Ascend C 算子编程概述

HUAWEI

持续打造极致性能、极简易用的全场景人工智能平台

全栈自主创新
使能行业应用

开放工具支撑应用开发
加速应用落地

开源框架使能模型开发
加速科研创新

开放算子使能伙伴开发
加速计算性能

开放模组主板
使能伙伴开发

行业应用　互联网、金融、电信、智慧城市、教育、能源、交通、制造等更多行业

应用使能

ModelArts　HiAI Service　第三方平台

MindX 昇腾应用使能

- MindX DL 深度学习使能
- MindX Edge 智能边缘使能
- ModelZoo 优选模型库
- MindX SDK 行业SDK

AI框架

[M]S 昇思 MindSpore
最佳匹配昇腾AI处理器算力的全场景AI框架

PyTorch / TensorFlow等第三方框架
可基于第三方框架进行模型训练和推理

异构计算框架

CANN
统一异构计算架构，释放昇腾硬件澎湃算力

全流程开发工具链 MindStudio

管理运维工具 FusionDirector / SmartKit

昇腾社区 hiascend.com

系列硬件

端　边　云

Atlas及伙伴系列硬件

HUAWEI

什么是Ascend C 算子

Ascend C是CANN针对算子开发场景推出的编程语言，原生支持C和C++标准规范，兼具开发效率和运行性能。基于Ascend C编写的算子程序，通过编译器编译和运行时调度，运行在昇腾AI处理器上。使用Ascend C，开发者可以基于昇腾AI硬件，高效的实现自定义的创新算法。使用Ascend C编程语言开发的算子我们称之为Ascend C算子。

使用Ascend C开发自定义算子的优势：

- ✓遵循C/C++编程规范，匹配用户开发习惯
- ✓自动并行调度，获得最优执行性能
- ✓结构化核函数编程，简化算子开发逻辑
- ✓CPU/NPU孪生调试，提升算子调试效率

什么场景需要开发自定义算子

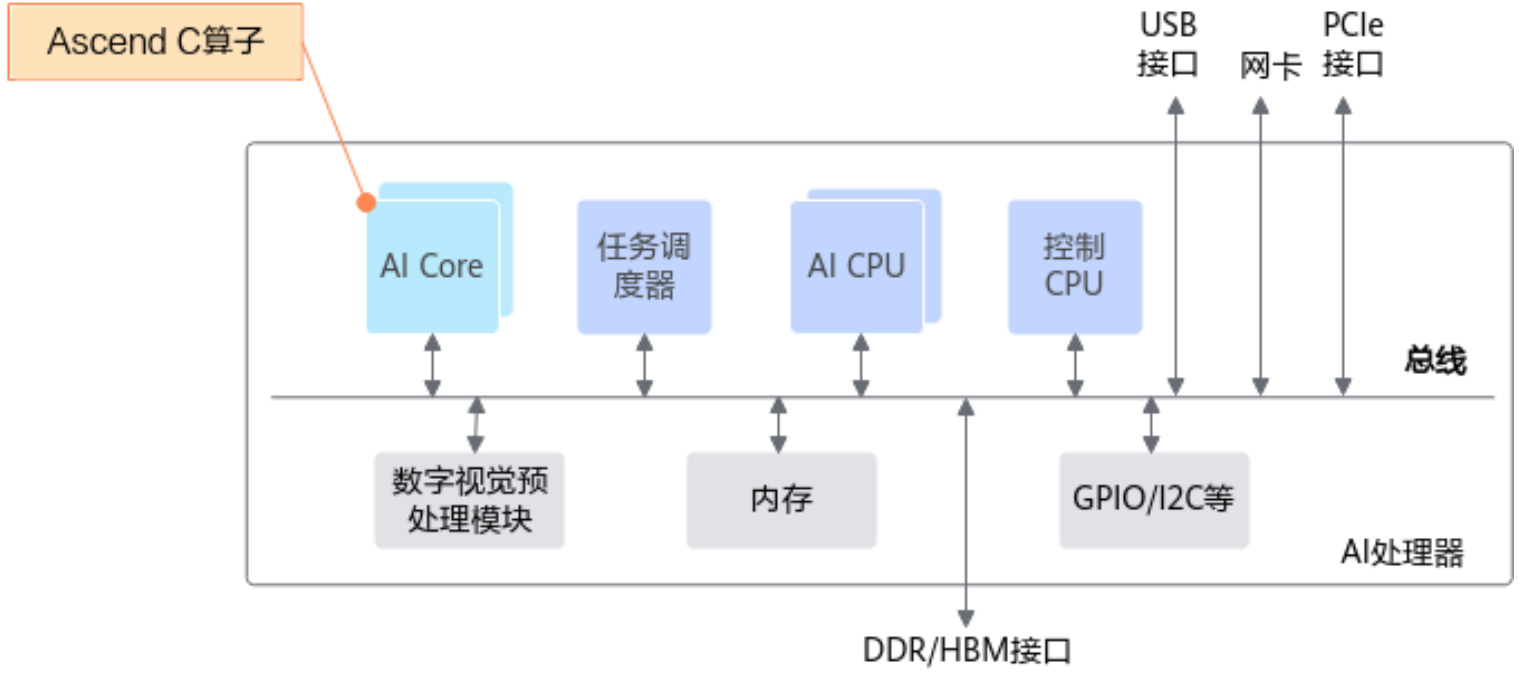
一般场景下，开发者无需自己开发算子，但若遇到以下场景，开发者需要考虑进行自定义算子的开发。

- 推理场景场景下，将第三方框架模型（例如TensorFlow、Caffe、ONNX等）使用ATC工具转换为适配昇腾平台的离线模型时遇到了不支持的算子。
- 推理场景下，若应用程序中的某些逻辑涉及到数学运算（例如查找最大值，进行数据类型转换等），开发者可以将这些操作通过自定义算子的方式进行实现，然后在应用程序中对算子进行调用，从而利用昇腾AI处理器进行加速。

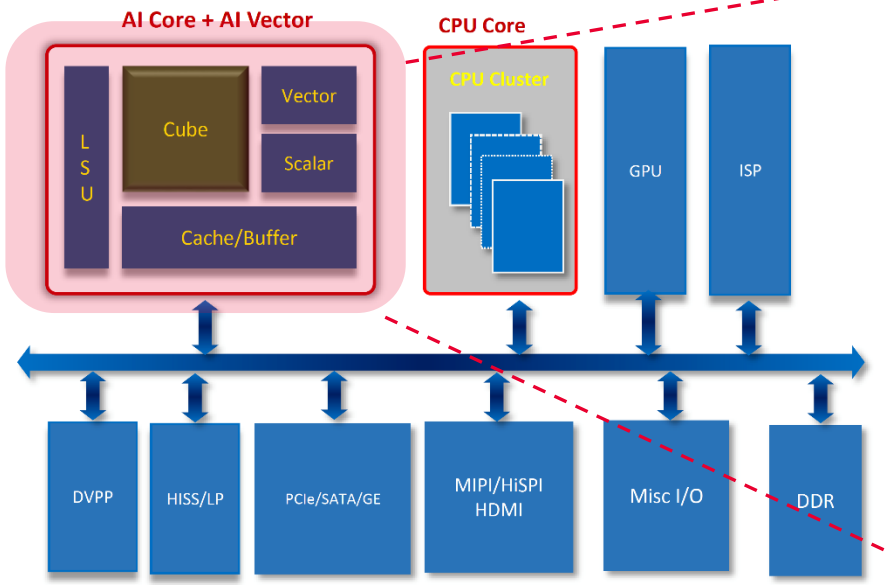
  例如，针对一个分类应用，我们想从分类模型的推理结果中查找可能性最大的前5个标识，则可以实现一个查找最大值的算子（例如ArgMax），后续就可以直接通过AscendCL接口调用此算子实现对推理结果的后处理。
- 训练场景下，将第三方框架（例如TensorFlow、PyTorch等）的网络训练脚本迁移到昇腾平台时遇到了不支持的算子。
- 网络调优时，发现某算子性能较低，想重新开发一个高性能算子替换性能较低的算子。

昇腾（Ascend）AI处理器

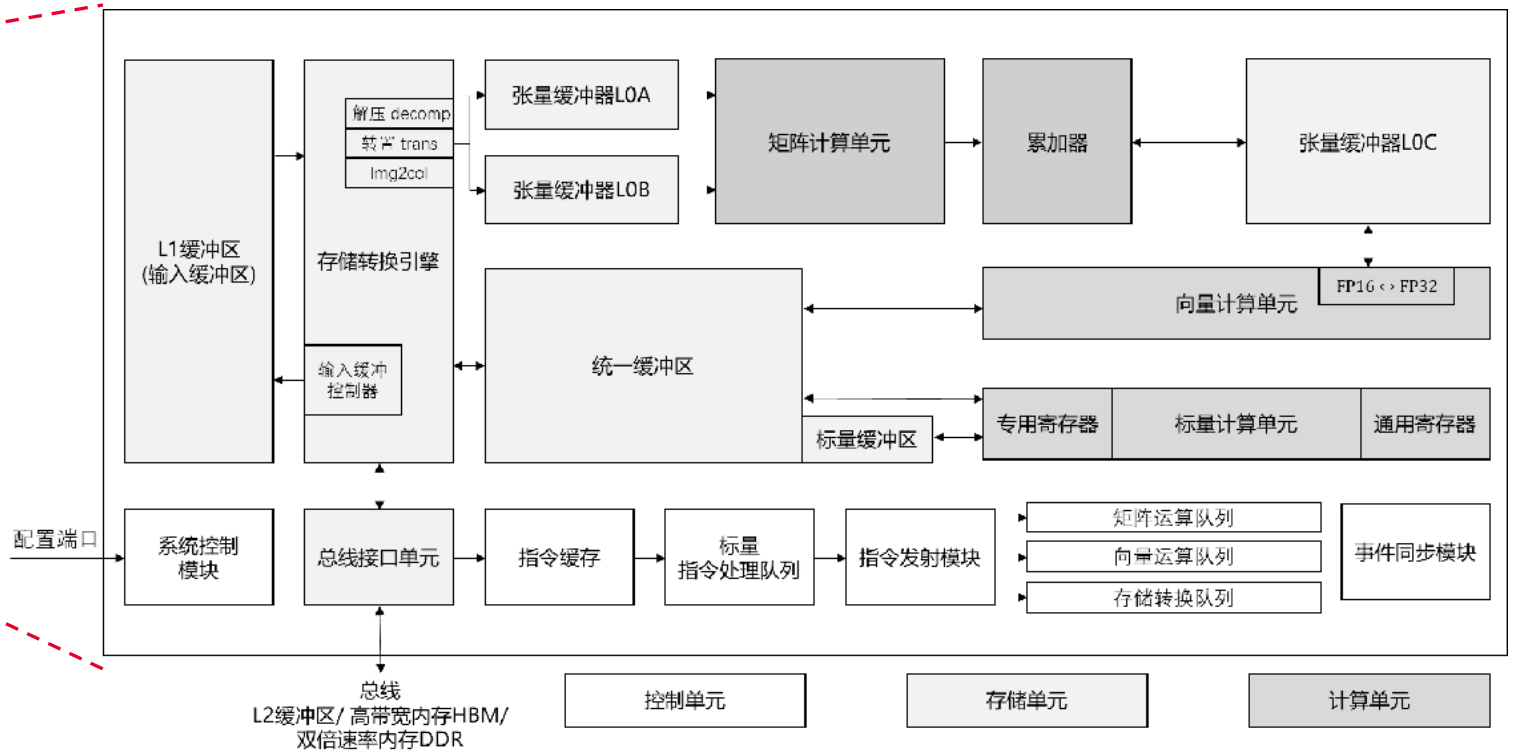
- 昇腾AI处理器芯片的逻辑架构
  - 芯片系统控制处理器（Control CPU）
  - 面向计算密集型任务的AI 计算核心（AI Core）
  - 面向非矩阵计算任务的AI处理器（AI CPU）
  - 层次化的片上系统缓存/缓冲区
  - 数字视觉预处理模块（Digital Vision Pre-Processing，DVPP）
  - I/O接口。



AI Core的逻辑架构



昇腾AI处理器



AI Core

- AI Core是昇腾AI处理器的计算核心，采用华为自研的达芬奇架构，通常也被叫做DaVinci Core
- 根据不同的处理器版本，AI Core里的计算、存储和带宽资源有不同的规格

达芬奇架构主要部分：

- 计算单元： 包含三种基础计算资源（矩阵计算单元、向量计算单元、标量计算单元）
- 存储系统： AI Core的片上存储单元和相应的数据通路构成了存储系统。
- 控制单元：整个计算过程提供了指令控制，相当于AI Core的司令部，负责整个AI Core的运行。

HUAWEI

AI Core内部并行计算架构抽象

使用**Ascend C编程语言**开发的算子运行在AI Core上，AI Core是**昇腾AI处理器**中的计算核心

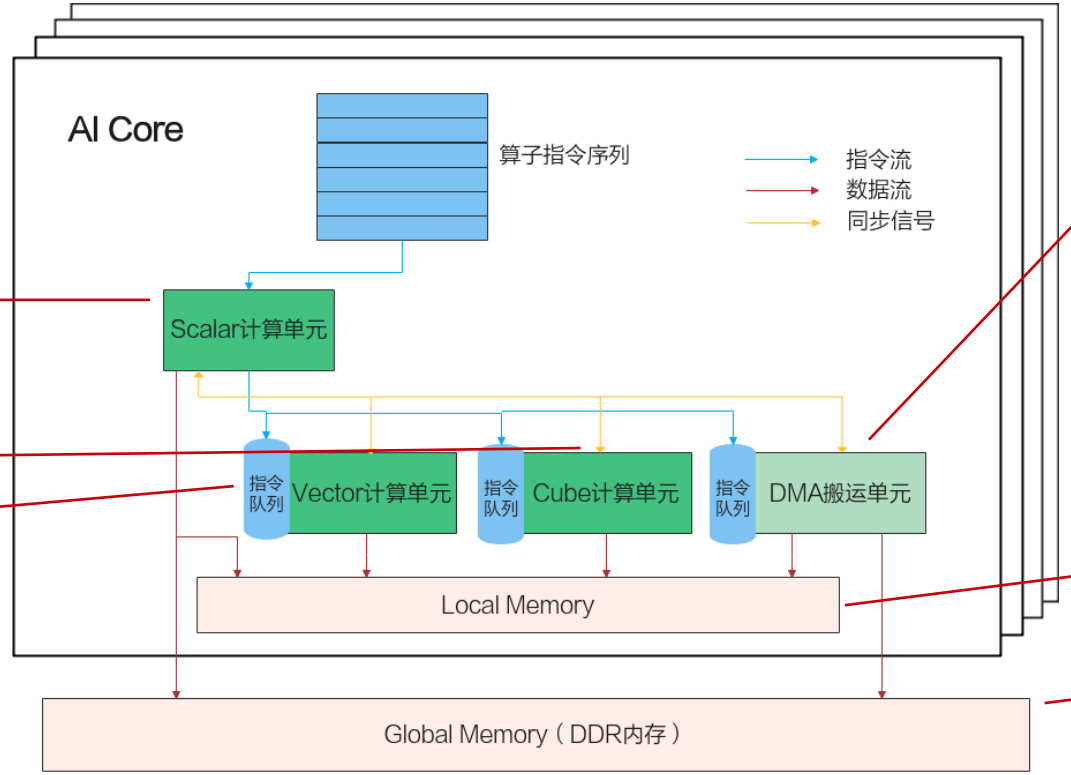
一个AI处理器内部有多个AI Core，AI Core中包含**计算单元、存储单元、搬运单元**等核心组件

计算单元包括了三种基础计算资源

① **Scalar计算单元：**执行地址计算、循环控制等标量计算工作，并把向量计算、矩阵计算、数据搬运、同步指令发射给对应单元执行

② **Cube计算单元：**负责执行矩阵运算

③ **Vector计算单元：**负责执行向量运算



搬运单元负责在Global Memory和Local Memory之间搬运数据，包含搬运单元MTE2（Memory Transfer Engine，数据搬入单元），MTE3（数据搬出单元）

存储单元为AI Core的内部存储，统称为Local Memory

与此相对应，AI Core的外部存储称之为Global Memory

AI Core内部并行计算架构抽象示意图

AI Core内部并行计算架构抽象

异步指令流

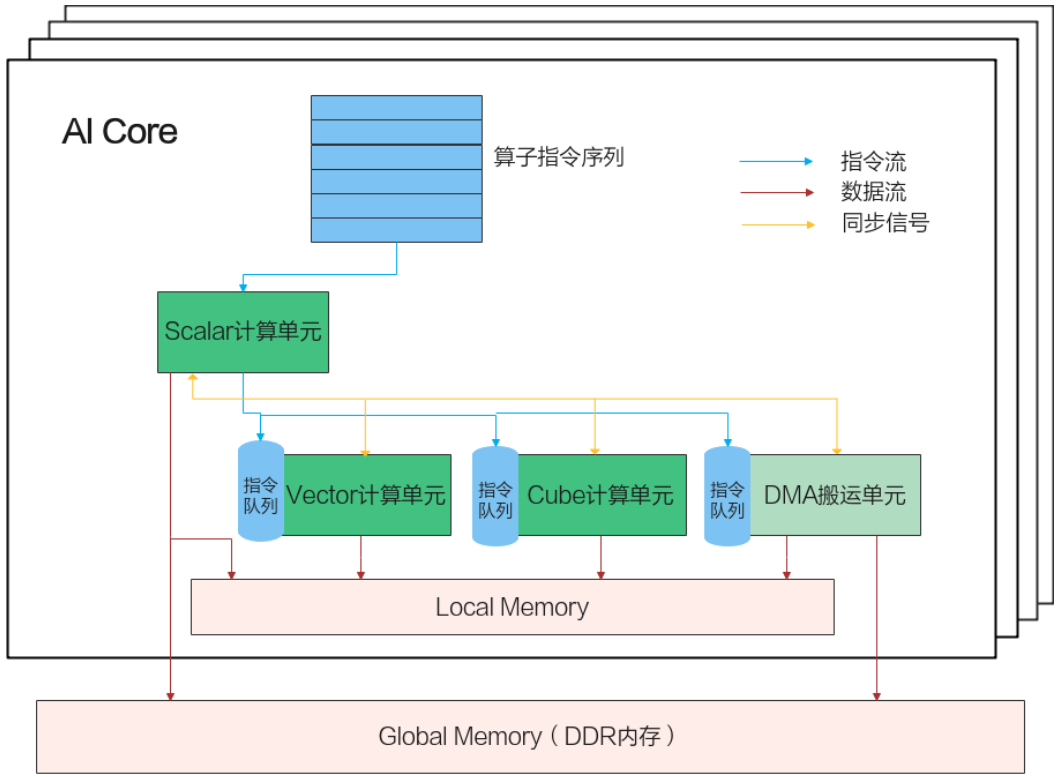
Scalar计算单元读取指令序列，并把向量计算、矩阵计算、数据搬运指令发射给对应单元的指令队列，向量计算单元、矩阵计算单元、数据搬运单元异步的并行执行接收到的指令

同步信号流

指令间可能会存在依赖关系，为了保证不同指令队列间的指令按照正确的逻辑关系执行，Scalar计算单元也会给对应单元下发同步指令

计算数据流

DMA搬入单元把数据搬运到Local Memory，Vector/Cube计算单元完成数据计算，并把计算结果写回Local Memory，DMA搬出单元把处理好的数据搬运回Global Memory



AI Core内部并行计算架构抽象示意图

Thank you.

昇腾开发者社区

http://hiascend.com

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