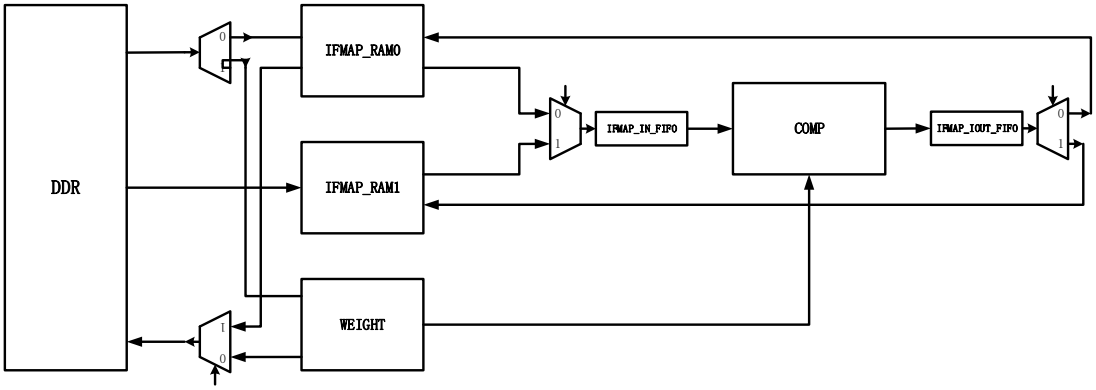
加速器方案（草稿）

特点：

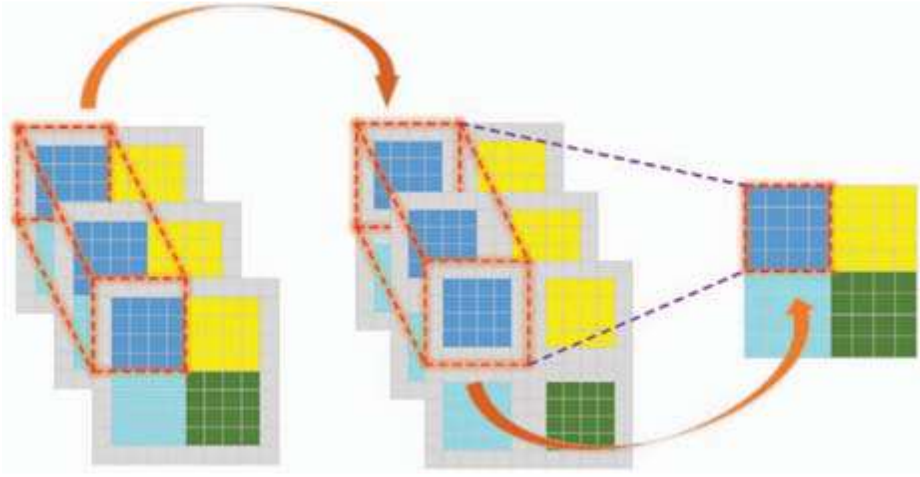
- 权重的零值压缩（降低存储空间、减少访存次数）
- 将为零权重的计算资源分配给非零权重（提高计算单元利用率、提升计算速度、降低功耗）
- 固定权重，轮换输入特征图（减少数据搬移）
- 输入特征图为零值跳过计算，直接得到输出结果 0（降低功耗）
- 输入特征图划块并连续计算卷积层（从第一层连续计算到最后一层卷积，降低访存）

架构简图：



数据流：

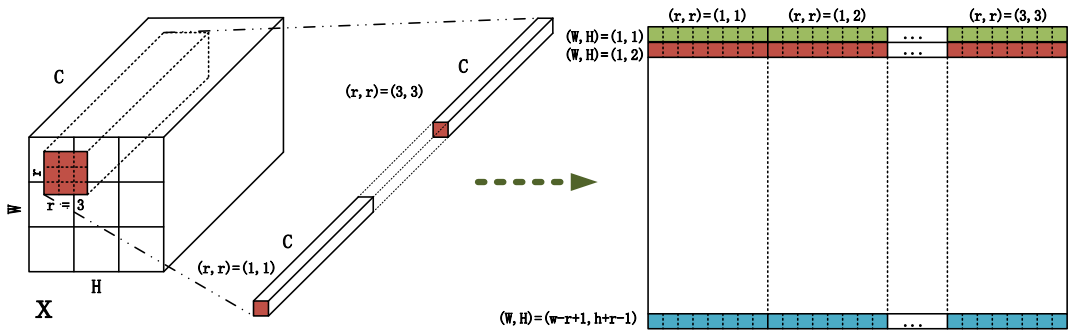
- 将输入特征图划块
  - 利用输入特征图在不同坐标的不相关性，将大图划成小块，并对数据从第一层连续做到最后一层



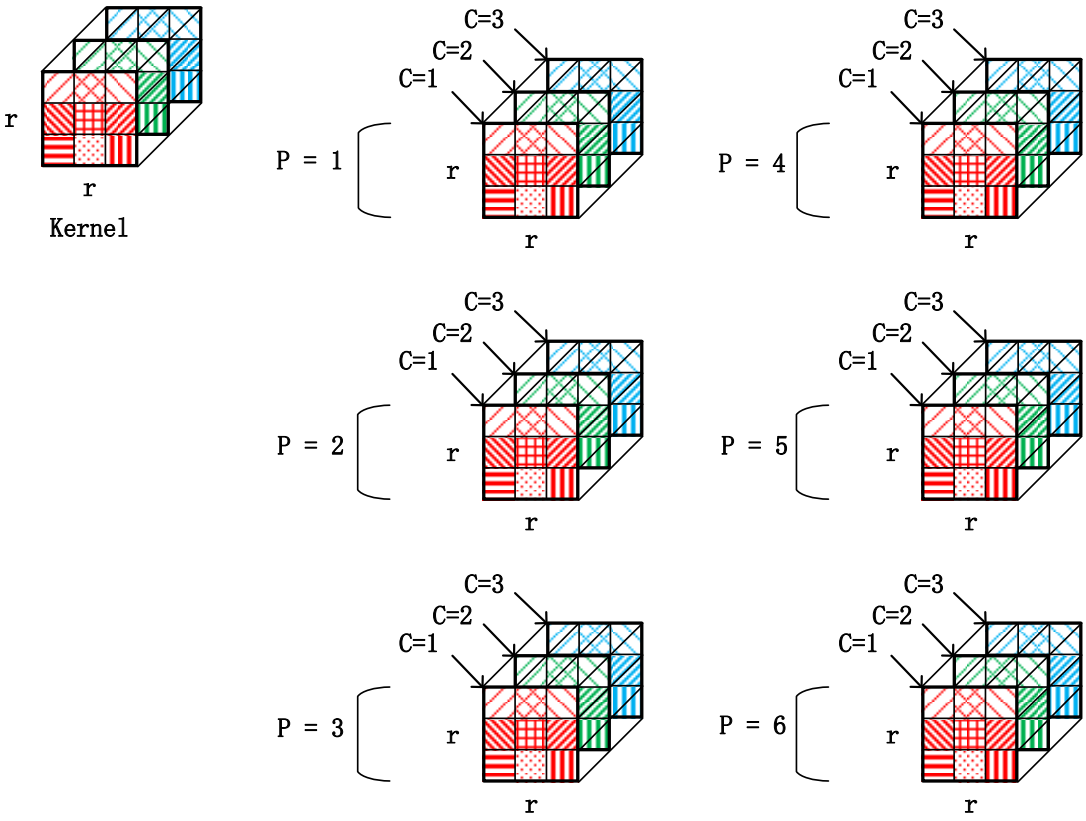
由于 cnn 后一层对前一层有依赖，而在输入特征图像上，不同坐标位置的像素是无关的。因此考虑将输入特征图像进行划块。比如说，将上述输入特征图划成 4 块，每一块周围打一圈 padding。

- 输入特征图块打平为二维矩阵

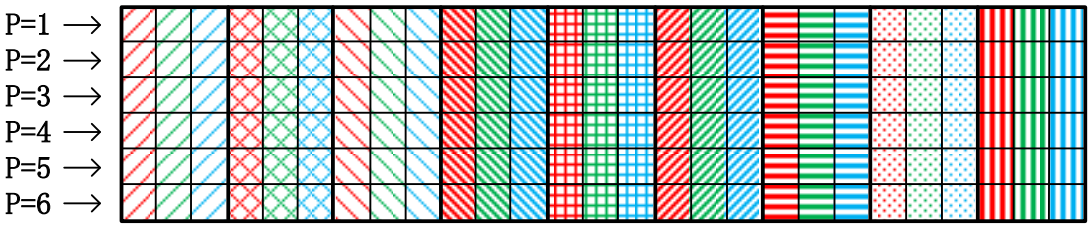
  将每一块划分出来的输入特征图进行打平，如下图所示



- 权重打平为二维矩阵

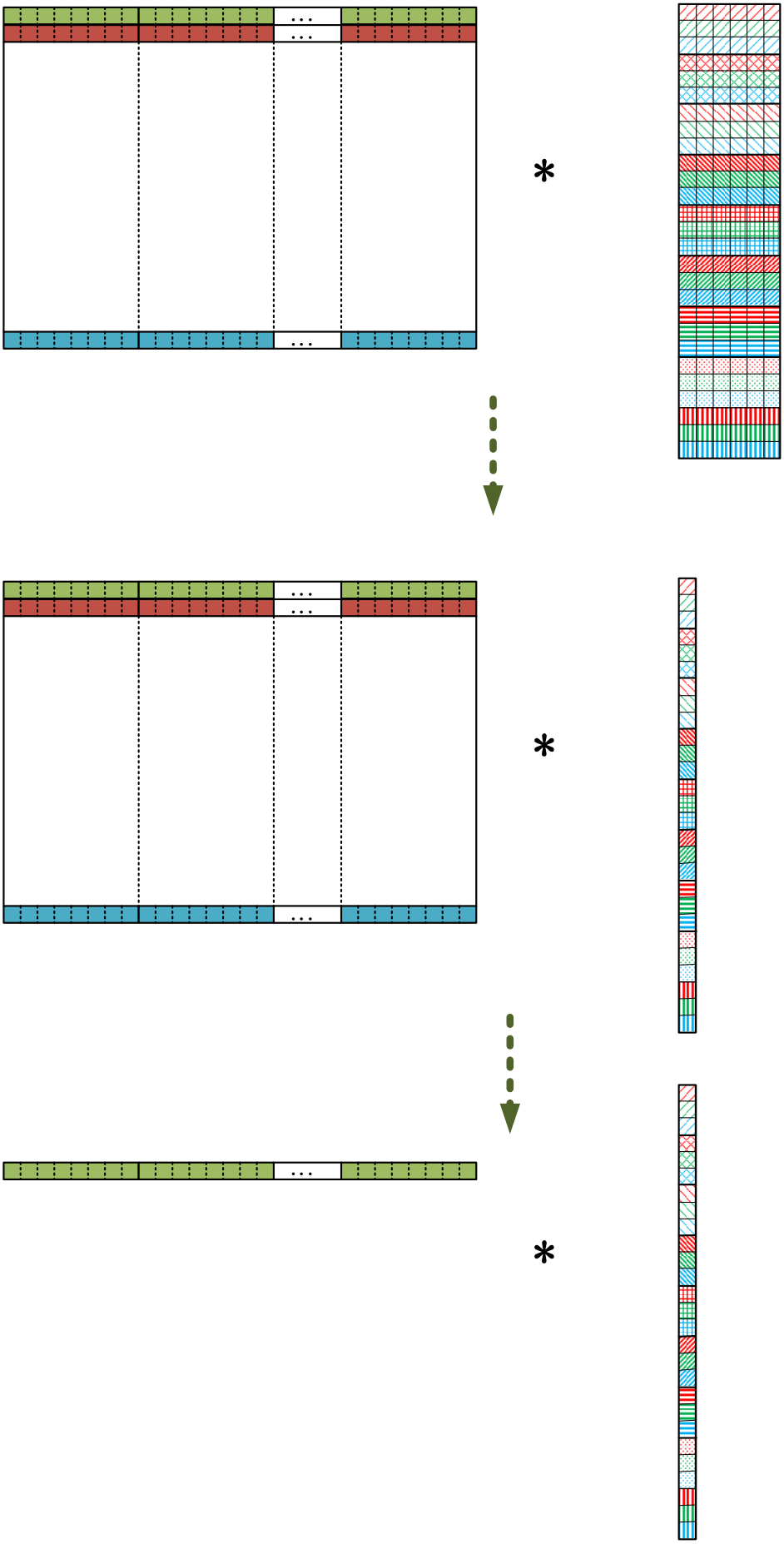


打平为二维矩阵的形式



Weight

- 矩阵乘法-》矩阵向量乘-》向量乘法



- 向量内积运算单元，权重为 0 的计算被取消掉，分配给其他计算使用