Parallel Homework #3

刘康来

2021 年 5 月 18 日

图 1: Here is the hardware's information

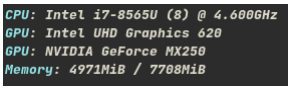


图 2: About the GPU and CUDA

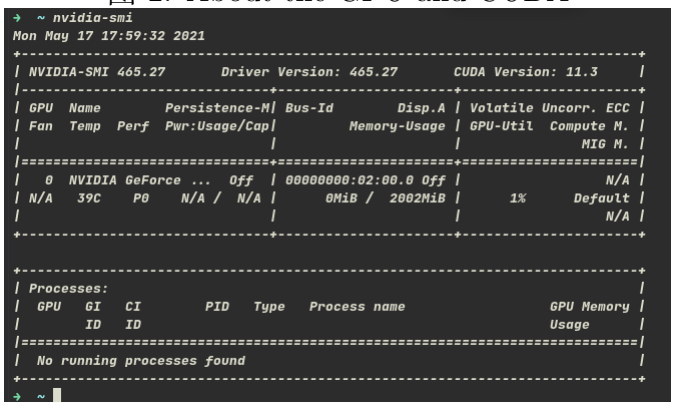
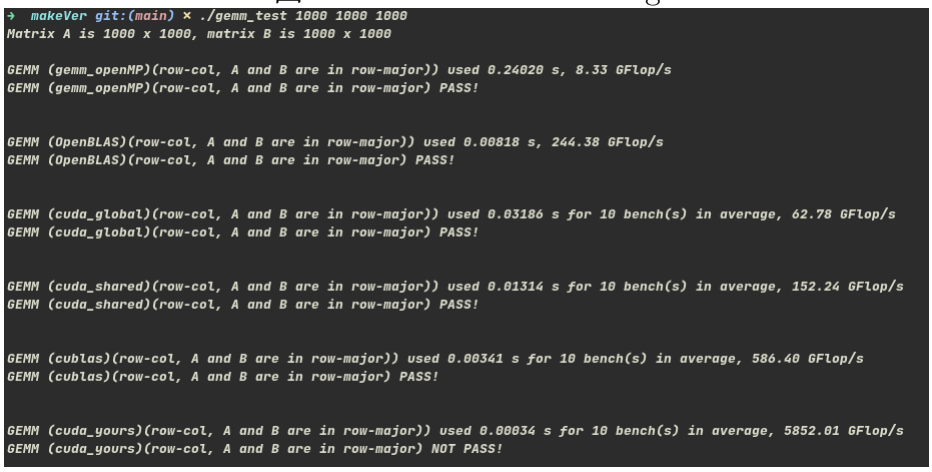
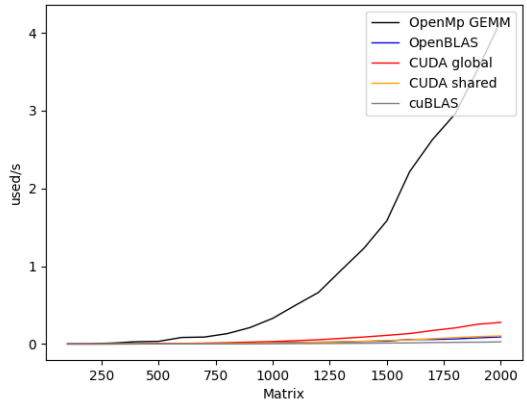
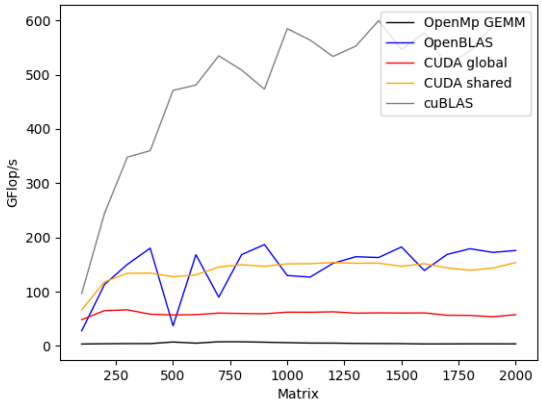


图 3: Here is the test runnig



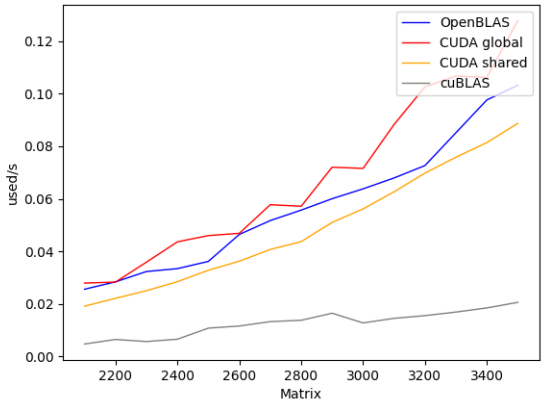


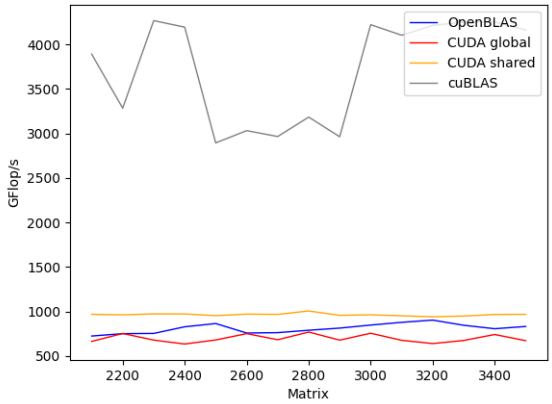


As you can see, the cuBLAS is the fastest, than is OpenBLAS, CUDA using shared memory, CUDA using global memory and basic code using OpenMP.

For better comparison, I use the given remote device to run the code(the Matrix from 2100 to 3500).

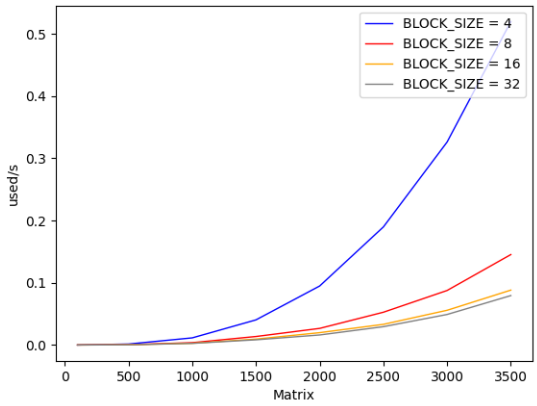
Here is the result:

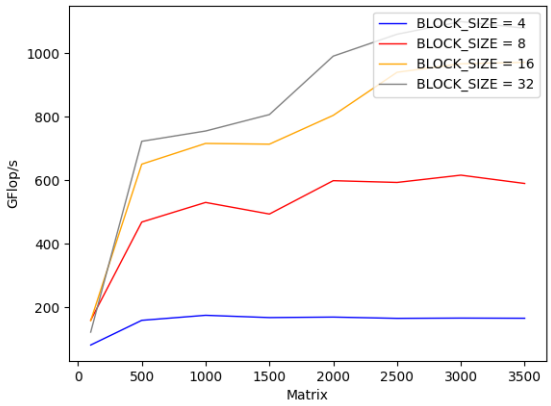




The CUDA GEMM code optimized by yourself with your own techniques...

In theory, using more blocks is better, but cost more memory. I try the BLOCK_SIZE 4, 8, 16, 32 to run the code, here is the performance:





The BLOCK_SIZE for 32 is the faster. Although the increase is not obvious compared to 32, but as the width of Matrix increases, it will do more. And I try the BLOCK_SIZE for 64, but it doesn't work, so a block have up to 1024 (32*32) threads (or more) in the device RTX 2060 SUPER?

That's all

End!