|  |  |  |  |
| --- | --- | --- | --- |
| EE557 Project4 | | | |
| Project name | SA4 | | |
| Document ref |  | | |
| Version |  | | |
| Release date |  | | |
| Author | Fei | | |
| Classification |  | | |
| Distribution List |  | | |
| Approved by | Name | Signature | Date |
|  |  |  |  |

**Fei Wu**

**6897429283**

**wufei@usc.edu**



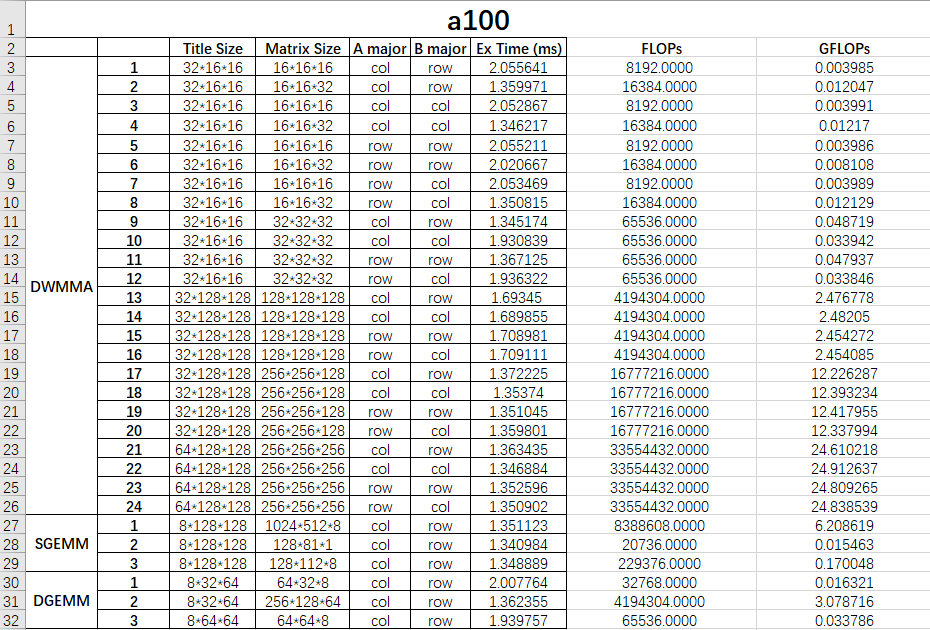
**Ming Hsieh Department of Electrical Engineering**

**University of Southern California, Los Angeles, CA 90089**

**Fall 2022**

**EE 557**

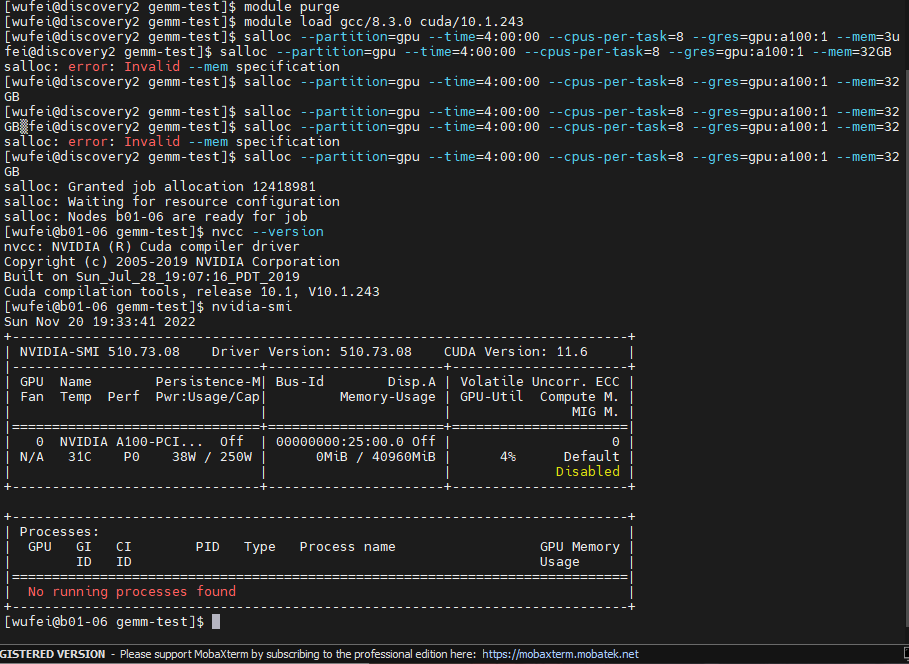
# Tables



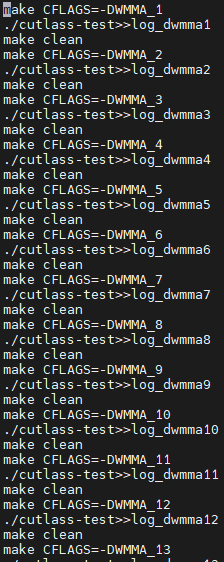


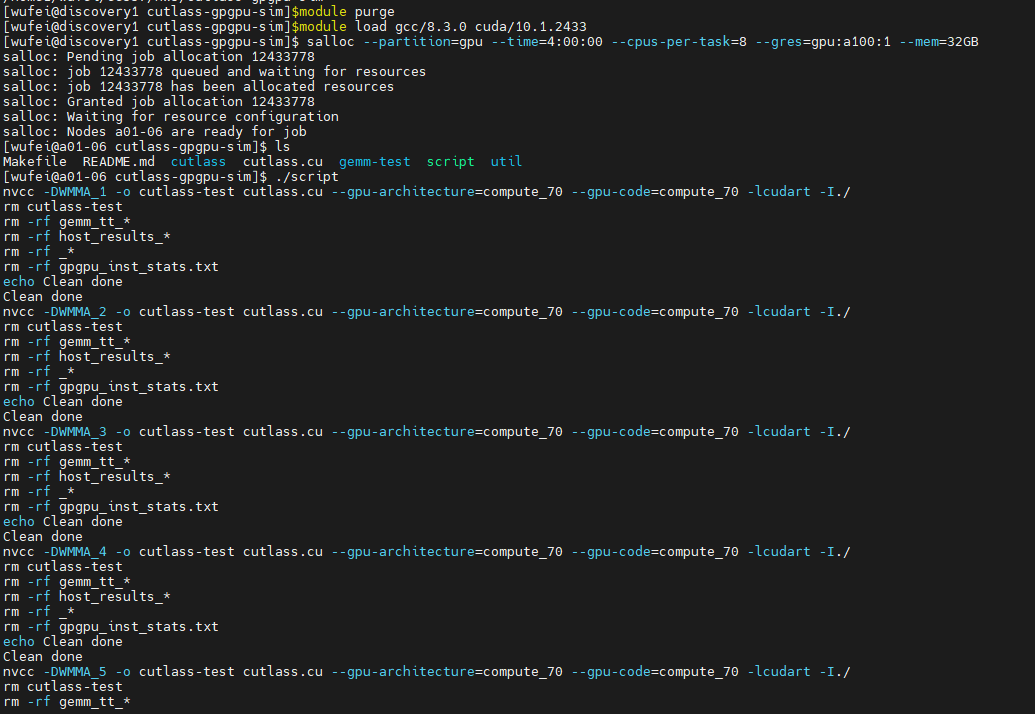
# Report

Set the GPU machine in the CARC system:

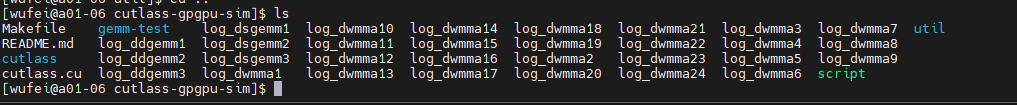


Check the script and check the output files which are going to be generated:

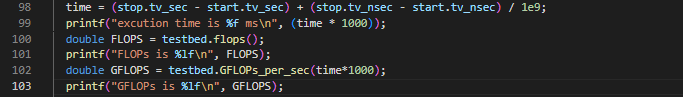




Check all the generated log files:

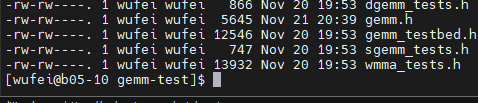


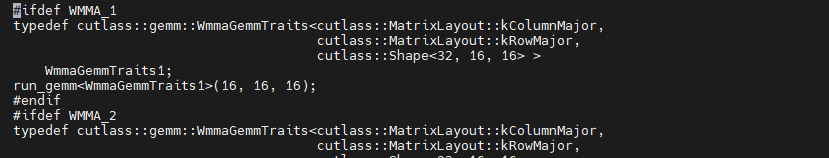
Modified the gemm.h to get the execution time and GFLOPs:



Check the complete coding inside the submission.

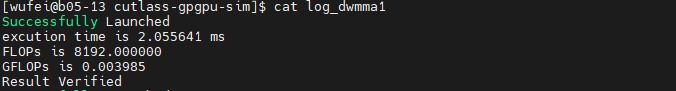
Check the three test code and record the parameters like tile size, matrix size, A major, B major.

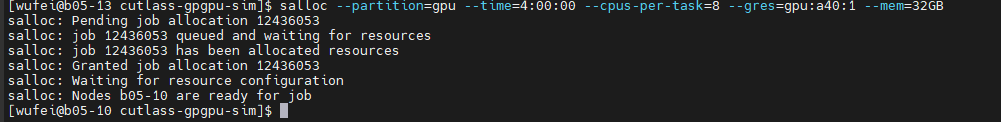


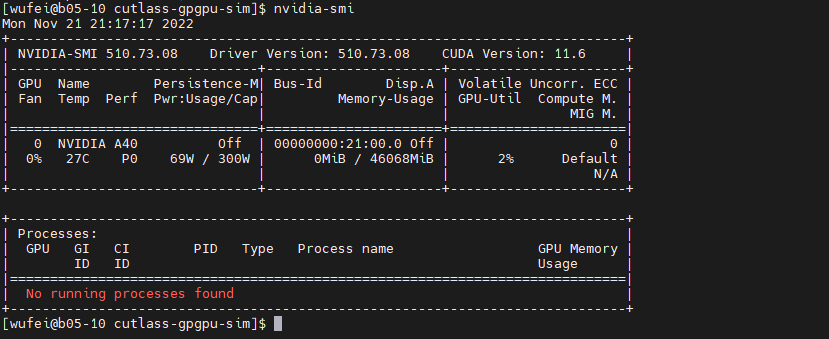


Run the script again and check the log:

Sample

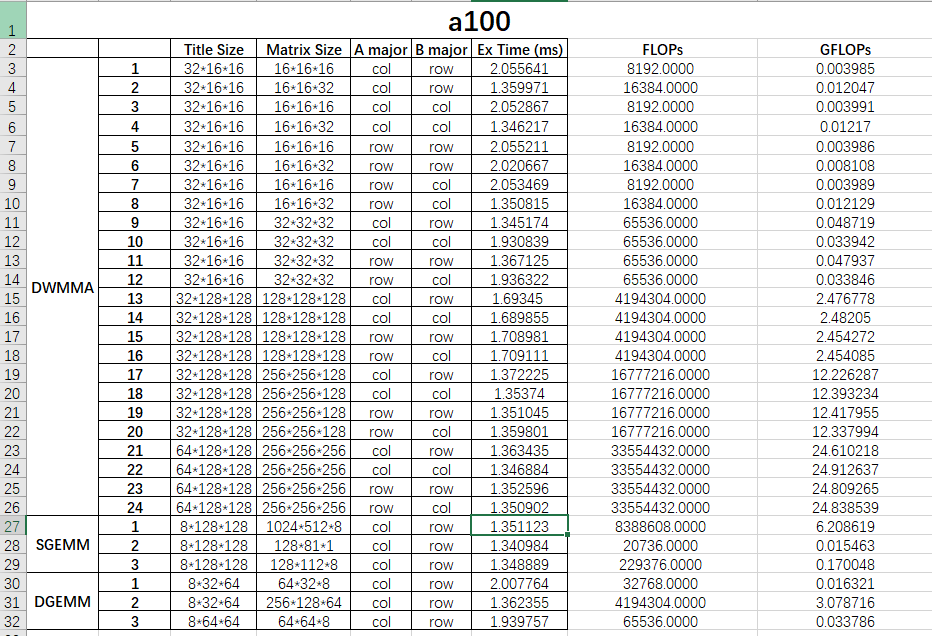


Reset the real GPU on the CARC to a40 with the similar command and rerun the script again:



get all the required parameters and make the table.





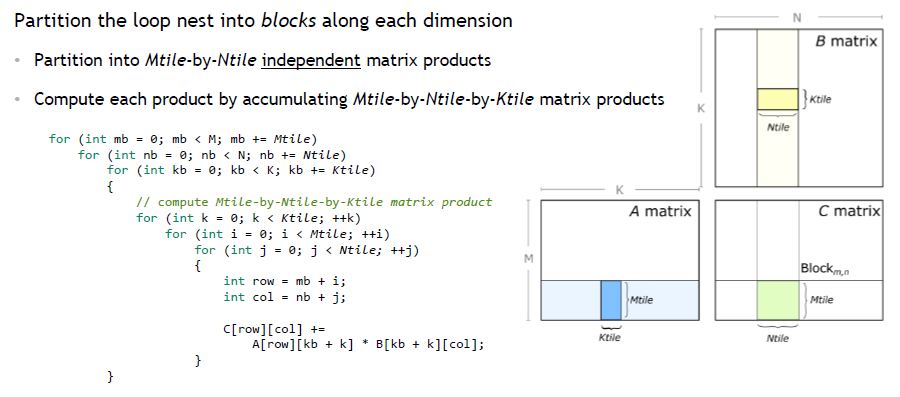
Check all the result inside the excel sheets submitted.

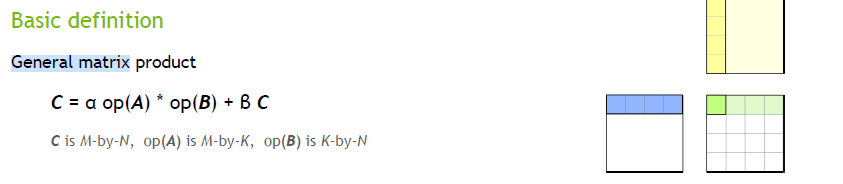
Analysis:

What is the tile size, matrix size, and column/row-major of the input matrices for each of the test?

ANS: The column/column major or row/row major will increase execution time. The pipeline stalls due to memory is not affected by the column/row-major order a lot.

Tile is the parameter shows that how the matric be partitioned into independent matric products in each dimension and the matrix is the products shows in three dimensions. And the matric a and b is the layout of the matrix defined in the cutlass.





Your analysis to the results, including the comparison of the two GPU performance and the influence of GEMM configurations to the computation performance. You may use tables and charts to assist your analysis.

ANS: In conclusion, the performance of a40 is much better than a100 with the comparation based on the GFLOPs.

As for the GEMM configuration, the GPU with a larger matrix size or saying with a lager c matrix size or K will have a higher GFLOPs; while with the same size configuration, the major A/B do not affect much on the performance; the GPU with the larger matrix A/B will perform better; the performance is growing in a linear way with the parameter matrix C /K; and the tile size and matrix size should be matched or saying that the matrix should not be partitioned too much so that the performance can be promised, which means the tile size should be smaller than matrix to gain a better performance.