多任务脉动阵列调研：

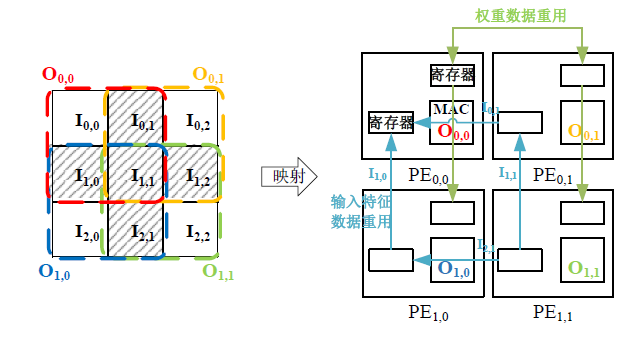
一类可配置的脉动 主要改PE结构

√异构SA：

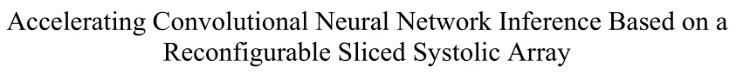


16x16 针对小尺寸和深度卷积；scaleout+交叉开关 + 多方向 + 异构PE

新数据流，单输出通道OS-S 输入垂直传输（利用计算时闲置的输出数据路径）



√切片SA：



16x16 3x3filter 阵列只有三行有效，并行多个任务，

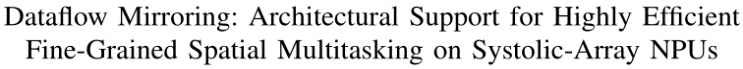
tile-column sliding dataflow, tc1、tc2、tc3和滤波器的卷积得第一个点，而tc2、tc3、tc4和得到第二个点。tc2和tc3是重用的，通过寄存器移位来读取输入图，四级流水

对于每个CNN层，不同数量的输入节点（Ci）设置不同的切片模式。

图形用户界面, 应用程序

描述已自动生成

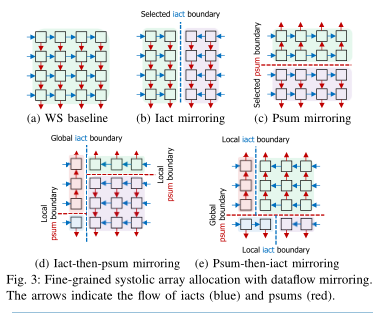
√DATA-mirror：



256x256 全方位PE间网络 任意位置划分四个任务；

life-time counter；

支持在NN进入和退出时动态重新分配硬件资源，并支持抢占



√FlexSA：



128x128 可以在运行时重新配置结构，适应剪枝DNN的工作负载；四个core独立或协同工作。

图形用户界面, 图示

中度可信度描述已自动生成



一类多线程但是不是细粒度SA切分的 异构

多线程技术：

√MPNA：

文本

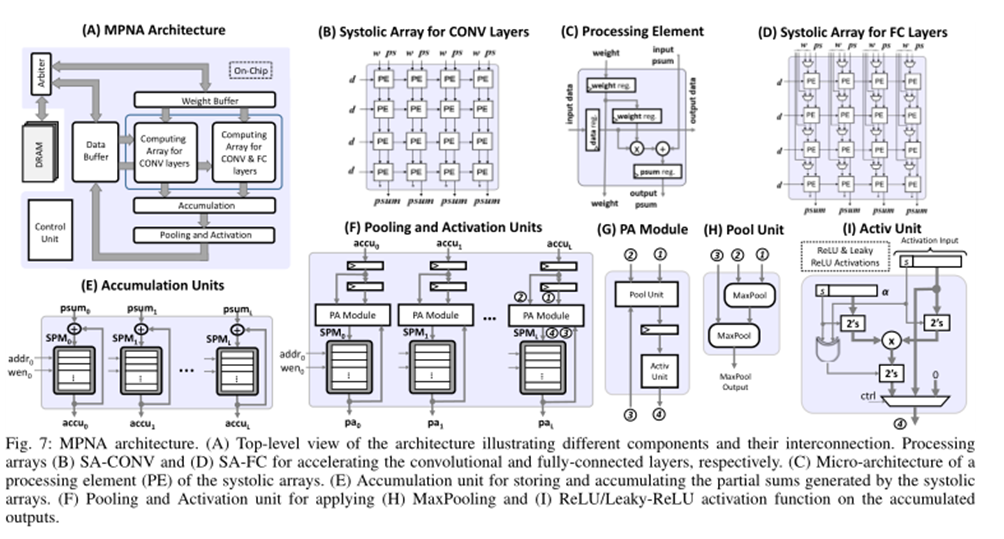
中度可信度描述已自动生成

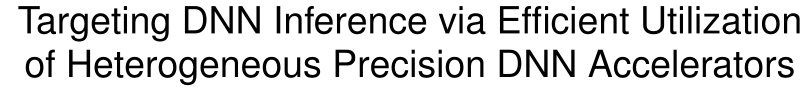
8x8异构SA,不同类型网络层效率不同

SA-CONV负责卷积层的计算密集型任务，

SA-FC负责处理存储密集型任务具有从权重buff到每个PE的专用连接。这使系统能够在每个时钟周期更新PE中的权重

还对几种可能的pattern 的数据流进行了优化 数据重用





图示

描述已自动生成



图示

中度可信度描述已自动生成

√AI-MT：

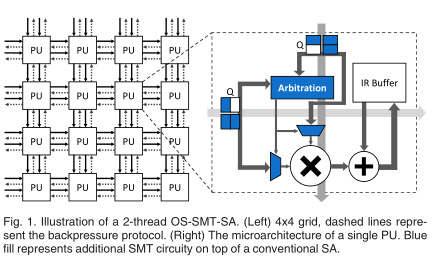


128x128 不同NN创建细粒度的计算和访存任务，添加记分牌 硬件调度器 进行多任务调度 分时使用SA

表格

描述已自动生成

√SMT-SA 多线程，类似于在通用处理器中实现的SMT，允许在遇到可跳过的计算时（零），切换到另一进程，隐藏等待有效计算到来的开销。



√PREMA 可抢占NPU和预测 多任务DNN调度算法

√Planaria：128x128 粗粒度分配 4个64

在运行时动态分解成小的DNN egine为多任务加速，主要是设计了全向SA和专用互联，和相应的任务调度算法

子阵列需要两个6位寄存器，一个保持当前配置状态，另一个预保持下一个状态。六位足以重新配置每个子阵列及其方向/总线。两位确定输入激活和部分和的方向。每个子阵列可以潜在地连接到四个其他子阵列，这些子阵列可以位于相邻的裂变吊舱中，由四位决定。

图示

描述已自动生成

√多pod结构



互联：SIGMA