计算脉动矩阵乘法拍数：



对于一个MxK和KxN的矩阵乘法，部署在KxK的脉动阵列上时，计算花费的拍数（cycles）为 M + K + N - 2；

考虑如图规模为MxK的稀疏矩阵A，α为非零行的占比，将非零行划分为nseg个分组，每个分组为m1、m2、m3、……，其中每个分组内非零列的占比为βi，非零列的数量为seg[i]。稠密矩阵B为KxN，脉动阵列的规模为SAxSA。

对于OS数据流，分块计算时采用如下方式：

令mi = SA，矩阵A被划分为nseg个SA x seg[i]的子矩阵，B矩阵划分为⌈seg[i]/SA⌉个seg[i] x N的子矩阵；得到计算拍数为：

先考虑MKN能被SA整除的情况，同时nseg = M·α/SA，seg[i] = K·βi，进一步化简为：

计算拍数受到α和β的影响，其中βi提供了并行的空间。

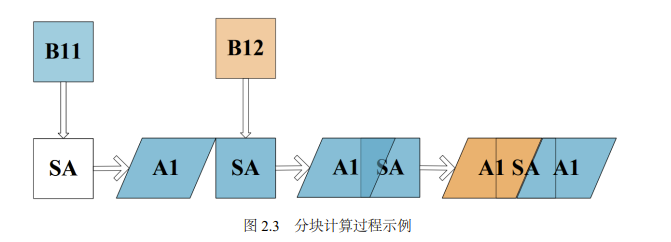
由于0< βi，α ≤1， 可求得COS的上界为：

对于AS数据流，首先将A中分出若干个SA x SA规模的子矩阵，B矩阵对应分出SA x N个子矩阵，根据分块方式可以得到计算拍数为：

与OS类似，进一步可以得到CAS的上界为：

同理，可以得到CBS为：

至此，可以发现三种模式的上界表达式中与稀疏参数无关，等于两个稠密矩阵相乘的结果。但是对于稠密计算的矩阵，可以通过双buffer和流水线优化等方法，进一步减少计算拍数，可得到COS-D、CAS-D、CBS-D分别为：



同理可以得到优化后带稀疏参数的计算拍数为：