# EE557 Project4

| | | | |
|---|---|---|---|
| Project name | SA4 | | |
| Document ref | | | |
| Version | | | |
| Release date | | | |
| Author | Fei | | |
| Classification | [Document classification] | | |
| Distribution List | [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**

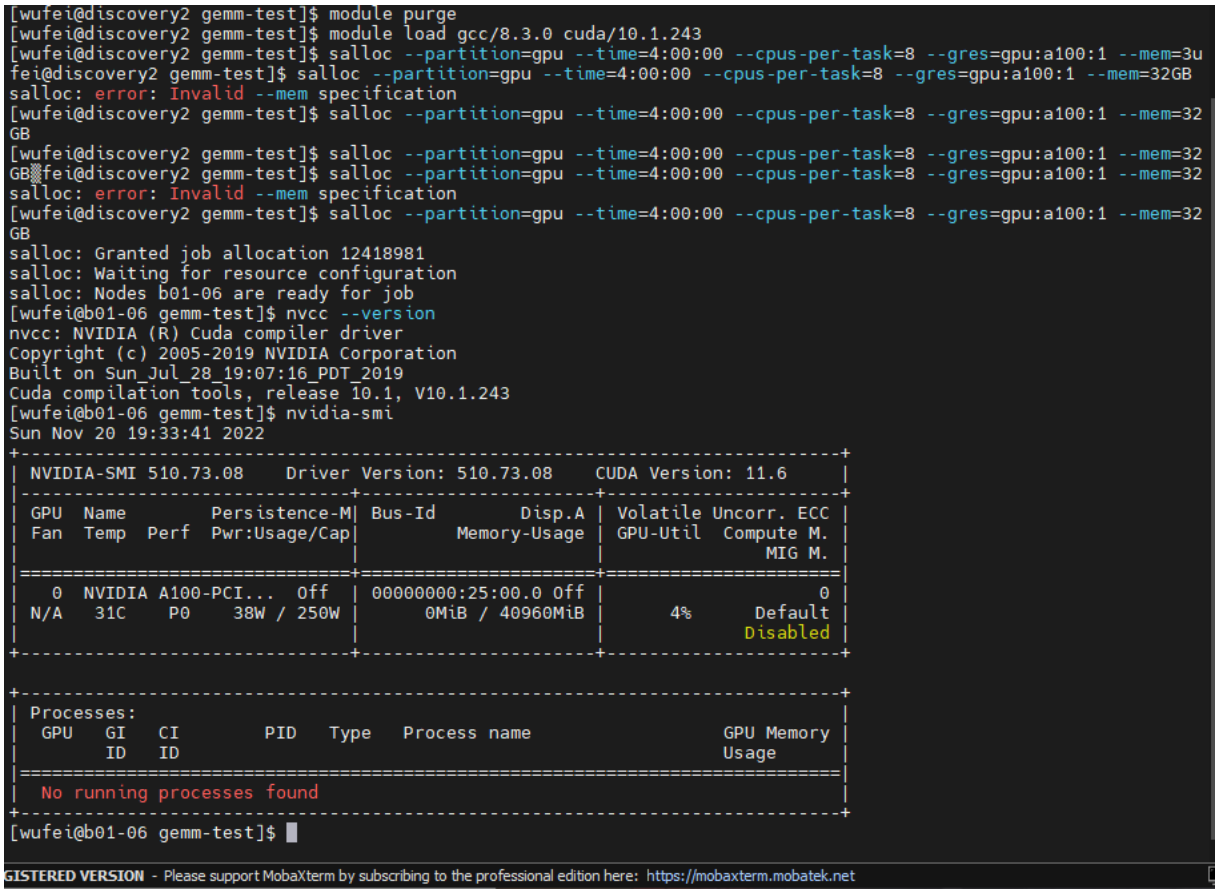
# 1 Tables

| a100 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Title Size | Matrix Size | A major | B major | Ex Time (ms) | FLOPs | GFLOPs |
| DWMMA | 1 | 32*16*16 | 16*16*16 | col | row | 2.055641 | 8192.0000 | 0.003985 |
| | 2 | 32*16*16 | 16*16*32 | col | row | 1.359971 | 16384.0000 | 0.012047 |
| | 3 | 32*16*16 | 16*16*16 | col | col | 2.052867 | 8192.0000 | 0.003991 |
| | 4 | 32*16*16 | 16*16*32 | col | col | 1.346217 | 16384.0000 | 0.01217 |
| | 5 | 32*16*16 | 16*16*16 | row | row | 2.055211 | 8192.0000 | 0.003986 |
| | 6 | 32*16*16 | 16*16*32 | row | row | 2.020667 | 16384.0000 | 0.008108 |
| | 7 | 32*16*16 | 16*16*16 | row | col | 2.053469 | 8192.0000 | 0.003989 |
| | 8 | 32*16*16 | 16*16*32 | row | col | 1.350815 | 16384.0000 | 0.012129 |
| | 9 | 32*16*16 | 32*32*32 | col | row | 1.345174 | 65536.0000 | 0.048719 |
| | 10 | 32*16*16 | 32*32*32 | col | col | 1.930839 | 65536.0000 | 0.033942 |
| | 11 | 32*16*16 | 32*32*32 | row | row | 1.367125 | 65536.0000 | 0.047937 |
| | 12 | 32*16*16 | 32*32*32 | row | col | 1.936322 | 65536.0000 | 0.033846 |
| | 13 | 32*128*128 | 128*128*128 | col | row | 1.69345 | 4194304.0000 | 2.476778 |
| | 14 | 32*128*128 | 128*128*128 | col | col | 1.689855 | 4194304.0000 | 2.48205 |
| | 15 | 32*128*128 | 128*128*128 | row | row | 1.708981 | 4194304.0000 | 2.454272 |
| | 16 | 32*128*128 | 128*128*128 | row | col | 1.709111 | 4194304.0000 | 2.454085 |
| | 17 | 32*128*128 | 256*256*128 | col | row | 1.372225 | 16777216.0000 | 12.226287 |
| | 18 | 32*128*128 | 256*256*128 | col | col | 1.35374 | 16777216.0000 | 12.393234 |
| | 19 | 32*128*128 | 256*256*128 | row | row | 1.351045 | 16777216.0000 | 12.417955 |
| | 20 | 32*128*128 | 256*256*128 | row | col | 1.359801 | 16777216.0000 | 12.337994 |
| | 21 | 64*128*128 | 256*256*256 | col | row | 1.363435 | 33554432.0000 | 24.610218 |
| | 22 | 64*128*128 | 256*256*256 | col | col | 1.346884 | 33554432.0000 | 24.912637 |
| | 23 | 64*128*128 | 256*256*256 | row | row | 1.352596 | 33554432.0000 | 24.809265 |
| | 24 | 64*128*128 | 256*256*256 | row | col | 1.350902 | 33554432.0000 | 24.838539 |
| SGEMM | 1 | 8*128*128 | 1024*512*8 | col | row | 1.351123 | 8388608.0000 | 6.208619 |
| | 2 | 8*128*128 | 128*81*1 | col | row | 1.340984 | 20736.0000 | 0.015463 |
| | 3 | 8*128*128 | 128*112*8 | col | row | 1.348889 | 229376.0000 | 0.170048 |
| DGEMM | 1 | 8*32*64 | 64*32*8 | col | row | 2.007764 | 32768.0000 | 0.016321 |
| | 2 | 8*32*64 | 256*128*64 | col | row | 1.362355 | 4194304.0000 | 3.078716 |
| | 3 | 8*64*64 | 64*64*8 | col | row | 1.939757 | 65536.0000 | 0.033786 |

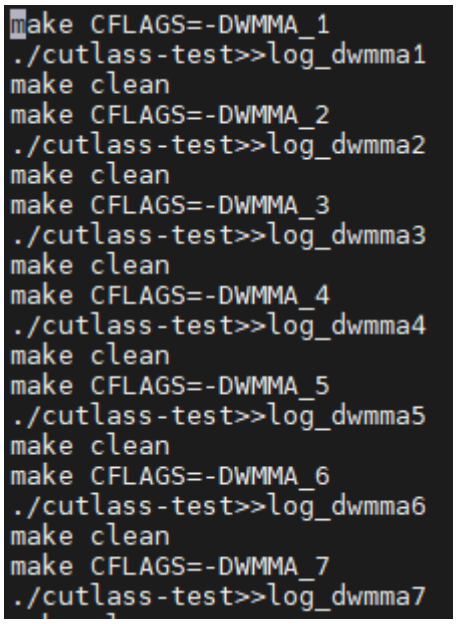
| a40 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Title Size | Matrix Size | A major | B major | Ex Time (ms) | FLOPs | GFLOPs |
| DWMMA | 1 | 32*16*16 | 16*16*16 | col | row | 1.044462 | 8192.0000 | 0.007843 |
| | 2 | 32*16*16 | 16*16*32 | col | row | 0.769658 | 16384.0000 | 0.021287 |
| | 3 | 32*16*16 | 16*16*16 | col | col | 0.680622 | 8192.0000 | 0.0120336 |
| | 4 | 32*16*16 | 16*16*32 | col | col | 0.779821 | 16384.0000 | 0.02101 |
| | 5 | 32*16*16 | 16*16*16 | row | row | 0.71148 | 8192.0000 | 0.011514 |
| | 6 | 32*16*16 | 16*16*32 | row | row | 0.77629 | 16384.0000 | 0.021106 |
| | 7 | 32*16*16 | 16*16*16 | row | col | 0.714236 | 8192.0000 | 0.01147 |
| | 8 | 32*16*16 | 16*16*32 | row | col | 0.803424 | 16384.0000 | 0.020393 |
| | 9 | 32*16*16 | 32*32*32 | col | row | 0.782515 | 65536.0000 | 0.08375 |
| | 10 | 32*16*16 | 32*32*32 | col | col | 0.791541 | 65536.0000 | 0.082795 |
| | 11 | 32*16*16 | 32*32*32 | row | row | 0.783696 | 65536.0000 | 0.083624 |
| | 12 | 32*16*16 | 32*32*32 | row | col | 0.785251 | 65536.0000 | 0.083395 |
| | 13 | 32*128*128 | 128*128*128 | col | row | 0.72180 | 4194304.0000 | 5.810863 |
| | 14 | 32*128*128 | 128*128*128 | col | col | 0.780555 | 4194304.0000 | 5.373489 |
| | 15 | 32*128*128 | 128*128*128 | row | row | 0.749086 | 4194304.0000 | 5.599229 |
| | 16 | 32*128*128 | 128*128*128 | row | col | 0.717948 | 4194304.0000 | 5.842072 |
| | 17 | 32*128*128 | 256*256*128 | col | row | 0.793625 | 16777216.0000 | 21.139979 |
| | 18 | 32*128*128 | 256*256*128 | col | col | 0.780471 | 16777216.0000 | 21.496271 |
| | 19 | 32*128*128 | 256*256*128 | row | row | 0.804676 | 16777216.0000 | 20.849654 |
| | 20 | 32*128*128 | 256*256*128 | row | col | 0.828801 | 16777216.0000 | 20.242755 |
| | 21 | 64*128*128 | 256*256*256 | col | row | 0.74797 | 33554432.0000 | 44.860666 |
| | 22 | 64*128*128 | 256*256*256 | col | col | 0.793095 | 33554432.0000 | 42.308213 |
| | 23 | 64*128*128 | 256*256*256 | row | row | 0.773748 | 33554432.0000 | 43.366099 |
| | 24 | 64*128*128 | 256*256*256 | row | col | 0.782936 | 33554432.0000 | 42.857184 |
| SGEMM | 1 | 8*128*128 | 1024*512*8 | col | row | 0.930532 | 8388608.0000 | 9.014852 |
| | 2 | 8*128*128 | 128*81*1 | col | row | 0.778397 | 20736.0000 | 0.026639 |
| | 3 | 8*128*128 | 128*112*8 | col | row | 0.758239 | 229376.0000 | 0.302511 |
| DGEMM | 1 | 8*32*64 | 64*32*8 | col | row | 0.930532 | 32768.0000 | 9.014852 |
| | 2 | 8*32*64 | 256*128*64 | col | row | 0.778397 | 4194304.0000 | 0.026639 |
| | 3 | 8*64*64 | 64*64*8 | col | row | 0.758239 | 65536.0000 | 0.302511 |

# 2 Report

Set the GPU machine in the CARC system:




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



```
[wufei@discovery1 cutlass-gpgpu-sim]$module purge
[wufei@discovery1 cutlass-gpgpu-sim]$module load gcc/8.3.0 cuda/10.1.2433
[wufei@discovery1 cutlass-gpgpu-sim]$ salloc --partition=gpu --time=4:00:00 --cpus-per-task=8 --gres=gpu:a100:1 --mem=32GB
salloc: Pending job allocation 12433778
salloc: job 12433778 queued and waiting for resources
salloc: job 12433778 has been allocated resources
salloc: Granted job allocation 12433778
salloc: Waiting for resource configuration
salloc: Nodes a01-06 are ready for job
[wufei@a01-06 cutlass-gpgpu-sim]$ ls
Makefile  README.md  cutlass  cutlass.cu  gemm-test  script  util
[wufei@a01-06 cutlass-gpgpu-sim]$ ./script
nvcc -DWMMA_1 -o cutlass-test cutlass.cu --gpu-architecture=compute_70 --gpu-code=compute_70 -lcudart -I./
rm cutlass-test
rm -rf gemm_tt_*
rm -rf host_results_*
rm -rf _*
rm -rf gpgpu_inst_stats.txt
echo Clean done
Clean done
nvcc -DWMMA_2 -o cutlass-test cutlass.cu --gpu-architecture=compute_70 --gpu-code=compute_70 -lcudart -I./
rm cutlass-test
rm -rf gemm_tt_*
rm -rf host_results_*
rm -rf _*
rm -rf gpgpu_inst_stats.txt
echo Clean done
Clean done
nvcc -DWMMA_3 -o cutlass-test cutlass.cu --gpu-architecture=compute_70 --gpu-code=compute_70 -lcudart -I./
rm cutlass-test
rm -rf gemm_tt_*
rm -rf host_results_*
rm -rf _*
rm -rf gpgpu_inst_stats.txt
echo Clean done
Clean done
nvcc -DWMMA_4 -o cutlass-test cutlass.cu --gpu-architecture=compute_70 --gpu-code=compute_70 -lcudart -I./
rm cutlass-test
rm -rf gemm_tt_*
rm -rf host_results_*
rm -rf _*
rm -rf gpgpu_inst_stats.txt
echo Clean done
Clean done
nvcc -DWMMA_5 -o cutlass-test cutlass.cu --gpu-architecture=compute_70 --gpu-code=compute_70 -lcudart -I./
rm cutlass-test
rm -rf gemm_tt_*
```

Check all the generated log files:

```
[wufei@a01-06 cutlass-gpgpu-sim]$ ls
Makefile    gemm-test    log_dsgemm1  log_dwmma10  log_dwmma14  log_dwmma18  log_dwmma21  log_dwmma3  log_dwmma7  util
README.md   log_ddgemm1  log_dsgemm2  log_dwmma11  log_dwmma15  log_dwmma19  log_dwmma22  log_dwmma4  log_dwmma8
cutlass     log_ddgemm2  log_dsgemm3  log_dwmma12  log_dwmma16  log_dwmma2   log_dwmma23  log_dwmma5  log_dwmma9
cutlass.cu  log_ddgemm3  log_dwmma1   log_dwmma13  log_dwmma17  log_dwmma20  log_dwmma24  log_dwmma6  script
[wufei@a01-06 cutlass-gpgpu-sim]$
```

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

```
 98    time = (stop.tv_sec - start.tv_sec) + (stop.tv_nsec - start.tv_nsec) / 1e9;
 99    printf("excution time is %f ms\n", (time * 1000));
100    double FLOPS = testbed.flops();
101    printf("FLOPs is %lf\n", FLOPS);
102    double GFLOPS = testbed.GFLOPs_per_sec(time*1000);
103    printf("GFLOPs is %lf\n", 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.

```
-rw-rw----. 1 wufei wufei   866 Nov 20 19:53 dgemm_tests.h
-rw-rw----. 1 wufei wufei  5645 Nov 21 20:39 gemm.h
-rw-rw----. 1 wufei wufei 12546 Nov 20 19:53 gemm_testbed.h
-rw-rw----. 1 wufei wufei   747 Nov 20 19:53 sgemm_tests.h
-rw-rw----. 1 wufei wufei 13932 Nov 20 19:53 wmma_tests.h
[wufei@b05-10 gemm-test]$
```

```
#ifdef WMMA_1
typedef cutlass::gemm::WmmaGemmTraits<cutlass::MatrixLayout::kColumnMajor,
                                      cutlass::MatrixLayout::kRowMajor,
                                      cutlass::Shape<32, 16, 16> >
    WmmaGemmTraits1;
run_gemm<WmmaGemmTraits1>(16, 16, 16);
#endif
#ifdef WMMA_2
typedef cutlass::gemm::WmmaGemmTraits<cutlass::MatrixLayout::kColumnMajor,
                                      cutlass::MatrixLayout::kRowMajor,
```

Run the script again and check the log:

Sample

```
[wufei@b05-13 cutlass-gpgpu-sim]$ cat log_dwmma1
Successfully Launched
excution time is 2.055641 ms
FLOPs is 8192.000000
GFLOPs is 0.003985
Result Verified
```

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

```
[wufei@b05-13 cutlass-gpgpu-sim]$ salloc --partition=gpu --time=4:00:00 --cpus-per-task=8 --gres=gpu:a40:1 --mem=32GB
salloc: Pending job allocation 12436053
salloc: job 12436053 queued and waiting for resources
salloc: job 12436053 has been allocated resources
salloc: Granted job allocation 12436053
salloc: Waiting for resource configuration
salloc: Nodes b05-10 are ready for job
[wufei@b05-10 cutlass-gpgpu-sim]$
```

```
[wufei@b05-10 cutlass-gpgpu-sim]$ nvidia-smi
Mon Nov 21 21:17:17 2022
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 510.73.08    Driver Version: 510.73.08    CUDA Version: 11.6     |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|                               |                      |               MIG M. |
|===============================+======================+======================|
|   0  NVIDIA A40          Off  | 00000000:21:00.0 Off |                    0 |
|  0%   27C    P0    69W / 300W |      0MiB / 46068MiB |      2%      Default |
|                               |                      |                  N/A |
+-------------------------------+----------------------+----------------------+

+-----------------------------------------------------------------------------+
| Processes:                                                                  |
|  GPU   GI   CI        PID   Type   Process name                  GPU Memory |
|        ID   ID                                                   Usage      |
|=============================================================================|
|  No running processes found                                                 |
+-----------------------------------------------------------------------------+
[wufei@b05-10 cutlass-gpgpu-sim]$
```

get all the required parameters and make the table.

## a40

| | | Title Size | Matrix Size | A major | B major | Ex Time (ms) | FLOPs | GFLOPs |
|---|---|---|---|---|---|---|---|---|
| DWMMA | 1 | 32*16*16 | 16*16*16 | col | row | 1.044462 | 8192.0000 | 0.007843 |
| | 2 | 32*16*16 | 16*16*32 | col | row | 0.769658 | 16384.0000 | 0.021287 |
| | 3 | 32*16*16 | 16*16*16 | col | col | 0.680622 | 8192.0000 | 0.0120336 |
| | 4 | 32*16*16 | 16*16*32 | col | col | 0.779821 | 16384.0000 | 0.02101 |
| | 5 | 32*16*16 | 16*16*16 | row | row | 0.71148 | 8192.0000 | 0.011514 |
| | 6 | 32*16*16 | 16*16*32 | row | row | 0.77629 | 16384.0000 | 0.021106 |
| | 7 | 32*16*16 | 16*16*16 | row | col | 0.714236 | 8192.0000 | 0.01147 |
| | 8 | 32*16*16 | 16*16*32 | row | col | 0.803424 | 16384.0000 | 0.020393 |
| | 9 | 32*16*16 | 32*32*32 | col | row | 0.782515 | 65536.0000 | 0.08375 |
| | 10 | 32*16*16 | 32*32*32 | col | col | 0.791541 | 65536.0000 | 0.082795 |
| | 11 | 32*16*16 | 32*32*32 | row | row | 0.783696 | 65536.0000 | 0.083624 |
| | 12 | 32*16*16 | 32*32*32 | row | col | 0.785251 | 65536.0000 | 0.083395 |
| | 13 | 32*128*128 | 128*128*128 | col | row | 0.72180 | 4194304.0000 | 5.810863 |
| | 14 | 32*128*128 | 128*128*128 | col | col | 0.780555 | 4194304.0000 | 5.373489 |
| | 15 | 32*128*128 | 128*128*128 | row | row | 0.749086 | 4194304.0000 | 5.599229 |
| | 16 | 32*128*128 | 128*128*128 | row | col | 0.717948 | 4194304.0000 | 5.842072 |
| | 17 | 32*128*128 | 256*256*128 | col | row | 0.793625 | 16777216.0000 | 21.139979 |
| | 18 | 32*128*128 | 256*256*128 | col | col | 0.780471 | 16777216.0000 | 21.496271 |
| | 19 | 32*128*128 | 256*256*128 | row | row | 0.804676 | 16777216.0000 | 20.849654 |
| | 20 | 32*128*128 | 256*256*128 | row | col | 0.828801 | 16777216.0000 | 20.242755 |
| | 21 | 64*128*128 | 256*256*256 | col | row | 0.74797 | 33554432.0000 | 44.860666 |
| | 22 | 64*128*128 | 256*256*256 | col | col | 0.793095 | 33554432.0000 | 42.308213 |
| | 23 | 64*128*128 | 256*256*256 | row | row | 0.773748 | 33554432.0000 | 43.366099 |
| | 24 | 64*128*128 | 256*256*256 | row | col | 0.782936 | 33554432.0000 | 42.857184 |
| SGEMM | 1 | 8*128*128 | 1024*512*8 | col | row | 0.930532 | 8388608.0000 | 9.014852 |
| | 2 | 8*128*128 | 128*81*1 | col | row | 0.778397 | 20736.0000 | 0.026639 |
| | 3 | 8*128*128 | 128*112*8 | col | row | 0.758239 | 229376.0000 | 0.302511 |
| DGEMM | 1 | 8*32*64 | 64*32*8 | col | row | 0.930532 | 32768.0000 | 9.014852 |
| | 2 | 8*32*64 | 256*128*64 | col | row | 0.778397 | 4194304.0000 | 0.026639 |
| | 3 | 8*64*64 | 64*64*8 | col | row | 0.758239 | 65536.0000 | 0.302511 |

## a100

| | | Title Size | Matrix Size | A major | B major | Ex Time (ms) | FLOPs | GFLOPs |
|---|---|---|---|---|---|---|---|---|
| DWMMA | 1 | 32*16*16 | 16*16*16 | col | row | 2.055641 | 8192.0000 | 0.003985 |
| | 2 | 32*16*16 | 16*16*32 | col | row | 1.359971 | 16384.0000 | 0.012047 |
| | 3 | 32*16*16 | 16*16*16 | col | col | 2.052867 | 8192.0000 | 0.003991 |
| | 4 | 32*16*16 | 16*16*32 | col | col | 1.346217 | 16384.0000 | 0.01217 |
| | 5 | 32*16*16 | 16*16*16 | row | row | 2.055211 | 8192.0000 | 0.003986 |
| | 6 | 32*16*16 | 16*16*32 | row | row | 2.020667 | 16384.0000 | 0.008108 |
| | 7 | 32*16*16 | 16*16*16 | row | col | 2.053469 | 8192.0000 | 0.003989 |
| | 8 | 32*16*16 | 16*16*32 | row | col | 1.350815 | 16384.0000 | 0.012129 |
| | 9 | 32*16*16 | 32*32*32 | col | row | 1.345174 | 65536.0000 | 0.048719 |
| | 10 | 32*16*16 | 32*32*32 | col | col | 1.930839 | 65536.0000 | 0.033942 |
| | 11 | 32*16*16 | 32*32*32 | row | row | 1.367125 | 65536.0000 | 0.047937 |
| | 12 | 32*16*16 | 32*32*32 | row | col | 1.936322 | 65536.0000 | 0.033846 |
| | 13 | 32*128*128 | 128*128*128 | col | row | 1.69345 | 4194304.0000 | 2.476778 |
| | 14 | 32*128*128 | 128*128*128 | col | col | 1.689855 | 4194304.0000 | 2.48205 |
| | 15 | 32*128*128 | 128*128*128 | row | row | 1.708981 | 4194304.0000 | 2.454272 |
| | 16 | 32*128*128 | 128*128*128 | row | col | 1.709111 | 4194304.0000 | 2.454085 |
| | 17 | 32*128*128 | 256*256*128 | col | row | 1.372225 | 16777216.0000 | 12.226287 |
| | 18 | 32*128*128 | 256*256*128 | col | col | 1.35374 | 16777216.0000 | 12.393234 |
| | 19 | 32*128*128 | 256*256*128 | row | row | 1.351045 | 16777216.0000 | 12.417955 |
| | 20 | 32*128*128 | 256*256*128 | row | col | 1.359801 | 16777216.0000 | 12.337994 |
| | 21 | 64*128*128 | 256*256*256 | col | row | 1.363435 | 33554432.0000 | 24.610218 |
| | 22 | 64*128*128 | 256*256*256 | col | col | 1.346884 | 33554432.0000 | 24.912637 |
| | 23 | 64*128*128 | 256*256*256 | row | row | 1.352596 | 33554432.0000 | 24.809265 |
| | 24 | 64*128*128 | 256*256*256 | row | col | 1.350902 | 33554432.0000 | 24.838539 |
| SGEMM | 1 | 8*128*128 | 1024*512*8 | col | row | 1.351123 | 8388608.0000 | 6.208619 |
| | 2 | 8*128*128 | 128*81*1 | col | row | 1.340984 | 20736.0000 | 0.015463 |
| | 3 | 8*128*128 | 128*112*8 | col | row | 1.348889 | 229376.0000 | 0.170048 |
| DGEMM | 1 | 8*32*64 | 64*32*8 | col | row | 2.007764 | 32768.0000 | 0.016321 |
| | 2 | 8*32*64 | 256*128*64 | col | row | 1.362355 | 4194304.0000 | 3.078716 |
| | 3 | 8*64*64 | 64*64*8 | col | row | 1.939757 | 65536.0000 | 0.033786 |

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?

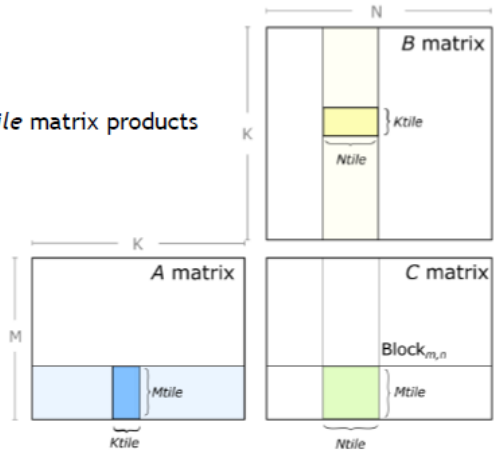
ANS: The column/column major or row/row major will increase execution time. The pipeline stalls due to memory is not affected by the column/row-major order a lot.
Tile is the parameter shows that how the matric be partitioned into independent matric products in each dimension and the matrix is the products shows in three dimensions. And the matric a and b is the layout of the matrix defined in the cutlass.

Partition the loop nest into *blocks* along each dimension

- Partition into *Mtile*-by-*Ntile* independent matrix products
- Compute each product by accumulating *Mtile*-by-*Ntile*-by-*Ktile* matrix products

```
for (int mb = 0; mb < M; mb += Mtile)
    for (int nb = 0; nb < N; nb += Ntile)
        for (int kb = 0; kb < K; kb += Ktile)
        {
            // compute Mtile-by-Ntile-by-Ktile matrix product
            for (int k = 0; k < Ktile; ++k)
                for (int i = 0; i < Mtile; ++i)
                    for (int j = 0; j < Ntile; ++j)
                    {
                        int row = mb + i;
                        int col = nb + j;

                        C[row][col] +=
                            A[row][kb + k] * B[kb + k][col];
                    }
        }
```

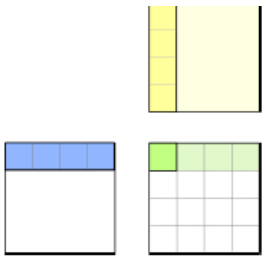



Basic definition

General matrix product

$$C = \alpha\ op(A) * op(B) + \beta\ C$$

$C$ is $M$-by-$N$, $op(A)$ is $M$-by-$K$, $op(B)$ is $K$-by-$N$



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.