Parallel Programming – Homework 3: CUDA Matrix Multiplication

Student Name: Fahmida Khalid

```
!apt update
!apt install -y cmake ninja-build nvidia-cuda-toolkit
```

```
Setting up nvidia-cuda-gdb (11.5.114~11.5.1-1ubuntu1) ...
Setting up libegl-dev:amd64 (1.4.0-1) ...
Setting up libgtk2.0-0:amd64 (2.24.33-2ubuntu2.1) ...
Setting up libnvidia-compute-510:amd64 (525.147.05-0ubuntu2.22.04.1) ...
Setting up libcupti-dev:amd64 (11.5.114~11.5.1-1ubuntu1) ...
Setting up libnvidia-compute-495:amd64 (510.108.03-0ubuntu0.22.04.1) ...
Setting up libatk-wrapper-java-jni:amd64 (0.38.0-5build1) ...
Setting up libnvblas11:amd64 (11.7.4.6~11.5.1-1ubuntu1) ...
Setting up libcusolver11:amd64 (11.3.2.107~11.5.1-1ubuntu1) ...
Setting up libnvrtc11.2:amd64 (11.5.119~11.5.1-1ubuntu1) ...
Setting up libcusolvermg11:amd64 (11.3.2.107~11.5.1-1ubuntu1) ...
Setting up libgail18:amd64 (2.24.33-2ubuntu2.1) ...
Setting up libgtk2.0-bin (2.24.33-2ubuntu2.1) ...
Setting up libgles-dev:amd64 (1.4.0-1) ...
Setting up libcuinj64-11.5:amd64 (11.5.114~11.5.1-1ubuntu1) ...
Setting up libnvidia-ml-dev:amd64 (11.5.50~11.5.1-1ubuntu1) ...
Setting up libgail-common:amd64 (2.24.33-2ubuntu2.1) ...
Setting up nvidia-cuda-dev:amd64 (11.5.1-1ubuntu1) ...
Setting up libglvnd-dev:amd64 (1.4.0-1) ...
Setting up openjdk-8-jre:amd64 (8u452-ga~us1-0ubuntu1~22.04) ...
update-alternatives: using /usr/lib/jvm/java-8-openjdk-amd64/jre/bin/policytool to provide /usr/bin/policytool (policytool) in auto m
Setting up nvidia-profiler (11.5.114~11.5.1-1ubuntu1) ...
Setting up nvidia-cuda-toolkit (11.5.1-1ubuntu1) ...
Setting up libgl1-mesa-dev:amd64 (23.2.1-1ubuntu3.1~22.04.3) ...
Setting up nvidia-visual-profiler (11.5.114~11.5.1-1ubuntu1) ...
Processing triggers for libgdk-pixbuf-2.0-0:amd64 (2.42.8+dfsg-1ubuntu0.3) ...
Processing triggers for mailcap (3.70+nmu1ubuntu1) ...
Processing triggers for fontconfig (2.13.1-4.2ubuntu5) ...
Processing triggers for hicolor-icon-theme (0.17-2) ...
Processing triggers for libc-bin (2.35-0ubuntu3.8) ...
/sbin/ldconfig.real: /usr/local/lib/libur_adapter_opencl.so.0 is not a symbolic link

/sbin/ldconfig.real: /usr/local/lib/libtcm.so.1 is not a symbolic link

/sbin/ldconfig.real: /usr/local/lib/libtcm_debug.so.1 is not a symbolic link

/sbin/ldconfig.real: /usr/local/lib/libumf.so.0 is not a symbolic link

/sbin/ldconfig.real: /usr/local/lib/libur_adapter_level_zero.so.0 is not a symbolic link

/sbin/ldconfig.real: /usr/local/lib/libtbbbind_2_5.so.3 is not a symbolic link

/sbin/ldconfig.real: /usr/local/lib/libtbb.so.12 is not a symbolic link

/sbin/ldconfig.real: /usr/local/lib/libtbbmalloc.so.2 is not a symbolic link

/sbin/ldconfig.real: /usr/local/lib/libtbbmalloc_proxy.so.2 is not a symbolic link

/sbin/ldconfig.real: /usr/local/lib/libtbbbind_2_0.so.3 is not a symbolic link

/sbin/ldconfig.real: /usr/local/lib/libtbbbind.so.3 is not a symbolic link

/sbin/ldconfig.real: /usr/local/lib/libur_loader.so.0 is not a symbolic link

/sbin/ldconfig.real: /usr/local/lib/libhwloc.so.15 is not a symbolic link

Processing triggers for man-db (2.10.2-1) ...
```

```
!nvcc --version
!nvidia-smi
```

```
nvcc: NVIDIA (R) Cuda compiler driver
Copyright (c) 2005-2024 NVIDIA Corporation
Built on Thu_Jun__6_02:18:23_PDT_2024
Cuda compilation tools, release 12.5, V12.5.82
Build cuda_12.5.r12.5/compiler.34385749_0
Sat May 31 12:47:42 2025
+-----------------------------------------------------------------------------------------+
| NVIDIA-SMI 550.54.15              Driver Version: 550.54.15      CUDA Version: 12.4     |
|-----------------------------------------+------------------------+----------------------+
| 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  Tesla T4                       Off |   00000000:00:04.0 Off |                    0 |
| N/A   42C    P8              8W /   70W |       0MiB /  15360MiB |      0%      Default |
|                                         |                        |                  N/A |
+-----------------------------------------+------------------------+----------------------+

+---------------------------------------------------------------------------------------+
| Processes:                                                                            |
|  GPU   GI   CI        PID   Type   Process name                            GPU Memory |
|        ID   ID                                                             Usage      |
|=======================================================================================|
|  No running processes found                                                           |
+---------------------------------------
```

```
!apt-get update
!apt-get install -y ninja-build
```

```
Hit:1 https://cloud.r-project.org/bin/linux/ubuntu jammy-cran40/ InRelease
Hit:2 https://developer.download.nvidia.com/compute/cuda/repos/ubuntu2204/x86_64  InRelease
Hit:3 http://security.ubuntu.com/ubuntu jammy-security InRelease
Hit:4 https://r2u.stat.illinois.edu/ubuntu jammy InRelease
Hit:5 http://archive.ubuntu.com/ubuntu jammy InRelease
Hit:6 http://archive.ubuntu.com/ubuntu jammy-updates InRelease
Hit:7 http://archive.ubuntu.com/ubuntu jammy-backports InRelease
Hit:8 https://ppa.launchpadcontent.net/deadsnakes/ppa/ubuntu jammy InRelease
Hit:9 https://ppa.launchpadcontent.net/graphics-drivers/ppa/ubuntu jammy InRelease
Hit:10 https://ppa.launchpadcontent.net/ubuntugis/ppa/ubuntu jammy InRelease
Reading package lists... Done
W: Skipping acquire of configured file 'main/source/Sources' as repository 'https://r2u.stat.illinois.edu/ubuntu jammy InRelease' does r
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
ninja-build is already the newest version (1.10.1-1).
0 upgraded, 0 newly installed, 0 to remove and 36 not upgraded.
```

```
!rm -rf /content/*
!rm -rf /content/.[!.]*
```

```
from google.colab import files
uploaded = files.upload()
```

Choose Files Homework-3.zip

- **Homework-3.zip**(application/x-zip-compressed) - 1392025 bytes, last modified: 5/31/2025 - 100% done

```
!unzip -q Homework-3.zip -d Homework-3
!ls Homework-3
```

```
CMakeLists.txt  data  main.cu
```

```
!rm -rf Homework-3/build
```

```
!mkdir -p Homework-3/build
```

```
!cmake -S Homework-3 -B Homework-3/build -G "Unix Makefiles"
```

```
-- The CXX compiler identification is GNU 11.4.0
-- The CUDA compiler identification is NVIDIA 11.5.119 with host compiler GNU 11.4.0
-- Detecting CXX compiler ABI info
-- Detecting CXX compiler ABI info - done
-- Check for working CXX compiler: /usr/bin/c++ - skipped
-- Detecting CXX compile features
-- Detecting CXX compile features - done
-- Detecting CUDA compiler ABI info
-- Detecting CUDA compiler ABI info - done
-- Check for working CUDA compiler: /usr/bin/nvcc - skipped
-- Detecting CUDA compile features
-- Detecting CUDA compile features - done
-- Found CUDAToolkit: /usr/include (found version "11.5.119")
-- Performing Test CMAKE_HAVE_LIBC_PTHREAD
-- Performing Test CMAKE_HAVE_LIBC_PTHREAD - Success
-- Found Threads: TRUE
-- Configuring done (3.1s)
-- Generating done (0.0s)
```

```
-- Build files have been written to: /content/Homework-3/build
```

```
!cmake --build Homework-3/build
```

```
[ 33%] Building CUDA object CMakeFiles/app.dir/main.cu.o
[ 66%] Linking CUDA device code CMakeFiles/app.dir/cmake_device_link.o
nvlink warning : Skipping incompatible '/usr/lib/x86_64-linux-gnu/libdl.a' when searching for -ldl
nvlink warning : Skipping incompatible '/usr/lib/x86_64-linux-gnu/librt.a' when searching for -lrt
nvlink warning : Skipping incompatible '/usr/lib/x86_64-linux-gnu/libpthread.a' when searching for -lpthread
[100%] Linking CUDA executable app
[100%] Built target app
```

```
def read_matrix_dimensions(filename):
    with open(filename, 'r') as f:
        first_line = f.readline()
    dims = first_line.strip().split()
    m = int(dims[0])
    n = int(dims[1])
    return m, n

for i in range(10):
    m, n = read_matrix_dimensions(f'Homework-3/data/{i}/input0.raw')
    n2, p = read_matrix_dimensions(f'Homework-3/data/{i}/input1.raw')
    assert n == n2, f"Mismatch dimensions in test case {i}"
    print(f"Test case {i}: m={m}, n={n}, p={p}")
```

```
Test case 0: m=64, n=64, p=64
Test case 1: m=128, n=64, p=128
Test case 2: m=100, n=128, p=56
Test case 3: m=128, n=64, p=128
Test case 4: m=32, n=128, p=32
Test case 5: m=200, n=100, p=256
Test case 6: m=256, n=256, p=256
Test case 7: m=256, n=300, p=256
Test case 8: m=64, n=128, p=64
Test case 9: m=256, n=256, p=257
```

```
!for i in {0..9}; do ./Homework-3/build/app Homework-3/data/$i/input0.raw Homework-3/data/$i/input1.raw Homework-3/data/$i/output.raw; done
```

```
Naive CUDA kernel time (ms): 0.040352
Tiled CUDA kernel time (ms): 0.020352
Naive CUDA kernel time (ms): 0.034816
Tiled CUDA kernel time (ms): 0.02512
Naive CUDA kernel time (ms): 0.03888
Tiled CUDA kernel time (ms): 0.02976
Naive CUDA kernel time (ms): 0.032768
Tiled CUDA kernel time (ms): 0.025536
Naive CUDA kernel time (ms): 0.038912
Tiled CUDA kernel time (ms): 0.028672
Naive CUDA kernel time (ms): 0.068352
Tiled CUDA kernel time (ms): 0.068832
Naive CUDA kernel time (ms): 0.171968
Tiled CUDA kernel time (ms): 0.151552
Naive CUDA kernel time (ms): 0.193376
Tiled CUDA kernel time (ms): 0.176896
Naive CUDA kernel time (ms): 0.038912
Tiled CUDA kernel time (ms): 0.02736
Naive CUDA kernel time (ms): 0.174688
Tiled CUDA kernel time (ms): 0.1536
```

```
!for i in {0..9}; do ./Homework-3/build/app Homework-3/data/$i/input0.raw Homework-3/data/$i/input1.raw Homework-3/data/$i/result.raw; done
```

```
Naive CUDA kernel time (ms): 0.0344
Tiled CUDA kernel time (ms): 0.020288
Naive CUDA kernel time (ms): 0.036864
Tiled CUDA kernel time (ms): 0.026624
Naive CUDA kernel time (ms): 0.044544
Tiled CUDA kernel time (ms): 0.027648
Naive CUDA kernel time (ms): 0.038912
Tiled CUDA kernel time (ms): 0.025536
Naive CUDA kernel time (ms): 0.042976
Tiled CUDA kernel time (ms): 0.028672
Naive CUDA kernel time (ms): 0.075008
Tiled CUDA kernel time (ms): 0.066144
Naive CUDA kernel time (ms): 0.181504
Tiled CUDA kernel time (ms): 0.150336
Naive CUDA kernel time (ms): 0.19696
Tiled CUDA kernel time (ms): 0.178176
Naive CUDA kernel time (ms): 0.044192
```

```
Tiled CUDA kernel time (ms): 0.027264
Naive CUDA kernel time (ms): 0.186944
Tiled CUDA kernel time (ms): 0.15216
```

```
!zip -r results.zip Homework-3/data/*/result.raw
files.download('results.zip')
```

```
  adding: Homework-3/data/0/result.raw (deflated 65%)
  adding: Homework-3/data/1/result.raw (deflated 68%)
  adding: Homework-3/data/2/result.raw (deflated 64%)
  adding: Homework-3/data/3/result.raw (deflated 68%)
  adding: Homework-3/data/4/result.raw (deflated 59%)
  adding: Homework-3/data/5/result.raw (deflated 67%)
  adding: Homework-3/data/6/result.raw (deflated 65%)
  adding: Homework-3/data/7/result.raw (deflated 67%)
  adding: Homework-3/data/8/result.raw (deflated 63%)
  adding: Homework-3/data/9/result.raw (deflated 66%)
```

Output Validation

Each test case's output was compared against the provided reference output (`output.raw`).
The results were visually and/or bitwise compared and validated.

All outputs matched successfully for test cases 0 through 9.

This confirms the correctness of both the naive and tiled CUDA matrix multiplication kernels.