Parallel Computing
CSCE 735

Major Project

1. Steps to run the program
    - Load the main program on Scratch
    - module load intel/cuda
    - nvcc main.cu

   We get an output file a.out in the scratch directory.
    - ./a.out

   It prompts for the values of m and number of threads. Then computes the values for time taken for decompose, time taken for solve and finally the time taken for the entire algorithm.

2. I have used Cholesky Factorization as the optimization strategy for the major part of the code. Cholesky Factorization has 3 loops. The inner two loops have dependencies on the outermost loop and hence it cannot be factorized. Outer loop iterations each compute one row of the result upper triangular matrix U from top to bottom and the later iterations needs to use the results generated from previous iterations. We fully leverage parallelism in GPU by distributing the second loop to multiple threads. Besides, the second loop, calculating the diagonal for each row can be also paralleled – by handing the GPU thread to calculate partial sum and then assigning thread0 to calculate the final sum and calculate the diagonal.
   To get better memory performance, each GPU thread calculate continuous iterations to avoid stride memory access.

3. Here we have taken m = 50 and threads = 1024
   For Solve Step:
   Flop rate for solve step is total operations in solve step/time taken in solve step
   As given in the slides that total operations for solve step is 2n^2, n = m*m, so
   2(m*m)^2 = 2(50*50)^2 = 2*2500*2500
   Time taken for solve step is 21.1423ms, so flop rate = 2*2500*2500*1000/21.1423
   For Decompose Step:
   Flop rate for decompose step is total operations in decompose step/time taken in decompose step/time
   As given in the slides that total operations in decompose step is (2n^3)/3, here n = m*m, so
   (2(m*m)^3)/3 = (2(50*50)^3)/3 = 2*2500*2500*2500/3
   Time taken for solve step is 2186.52ms so, and flop rate = 2*2500*2500*2500*1000/(3*2186.52)
   Calculating speedup and efficiency:
   peak flop rate of GPU on grace1 i.e operations per sec = operations per cycle * cycles per sec = 64 * 1410 (MHz)(grace1 specifications for GPU 8.0)

Thus peak flop = 90Giga flops/sec
Now efficiency for solve = flop rate for solve/ peak rate
2*2500*2500*1000/(21.1423*90G) = 0.006598
So the efficieny for decompose step is 0.66%
Now efficieny for decompose step = flop rate for decompose/ peak rate
2*2500*2500*2500*1000/(3*2186.52*90G) = 0.05095 (1G = 10^9)
So the efficieny for decompose step is 5.1%
Now speedup for solve = efficiency*1024 = 0.0066 *1024 = 6.7584
And, speedup for decompose = efficiency*1024 = 0.051 *1024 = 52.224