使用卷积神经网络的深度学习有着业界领先的精度在很多机器视觉的任务上，例如，目标检测，识别，分割。CNN中无论是训练还是推断，卷积大概占据了90%的处理时间，同时完整的卷积网络更加常用。为了实现业界领先的精度，CNN不仅需要大量的层，而且需要百万级别的卷积核和卷积核形状，例如卷积核大小，数量，通道数，在Fig14.5.1显示。

现在存在的加速器不能满足可配置需要来有效的支持有着不同形状的大规模CNN，同时使用移动GPU可能会十分昂贵。本文描述了一个加速器可以实现业界领先的精度同时有着很低的功耗（系统包括DRAM），通过使用两个关键的方法：1、有效的数据流和最小化数据移动2、利用数据的统计特性来最小化功耗通过跳过0来避免不必要的读取和计算，还有数据压缩来减小需要的内存带宽。

Fig14.5.2显示了是顶层架构和内存级别的加速器。数据迁移通过缓冲输入图片数据和卷积核权重、局部和在一个108KB SRAM缓冲器中来优化，这促使加载数据时暂时的复用。图片数据和卷积核权重从DRAM中读取到buf中，同时流进空间计算整列（允许交叠内存交互和计算）。数据流和复用允许系统实现很高的计算效率，甚至在内存连接处于一个很低的频率时。空间阵列计算内积在图片和卷积核之间，产生局部和（从阵列返回到buf然后，可选的ReLu和压缩），到DRAM。运行基于长度的压缩减少图片平均带宽2x。可配置的支持图片和卷积核大小不完全合适到空间阵列。空间阵列的代销和buf决定这些“路径”需要计算一个特定的层。没有使用的Pes被时钟门控

Fig14.5.3 展现了卷积核权重、图片值和局部和在阵列中的数据流，如果卷积核的高度等于阵列的行数，那么逻辑数据流如下：1、卷积核权重被喂给buf阵列左列（每个PE一个卷积核的行）然后卷积核权重从左行移动到右边。2、图片数据被喂给左列和下列的PE，然后斜上移动。3、局部和每个PE垂直向上移动，如果局部和在下一pass被被使用，他们被喂给。。。。。。。。。。。。。。。

为了最小化固定大小的阵列占用为了不同的形状，the映射可能需要折叠和裁剪如果形状太大或者太小