I read *Cuda By Example* by Sanders and Kandrot up to Chapter 7 to get an idea of how CUDA programming worked, and then proceeded to parallelize some algorithms so I can get first hand experience. I chose three algorithms: the SUMMA method for matrix multiplication (1), merge sort (2), and a most dominant peaks algorithm (MDPA) (3). For (1), I just experimented with the following three schemes given an arbitrary N x N matrix:

1. Have each block represent every element in the matrix

2. Have each block be a 2x2, sliding window that accumulates terms when doing the matrix multiply

3. Have each block be comprised of threads, where each thread in the block would be a single element in the matrix.

I implemented my own version of merge sort and is described as follows. Let us say that we have B blocks and T threads per block. Then, I divided the array A into B subarrays of size $\frac{length(A)}{B}$ – note that the last block was either slightly larger or smaller than the other blocks in order to fill up the portion of A that did not evenly divide by B. Then, each subarray A_i, $0 \leq i < B$ was further divided into subarrays A_{i_k}, $0 \leq k < T$ where A_{i_k} corresponded to the kth subarray of the ith subarray of A. Specifically, I gave each block a portion of the array and inside each block, each thread had a portion of that subarray. Each thread was then sorted using an iterative sorting algorithm; I used insertion sort to do this. Once the threads were sorted, each thread within a single block was pairwise reduced via merging with the adjacent thread until the entire subarray in the block was sorted. The execution of these steps lead to B sorted subarrays. The final step was to do the merging step again, this time with 1 block and B threads. After this step, the array A was sorted.

The MDPA is described in Liu et al. 2004 under the section "Regions of Dominance," second paragraph. Essentially, the algorithm is a way to extract the most relevant peaks from an autocorrelation surface of a periodic pattern; these peaks are used to construct the lattice that describes the shape of the pattern.

The MDPA was parallelized in the step of the code when the minimum distances, D_is are computed for each peak. This is because the preceding and ensuing step (sorting by height and by D_i, respectively) were not only already parallelized above, but they also could run efficiently on a serial machine in $\Theta(N \lg N)$ time. The parallelization for the distances is as follows. Let there be N peaks, and these peaks be described by the set $P = \{P_0, P_1, \dots, P_{N-1}\}$. Then let there be M blocks and N threads such that N is a power of 2. The idea is that a block B_i calculates D for a set of points P_{B_i}, where P_{B_i} is either empty or nonempty, starting at P_1. For example if $M = N - 1$, then $P_{B_i} = \{P_i\}$; if $M = \frac{N}{2} - 1$, then $P_{B_i} = \left\{P_i, P_{i+\frac{N}{2}-1}\right\}$. Each thread T_{k_i} in B_i represents a set of points $P_{T_{k_i}}$ that precede the current block's point P_{B_i}, where T_{k_i} is read as the k^{th} thread in the i^{th} block. If $N = i$, then $P_{T_{k_i}} = \{P_k\}$; if $N = \frac{i}{2} - 1$, then $P_{T_{k_i}} = \left\{P_k, P_{k+\frac{i}{2}-1}\right\}$. The idea is that given a block point P_i, each thread T_{k_i} evaluates to D_{k_i} where D_{k_i} is the k^{th} thread's minimum distance of its set of points with respect to P_i. Specifically if we label the set $P_{T_{k_i}} =$

$\{Q_0, Q_1, \dots Q_j\}$ where $Q_l \in P$, and if we define $dist(A, B)$ as returning the square of the Euclidian distance of points A and B, then $D_{k_i} = \min\left(dist(P_i, Q_0), dist(P_i, Q_1), \dots, dist(P_i, Q_j)\right)$. Each D_{k_i} is stored in the k^{th} index of a shared array called $minArray$. Note that for threads having an empty set of $P_{T_{k_i}}$ points, $minArray[k] = \infty$. After each element of $minArray$ is calculated, the array is pair-wise reduced until a single element remains in $minArray[0]$; this value is the minimum distance of point P_i. This process continues inside the block until we reach the end of its P_{B_i} set, for all the blocks.