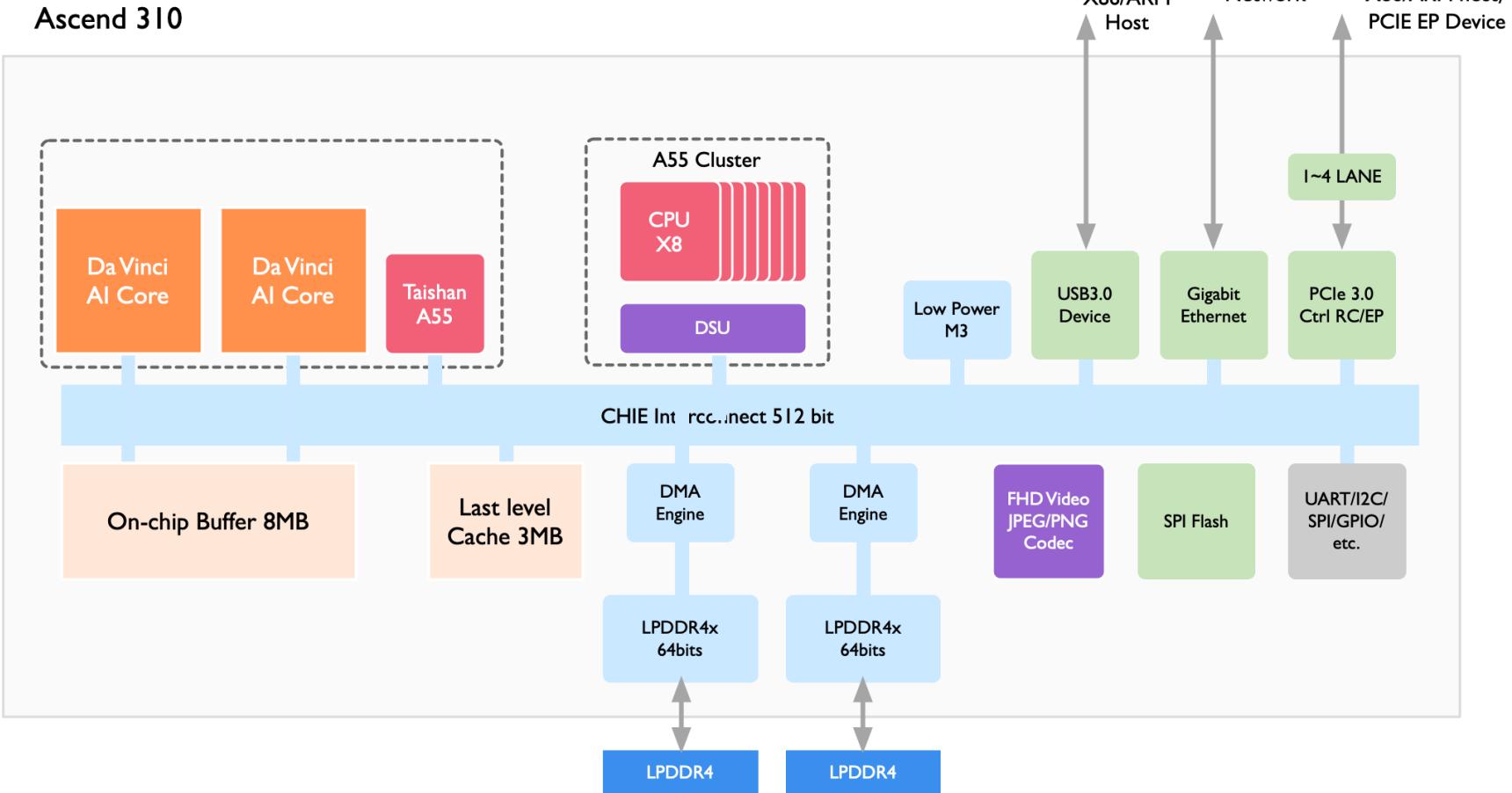
X TOPS / W

**3D:** 1 Cycle  
1个 $N^3$  3D Cube

1.1

# 310 SOC 架构

# Ascend 310 处理器逻辑架构

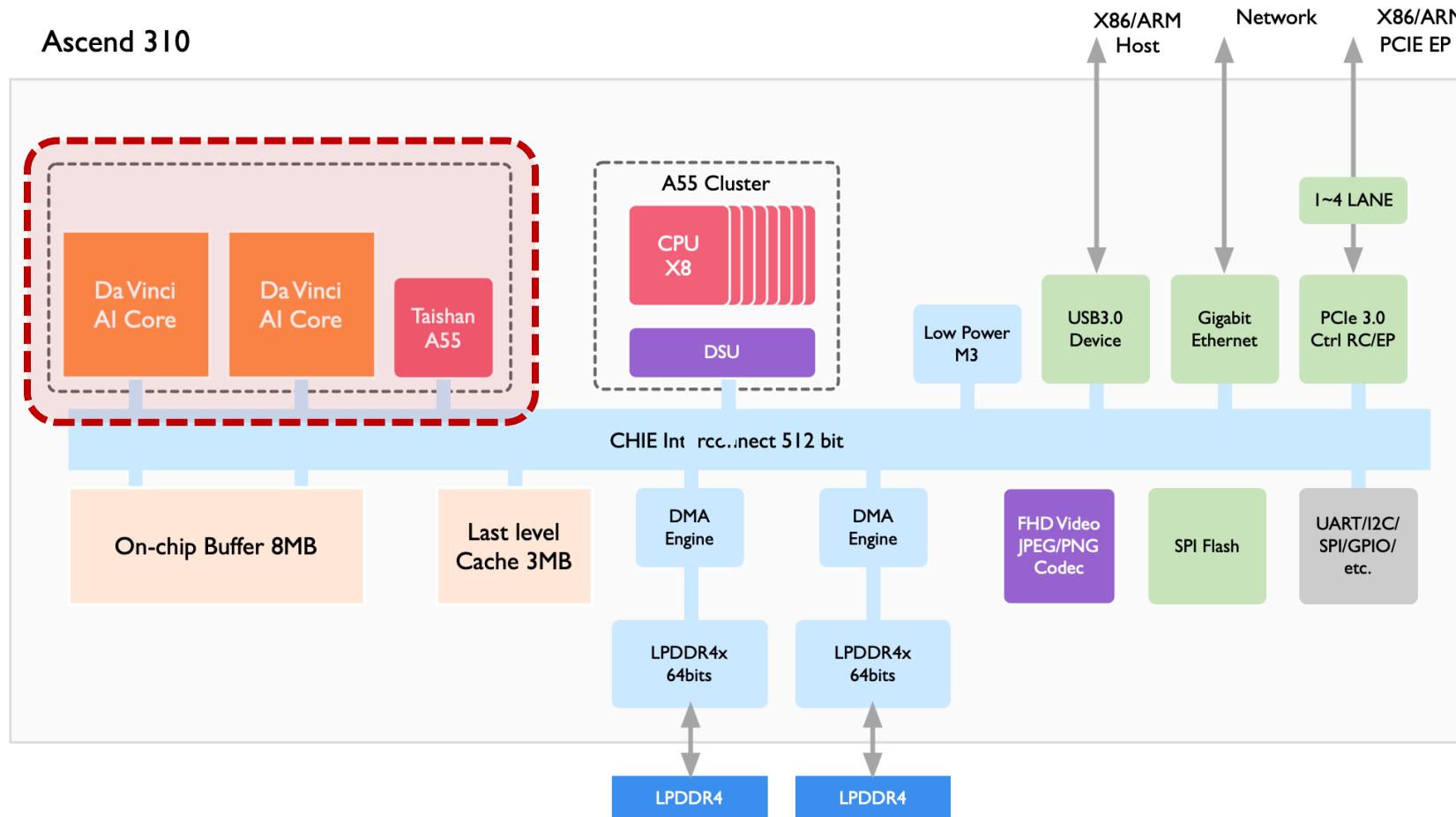


## 主要组成：

- AI Core/AI CPU
- Control CPU
- Cache/Buffer
- DVPP etc.

# Ascend 310 处理器逻辑架构

Ascend 310



## AI Core:

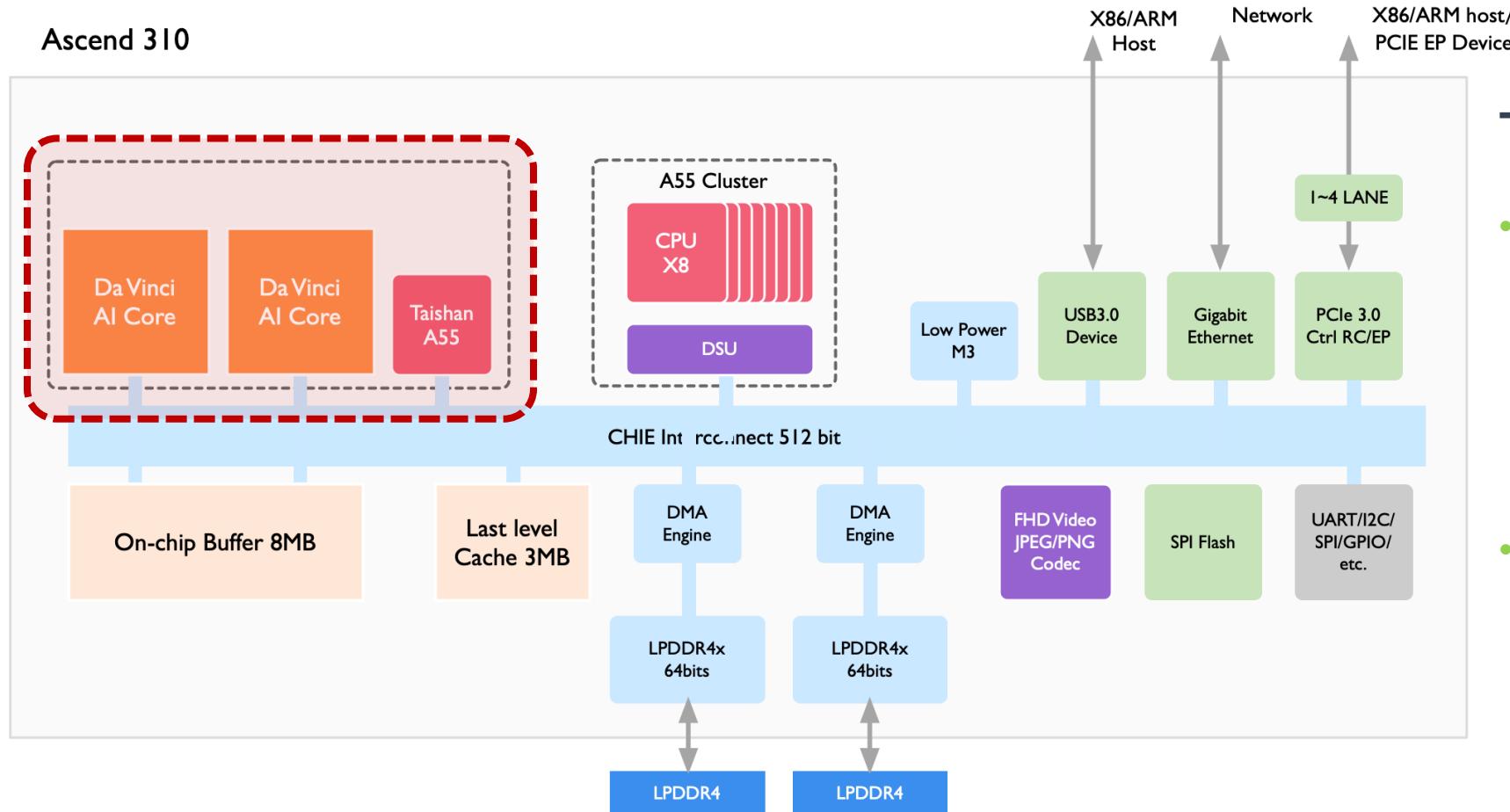
- 计算核心，负责执行矩阵、向量、标量计算密集的算子任务，采用达芬奇架构，图例集成 2 个 AI Core。

## AI CPU:

- 承担非矩阵类复杂计算，即负责执行不适合跑在 AI Core 上算子。

# Ascend 310 处理器逻辑架构

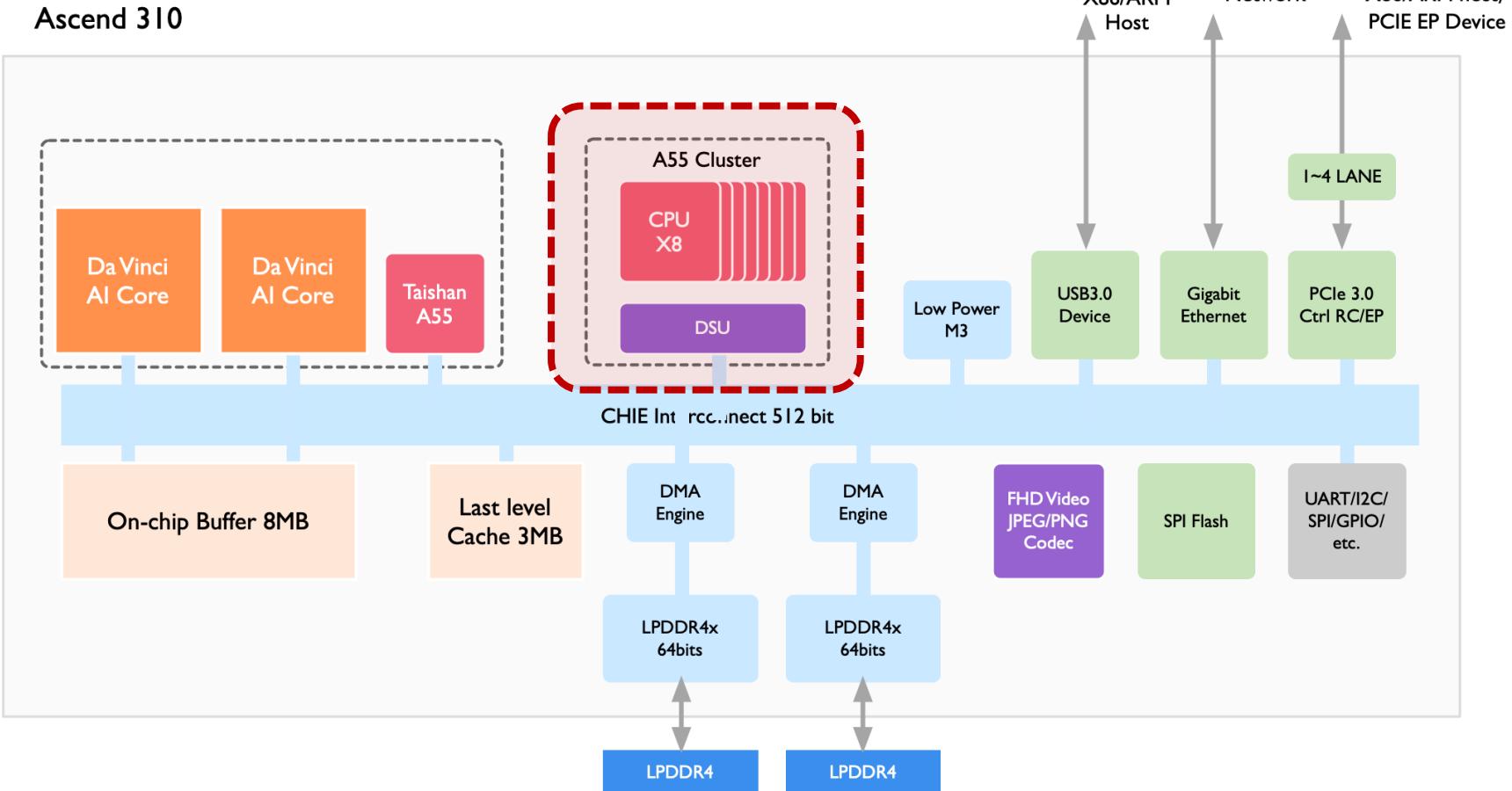
Ascend 310



## TS Core:

- 专用 CPU 作为任务调度器 (Task Scheduler, TS) , 以实现计算任务在 AI Core 上高效分配和调度;
- 该CPU专门服务于AI Core和AI CPU, 不承担任何其他的事物和工作。

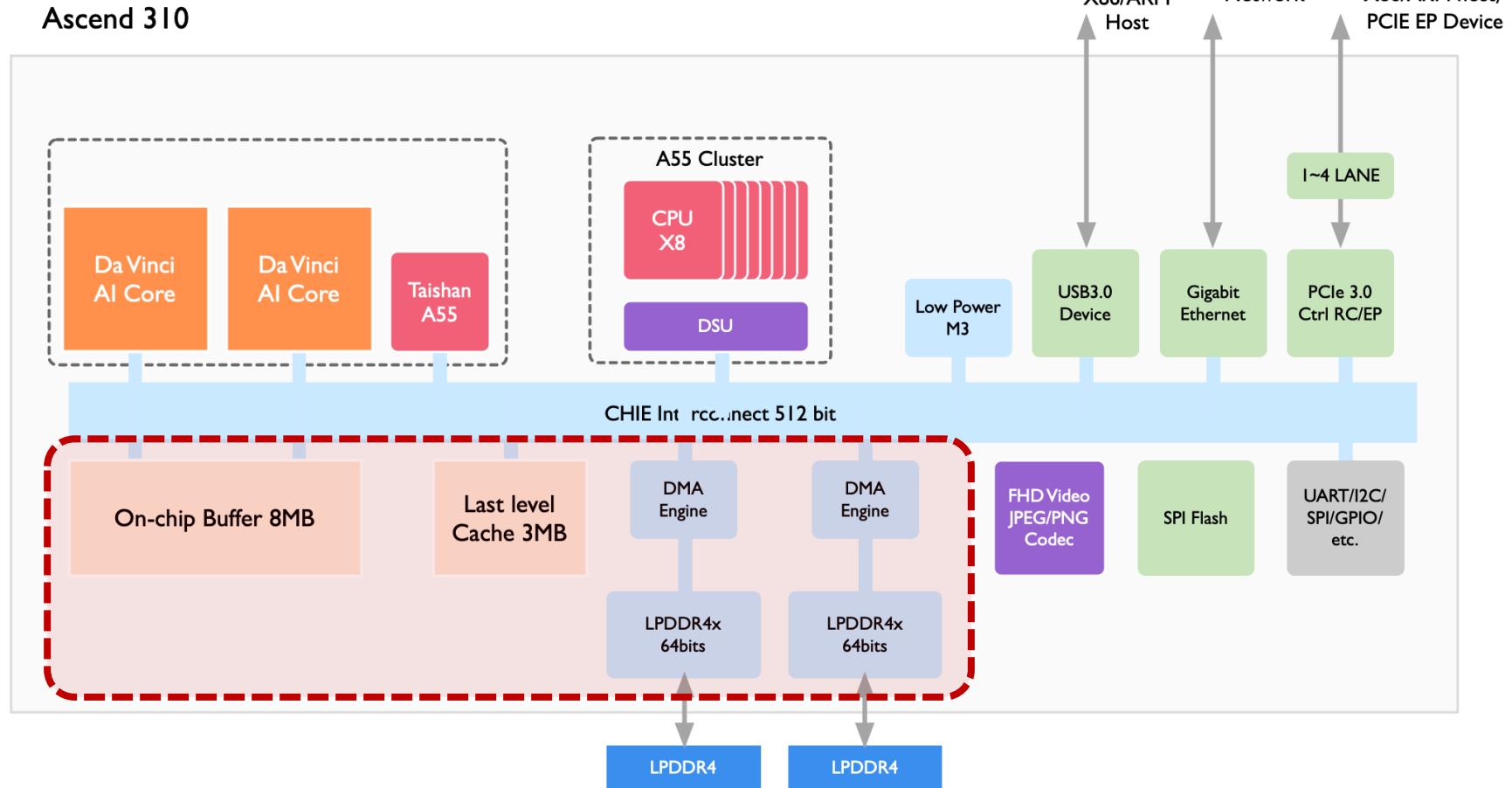
# Ascend 310 处理器逻辑架构



## ARM CPU 核心:

- 专用控制芯片整体运行的控制CPU。Cube or Vector 任务占用的CPU核数可由软件根据系统实际运行情况动态分配。

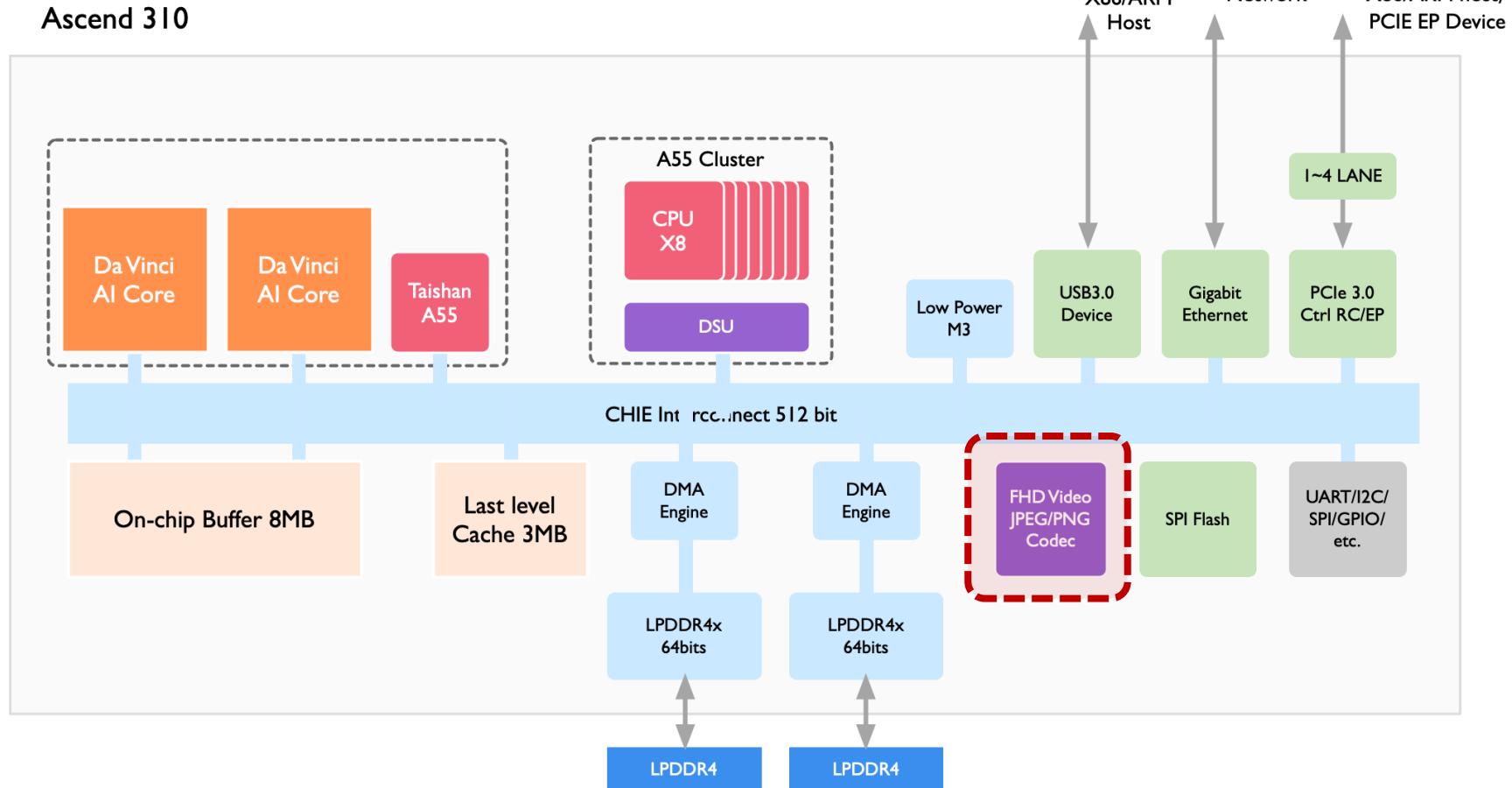
# Ascend 310 处理器逻辑架构



## Cache & Buffer:

- SOC片内层次化 memory 结构, AI core 内部两级 memory buffer;
- SOC片上 8MB L2 buffer, 专用于 AI Core、AI CPU, 提供高带宽、低延迟 memory 访问。
- 芯片集成 LPDDR4x 控制器, 提供更大容量 DDR 内存。

# Ascend 310 处理器逻辑架构

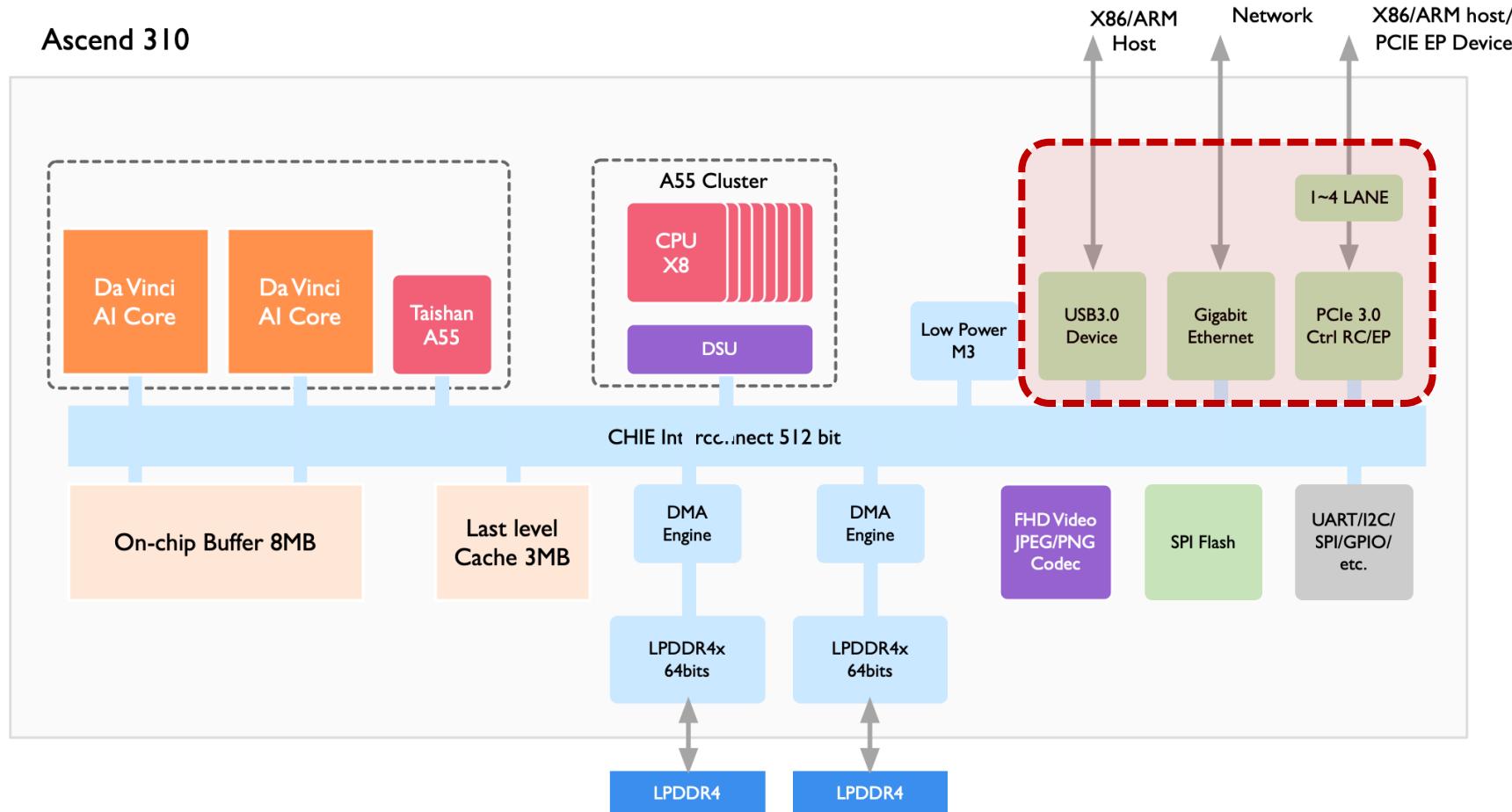


## DVPP:

- 数字视觉预处理子系统，完成图像视频编解码。
- 用于将从网络或终端设备获得 Image Pipeline Buffer，进行预处理以实现格式和精度转换等要求，之后提供给 AI 计算引擎。

# Ascend 310 处理器逻辑架构

Ascend 310



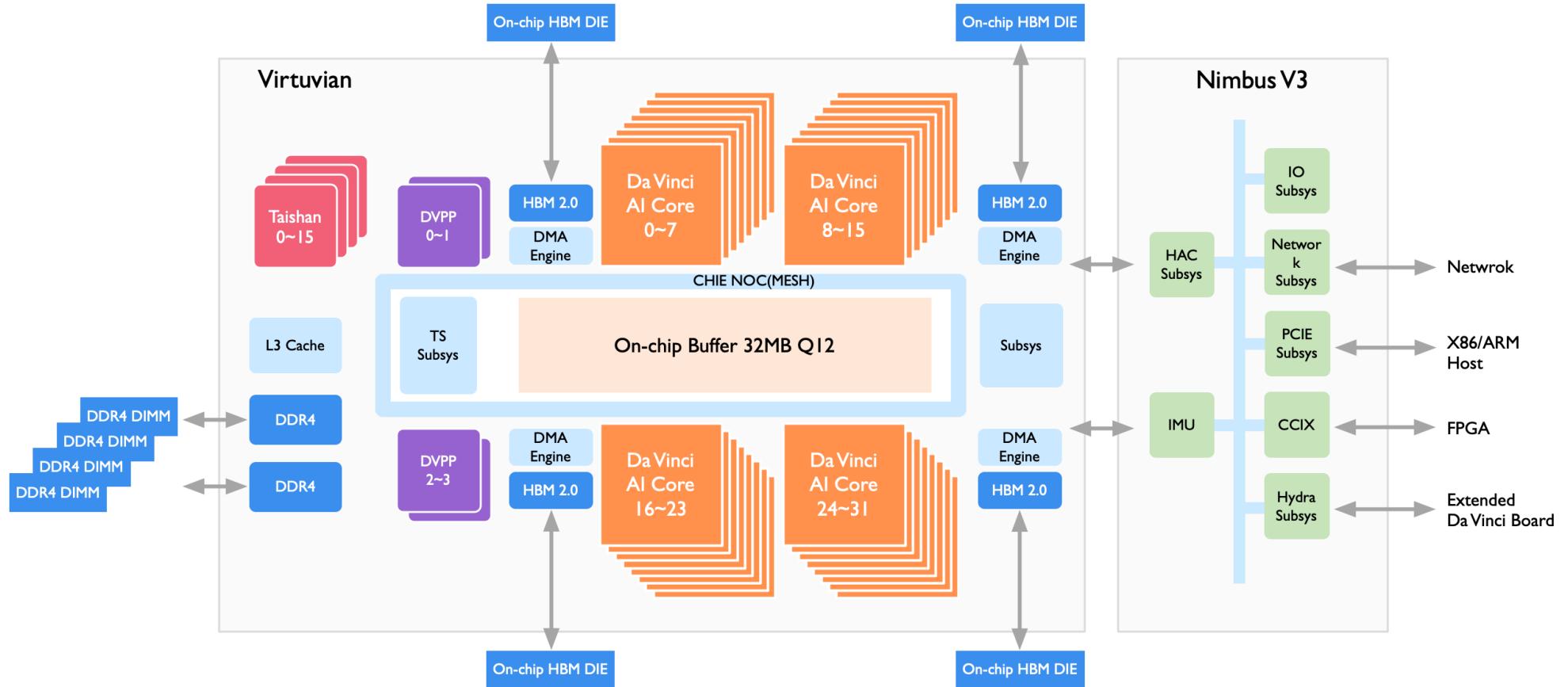
## 对外接口：

- 支持 PCIE3.0、RGM II、USB3.0 等高速接口；
- 以及 GPIO、UART、I2C、SPI 等低速接口。

# 1.2

# 910 SOC 架构

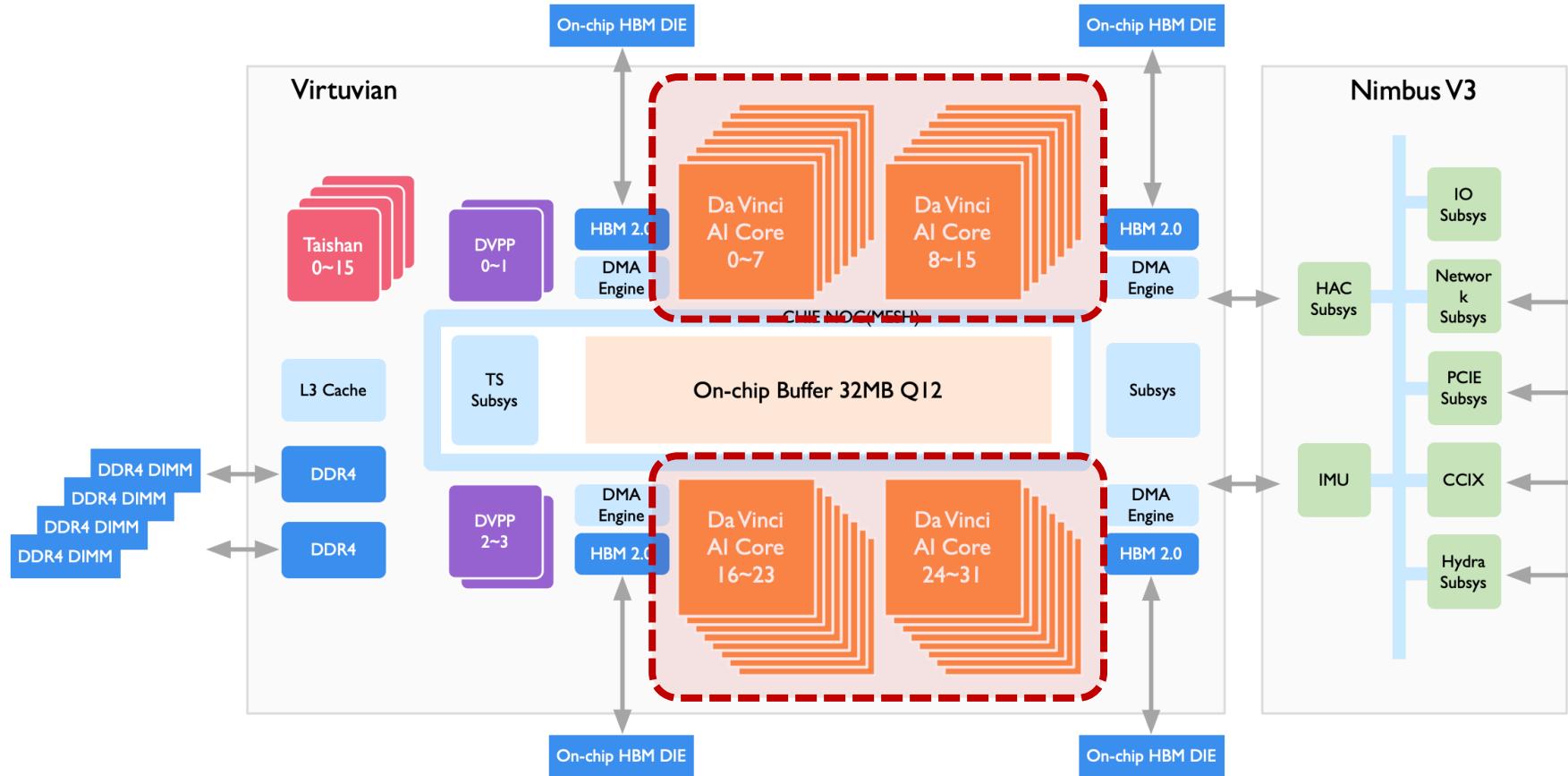
# Ascend 910 处理器逻辑架构



**主要组成：**

- AI Core
- ARM CPU
- DVPP
- Cache/Buffer
- 低速外设接口

# Ascend 910 处理器逻辑架构

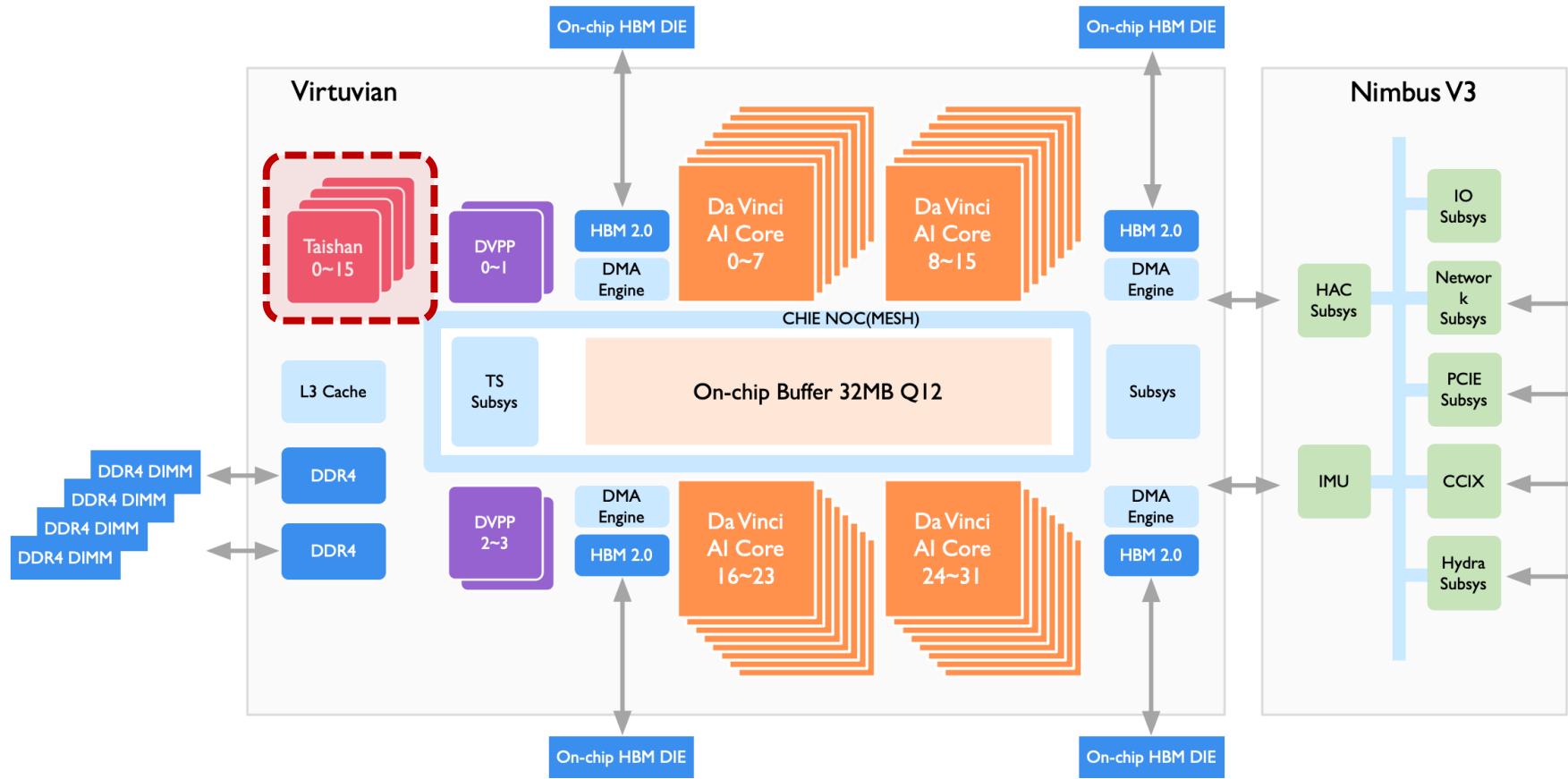


**AI Core:**

- 计算核心，负责执行矩阵 Cube、向量 Vector 计算密集型任务；

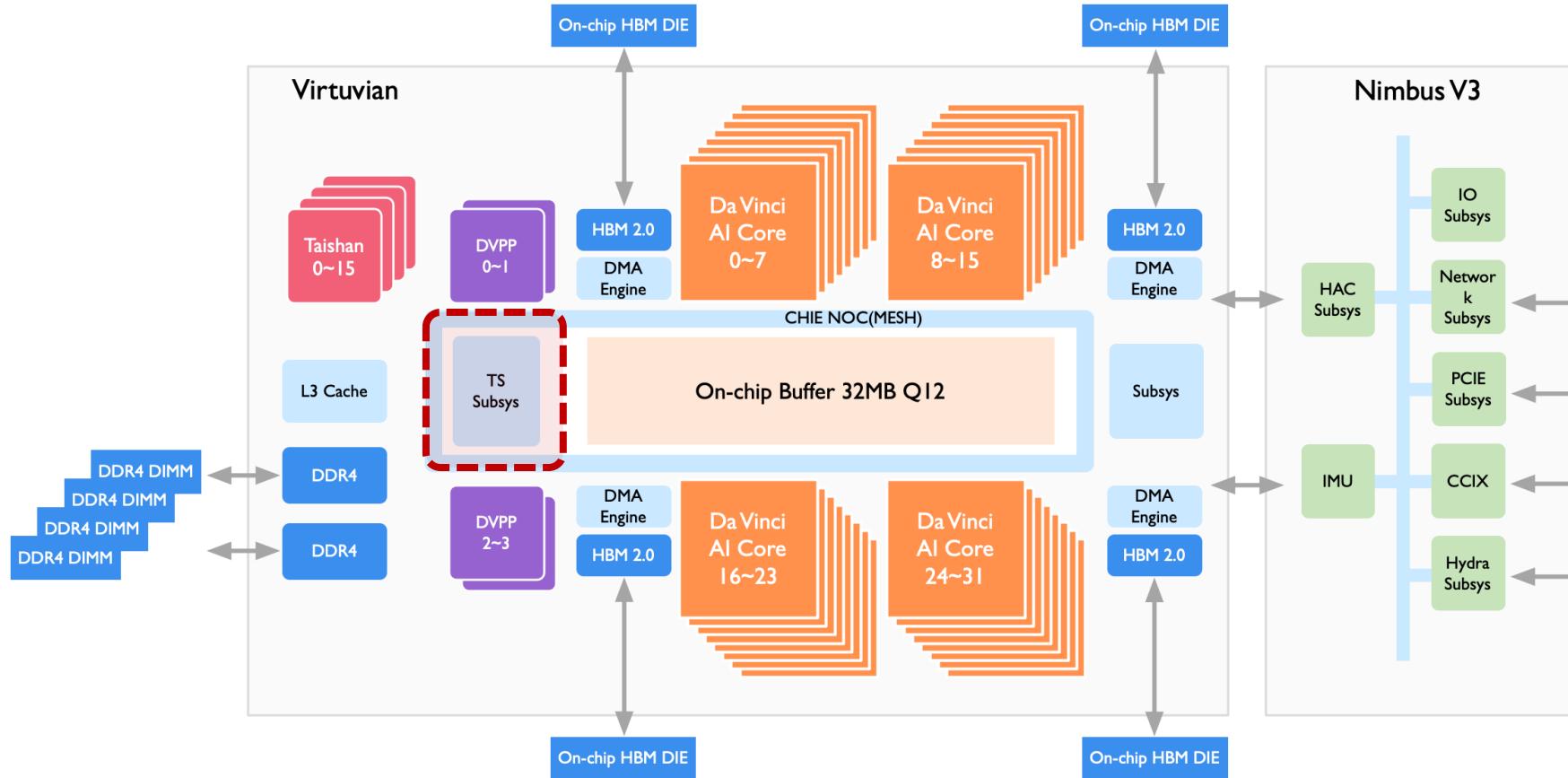
图例 Ascend 910 集成了 32 个 AI Core。

# Ascend 910 处理器逻辑架构



- **CPU 子系统:**
- TaishanVII 10 Core (4个Core构成一个Cluster) ;
- Taishan Core 一部分为AI CPU, 承担部分 AI 计算;
- 一部分部署为 Ctrl CPU, 负责 SoC 控制功能;
- 两类CPU占用核数由软件分配。

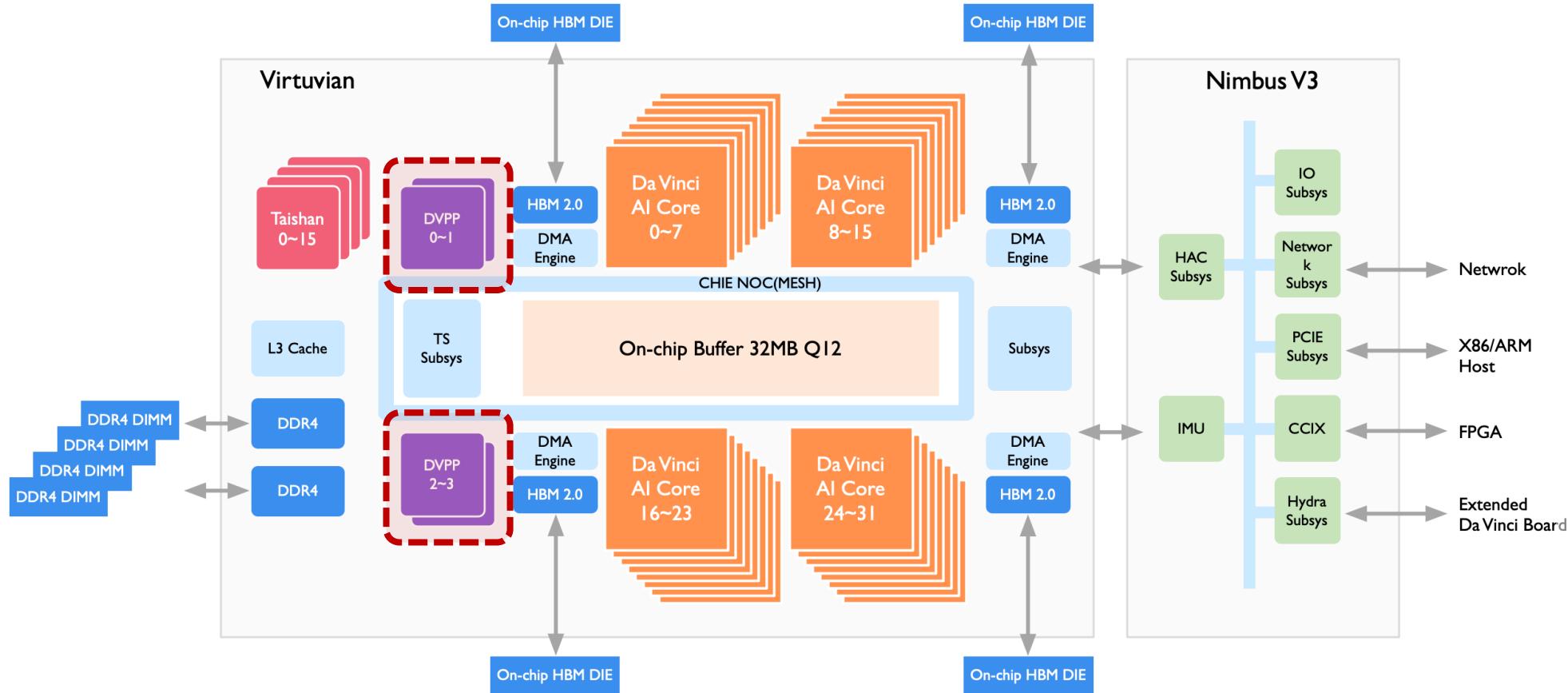
# Ascend 910 处理器逻辑架构



**TS CPU (Task Scheduler):**

- 4 核 A55 Cluster (ARMv8 64位架构)，负责任务调度，把算子任务切分之后，通过硬件调度器 (HWTS)，分发给 AI Core 或 AI CPU 执行计算。

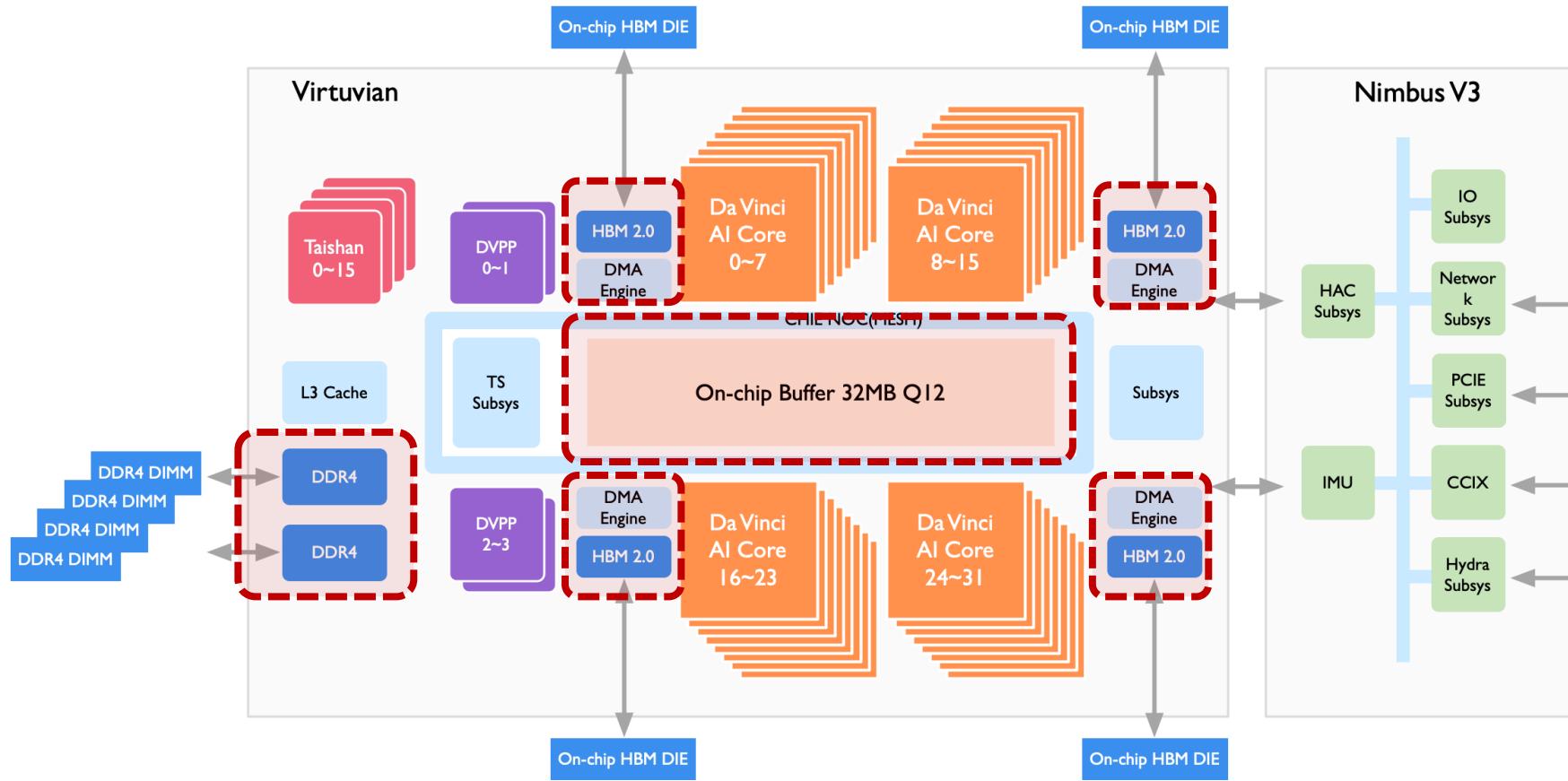
# Ascend 910 处理器逻辑架构



## DVPP:

- 数字视觉预处理子系统，完成图像视频编解码等预处理操作。

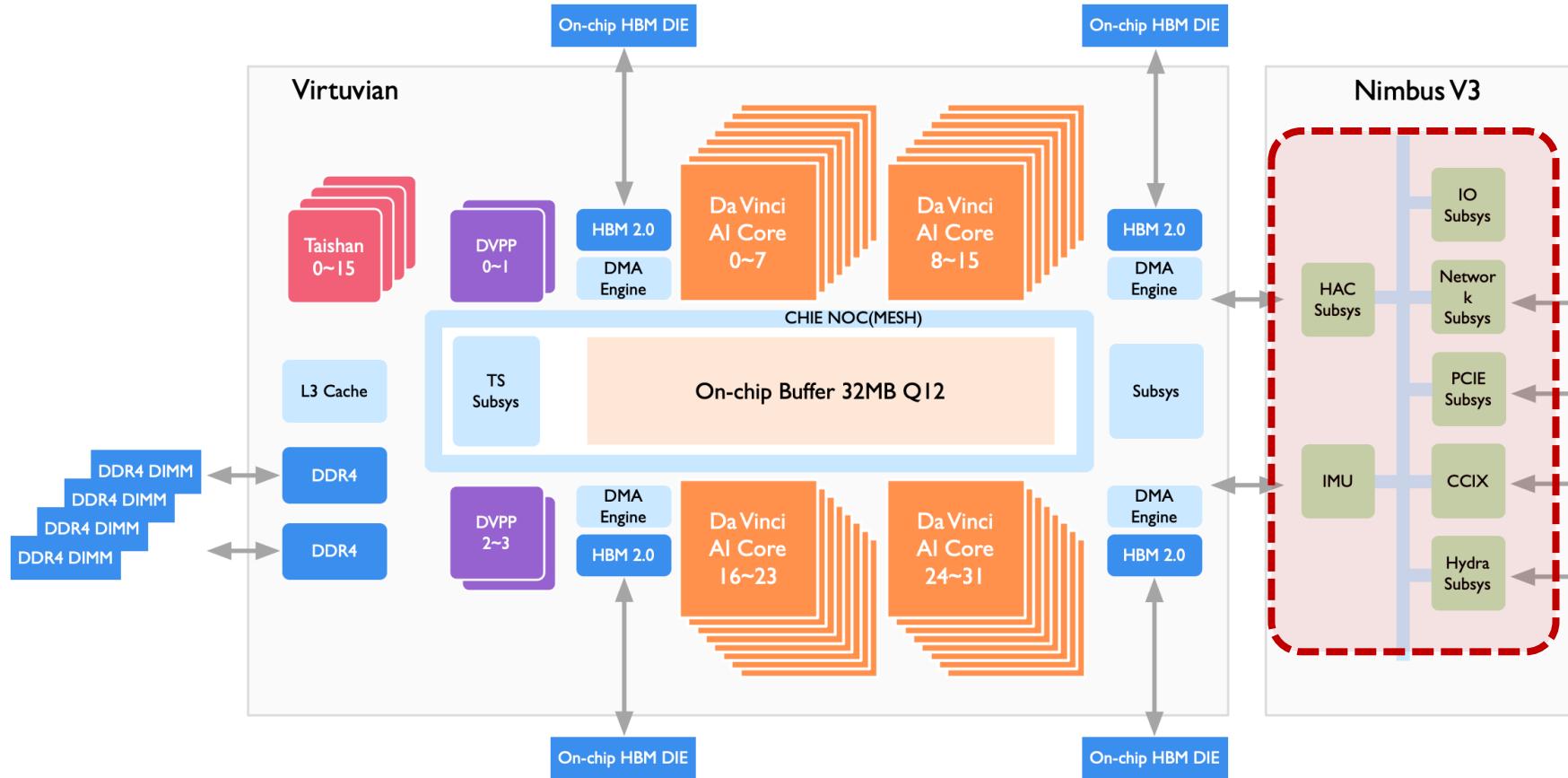
# Ascend 910 处理器逻辑架构



## Cache & Buffer:

- 片内层次化 memory 结构；
- AI Core 内部两级 memory buffer；
  - SOC 片上 64MB L2 buffer，  
专用于 AI Core、AI CPU，  
提供高带宽、低延迟的 memory 访问；
- Virtruvian 连接 N 个 HBM 颗粒；芯片集成 DDR 控制器，为芯片提供 DDR 内存。

# Ascend 910 处理器逻辑架构



## Nimbus:

- 提供x16 PCIe 接口和 Host CPU 对接；
- 提供100G NIC (支持RoCE V2协议) 用于跨服务器传递数据；
- 集成1个 ARM CPU 核，执行启动、功耗控制等硬件管理任务。

# 小结与思考

# 思考

- 那么多不同的训练和推理产品形态，对应的 AI 芯片到底有多少种？怎么组合？
- 你还想了解昇腾哪些内容呢？





# Thank you

把AI系统带入每个开发者、每个家庭、  
每个组织，构建万物互联的智能世界

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

Copyright © 2023 XXX Technologies Co., Ltd.  
All Rights Reserved.

The information in this document may contain predictive statements including, without limitation, statements regarding the future financial and operating results, future product portfolio, new technology, etc. There are a number of factors that could cause actual results and developments to differ materially from those expressed or implied in the predictive statements. Therefore, such information is provided for reference purpose only and constitutes neither an offer nor an acceptance. XXX may change the information at any time without notice.



Course [chenzomi12.github.io](https://chenzomi12.github.io)

GitHub [github.com/chenzomi12/DeepLearningSystem](https://github.com/chenzomi12/DeepLearningSystem)