Graeme Knowles
CSC196U

Assignment 3

ListScan:

1. Name 3 applications of parallel scan.
    a. Radix/Quick sorts
    b. Histograms
    c. Stream Compaction
2. How many iterations are being performed in the reduction step of your scan kernel in terms of the input length N? Explain.
    a. Each iteration the stride length doubles, so the number of iterations is Log_2N
3. How many floating operations are being performed in the reduction step of your scan kernel in terms of the input length N? Explain.
    a. Iterations do N/2 + N/4 + N/8 + …. adds.
    b. This is a known recurrence relation which adds to N.
4. How many iterations are being performed in the post-reduction reverse step of your scan kernel in terms of the input length N? Explain.
    a. $O(Log_2N)$ iterations
    b. Starting from iteration 1, each step adds N * 2 - 1 elements, until greater than half the number of elements.
5. How many floating operations are being performed in the reduction step of your scan kernel in terms of the input length N? Explain.
    a. Each iteration does n/2 - 1 adds
    b. The total number of adds is: $(n-2) - (\log_2(n)-1) = O(N)$
6. How many global memory reads are being performed by your kernel in terms of input length N? Explain.
    a. Assuming block size = 512
    b. Each global element is read into local memory. = N
    c. Block maximums are scanned (which reads them into local memory). = N / 512
    d. Block maximums are added to each element after the first. = N + N / 512 - 512
    e. 2N + 2N / 512 - 512 = **O(N)**
7. How many global memory writes are being performed by your kernel in terms of input length N? Explain.
    a. Assuming block size = 512
    b. Each global element is written to once after computing the scan for its block. = N
    c. Each block writes the maximum value for the block into global memory. = N/512
    d. The maximum values for the blocks are scanned. = N / 512
    e. Each element of each block after the first has the value for the block before it added. = N - 512
    f. 2N + 2N / 512 - 512 = **O(N)**
8. Describe what optimizations does the work-efficient scan kernel perform to achieve a performance speedup over that of the work-inefficient scan kernel.

    a. The work efficient kernel better uses intermediate results of the addition operations. By adding the values in a tree structure, no extra additions are performed, while still allowing the additions to be done in parallel.
    b. A better thread to data mapping improves control divergence.

Histogram:

1. Describe how does the privatization technique improve the performance of the histogram kernel.
    a. When a bin of the histogram is incremented, it has to be done atomically, which is done in sequential order. When each thread block is privatized to work on its own copy of the histogram, the number of atomic operations that can be done in parallel increases to the number of thread blocks. This increase in parallelization is the source of the performance speedup.
2. How many global memory reads are being performed by your histogram kernel in terms of the input length N? Explain.
    a. N global reads are performed for an input of length N.
    b. For small number of elements, there are more threads than elements, but these extra threads don't participate in reading/incrementing the bins, every element is read once.
    c. For large number of elements, each thread will read more than one element, but each element is only read once. This is because the blocks read strides of elements at a time.
3. How many global memory writes are being performed by your histogram kernel in terms of the grid size (number of blocks launched) G and NUM_BINS? Explain.
    a. Number of global writes is equal to G * NUM_BINS.
    b. Each block has its own local copy of the histogram, and at the end of the kernel, the local histograms have to be combined into the single output histogram. Each block adds its local bin value to the corresponding global bin value once.
4. How many atomic operations are being performed by your histogram kernel in terms of the input length N, the grid size G, and NUM_BINS? Explain.
    a. Num Atomic Operations = N + G * NUM_BINS
    b. One add for each input element, incrementing the local bin value.
    c. One add for every element in the bin adding the local bin to the global bin, which is done a separate time for every block.
5. For the histogram kernel, what contentions would you expect if every element in the array has the same value?
    a. Every thread in each block would contend with the other threads in its block to increment the local bin.
    b. Each block would contend with the other blocks to increment the global bin, with a worst case scenario of Number of Blocks * Number of Bins contentions.
6. For the histogram kernel, what contentions would you expect if every element in the input array has a random value?

a. For computing the block local bins, it would depend on the size of the input, but for a large input, the number of contentions would be on average (Block size / Number of bins). Each element has a 1/(block size)^2 chance of contending with any other single element.
b. Then for building the final histogram, it would be the same as in 5, each block would contend with the other blocks to increment each element in the global bin.