Low Power Method in DLA

Section 1 : mechanism

DVAFS 是一种电路级的低功耗技术，它是在之前的DAS, DVAS基础上提出的一种在保证恒定吞吐率的基础上，通过对于单元活动率α，电源电压V，以及系统频率f进行实时调节从而来降低功耗，提升能效的方法。

再展开之前需要回顾以下电路的基本知识：

1. 电路动态功耗公式：

…………………………………………… (1)

1. 反相器延迟计算公式：

…………………… (2)

DAS, DVAS, DVAFS这种方法进行的功耗调节可以分为两个部分，一部分是计算模块，也就是可以scaling的模块，例如adders， multipliers。另外一部分就是不可scaling的模块，例如memory、control logic等。

以下将以一个乘法器为例，来说明DAS，DVAS，DVAFS三种方法的原理，乘法器电路结构如图，以4x4 bit的乘法器（进位保留乘法器）为例：



无论是DAS, DVAS, DVAFS他们所作的调节方式都是在低精度所带来的误差能够容忍的情况下进行讨论的。

1. DAS：Dynamic Accuracy Scaling

这是一种在电路中的自然现象，例如上图所示，如果是全精度的乘法运算的话，那么整个乘法器中的所有单元都是处于active的状态，但是如果精度降到2bit，那么将会有大部分的单元处于non-active状态，如下图所示：



如图所示，当精度降低之后，仅有25%的单元处于active状态，也就是会产生的动态功耗减少了75%。所以此时在DAS情况下的功耗可以按照下列式子计算：

………………………….. (3)

1. DVAS：Dynamic Votage Accuracy Scaling

在DAS的基础上进一步对低精度计算性质的利用，当精度降低的时候，乘法器的关键路径变得更短了，当为2bit的时候的关键路径大约变成了4bit情况下的1/4。由于关键路径变短，则乘法器能够运行的最大频率上升，在这种情况下由公式(2)可知，可以通过电压来补偿频率。也就是说，在保证恒定吞吐率的情况下，可以通过降低电源电压的方式，来使得正时钟裕量降为0。此时由于电源电压降低，所以动态功耗自然降低，此时的电路功耗如图所示：

……………………….. (4)

但是需要注意的是，这里仅仅是因为计算模块的关键路变短，所以只有计算模块的时钟裕量增加，所以仅仅能够对计算模块进行降压，此时电路将出现两个power domain。



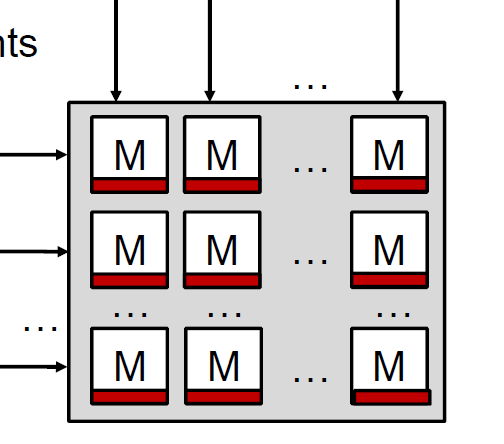
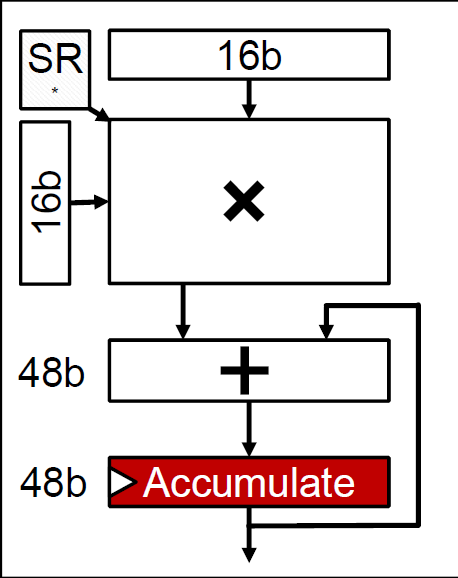
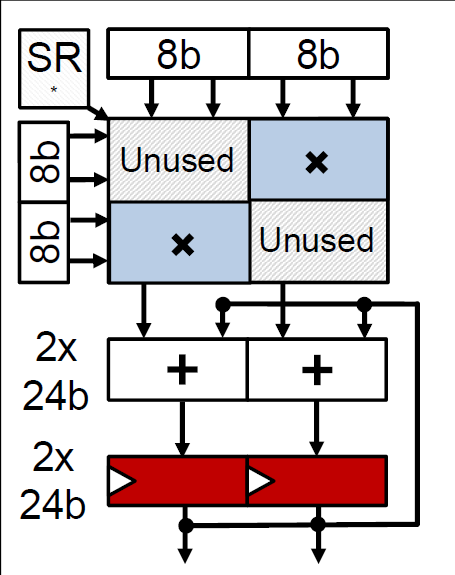
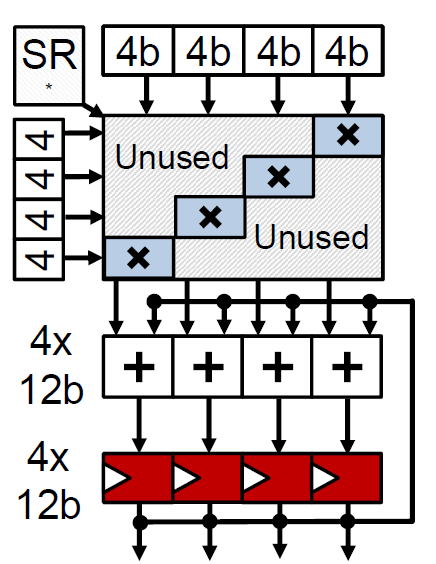
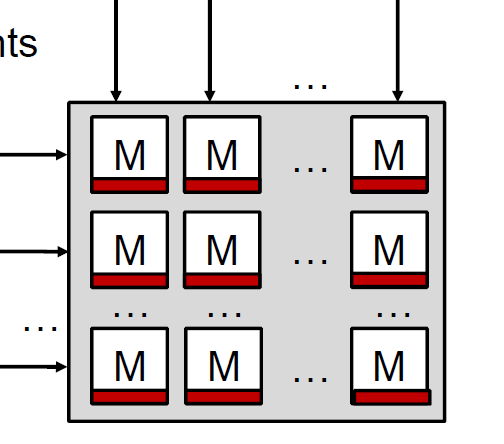
1. DVAFS：Dynamic Votage Accuracy Frequency Scaling



考虑到在低精度情况下，乘法器中有很多单元并没有使用，于是希望能够重利用这些non-active的单元。现在对于最大精度为4bit的乘法器来说，当精度为2bit的时候，如果重利用那么这个乘法器阵列最大能够同时产生2个计算结果。如果我们设定吞吐率是一样的情况，那么我们就可降低频率来使得功耗进一步降低，频率降低由两个好处：1. 非计算逻辑的频率也能够的到降低；2. 由于频率降低，那么计算逻辑的时钟裕量又得到提升，所以我们能够进一步通过降低电源电压来补偿这部分裕量，使得裕量为零。所以在DVAFS中的计算逻辑的电源电压比DVAS中更低，此时的功耗表达式为：

……………….. (5)

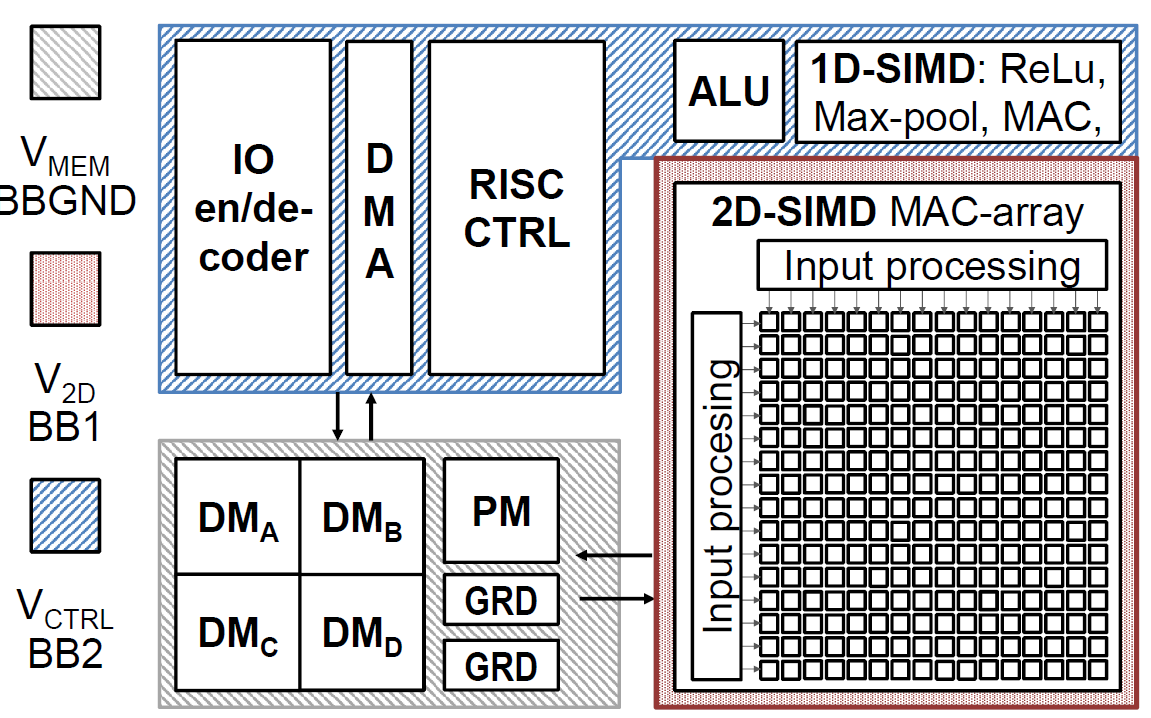
Section 2: experiment results

对于神经网络加速器来说，其实现原理就不过多讨论了，在神经网络加速其中，往往当精度降低的时候，memory的带宽以及其数据搬运带来的功耗会成为其

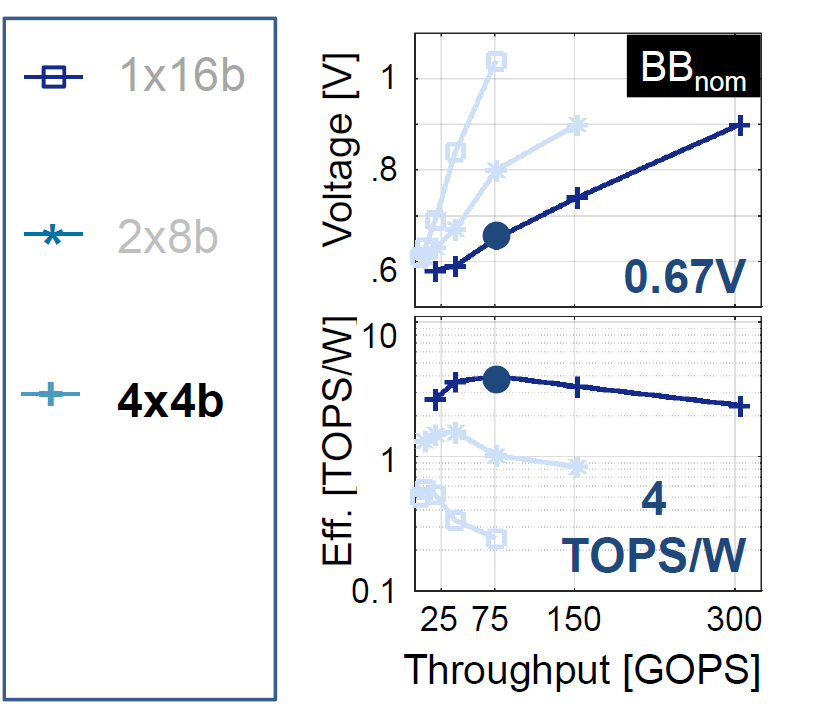
瓶颈，所以使用这种DVAFS可以在一个恒定memory带宽的情况下提升能效。这里以论文中的结构进行简单介绍，然后从其对比实验结果分析DVAFS对于功耗和能效的提升。

可以看到对于不同精度下的每一个计算单元MAC，根据不同的精度一次性能够处理的数据如上图所示，整个加速器的实现方式如图：

（对于一般普通的DLA而言，通过在精度降低情况下，增加固定带宽中所含有的数据的量是比较少见的，因为则会增加很多硬件的开销，包括乘累加单元，以及后级的adder tree。这种方式会带来由于吞吐增加而导致的memory 带宽不匹配的结果，就会造成需要大量的buffer来解决这一问题，则会增加硬件资源和延迟以及功耗）



根据不同模块功能的不同分为不同的power domain，结合公式(5)，以28nm FDSOI, AlexNet进行实验，得到如下图所示的结果



如图，展示了在不同精度情况下，恒定performance下的efficiency与voltage的变化，可以看到这种方式能够非常有效的降低其功耗，从而带来能效的提升。