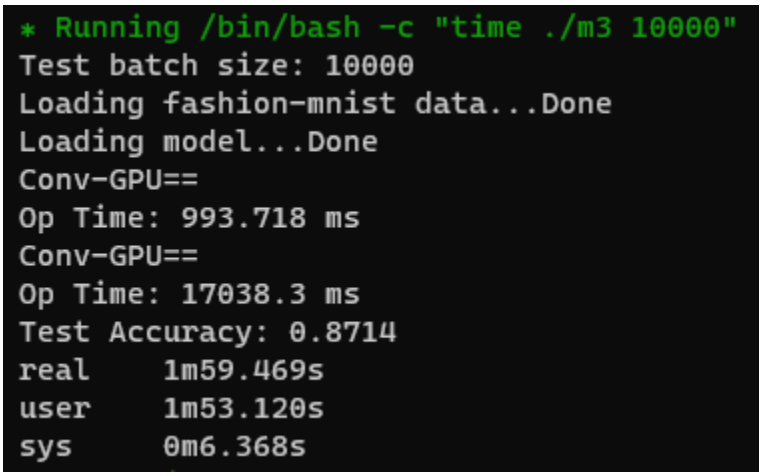
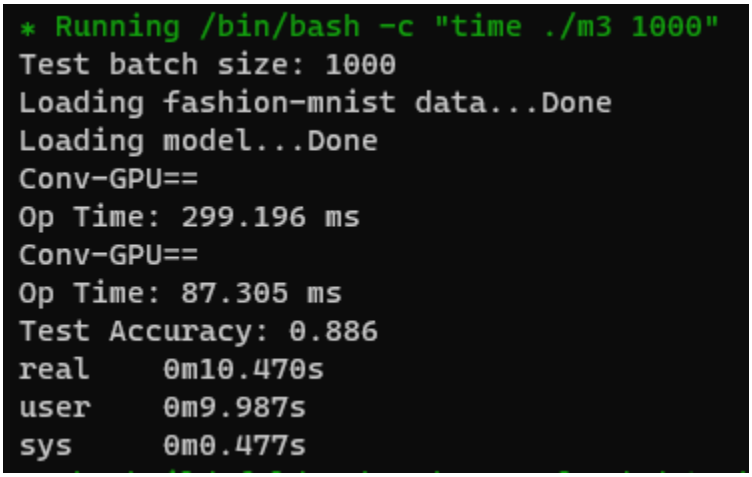
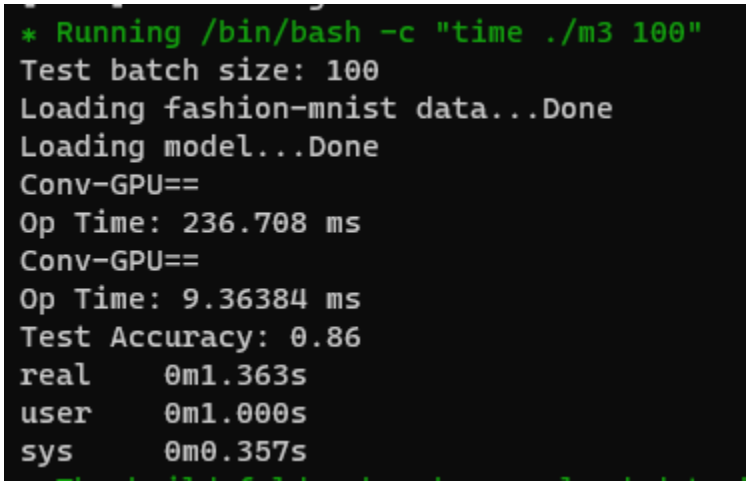
Jimmy Zhang
ECE 408
PM3

Timing on three different datasets

Runtimes for batch sizes of 10000,1000,100 respectively:







The runtimes increase roughly linearly to batch size.

Nsys output

using batch size of 10000:

```
Time(%)      Total Time       Calls         Average         Minimum         Maximum  Name
-------  --------------  ----------  --------------  --------------  --------------  ------------------------------
   91.7      1751147021           8     218893377.6           51995       644411262  cudaMemcpy
    8.1       155373157           6      25895526.2           69284       152583963  cudaMalloc
    0.1         2289163           6        381527.2           59995          943682  cudaFree
    0.0          255003           2        127501.5           32593          222410  cudaLaunchKernel
Generating CUDA Kernel Statistics...

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

Time(%)      Total Time   Instances         Average         Minimum         Maximum  Name
-------  --------------  ----------  --------------  --------------  --------------  ------------------------------
  100.0       270702489           2     135351244.5        49246769       221455720  conv_forward_kernel


CUDA Memory Operation Statistics (nanoseconds)

Time(%)      Total Time  Operations         Average         Minimum         Maximum  Name
-------  --------------  ----------  --------------  --------------  --------------  ------------------------------
   68.0      1004867286           2     502433643.0       422260075       582607211  [CUDA memcpy DtoH]
   32.0       472437472           6      78739578.7            1216       231922240  [CUDA memcpy HtoD]


CUDA Memory Operation Statistics (KiB)

            Total      Operations            Average            Minimum            Maximum  Name
-----------------  --------------  -----------------  -----------------  -----------------  ------------------------
        2261419.0               6           376903.0              0.766          1000000.0  [CUDA memcpy HtoD]
        1722500.0               2           861250.0         722500.000          1000000.0  [CUDA memcpy DtoH]
Generating Operating System Runtime API Statistics...
Operating System Runtime API Statistics (nanoseconds)
```

```
Time(%)      Total Time       Calls         Average         Minimum         Maximum  Name
-------  --------------  ----------  --------------  --------------  --------------  --------------------------------
   33.3     98768928981        1001      98670258.7           28444       100197614  sem_timedwait
   33.3     98676362545        1000      98676362.5           41797       100410637  poll
   22.3     66124725212           2   33062362606.0     24185976285     41938748927  pthread_cond_wait
   11.0     32507714645          65     500118686.8       500065007       500182092  pthread_cond_timedwait
    0.1       186652509         856        218052.0            1005       110824427  ioctl
    0.0        17285399        9072          1905.4            1239           18262  read
    0.0         3015576          98         30771.2            1004         1364548  mmap
    0.0         1073960         101         10633.3            4474           24503  open64
    0.0          287723           5         57544.6           39719           78514  pthread_create
    0.0           82582          18          4587.9            1267           20296  munmap
    0.0           78362          15          5224.1            2382           10710  write
    0.0           71152          24          2964.7            1063            9220  fopen
    0.0           68481           3         22827.0            7809           52632  fgets
    0.0           52153           3         17384.3            3232           29038  fopen64
    0.0           46750           7          6678.6            3327            8875  fflush
    0.0           30429           5          6085.8            4432            7126  open
    0.0           20073           3          6691.0            1166            9940  fcntl
    0.0           19574           3          6524.7            6007            6992  pipe2
    0.0           18481           9          2053.4            1140            6668  fclose
    0.0           14121           2          7060.5            6815            7306  socket
    0.0           13638           3          4546.0            1394            7856  fwrite
    0.0            7554           2          3777.0            3038            4516  pthread_cond_signal
0.0             6808           1          6808.0            6808            6808  connect
    0.0             1707           1          1707.0            1707            1707  bind
Generating NVTX Push-Pop Range Statistics...
NVTX Push-Pop Range Statistics (nanoseconds)
```

cudaMalloc and cudaMemcpy count for 87%, 12.9% of time consumed by API calls, and in total 99.9% of the time.

The conv_forward_kernel takes 100% of the runtime as expected.

API calls are call to functions in the built in cuda framework; they are called by the host and performed on the device. The kernel however is defined by the programmer and is not strictly part of the CUDA api, although it may contain calls to the CUDA API.

Nsight-Compute output

batch size of 10000: