**MobileNet V2 FPGA实现方案**

参考论文，本设计采用338个双通道WS DSP。目的是实现MobileNet V2模型。该模型主要分为CONV、DW、PW、SC、AVGPOOL与FC六种类型的处理层。

首先关注CONV、DW与PW的DSP分配。本模块DSP由中央控制系统进行动态分配，根据算法的不同调节它的输出端接到加法树的不同端口，或是输入端接到不同的BRAM输出上面。

首先分析可以implement parallelism进行performance提升的空间。

DW：

由于MobileNet V2只涉及到3\*3卷积，将3\*3窗口视为一个element，一个element中的全部运算可以进行并行化。进一步地，各个channel的运算可以进行并行化。规定：**在DW的parallel中，一个小周期仅处理两个element(双通道的DSP匹配双通道的BRAM)，定义该小周期的结束信号为element\_end，则当收到element\_end信号时，Cin中37个channel里的element被处理完毕。**

**故本模块采用37\*2个BRAM进行ftmap输入输出的存储，一组用于读取输入，另一组用于存放输出。**

PW：Pointwise的并行化首先在于对于一个point而言，其值可以被broadcast到各个核上以求得输出，即在PW阶段，ftmap是不变的，变的是各个核参数，核的多少决定了输出ftmap的channel数。拟采用systolic array的方式处理PW运算，故首先一个point对应的Cin可以并行化处理，输出Partial Sum(PS)，多个PS相加可以得到当前point在一个核下输出的值。由于BRAM个数的限制，一次最多只能读取Cin中最多37个输入，这些输入称为group，将被送入加法树，累加之后输出ACC，并发送pw\_channel\_group\_valid(PCGV)。在下一个clk里，下一个point的channel group被送进DSP array，在一个小周期后，完成所有point在当前group的运算，运算结果输入ACC，发送信号pw\_frame\_valid(PFV)。更新权重后，开始处理下一个group，所有group处理完毕后，发送pw\_out\_channel\_valid(POCV)，更新权值。**因DSP数量的限制，一次仅能并行处理9个核，故systolic array的深度为9，所有Cout计算完毕后，发送信号pw\_ftmout\_valid。由于DSP的双通道，一次可完成两个point的运算。**