CS516: Parallelization of Programs
Hands on Session - 3

Date: 08-Feb-2024

Notes:

1. The goal of this session is to make you familiar with GPU shared memory optimizations. You should submit the solutions on canvas by Feb 9, 2024.

Task-1:

Write the matrix-matrix multiplication implementations discussed in the class to understand the performance differences. You can assume M, N are of same size WIDTH*WIDTH, and the result matrix is P. You can take the code from the slides directly.

1. Implement the sequential version
2. Parallelize the multiplication using global memory and multiple thread blocks
3. Use shared memory with the help for tiling

Execute the above three implementations by experimenting with WIDTH*WIDTH as 1024*1024, 2048*2048, 4096*4096, 8192*8192, 16384*16384 and compute the speed ups w.r.t sequential implementation. Write your observations.

Task-2:

Implement the reduction operation discussed in the class under the following scenarios. You can again take the code from the slides directly.

1. Using shared memory without any optimization
2. Using shared memory by reducing warp divergence
3. Using shared memory by reducing warp divergence and shared memory bank conflicts.

Understand the performance difference by executing with the array size for various values as 2^{22}, 2^{24}, 2^{26}. Note the observed speed ups.