Assignment2 on CUDA Programming

1. There are many problems, which can be effectively solved with CUDA, are suitable for a novice as exercise. Here are two of them. a) Matrix multiplication, also known as matrix product and the multiplication of two matrices, produces a single matrix. PROBLEM: Please write CUDA kernel functions to effectively compute matrix multiplication of two matrices . b) Given an array, computing the sum of all the elements of that array is very easy, the question is how to ensure that the calculation is fast. PROBLEM: Please write CUDA kernel functions to effectively compute the sum of an array.
2. To achieve the maximum performance, there are many tricks based on the processor and memory architecture. The memory hierarchy of CUDA architecture brings us global memory (including constant memory, texture memory and surface memory), shared memory and registers. You must realize that shared memory should be used, if possible, to avoid high latency accessing global memory. How can we achieve the extreme performance? You can optimize in the following ways or using any other method that makes sense: a. using block matrix multiplication b. avoiding bank conflicts in shared memory c. data prefetching d. resolving warp divergent e. loop unrolling

NOTICE:

1. You' d better use Ubuntu and you have to write a makefile to compile your code. You also need to write a README file to explain how you optimize the problems. If you don’t have NVIDIA GPU environment, please contact TA at niugeng2023@sjtu.edu.cn or by Wechat group.
2. What you write should be standalone functions to perform the computation, i.e., everybody can reuse your function to do the similar job with variable configurations.
3. Send your final version on canvas . You should archive your source code and makefile with StudentID_Name_HW2.tar.gz(or any archive file types). Do not include binary file.
4. Should you have any questions, please feel free to contact TA .