智能芯片设计

第5章 智能芯片架构设计

华中科技大学 人工智能与自动化学院
多谱信息智能处理技术全国重点实验室

本章内容

5.1 时域与空域计算架构

5.2 DianNao系列智能芯片架构

5.3 Thinker智能芯片架构

5.4 DNPU智能芯片架构

5.5 A310智能芯片架构

5.5 A310智能芯片架构

□ AI芯片—华为昇腾310/910

Ascend 310 —面向推理

- 架构: 达芬奇
- 半精度 (FP16): 8 TFLOPS
- 整数精度 (INT8) : 16 TOPS
- 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
- 12nm制程

Ascend 910 —面向训练

- 架构: 达芬奇
- 半精度 (FP16): 256 TFLOPS
- 整数精度 (INT8) : 512 TOPS
- 128 通道 全高清 视频解码器 – H.264/265;
- 最大功耗: 350W
- 7nm制程

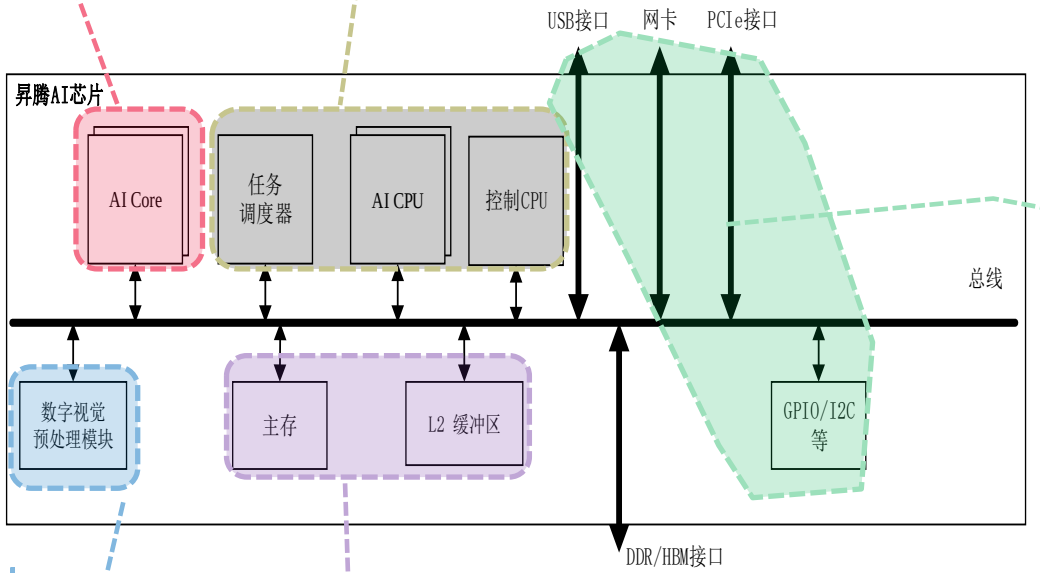
5.5 A310智能芯片架构

AI Core

- DSA达芬奇架构
- 矩阵/向量/标量计算
- 集成了2个AI Core

ARM CPU核心

- 8个ARM A55，用作AI/控制 CPU，执行不适合AI Core上的计算（承担非矩阵类复杂计算），两类数可动态分配
- 1个专用CPU作为任务调度器（TS），计算任务在AI Core上的高效分配和调度，该CPU专门服务于AI Core和AI CPU



对外接口

- PCIE/RGMII/USB
- UART/I2C/SPI

DVPP

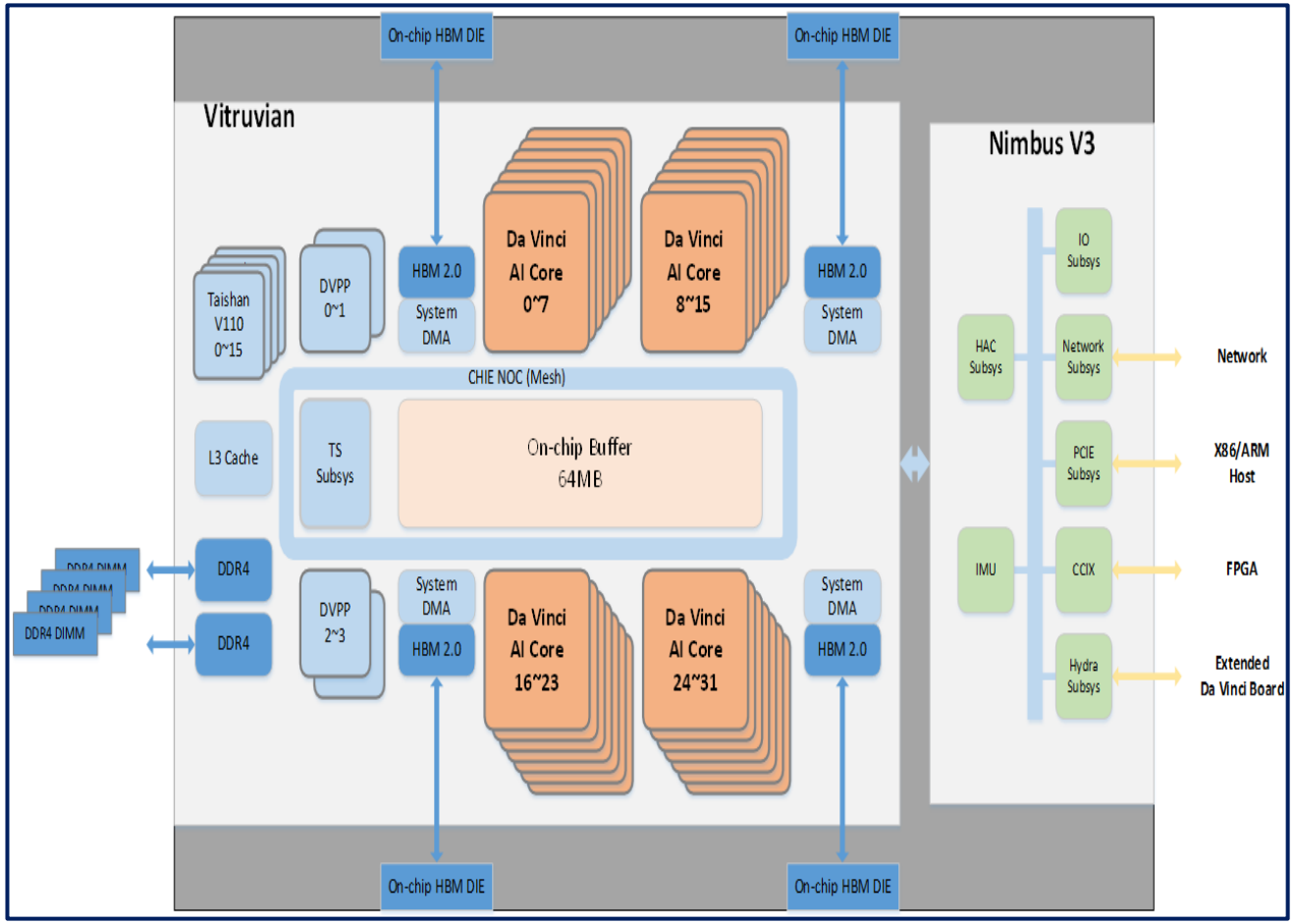
- 数字视觉预处理子系统
- 图像视频编解码
- 格式和精度转换

Cache & Buffer

- AI core内部L0+L1两级访存
- 8MB L2 buffer
- LPDDR4x控制器

5.5 Ascend 910处理器整体架构

- Ascend 910处理器主要架构组成
  - ◆AI数据处理子系统（AI Core）
  - ◆计算子系统（CPU）
  - ◆图像视频处理子系统（DVPP）
  - ◆存储子系统（层次化的片上系统缓存Cache或缓冲区Buffer）
  - ◆低速外设接口（Nimbus外部通信模块）



5.5 Ascend 910处理器整体架构

CPU子系统

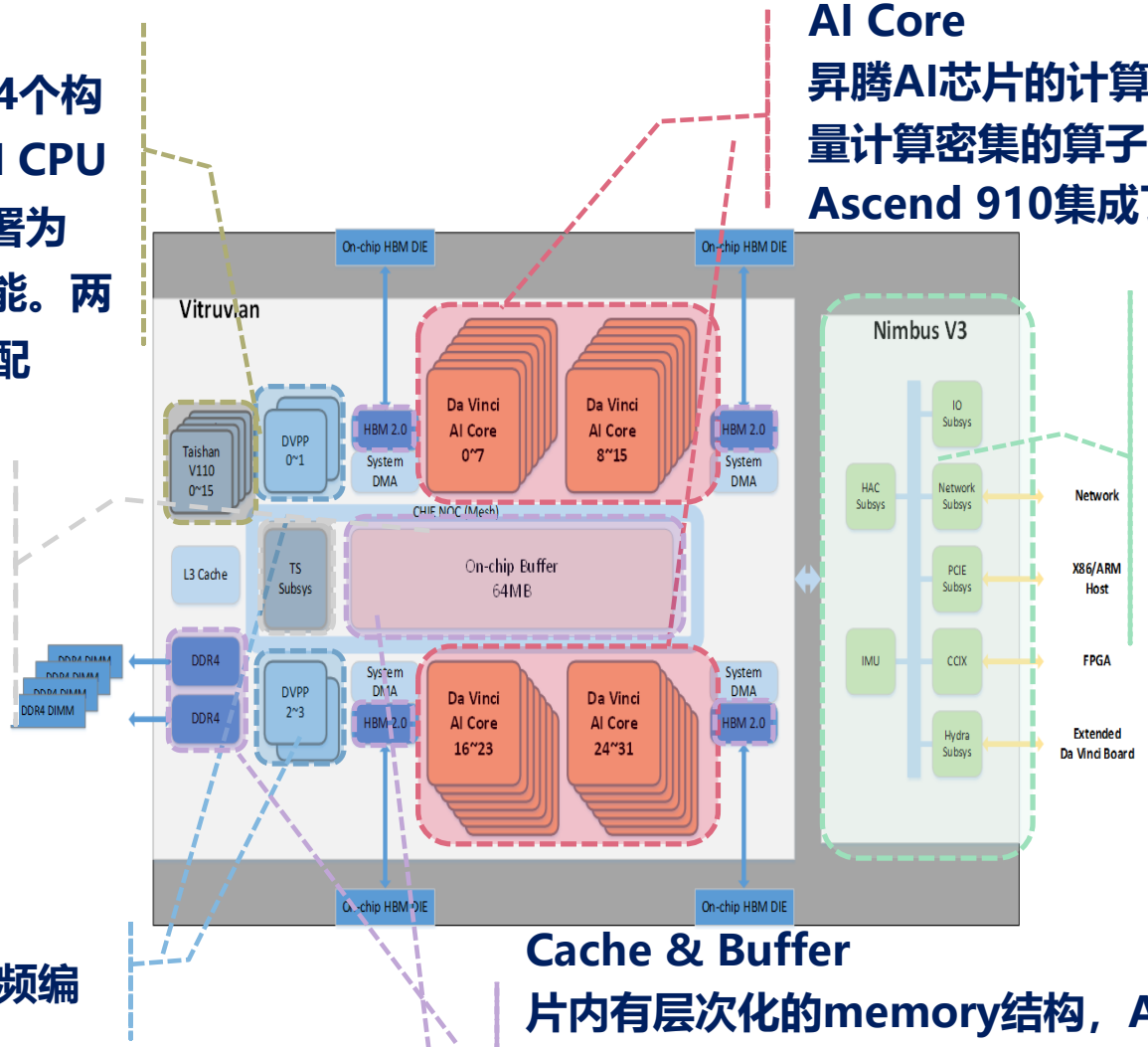
集成16个TaishanV110 Core（4个构成一个Cluster）。部分部署为AI CPU，承担部分AI计算功能；部分部署为Ctrl CPU，负责整SoC的控制功能。两类CPU占用的CPU核数由软件分配

TS CPU

独立的4核A55 Cluster 负责任务调度，算子任务切分之后，通过硬件调度器（HWTS），分发给AI Core或AI CPU

DVPP

数字视觉预处理子系统，图像视频编解码等预处理操作



AI Core

昇腾AI芯片的计算核心，主要负责执行矩阵、向量计算密集的算子任务，采用达芬奇架构。Ascend 910集成了32个AI Core。

Nimbus

提供x16 PCIe 4.0接口，和Host CPU对接，100G NIC（支持ROCE V2协议）用于跨服务器传递数据；集成1个A53 CPU核，执行启动、功耗控制等管理任务

Cache & Buffer

片内有层次化的memory结构，AI Core内部有两级memory buffer，还有64MB L2 buffer，专用于AI Core、AI CPU，提供高带宽、低延迟的访存。Virtruvian连接4个HBM 2.0颗粒，总计32GB，还集成DDR 4.0控制器，提供DDR内存

5.5 A310智能芯片架构

口 计算

AI Core核心计算模块分为张量、向量和标量三类计算单元，完成不同类型的数据计算，Cube计算单元是设计亮点，支持16*16运算，每个cycle完成4096次MAC运算

口 存储

按读写速度分为GPR/SPR(通用、特殊功能寄存器)，L0 buffer，L1 buffer、Unified buffer以及L2 buffer/HBM/DDR

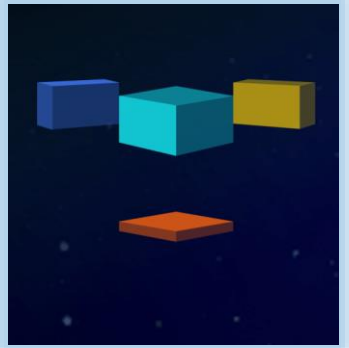
MTE （Memory Transform Engine）是内存传输引擎，主要完成AI Core内部数据在不同Buffer之间的读写管理及一些格式转换的操作

口 控制

负责协调AI Core的整体控制流，参数配置和功耗控制等。其中标量指令处理队列主要实现控制指令的译码，当指令译码完成之后，指令发射模块根据指令的类型，会分发到相应的运算队列去执行计算（矩阵运算队列、向量运算队列和存储转换队列）

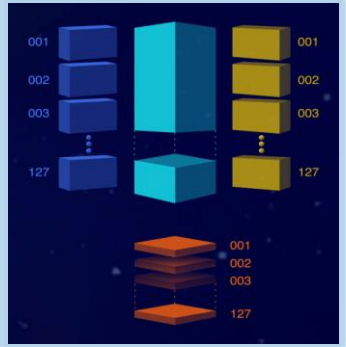
5.5 A310智能芯片—达芬奇架构

标量计算



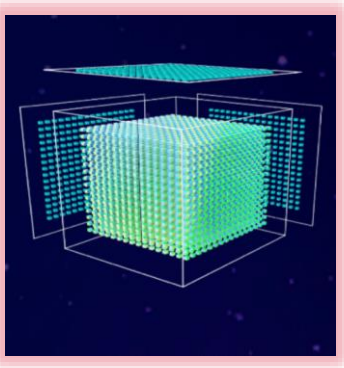
0.00X TOPS / W

向量计算



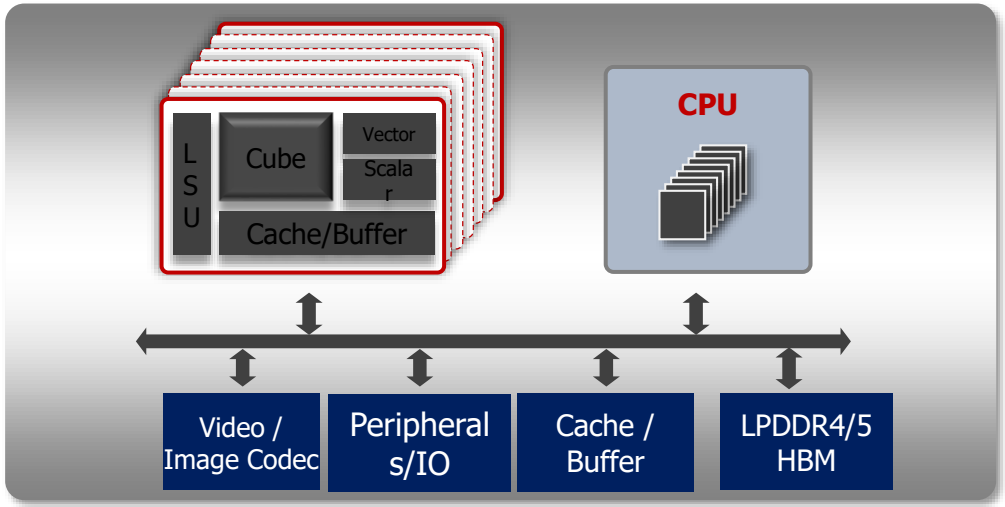
0.X TOPS / W

张量计算



X TOPS / W

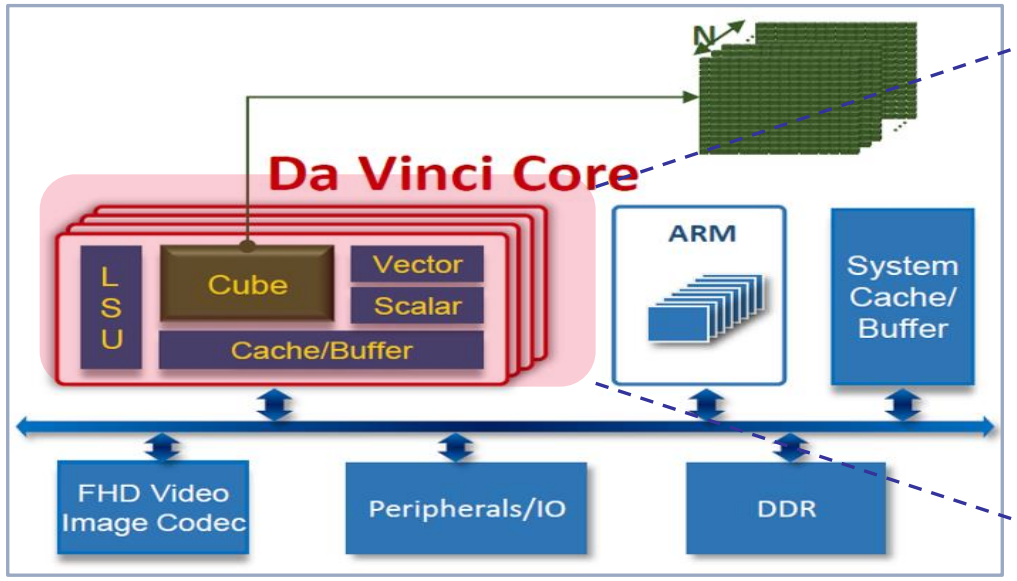
单周期完成
4096个FP16
MAC 运算

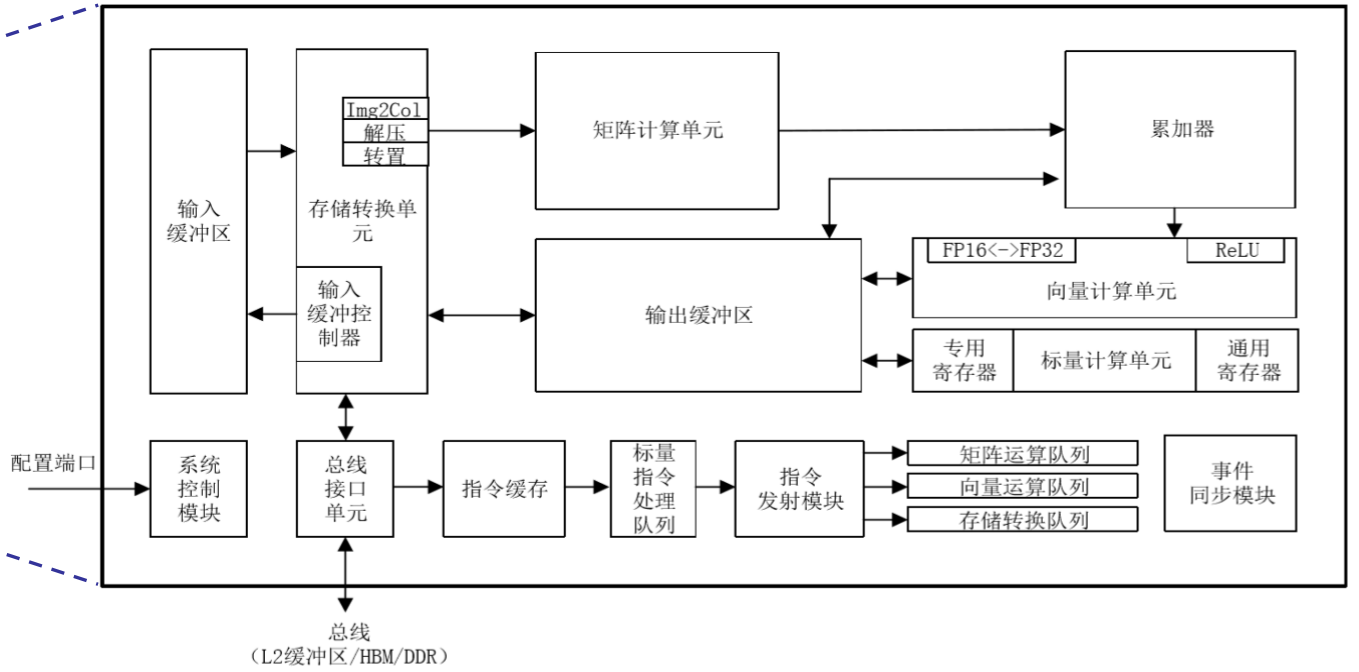


3D Cube:16^3三维弹性立方体

- 高算力：可用一条指令完成4096个FP16 MAC 运算
- 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

5.5 A310智能芯片— AI Core





- 计算单元：矩阵、向量、标量
- 存储系统： AI Core片上存储和相应的数据通路构成了存储系统
- 控制单元：为整个计算过程提供了指令控制，负责整个AI Core的运行

- 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
- Ascend310/910，AI Core里的计算、存储和带宽资源有不同的规格

5.5 A310智能芯片— 计算单元

- 三种基础计算资源：矩阵、向量和标量计算单元，对应矩阵、向量和标量计算

  - 标量（Scalar）：由单独一个数组成
  - 向量（Vector）：由一组一维有序数组成，每个数由一个索引标识
  - 矩阵（Matrix）：由一组二维有序数组成，每个数由两个索引标识
  - 张量（Tensor）：由一组n维有序数组成，每个数由n个索引标识

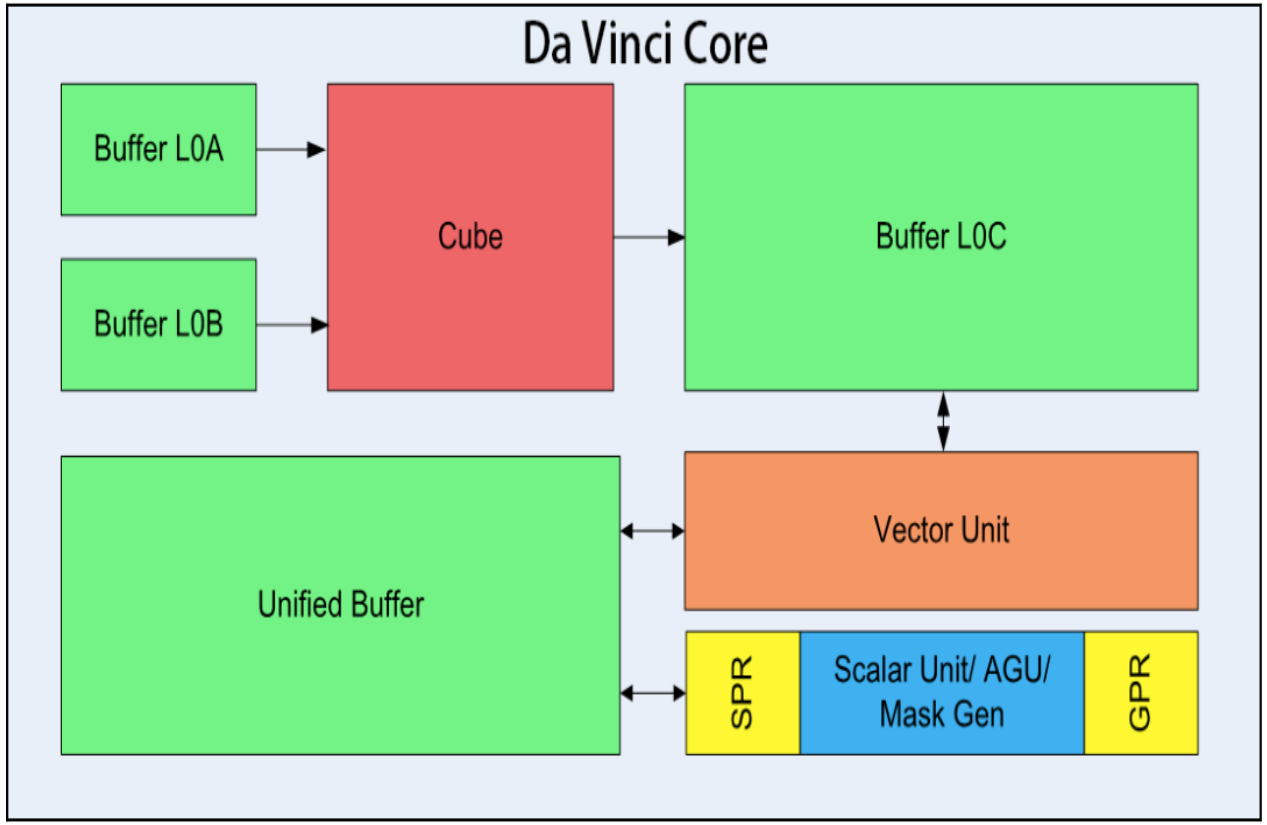
矩阵乘运算中，标量、向量、矩阵算力密度依次增加

达芬奇架构设计的核心：以最小的计算代价增加矩阵乘的算力，实现更高的AI能效

5.5 A310智能芯片— 计算单元

- 三种基础计算资源：矩阵、向量和标量计算单元，各司其职

- 3D Cube矩阵乘法单元：算力担当，用于完成矩阵乘，Buffer L0A、L0B和L0C则用于存储输入矩阵和输出矩阵数据，负责向Cube计算单元输送数据和存放计算结果。

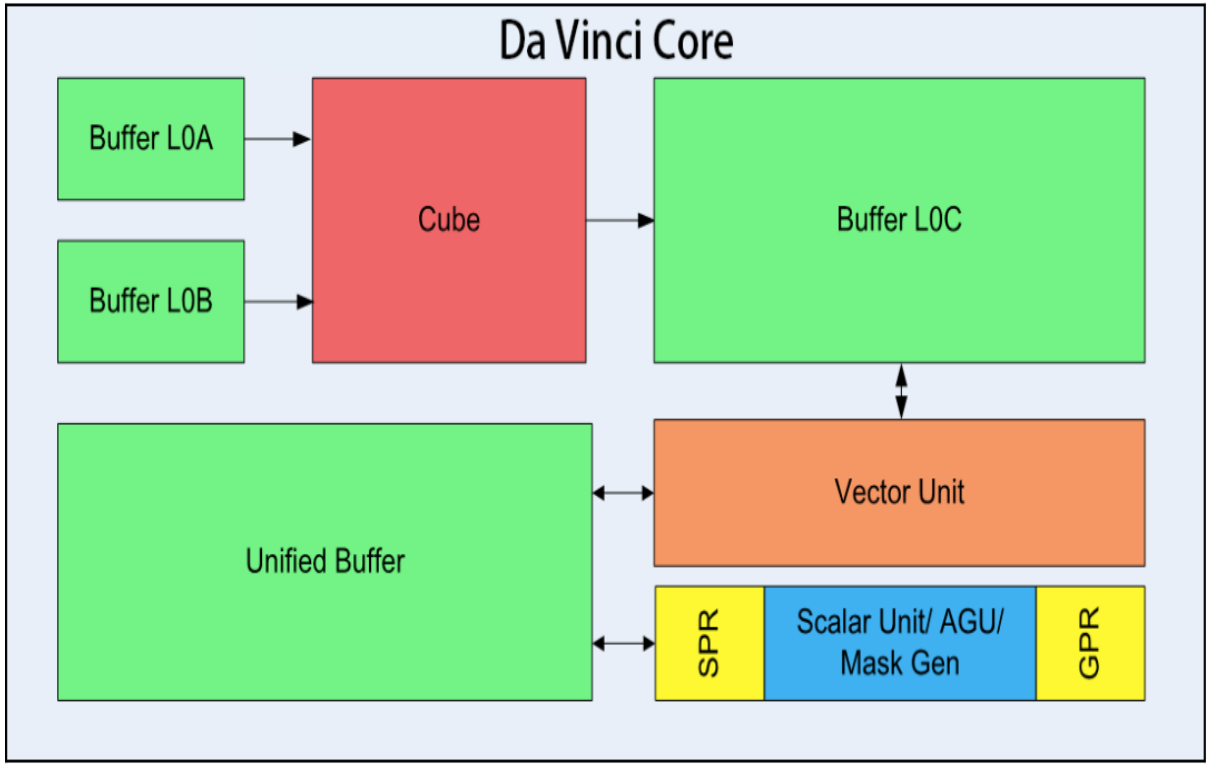


5.5 A310智能芯片— 计算单元

- 三种基础计算资源：矩阵、向量和标量计算单元，各司其职

■ Vector向量计算单元：灵活的多面手

虽然Cube的算力很强大，但只能完成矩阵乘运算，还有很多计算类型要依靠Vector向量计算单元来完成。Vector的指令相对来说非常丰富，可以覆盖各种基本的计算类型和许多定制的计算类型。

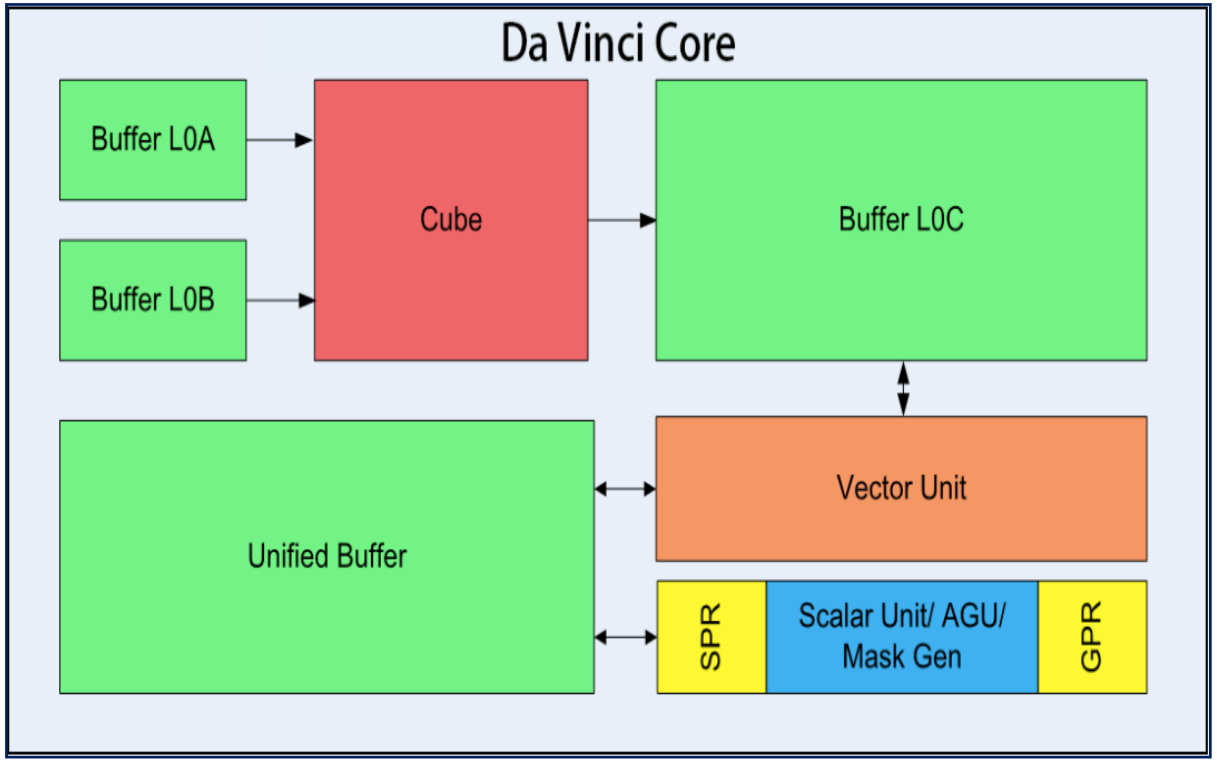


5.5 A310智能芯片— 计算单元

- 三种基础计算资源：矩阵、向量和标量计算单元，各司其职

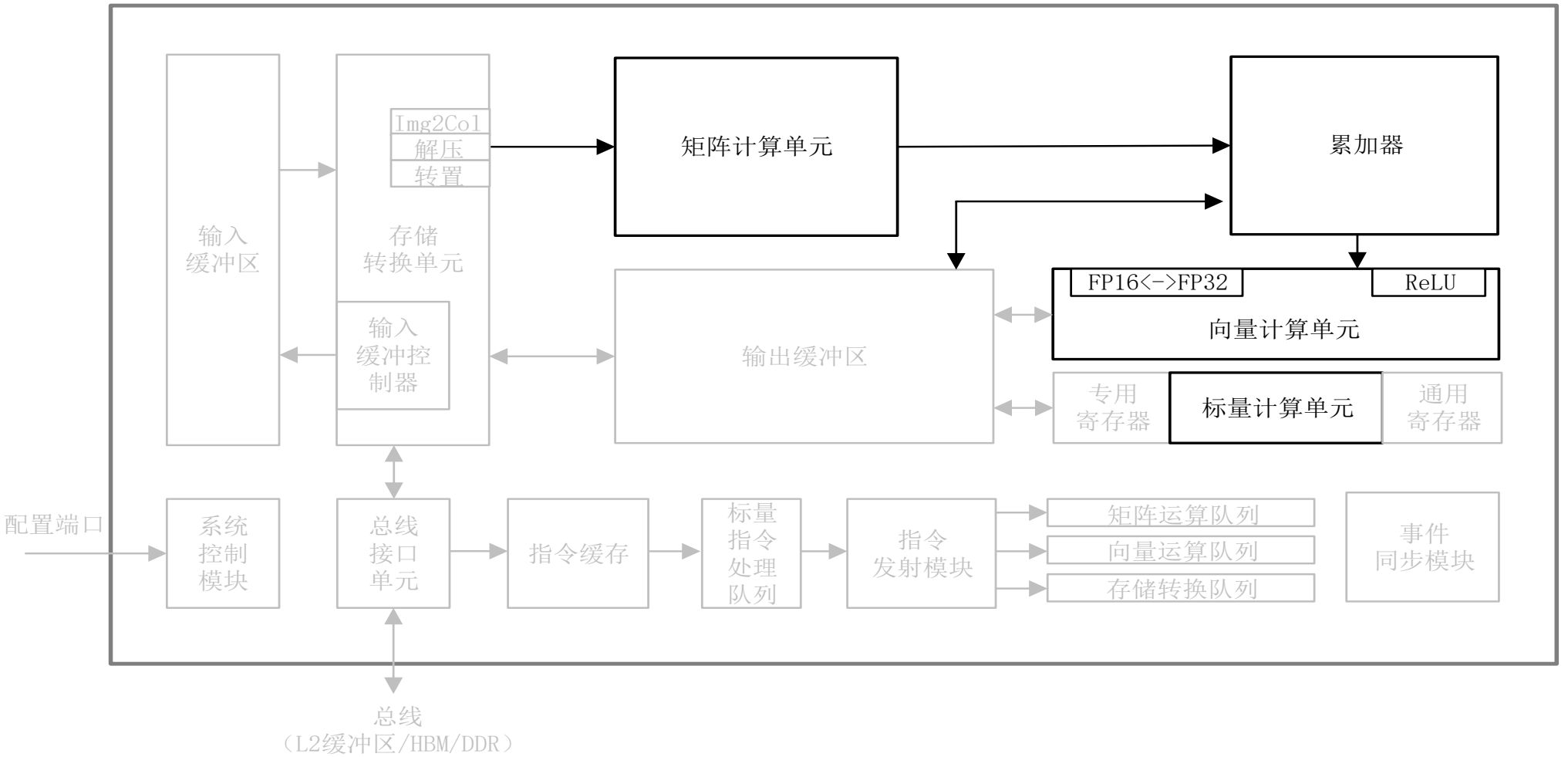
■ Scalar标量计算单元：流程控制的管家

主要负责AI Core的标量运算，功能上可以看作一个小CPU，完成整个程序的循环控制，分支判断，为Cube和Vector提供数据地址和参数计算，并能实现基本的算术运算。



5.5 A310智能芯片— 计算单元

- 三种基础计算资源：矩阵、向量和标量计算单元，对应矩阵、向量和标量计算



5.5 A310智能芯片—计算单元

- 三种基础计算资源：矩阵、向量和标量计算单元，对应矩阵、向量和标量计算

◆矩阵计算单元（Cube Unit）：

矩阵计算单元和累加器主要完成矩阵相关运算。一个指令完成一个fp16的 16x16矩阵乘（4096）；如果是int8输入，则一拍完成 16x32 与 32x16 矩阵乘（8192）

◆向量计算单元（Vector Unit）：

实现向量和标量，或双向量之间的计算，功能覆盖各种基本的计算类型和许多定制的计算类型，主要包括FP16/ FP32/Int32/Int8等数据类型的计算。一拍可以完成两个128长度fp16类型的向量相加/乘， 或者64个fp32/int32类型的向量相加/乘

◆标量计算单元（Scalar Unit）：

相当于一个微型CPU，控制整个AI Core的运行

◆累加器：

把当前矩阵乘的结果与前次计算的中间结果相加， 可以用于完成卷积中加bias操作

5.5 A310智能芯片—计算加速原理

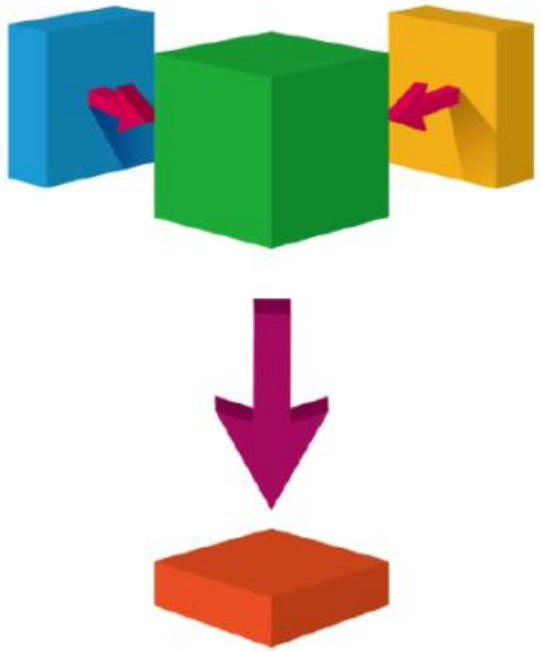
Scalar Unit
Full Flexibility Computation

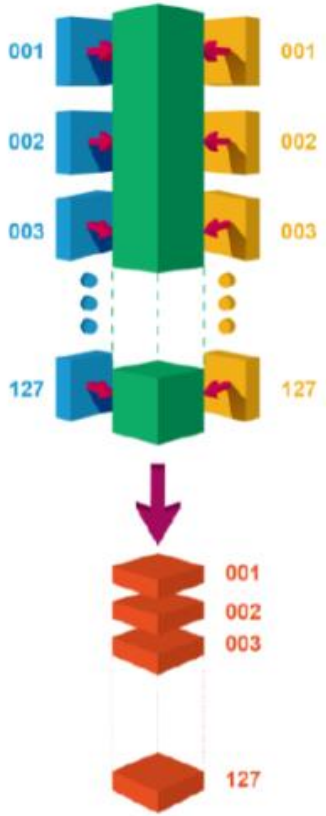
+

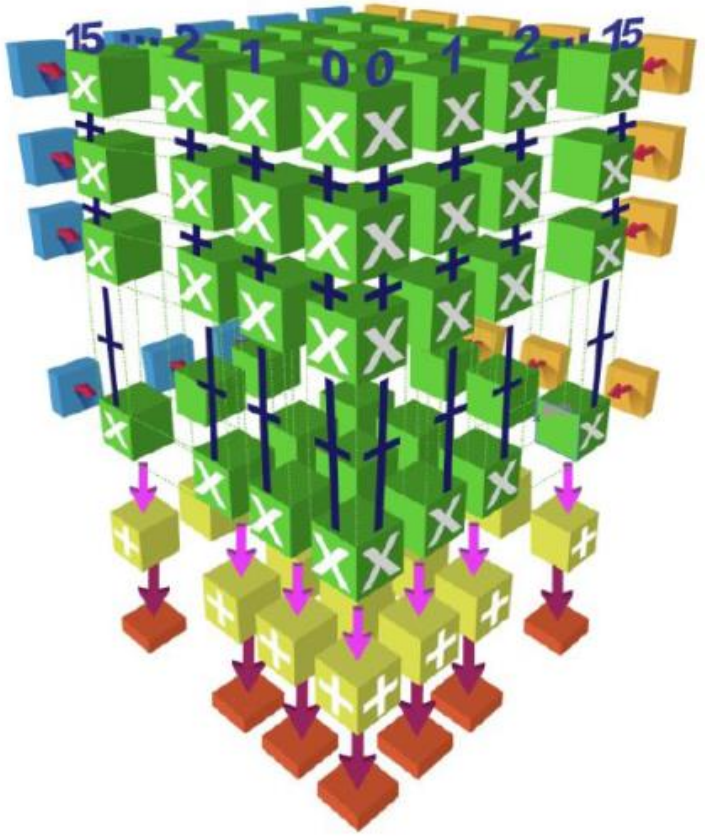
Vector Unit
Rich & Efficient Operations

+

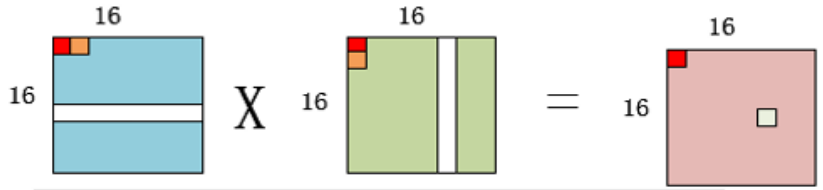
Cube Unit
High Intensity Computation







5.5 A310智能芯片—计算加速原理




```
float a[16][16], b[16][16], c[16][16];
```

CPU:

```
for(int i=0; i< 16; i++)
    for (int j=0; j<16; j++)
        for(int k=0; k<16; k++) {
            c[i][j] += a[i][k] * b[k][j];
        }
```

Vector:

```
for(int i=0; i<16; i++)
    for ( int j =0; j<16; j++) {
        c[i][j] = a[i][:] *+  b[:][j]
    }
```

CUBE:

```
CUBE:  c[:][:] = a[:][:] X b[:][:]
```

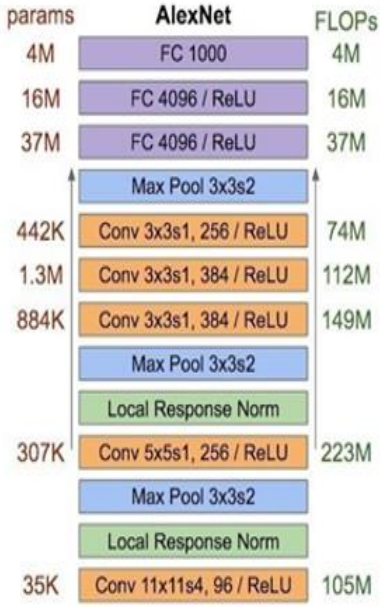
算力密度高

灵活

Cycle=16*16*16*2 = 8192
DataNum per cycle: Rd 2, Wr 1

Cycle=16*16 = 256
DataNum per cycle: Rd 2*16, Wr 16

Cycle=1
DataNum per cycle: Rd 2*16*16
Wr:16*16



AlexNet模型每层

每秒浮点运算次数及参数数量

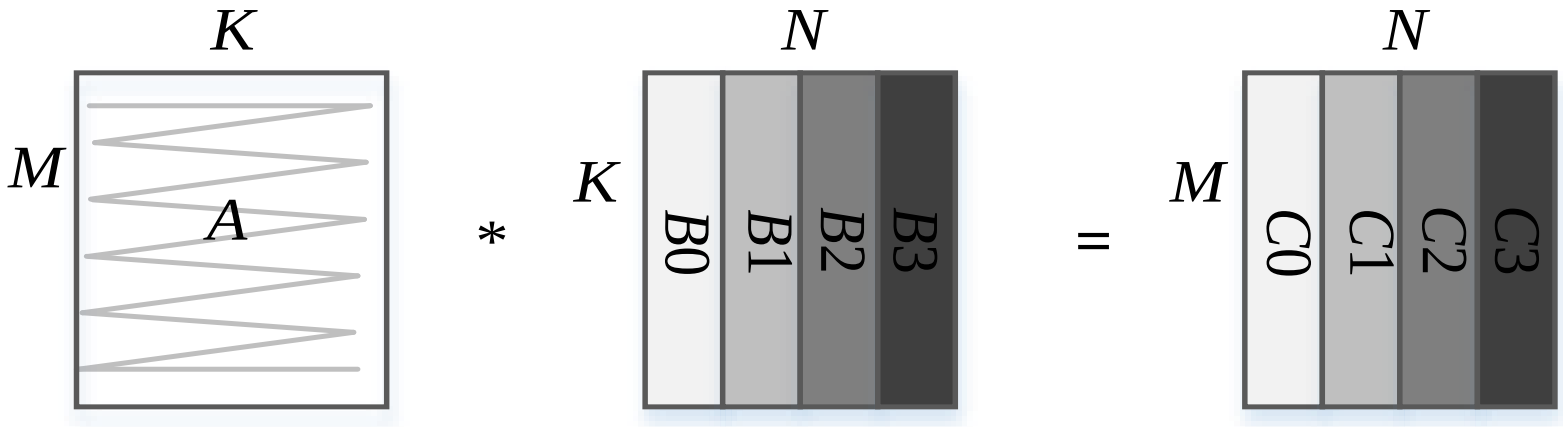
CNN经典模型的内存，计算量和参数数量对比			
	AlexNet	VGG16	Inception-v3
模型内存(MB)	> 200	> 500	90-100
参数（百万）	60	138	23.2
计算量（百万）	720	15300	5000

示例：矩阵a和矩阵b之间的乘法运算c=a*b，在不同计算单元中实现该矩阵乘，其复杂度和运算效率差别大

99%以上计算都是矩阵乘

5.5 A310智能芯片—计算单元

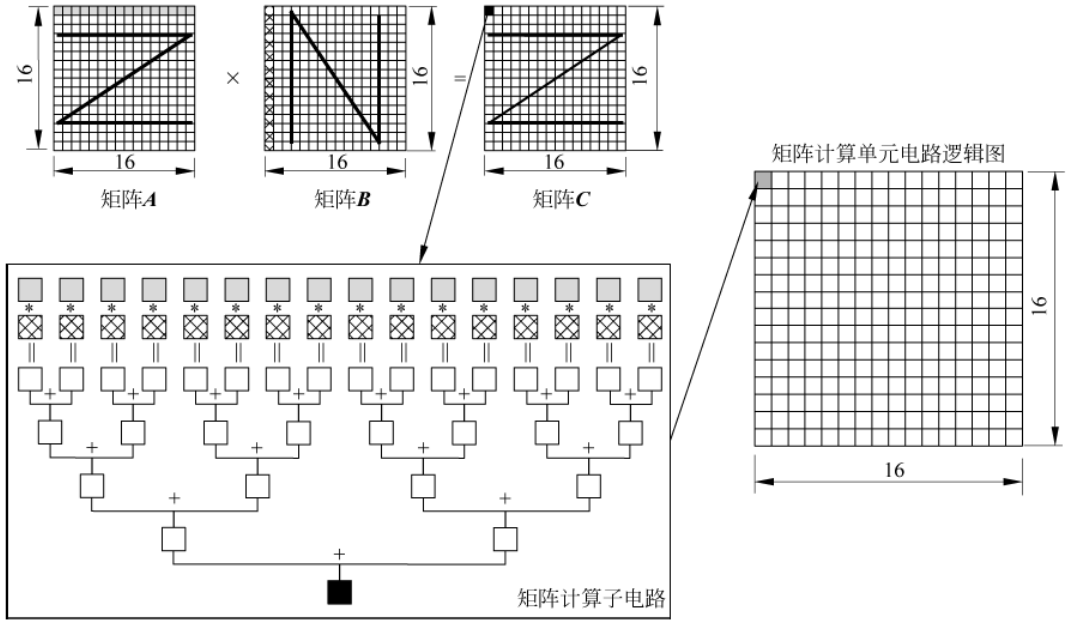
矩阵分块计算



- 矩阵较大时，由于片上计算和存储资源有限，需要对矩阵进行分块平铺处理（Tiling）
- 受限于片上缓存的容量，当一次难以装下整个矩阵B时，划分成为B0、B1、B2和B3等多个子矩阵
- 每个子矩阵的大小都可以适合一次性存储到片上缓存，并与矩阵A计算得到结果子矩阵
- 充分利用数据的局部性，尽可能把缓存中的子矩阵数据重复使用完毕，并得到所有相关的子矩阵结果后，再读入新的子矩阵开始新的周期
- 如此往复可以依次将所有的子矩阵都一一搬运到缓存中，并完成整个矩阵计算的全过程，最终得到结果矩阵C

5.5 A310智能芯片—计算单元

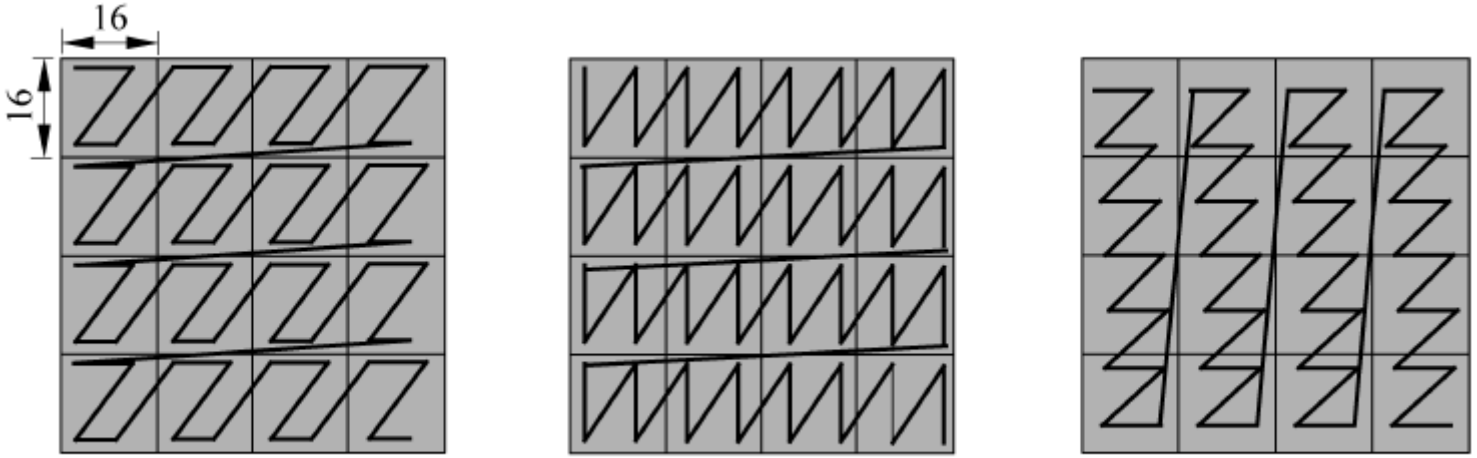
矩阵单元计算方式



- 在矩阵相乘运算中，矩阵C的第一元素由矩阵A的第一行的16个元素和矩阵B的第一列的16 个元素由矩阵计算单元子电路进行16次乘法和15次加法运算得出。矩阵计算单元中共有256个矩阵计算子电路，可以由一条指令并行完成矩阵C的256个元素计算。

5.5 A310智能芯片—计算单元

矩阵单元计算方式



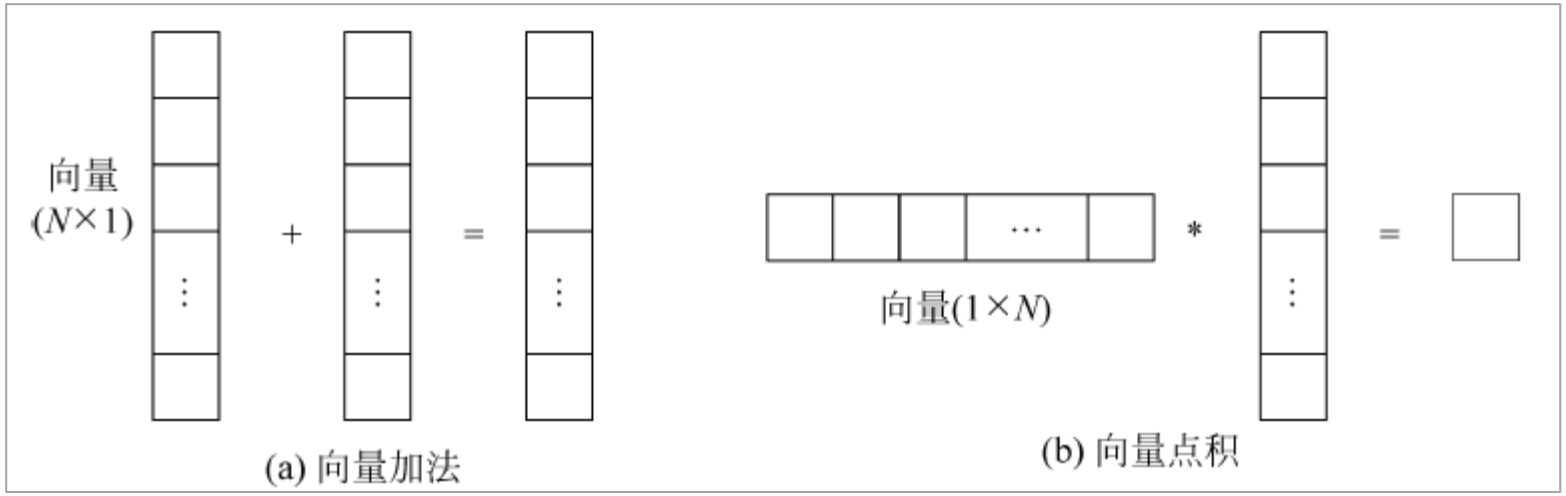
矩阵*A*　　矩阵*B*　　矩阵*C*

- 矩阵A展示的切割和排序方式称作“大Z小Z”，直观地看就是矩阵A的各个分块之间按照行的顺序排序，称为“大Z”方式；而每个块的内部数据也是按照行的方式排列，称为“小Z”方式。
- 矩阵B的各个分块之间按照行排序，而每个块的内部按照列排序，称为“大Z小 N”的排序方式。
- 昇腾AI处理器内部专用电路实现将如此排列的A、B矩阵相乘得到结果矩阵C，而矩阵C将会呈现出各个分块之间按照列排序，而每个块内部按照行排序的格式，称为“大N小Z”的排列方式。

5.5 A310智能芯片—计算单元

向量计算单元

- AI Core中的向量计算单元主要负责完成和向量相关的运算，能够实现向量和标量，或双向量之间的计算，功能覆盖各种基本和多种定制的计算类型，主要包括FP32、FP16、INT32和INT8等数据类型的计算
- 向量计算单元可以快速完成两个FP16类型的向量相加或者相乘。向量计算单元的源操作数和目的操作数通常都保存在输出缓冲器中。对向量计算单元而言，输入的数据可以不连续，这取决于输入数据的寻址模式



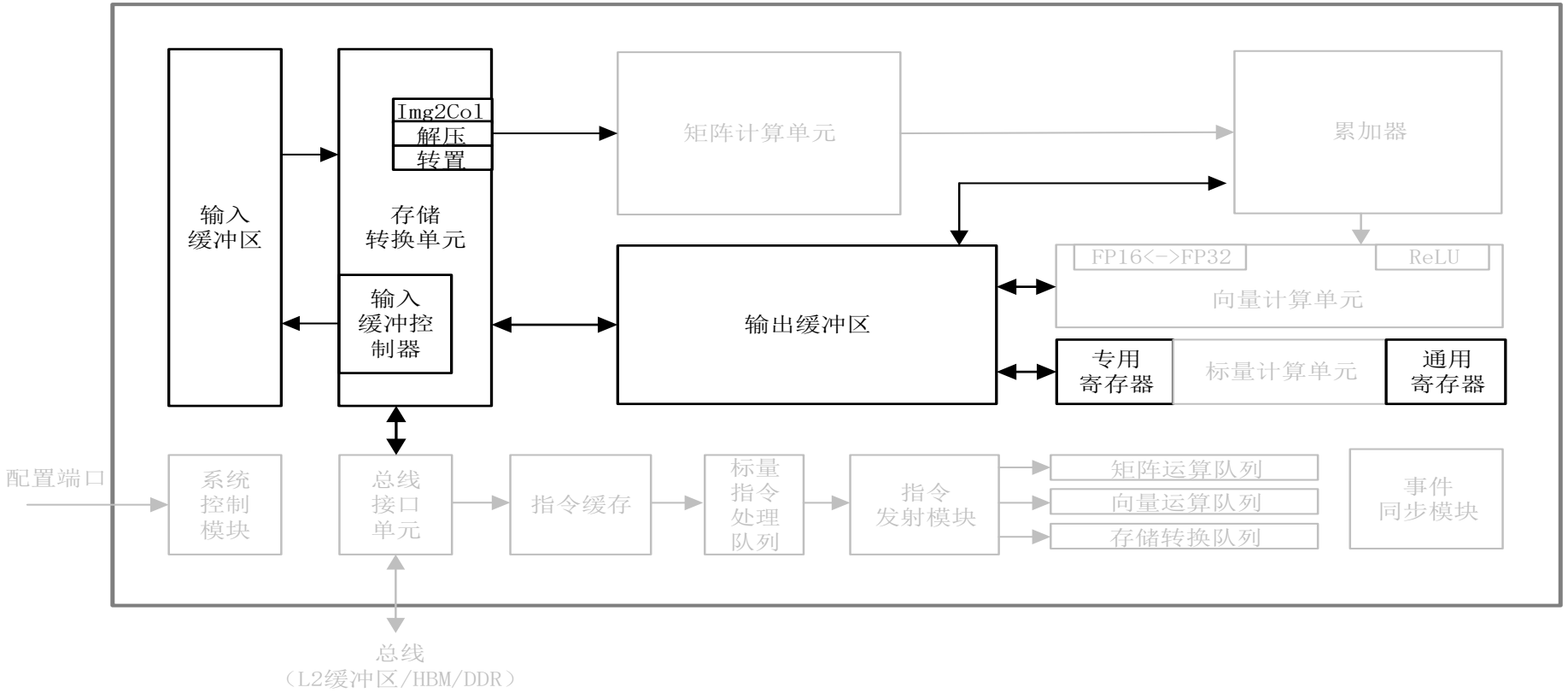
5.5 A310智能芯片—计算单元

标量计算单元

- 标量计算单元负责完成AI Core中与标量相关的运算。它相当于一个微型CPU，控制整个AI Core的运行
- 标量计算单元可以对程序中的循环进行控制，可以实现分支判断，其结果可以通过在事件同步模块中插入同步符的方式来控制AI Core中其它功能性单元的执行流水
- 为矩阵计算单元或向量计算单元提供数据地址和相关参数的计算，并且能够实现基本的算术运算。其它复杂度较高的标量运算则由专门的AI CPU通过算子完成
- 周围配备了多个通用寄存器（General Purpose Register，GPR）和专用寄存器（Special Purpose Register，SPR）。GPR用于变量或地址的寄存，为算术逻辑运算提供源操作数和存储中间计算结果；SPR为了支持指令集中一些指令的特殊功能，一般不可以直接访问，只有部分可以通过指令读写

5.5 A310智能芯片—存储系统

- AI Core采用了大容量的片上缓冲区设计，通过增大的片上缓存数据量来减少数据从片外存储系统搬运到AI Core中的频次，从而可以降低数据搬运过程中所产生的功耗，有效控制了整体计算的能耗



存储单元和相应的数据通路，构成了AI Core的存储系统

5.5 A310智能芯片—存储系统

□ 高效访存与数据通路架构

◆ 多层级访存架构

存储控制单元：数据读写管理/格式转换

输入缓冲区：暂存频繁重复使用的数据

输出缓冲区：存放NN每层中间结果

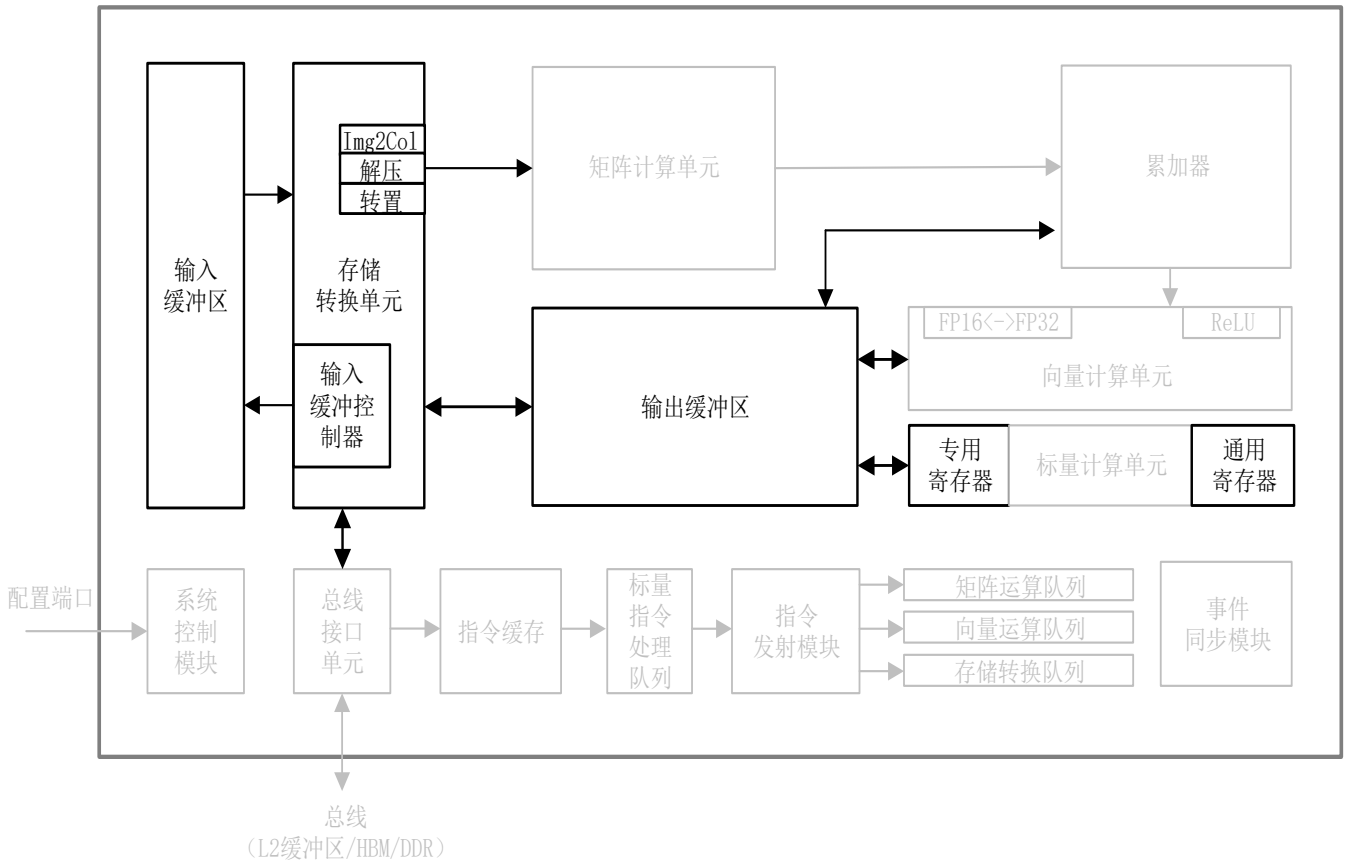
寄存器： 标量计算单元使用

◆ 数据通路——多进单出

多输入路径提高数据流入效率

单输出路径节约芯片硬件资源

CNN计算特性：多卷积核+多IFM→OFM

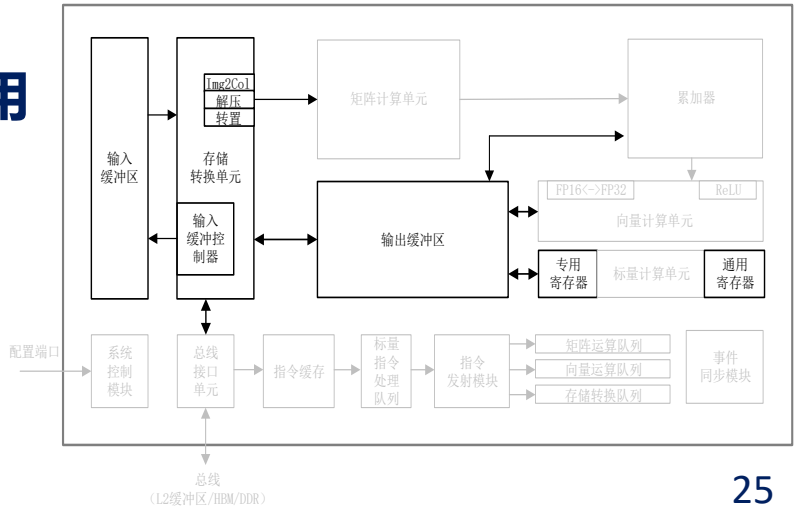


预存预取降低外存访问频次，高效通路提高数据流动效能

5.5 A310智能芯片—存储系统

□存储单元由存储控制单元、缓冲区和寄存器组成

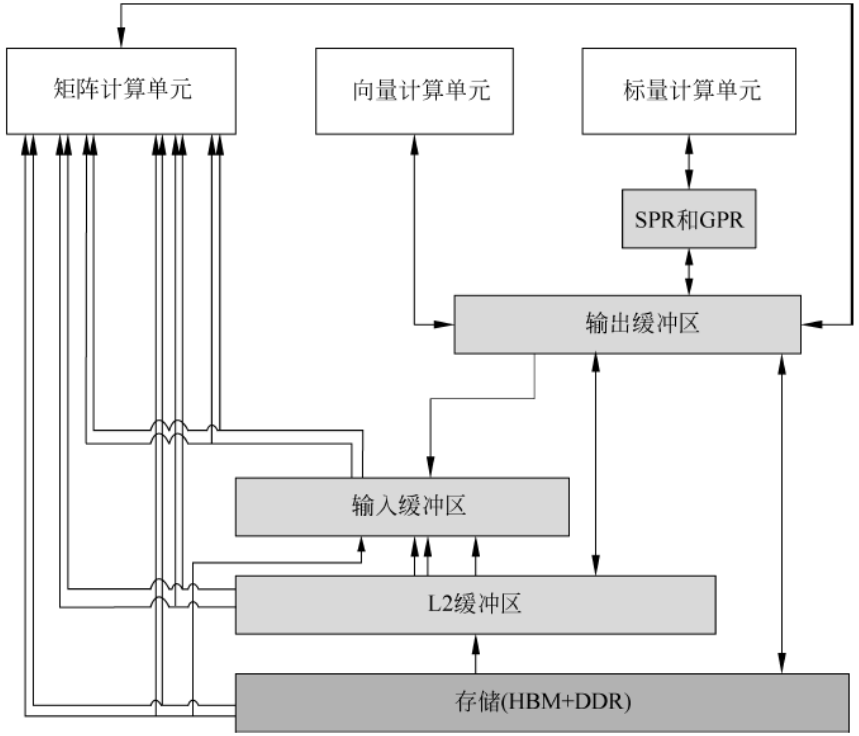
- ◆存储控制单元：通过总线接口直接访问AI Core之外的更低层级的缓存，也可以直通到DDR或HBM直接访问内存。还设置了存储转换单元，作为AI Core内部数据通路的传输控制器，负责AI Core内部数据在不同缓冲区之间的读写管理，完成格式转换操作，如补零，Img2Col，转置、解压缩等
- ◆输入缓冲区：暂存需要频繁重复使用的数据，避免频繁外部读取，减少总线访问频次的同时降低总线拥堵风险，节省功耗、提高性能
- ◆输出缓冲区：存放神经网络中每层计算的中间结果，进入下一层计算时方便获取数据。相比通过总线读取数据，通过输出缓冲区可以大大提升计算效率
- ◆寄存器： AI Core中的各类寄存器资源，主要是标量计算单元在使用



5.5 A310智能芯片—存储系统

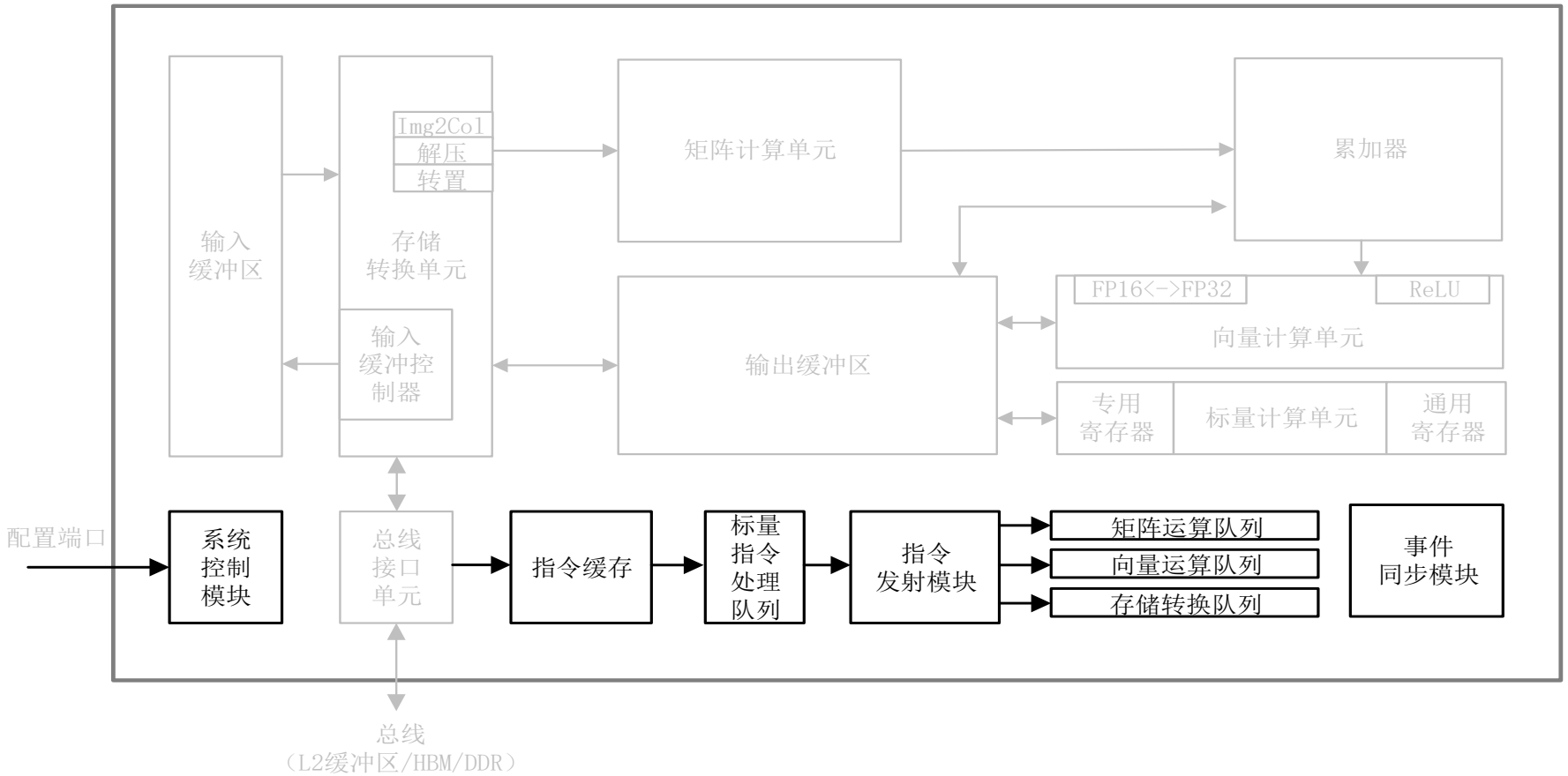
口数据通路：是指AI Core在完成一次计算任务时，数据在其中的流通路径

达芬奇架构数据通路的特点是多进单出，神经网络计算中，输入数据种类繁多、数量巨大，通过并行输入方式提高数据流入效率；与此相反，多种输入数据处理完成后往往只生成输出特征矩阵，数据种类相对单一，单输出的数据通路，可以节约硬件资源



5.5 A310智能芯片—控制单元

- 控制单元主要组成部分为系统控制模块、指令缓存、标量指令处理队列、指令发射模块、矩阵运算队列、向量运算队列、存储转换队列和事件同步模块



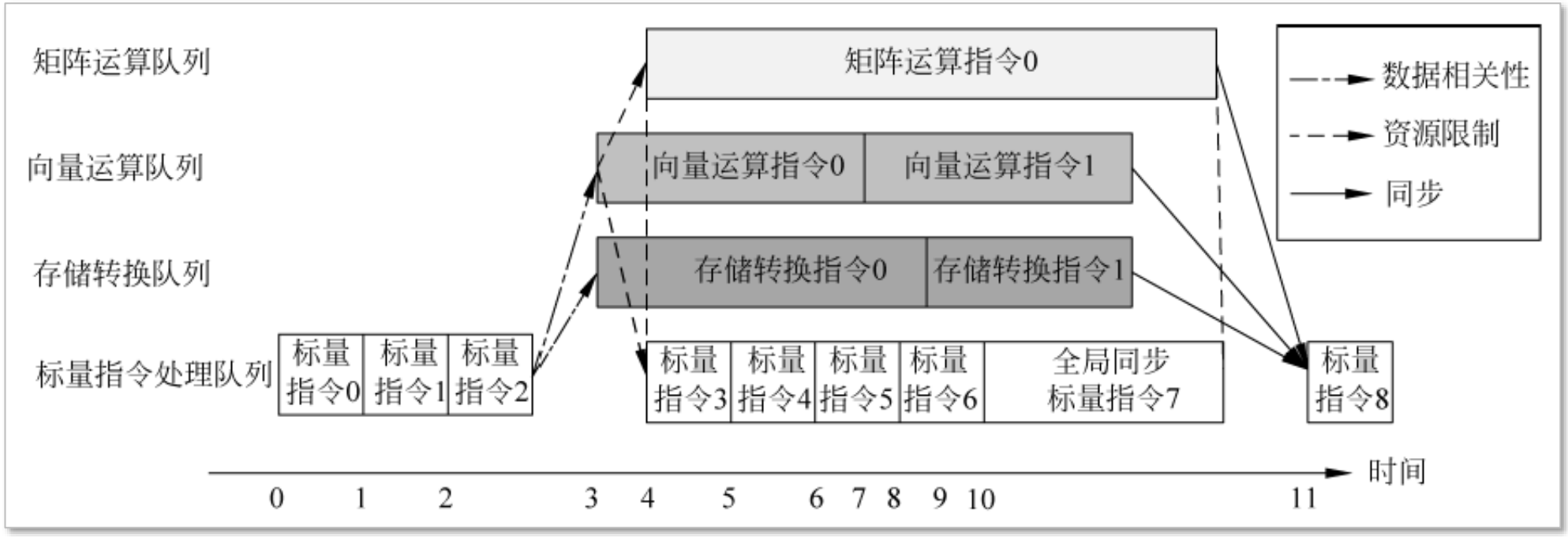
5.5 A310智能芯片—控制单元

- 系统控制模块：初始化AI Core配置；控制任务块（AI Core最小计算任务粒度）的执行进程，在任务块执行完成后，处理中断和申报状态，如果执行过程出错，把执行错误状态报告给任务调度器
- 指令缓存：在指令执行过程中，预取后续指令，并一次读入多条指令进入缓存，提升指令执行效率
- 标量指令处理队列：指令解码后导入标量队列，实现地址解码与运算控制，这些指令包括矩阵计算指令、向量计算指令以及存储转换指令等
- 指令发射模块：读取标量指令队列中配置好的指令地址和参数解码，根据指令类型分别发送到对应的指令执行队列中，标量指令驻留在标量指令处理队列中进行后续执行
- 指令执行队列：指令执行队列由矩阵运算队列、向量运算队列和存储转换队列组成，不同的指令进入相应的运算队列，队列中的指令按进入顺序执行
- 事件同步模块：时刻控制每条指令流水线的执行状态，并分析不同流水线的依赖关系，从而解决指令流水线之间的数据依赖和同步的问题

5.5 A310智能芯片—控制单元

□ 指令流水线执行流程

事件同步模块控制4条指令流水线正常执行



3.4 昇腾AI计算系统

■ 基于昇腾AI处理器的计算服务器

AI服务器

应用场景

政府	医疗	公共安全	制 造
OCR识别	图像识别	人脸识别 轨迹跟踪	质量检测

3.4 昇腾AI计算系统

■ 基于昇腾AI处理器的边端产品

AI边端产品

Atlas 200(加速模块)	Atlas 200DK(开发套件)	Atlas 300(PCIE加速卡)	Atlas 500(智能小站)
• 16TOPS INT8 @ 9.5W • 16路高清视频实时分析 • 4GB/8GB内存 | PCIe 3.0 x 4 • 工作温度：-25℃~+80℃ • 尺寸：52 x 38 x 10.2 mm	• 16TOPS INT8 @ 24W • 1*GE网口 | 1* SD卡插槽 • 8GB内存 • 工作温度：0℃~45℃ • 尺寸：125 × 80 × 24 mm	• 64TOPS INT8 @ 67W • 64路高清视频实时分析 • 32GB内存 | 204.8GB/s带宽 • PCIe 3.0 x16，半高半长卡	• 16 TOPS INT8 @ 25-40W • 支持WiFi & LTE • 16路高清视频实时分析 • 无风扇设计 | –40°C 至 +70°C

应用场景

视频分析 | OCR | 语音识别 |

精准营销 | 医疗影像分析

昇腾A310智能视觉算力

A310智能视觉算力优势明显

算法模型	A310		Hi3559A		TX2	
	耗时ms	fps	耗时ms	fps	耗时ms	fps
googlenet	14.35	278.7	22.08	181.2	35.72	112.0
resnet50	5.96	670.7	51.56	77.6	82.56	48.4
vggnet16	16.36	244.5	140.46	28.5	221.68	18.0
mobilenet	3.78	1059.3	15.10	258.6	24.80	161.3