EE 451: Parallel and Distributed Computation

PA7b — Spring 2021

Due date: Monday 19th April 2021 11:59 PM

1. **Examples**

   The `hello.cu` contains the CUDA implementation of HelloWorld.

   1. Login to HPC
   2. Setup MPI toolchain:

      ```
      module purge
      module load gcc/8.3.0 cuda/10.1.243
      ```

   3. Compile

      ```
      nvcc -O3 -arch=sm_20 hello.cu
      ```

   4. Run

      ```
      srun -n1 --gres=gpu:1  -t1 ./a.out
      ```

      The option `-t` specifies the limit of run time. Setting it as a small number will get your program scheduled earlier. For more information on `srun` options, you can use `man srun` to find out.

   5. Profile (optional)

      ```
      srun -n1 --gres=gpu:p100:1 --partition=debug nvprof ./a.out
      ```

   6. Allocate a machine

      ```
      salloc -n1 --gres=gpu:1 --mem=16G -t10
      // After the allocation, you will log on the machine and have
         10 minutes to perform multiple operations
      ./a.out
      // edit, compile, and run again without waiting for a new
         allocation
      ./a.out
      ./a.out
      ```


Youwei Zhuo: youweizh@usc.edu

2. (100 points) 1024×1024 matrix multiplication using these two approaches.

   - Approach 1 (unoptimized implementation using global memory only) :
     - Name this program as 'mm1.cu'
     - The value of each element of A is 1
     - The value of each element of B is 2
     - Thread block configuration: 16×16
     - Grid configuration: 64×64
     - After computation, print the value of $C[451][451]$
   - Approach 2 (block matrix multiplication using shared memory) :
     - Name this program as 'mm2.cu'
     - The value of each element of A is 1
     - The value of each element of B is 2
     - Thread block configuration: 32×32
     - Grid configuration: 32×32
     - More details of this algorithm can be found in the paper 'Matrix Multiplication with CUDA'.
     - After computation, print the value of $C[451][451]$

   Measure the execution time of the kernel of Approach 1 and Approach 2, respectively.

   Submission Instructions: Submit your code, screenshots, and a performance report as described above.

Youwei Zhuo: youweizh@usc.edu