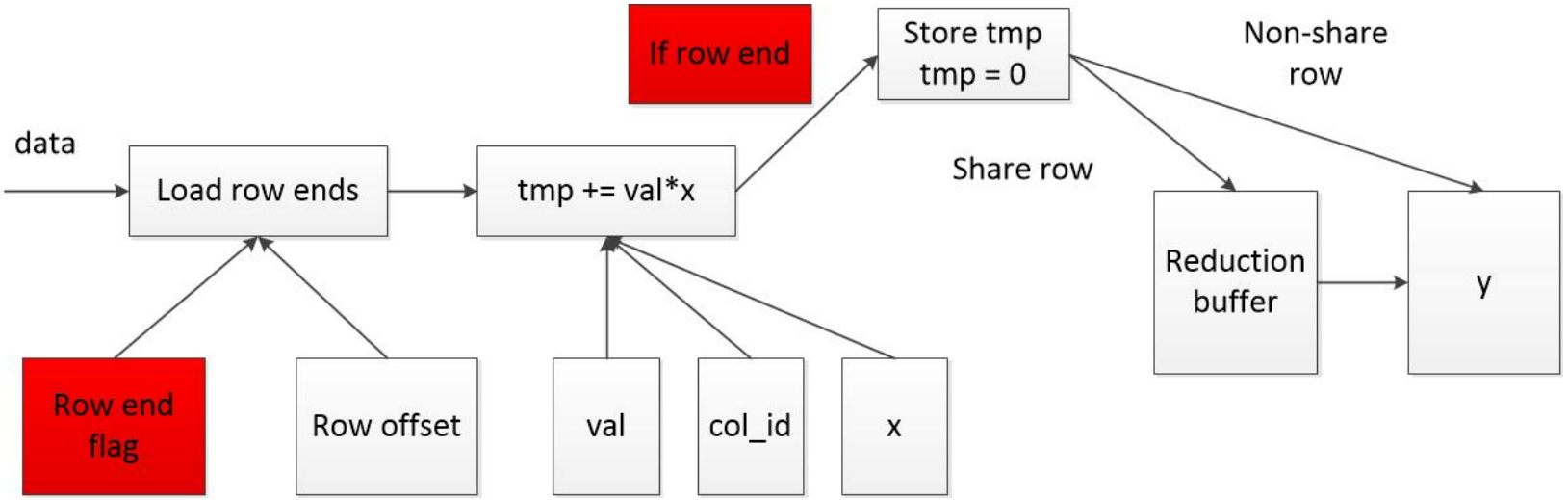
第一周周工作报告

报告人：黄一智
指导老师：李仁发教授、刘彦老师
2017-09-08

本周主要工作：

- 文献总结：主要针对需要对比的三个文献

- 文献一： CSR5: An Efficient Storage Format for Cross-Platform Sparse Matrix-Vector Multiplication
- 出处：ICS’15
- 核心思想：GPU每个线程处理N个连续非零项（等数据量处理，保证线程对x数据的绝对平衡）



- 优势：1、保证数据的绝对平衡，在tmp += val * x操作上不存在warp divergence；2、在矩阵数据访问上是coalesced access；

- 核心问题：1、需要事先计算row end flag；2、存在大量线程处理的数据不满一行，存在大量线程处理的数据多于一行。不满一行需要多线程规约，多于一行则多次回写结果内存，reduce操作和仿存存在大量warp divergence

- 文献二：Structural Agnostic SpMV: Adapting CSR-Adaptive for Irregular Matrices
- 出处：HIPC’15
- 核心思想：依据非零行的行长采用不同的策略（CSR-stream 、CSR-vector、CSR-scalar），其中CSR-Stream策略的核心是将非零元素切割成近似相等的块，用于保证小行不跨块。
- 优势：三种策略的适应领域是和非零行长相关的，自适应不同行长矩阵
- 劣势：1、为了防止仿存冲突，切分长行到块中使用了锁机制，当长行非常多时，开销很大。2、串行部分的预处理带来了很大的开销。

- 文献三： Merge-based Parallel Sparse Matrix-Vector Multiplication
- 核心思想：分配给每个线程的非零元素数目和写入输出向量数据数目之和相等（并非计算数据上的绝对均衡，而是一种相对均衡）。
- 优势：1、块中输出行的数量是有限的，使得输出数据可以存储在共享存储中。2、对空行进行了隐式处理，不影响负载均衡
- 劣势：1、计算每个线程要处理的行数据的位置；2、存在跨行问题。

总结

- 三种方法采用不同策略解决CSR格式行不规则导致负载不均衡的问题。
- 三种策略都引入了额外的计算以及额外的开销，开销与GPU结构、策略之间的联系与影响是关注的重点。