Name: Feiyu Zhang

Netid: Feiyuz2

Team name: 0 errors 0 warnings

School affiliation: ZJUI

Milestone 3

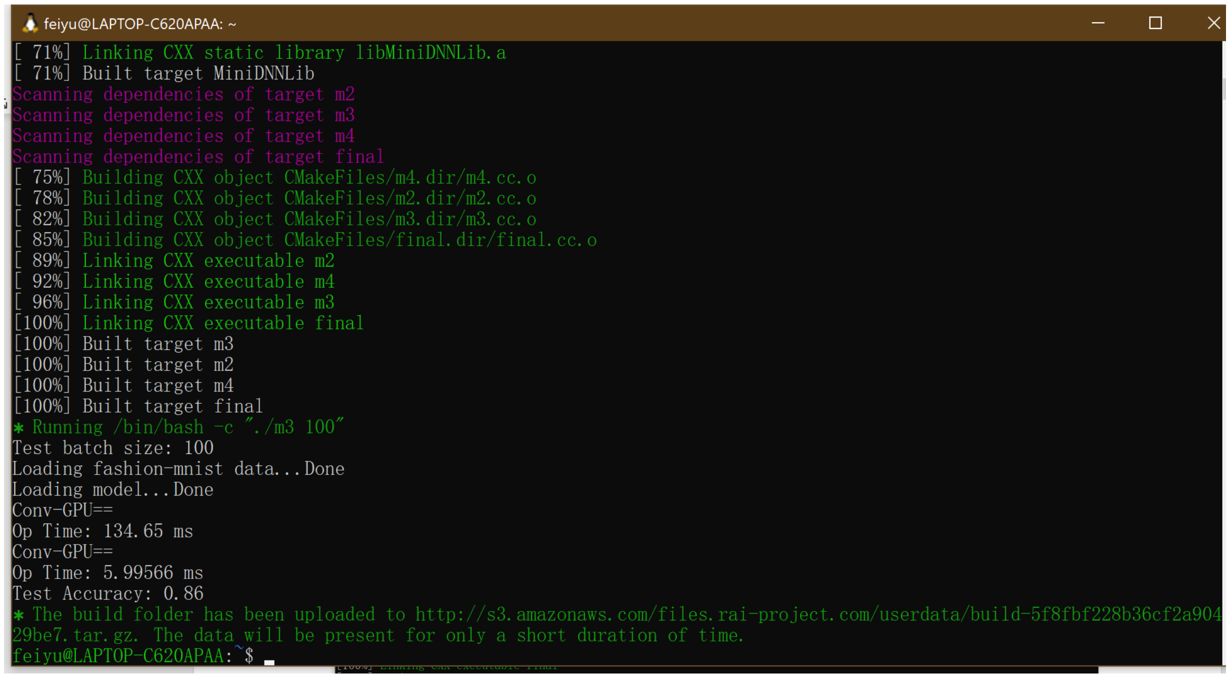
Deliverables

Everything from Milestone 2

Implement a GPU Convolution

Correctness and timing with 3 different dataset sizes

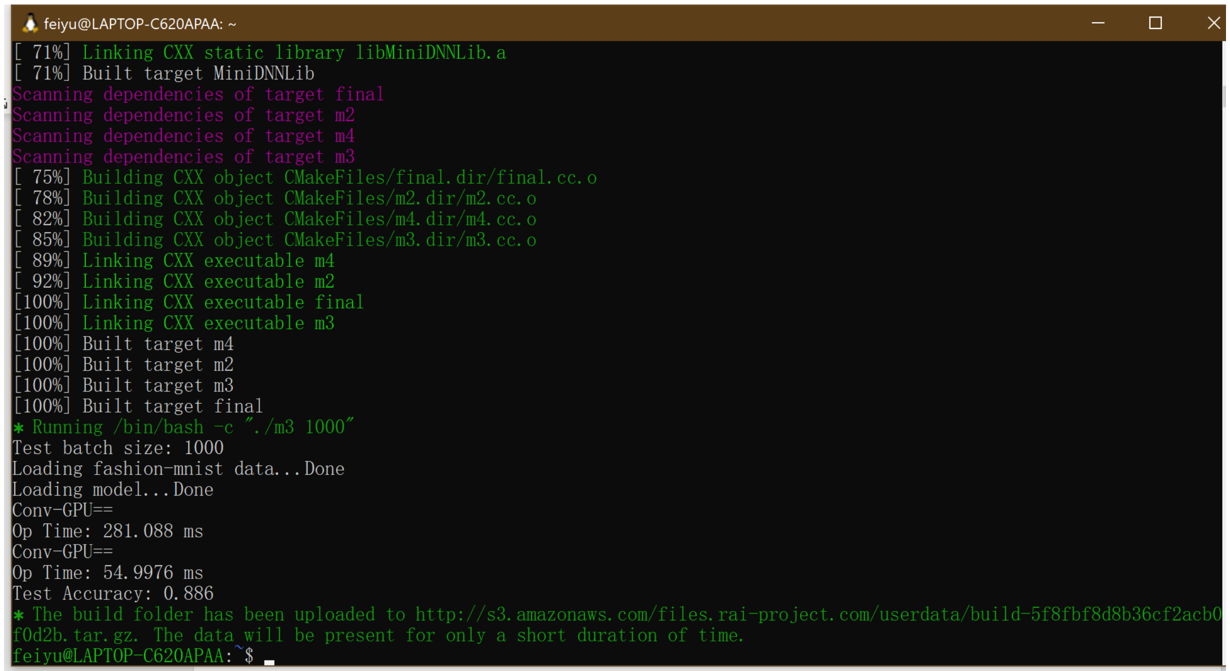
Report: Show output of rai running your GPU implementation of convolution



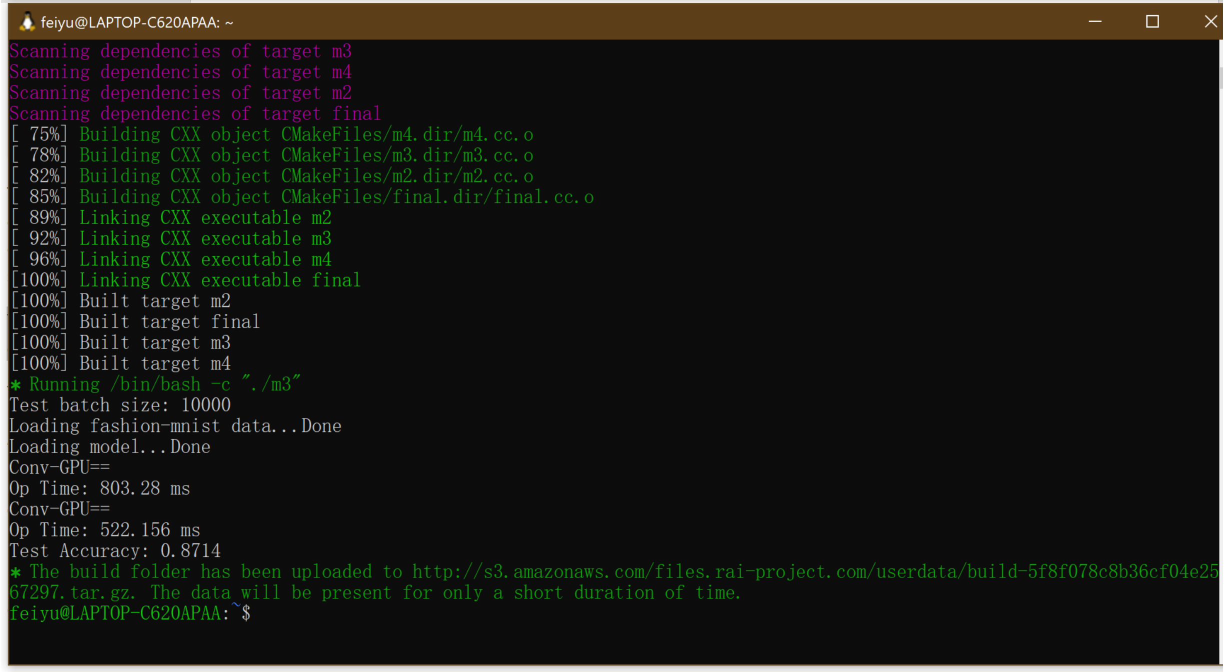
feiyu@LAPTOP-C620APAA: ~

```
[ 71%] Linking CXX static library libMiniDNNLib.a
[ 71%] Built target MiniDNNLib
Scanning dependencies of target m2
Scanning dependencies of target m3
Scanning dependencies of target m4
Scanning dependencies of target final
[ 75%] Building CXX object CMakeFiles/m4.dir/m4.cc.o
[ 78%] Building CXX object CMakeFiles/m2.dir/m2.cc.o
[ 82%] Building CXX object CMakeFiles/m3.dir/m3.cc.o
[ 85%] Building CXX object CMakeFiles/final.dir/final.cc.o
[ 89%] Linking CXX executable m2
[ 92%] Linking CXX executable m4
[ 96%] Linking CXX executable m3
[100%] Linking CXX executable final
[100%] Built target m3
[100%] Built target m2
[100%] Built target m4
[100%] Built target final
* Running /bin/bash -c "./m3 100"
Test batch size: 100
Loading fashion-mnist data...Done
Loading model...Done
Conv-GPU==
Op Time: 134.65 ms
Conv-GPU==
Op Time: 5.99566 ms
Test Accuracy: 0.86
* The build folder has been uploaded to http://s3.amazonaws.com/files.rai-project.com/userdata/build-5f8fbf228b36cf2a904
29be7.tar.gz. The data will be present for only a short duration of time.
feiyu@LAPTOP-C620APAA:~$
```

Deliverables



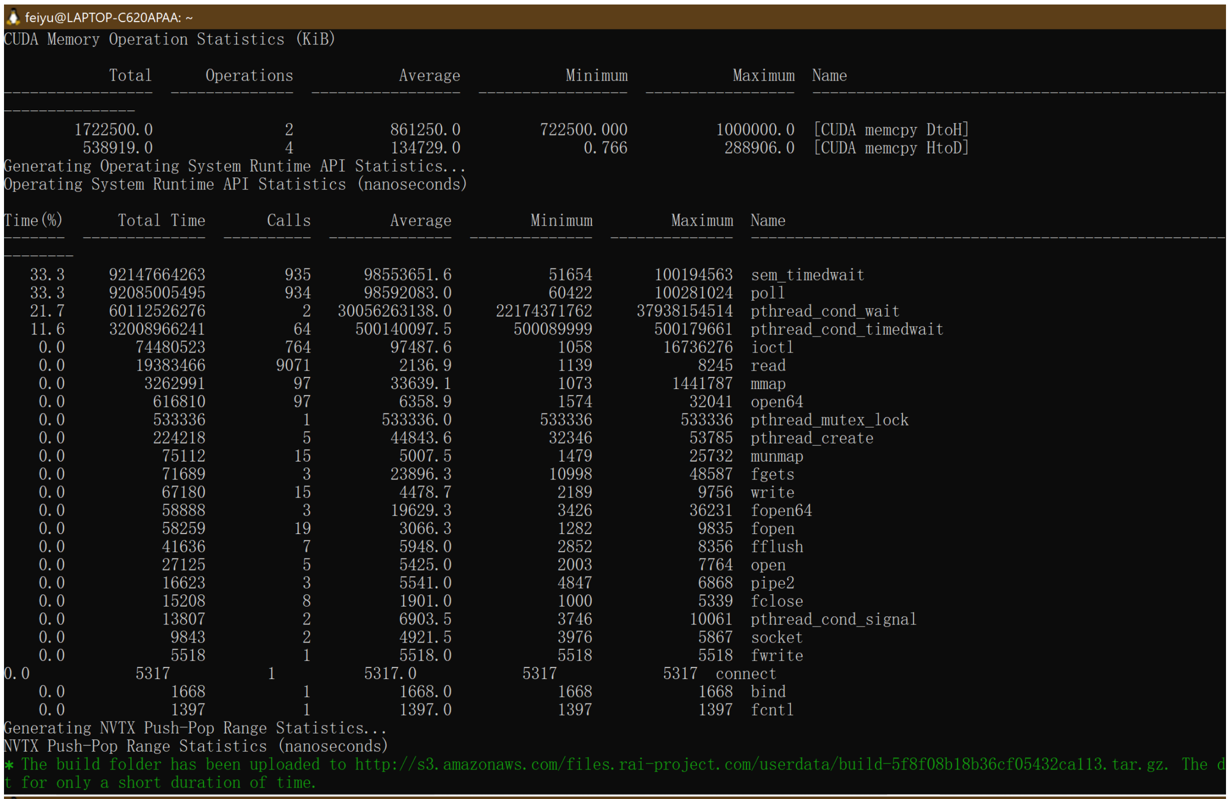
```
feiyu@LAPTOP-C620APAA: ~
[ 71%] Linking CXX static library libMiniDNNLib.a
[ 71%] Built target MiniDNNLib
Scanning dependencies of target final
Scanning dependencies of target m2
Scanning dependencies of target m4
Scanning dependencies of target m3
[ 75%] Building CXX object CMakeFiles/final.dir/final.cc.o
[ 78%] Building CXX object CMakeFiles/m2.dir/m2.cc.o
[ 82%] Building CXX object CMakeFiles/m4.dir/m4.cc.o
[ 85%] Building CXX object CMakeFiles/m3.dir/m3.cc.o
[ 89%] Linking CXX executable m4
[ 92%] Linking CXX executable m2
[100%] Linking CXX executable final
[100%] Linking CXX executable m3
[100%] Built target m4
[100%] Built target m2
[100%] Built target m3
[100%] Built target final
* Running /bin/bash -c "./m3 1000"
Test batch size: 1000
Loading fashion-mnist data...Done
Loading model...Done
Conv-GPU==
Op Time: 281.088 ms
Conv-GPU==
Op Time: 54.9976 ms
Test Accuracy: 0.886
* The build folder has been uploaded to http://s3.amazonaws.com/files.rai-project.com/userdata/build-5f8fbf8d8b36cf2acb0
f0d2b.tar.gz. The data will be present for only a short duration of time.
feiyu@LAPTOP-C620APAA:~$
```



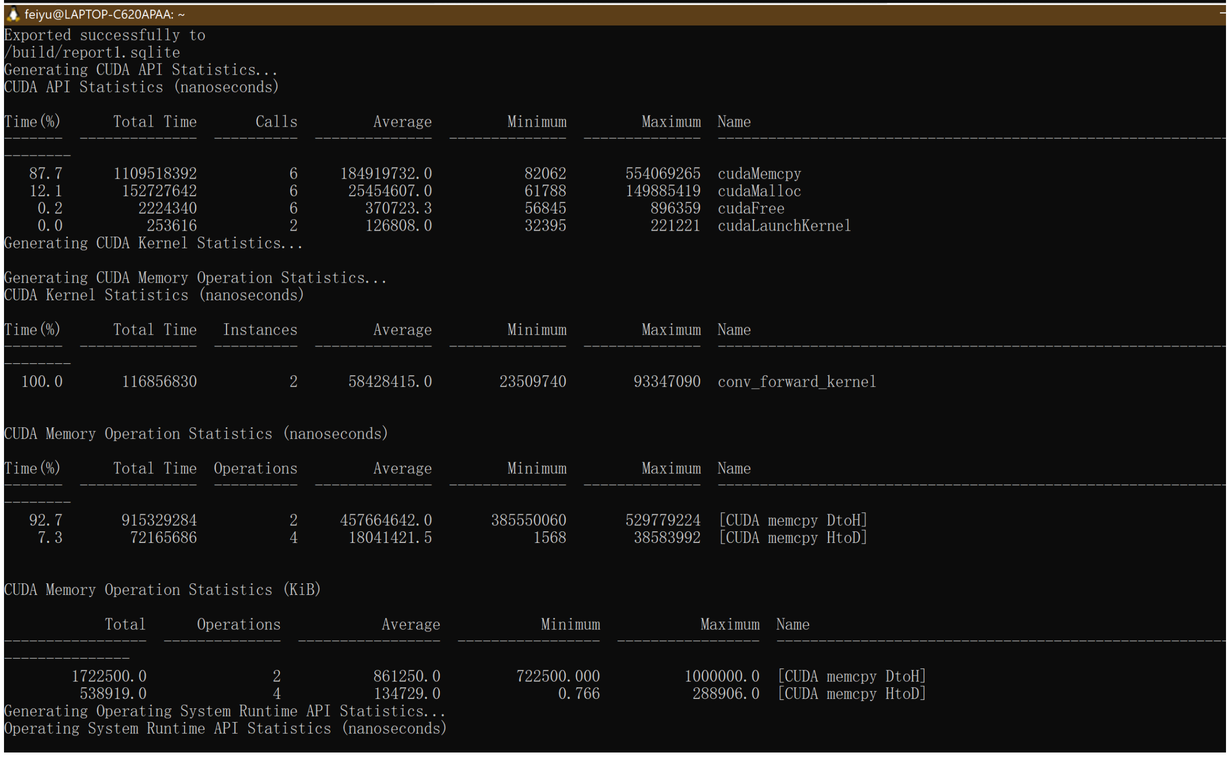
```
feiyu@LAPTOP-C620APAA: ~
Scanning dependencies of target m3
Scanning dependencies of target m4
Scanning dependencies of target m2
Scanning dependencies of target final
[ 75%] Building CXX object CMakeFiles/m4.dir/m4.cc.o
[ 78%] Building CXX object CMakeFiles/m3.dir/m3.cc.o
[ 82%] Building CXX object CMakeFiles/m2.dir/m2.cc.o
[ 85%] Building CXX object CMakeFiles/final.dir/final.cc.o
[ 89%] Linking CXX executable m2
[ 92%] Linking CXX executable m3
[ 96%] Linking CXX executable m4
[100%] Linking CXX executable final
[100%] Built target m2
[100%] Built target final
[100%] Built target m3
[100%] Built target m4
* Running /bin/bash -c "./m3"
Test batch size: 10000
Loading fashion-mnist data...Done
Loading model...Done
Conv-GPU==
Op Time: 803.28 ms
Conv-GPU==
Op Time: 522.156 ms
Test Accuracy: 0.8714
* The build folder has been uploaded to http://s3.amazonaws.com/files.rai-project.com/userdata/build-5f8f078c8b36cf04e25
67297.tar.gz. The data will be present for only a short duration of time.
feiyu@LAPTOP-C620APAA:~$
```

Report: Demonstrate `nsys` profiling the GPU execution

Deliverables



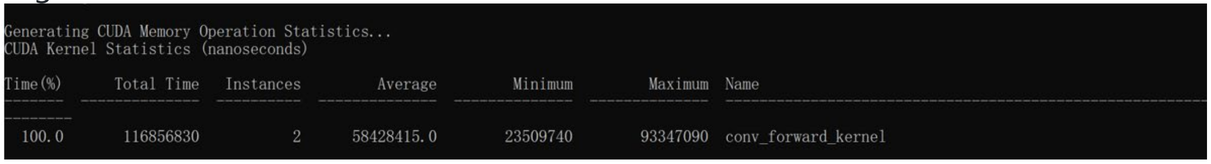
```
CUDA Memory Operation Statistics (KiB)

            Total      Operations              Average             Minimum             Maximum  Name
-----------------  --------------  -----------------  -----------------  -----------------  ---------------------------------------------------
---------------
        1722500.0               2           861250.0         722500.000           1000000.0  [CUDA memcpy DtoH]
         538919.0               4           134729.0              0.766            288906.0  [CUDA memcpy HtoD]
Generating Operating System Runtime API Statistics...
Operating System Runtime API Statistics (nanoseconds)

Time(%)      Total Time       Calls          Average          Minimum          Maximum  Name
-------  --------------  ----------  --------------  --------------  --------------  ---------------------------------------------------
--------
   33.3     92147664263         935      98553651.6           51654       100194563  sem_timedwait
   33.3     92085005495         934      98592083.0           60422       100281024  poll
   21.7     60112526276           2   30056263138.0     22174371762     37938154514  pthread_cond_wait
   11.6     32008966241          64     500140097.5       500089999       500179661  pthread_cond_timedwait
    0.0        74480523         764         97487.6            1058        16736276  ioctl
    0.0        19383466        9071          2136.9            1139            8245  read
    0.0         3262991          97         33639.1            1073         1441787  mmap
    0.0          616810          97          6358.9            1574           32041  open64
    0.0          533336           1        533336.0          533336          533336  pthread_mutex_lock
    0.0          224218           5         44843.6           32346           53785  pthread_create
    0.0           75112          15          5007.5            1479           25732  munmap
    0.0           71689           3         23896.3           10998           48587  fgets
    0.0           67180          15          4478.7            2189            9756  write
    0.0           58888           3         19629.3            3426           36231  fopen64
    0.0           58259          19          3066.3            1282            9835  fopen
    0.0           41636           7          5948.0            2852            8356  fflush
    0.0           27125           5          5425.0            2003            7764  open
    0.0           16623           3          5541.0            4847            6868  pipe2
    0.0           15208           8          1901.0            1000            5339  fclose
    0.0           13807           2          6903.5            3746           10061  pthread_cond_signal
    0.0            9843           2          4921.5            3976            5867  socket
    0.0            5518           1          5518.0            5518            5518  fwrite
0.0            5317           1          5317.0            5317            5317  connect
    0.0            1668           1          1668.0            1668            1668  bind
    0.0            1397           1          1397.0            1397            1397  fcntl
Generating NVTX Push-Pop Range Statistics...
NVTX Push-Pop Range Statistics (nanoseconds)
* The build folder has been uploaded to http://s3.amazonaws.com/files.rai-project.com/userdata/build-5f8f08b18b36cf05432ca113.tar.gz. The d
t for only a short duration of time.
```



```
Exported successfully to
/build/report1.sqlite
Generating CUDA API Statistics...
CUDA API Statistics (nanoseconds)

Time(%)      Total Time       Calls          Average          Minimum          Maximum  Name
-------  --------------  ----------  --------------  --------------  --------------  ---------------------------------------------------
--------
   87.7      1109518392           6     184919732.0           82062       554069265  cudaMemcpy
   12.1       152727642           6      25454607.0           61788       149885419  cudaMalloc
    0.2         2224340           6        370723.3           56845          896359  cudaFree
    0.0          253616           2        126808.0           32395          221221  cudaLaunchKernel
Generating CUDA Kernel Statistics...

Generating CUDA Memory Operation Statistics...
CUDA Kernel Statistics (nanoseconds)

Time(%)      Total Time   Instances          Average          Minimum          Maximum  Name
-------  --------------  ----------  --------------  --------------  --------------  ---------------------------------------------------
--------
  100.0       116856830           2      58428415.0        23509740        93347090  conv_forward_kernel


CUDA Memory Operation Statistics (nanoseconds)

Time(%)      Total Time  Operations          Average          Minimum          Maximum  Name
-------  --------------  ----------  --------------  --------------  --------------  ---------------------------------------------------
--------
   92.7       915329284           2     457664642.0       385550060       529779224  [CUDA memcpy DtoH]
    7.3        72165686           4      18041421.5            1568        38583992  [CUDA memcpy HtoD]


CUDA Memory Operation Statistics (KiB)

            Total      Operations              Average             Minimum             Maximum  Name
-----------------  --------------  -----------------  -----------------  -----------------  ---------------------------------------------------
---------------
        1722500.0               2           861250.0         722500.000           1000000.0  [CUDA memcpy DtoH]
         538919.0               4           134729.0              0.766            288906.0  [CUDA memcpy HtoD]
Generating Operating System Runtime API Statistics...
Operating System Runtime API Statistics (nanoseconds)
```

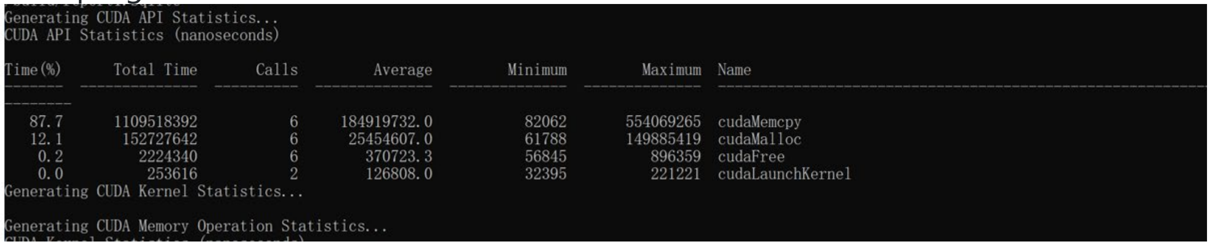
Deliverables

Report: Include a list of all kernels that collectively consume more than 90% of the program time.



```
Generating CUDA Memory Operation Statistics...
CUDA Kernel Statistics (nanoseconds)

Time(%)    Total Time    Instances      Average         Minimum        Maximum    Name
-------  -------------  ----------  -------------  -------------  -------------  ----------------------------------------------------------------
--------
  100.0     116856830           2     58428415.0        23509740       93347090   conv_forward_kernel
```

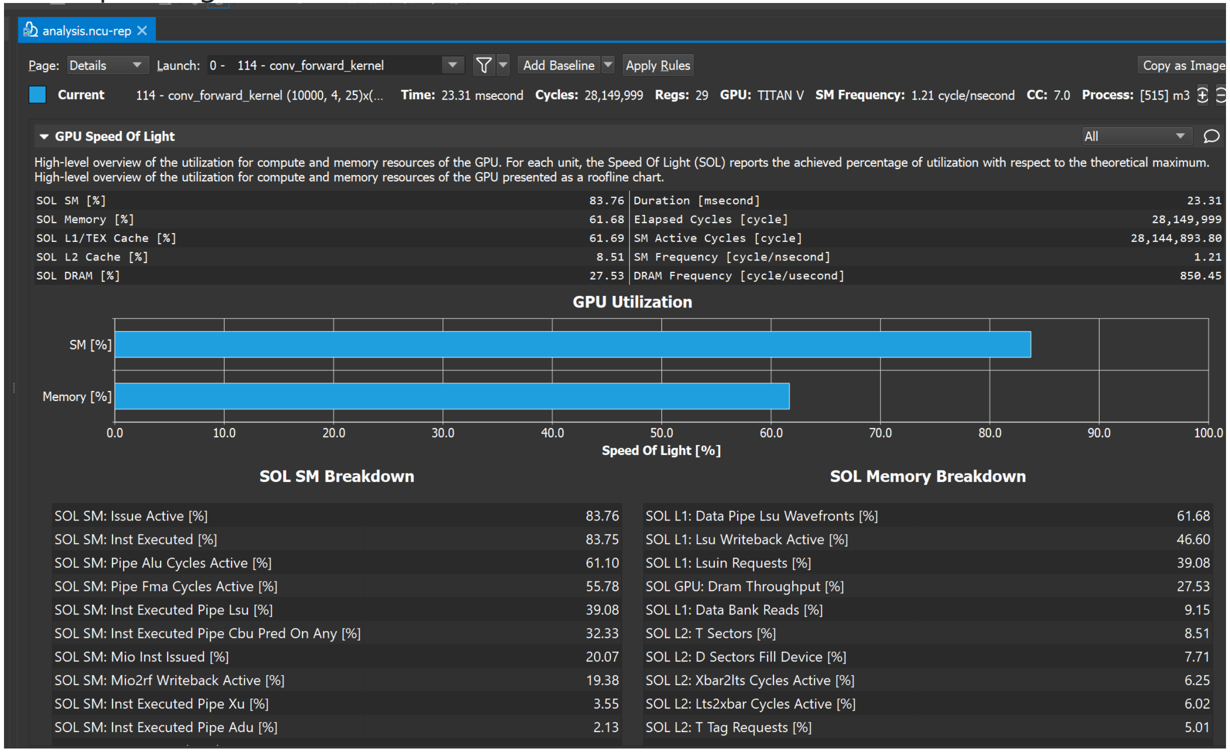
Report: Include a list of all CUDA API calls that collectively consume more than 90% of the program time.



```
Generating CUDA API Statistics...
CUDA API Statistics (nanoseconds)

Time(%)    Total Time      Calls        Average         Minimum        Maximum    Name
-------  -------------  ----------  -------------  -------------  -------------  ----------------------------------------------------------------
--------
   87.7    1109518392           6    184919732.0          82062      554069265   cudaMemcpy
   12.1     152727642           6     25454607.0          61788      149885419   cudaMalloc
    0.2       2224340           6       370723.3          56845         896359   cudaFree
    0.0        253616           2       126808.0          32395         221221   cudaLaunchKernel
Generating CUDA Kernel Statistics...

Generating CUDA Memory Operation Statistics...
```

Report: Include an explanation of the difference between kernels and API calls
API calls are provided and kernels are written by us.
API calls are executed sequentially and kernels are executed parallelly.

Report: Screenshot of the GPU SOL utilization in Nsight-Compute GUI for your kernel profiling data



Current 114 - conv_forward_kernel (10000, 4, 25)x(... **Time:** 23.31 msecond **Cycles:** 28,149,999 **Regs:** 29 **GPU:** TITAN V **SM Frequency:** 1.21 cycle/nsecond **CC:** 7.0 **Process:** [515] m3

GPU Speed Of Light

High-level overview of the utilization for compute and memory resources of the GPU. For each unit, the Speed Of Light (SOL) reports the achieved percentage of utilization with respect to the theoretical maximum.
High-level overview of the utilization for compute and memory resources of the GPU presented as a roofline chart.

SOL SM [%]	83.76	Duration [msecond]	23.31
SOL Memory [%]	61.68	Elapsed Cycles [cycle]	28,149,999
SOL L1/TEX Cache [%]	61.69	SM Active Cycles [cycle]	28,144,893.80
SOL L2 Cache [%]	8.51	SM Frequency [cycle/nsecond]	1.21
SOL DRAM [%]	27.53	DRAM Frequency [cycle/usecond]	850.45

SOL SM Breakdown		SOL Memory Breakdown	
SOL SM: Issue Active [%]	83.76	SOL L1: Data Pipe Lsu Wavefronts [%]	61.68
SOL SM: Inst Executed [%]	83.75	SOL L1: Lsu Writeback Active [%]	46.60
SOL SM: Pipe Alu Cycles Active [%]	61.10	SOL L1: Lsuin Requests [%]	39.08
SOL SM: Pipe Fma Cycles Active [%]	55.78	SOL GPU: Dram Throughput [%]	27.53
SOL SM: Inst Executed Pipe Lsu [%]	39.08	SOL L1: Data Bank Reads [%]	9.15
SOL SM: Inst Executed Pipe Cbu Pred On Any [%]	32.33	SOL L2: T Sectors [%]	8.51
SOL SM: Mio Inst Issued [%]	20.07	SOL L2: D Sectors Fill Device [%]	7.71
SOL SM: Mio2rf Writeback Active [%]	19.38	SOL L2: Xbar2lts Cycles Active [%]	6.25
SOL SM: Inst Executed Pipe Xu [%]	3.55	SOL L2: Lts2xbar Cycles Active [%]	6.02
SOL SM: Inst Executed Pipe Adu [%]	2.13	SOL L2: T Tag Requests [%]	5.01

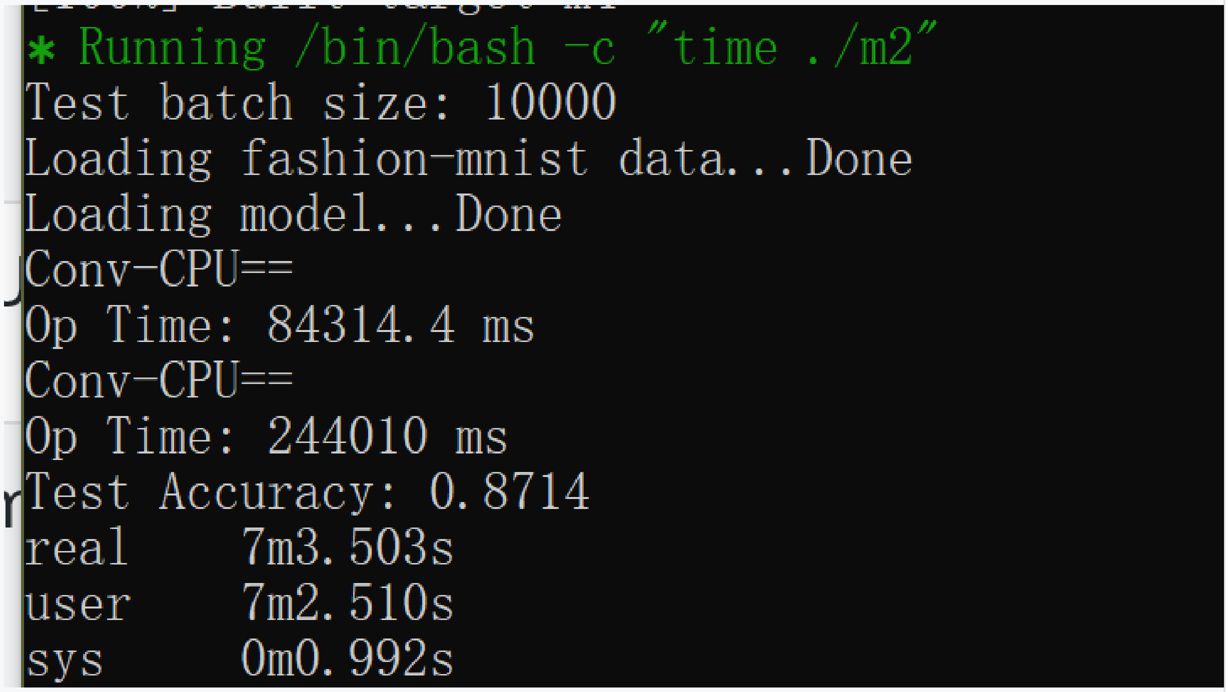
Deliverables

Use `rai -p <project folder> --queue rai_amd64_ece408 --submit=m3` to mark your job for grading

Milestone 2

Report: Show output of rai running Mini-DNN on the CPU (CPU convolution implemented) for batch size of 10k images

```
✱ Running /bin/bash -c "time ./m2"

Test batch size: 10000

Loading fashion-mnist data...Done

Loading model...Done

Conv-CPU==

Op Time: 84314.4 ms

Conv-CPU==

Op Time: 244010 ms

Test Accuracy: 0.8714

real    7m3.503s

user    7m2.510s

sys     0m0.992s
```



```
* Running /bin/bash -c "time ./m2"
Test batch size: 10000
Loading fashion-mnist data...Done
Loading model...Done
Conv-CPU==
Op Time: 84314.4 ms
Conv-CPU==
Op Time: 244010 ms
Test Accuracy: 0.8714
real    7m3.503s
user    7m2.510s
sys     0m0.992s
```

Report: List Op Times (CPU convolution implemented) for batch size of 10k images

Op Times:

Conv-CPU==

Op Time: 84314.4 ms

Conv-CPU==

Op Time: 244010 ms

Report: List whole program execution time (CPU convolution implemented) for batch size of 10k images

Whole execution time: user+sys=7m2.510s+0m0.992s=7m3.502s