

CSCI-GA.3033-004
Graphics Processing Units (GPUs): Architecture and Programming
Homework Assignment 2

1. Assume the following piece of code is running on G80 (We saw many examples about G80 and its properties in class.):

```
1 #define VECTOR_N 1024
2 #define ELEMENT_N 256
3 const int DATA_N      = VECTOR_N * ELEMENT_N;
4 const int DATA_SZ     = DATA_N * sizeof(float);
5 const int RESULT_SZ    = VECTOR_N * sizeof(float);
6 ...
7 float *d_A, *d_B, *d_C;
8 ...
9 cudaMalloc((void **)&d_A, DATA_SZ);
10 cudaMalloc((void **)&d_B, DATA_SZ);
11 cudaMalloc((void **)&d_C, RESULT_SZ);
12 ...
13 scalarProd<<<VECTOR_N, ELEMENT_N>>>(d_C, d_A, d_B, ELEMENT_N);
14
15 __global__ void
16 scalarProd(float *d_C, float *d_A, float *d_B, int ElementN)
17 {
18     __shared__ float accumResult[ELEMENT_N];
19     //Current vectors bases
20     float *A = d_A + ElementN * blockIdx.x;
21     float *B = d_B + ElementN * blockIdx.x;
22     int tx = threadIdx.x;
23
24     accumResult[tx] = A[tx] * B[tx];
25
26     for(int stride = ElementN / 2; stride > 0; stride >>= 1)
27     {
28         __syncthreads();
29         if(tx < stride)
30             accumResult[tx] += accumResult[stride + tx];
31     }
32     d_C[blockIdx.x] = accumResult[0];
33 }
```

- a.[1] How many threads are there in total?
 - b.[1] How many threads are there in a warp?
 - c. [1] How many threads are there in a block?
 - d. [1] How many global memory loads and stores are done for each thread?
 - e. [1] How many accesses to shared memory are done for each block?
 - f. [2] How many iterations of the for loop (Line 23) will have branch divergence?
Show your derivation.
 - g. [3] Identify an opportunity to significantly reduce the bandwidth requirement on the global memory. How would you achieve this? How many accesses can you eliminate?
-
2. [3] What factors can make two threads corresponding to two different warps but of the same block take different amount of time to finish? To get full credit, write at least 3 factors.
 3. [2] What is the difference between shared memory and L1 cache?
 4. [2] Can memory be coalesced for threads in a warp yet not-coalesced for threads in a different warp of the same block?
 - 5.[3] Suppose you want to write a kernel that operates on an image of size 400x900 pixels. You want to assign one thread to each pixel. Your thread blocks are square, in geometry, and try to use the maximum number of threads per block possible on the device. The maximum number of threads per block is 1024. How would you select the grid dimensions and block dimensions of your kernel?