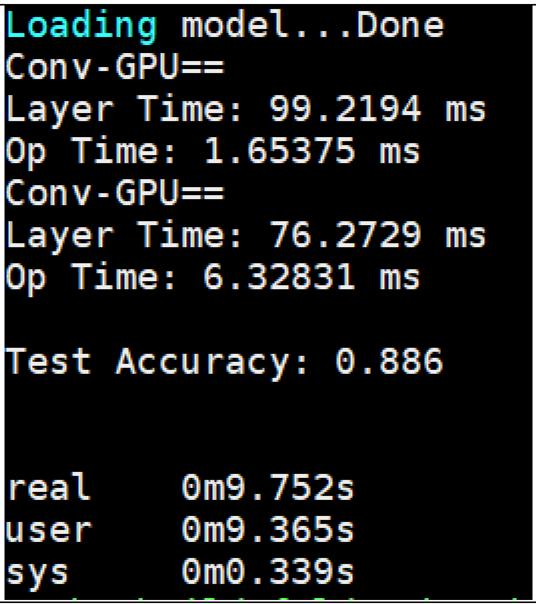
Name: *Ding-Hsiang Huang*
NetID: *dhhuang3*
Section: *AL2*

ECE 408/CS483 Milestone 2 Report

1. Show output of rai running Mini-DNN on the basic GPU convolution implementation for batch size of 1k images. This can either be a screen capture or a text copy of the running output. Please do not show the build output. (The running output should be everything including and after the line "*Loading fashion-mnist data...Done*").



```
Loading model...Done
Conv-GPU==
Layer Time: 99.2194 ms
Op Time: 1.65375 ms
Conv-GPU==
Layer Time: 76.2729 ms
Op Time: 6.32831 ms

Test Accuracy: 0.886


real    0m9.752s
user    0m9.365s
sys     0m0.339s
```

2. For the basic GPU implementation, list Op Times, whole program execution time, and accuracy for batch size of 100, 1k, and 10k images.

Batch Size	Op Time 1	Op Time 2	Total Execution Time	Accuracy
100	*0.176167ms*	*0.637506 ms*	*0m1.245s*	*0.86*
1000	*1.65375ms*	*6.32831ms*	*9.752s*	*0.886*
10000	*16.1505 ms*	*63.687 ms*	*1m35.159s*	*0.8714*

3. List all the kernels that collectively consumed more than 90% of the kernel time and what percentage of the kernel time each kernel did consume (start with the kernel that consumed the most time, then list the next kernel, until you reach 90% or more).

conv_forward_kernel 100.0%

4. List all the CUDA API calls that collectively consumed more than 90% of the API time and what percentage of the API time each call did consume (start with the API call that consumed the most time, then list the next call, until you reach 90% or more).

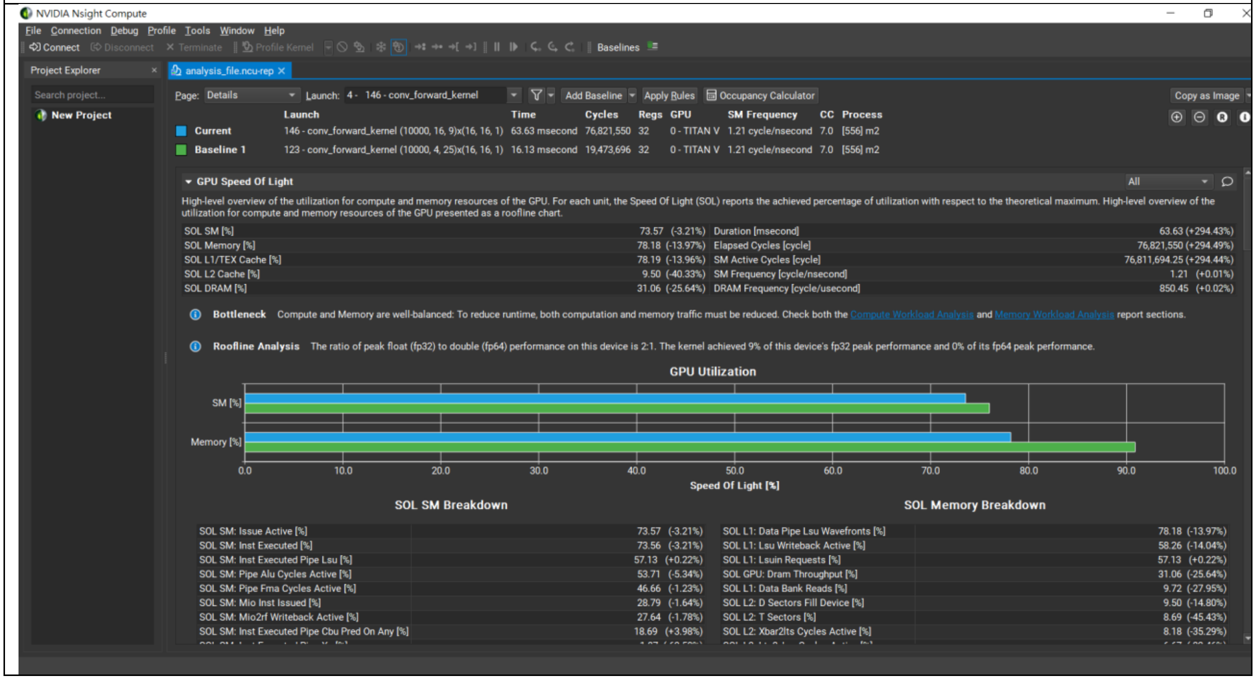
cudaMemcpy 77.4%
cudaMalloc 17.8%

5. Explain the difference between kernels and CUDA API calls. Please give an example in your explanation for both.

Kernels are code that executed in device(kernel) like" conv_forward_kernel", which contained the calculation that executed in parallel.

CUDA API calls are those code that calling between host and device. CudaMemcpy, cudaMalloc, cudaDeviceSynchronize… all of them are calls made by the code into the CUDA driver.

6. Show a screenshot of the GPU SOL utilization



Page: Details Launch: 4 - 146 - conv_forward_kernel

	Launch	Time	Cycles	Regs	GPU	SM Frequency	CC	Process
Current	146 - conv_forward_kernel (10000, 16, 9)x(16, 16, 1)	63.63 msecond	76,821,550	32	0 - TITAN V	1.21 cycle/nsecond	7.0	[556] m2
Baseline 1	123 - conv_forward_kernel (10000, 4, 25)x(16, 16, 1)	16.13 msecond	19,473,696	32	0 - TITAN V	1.21 cycle/nsecond	7.0	[556] m2

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 [%]	73.57 (-3.21%)	Duration [msecond]	63.63 (+294.43%)
SOL Memory [%]	78.18 (-13.97%)	Elapsed Cycles [cycle]	76,821,550 (+294.49%)
SOL L1/TEX Cache [%]	78.19 (-13.96%)	SM Active Cycles [cycle]	76,811,694.25 (+294.44%)
SOL L2 Cache [%]	9.50 (-40.33%)	SM Frequency [cycle/nsecond]	1.21 (+0.01%)
SOL DRAM [%]	31.06 (-25.64%)	DRAM Frequency [cycle/usecond]	850.45 (+0.02%)

Bottleneck Compute and Memory are well-balanced: To reduce runtime, both computation and memory traffic must be reduced. Check both the Compute Workload Analysis and Memory Workload Analysis report sections.

Roofline Analysis The ratio of peak float (fp32) to double (fp64) performance on this device is 2:1. The kernel achieved 9% of this device's fp32 peak performance and 0% of its fp64 peak performance.

SOL SM Breakdown

SOL SM: Issue Active [%]	73.57 (-3.21%)
SOL SM: Inst Executed [%]	73.56 (-3.21%)
SOL SM: Inst Executed Pipe Lsu [%]	57.13 (+0.22%)
SOL SM: Pipe Alu Cycles Active [%]	53.71 (-5.34%)
SOL SM: Pipe Fma Cycles Active [%]	46.66 (-1.23%)
SOL SM: Mio Inst Issued [%]	28.79 (-1.64%)
SOL SM: Mio2rf Writeback Active [%]	27.64 (-1.78%)
SOL SM: Inst Executed Pipe Cbu Pred On Any [%]	18.69 (+3.98%)

SOL Memory Breakdown

SOL L1: Data Pipe Lsu Wavefronts [%]	78.18 (-13.97%)
SOL L1: Lsu Writeback Active [%]	58.26 (-14.04%)
SOL L1: Lsuin Requests [%]	57.13 (+0.22%)
SOL GPU: Dram Throughput [%]	31.06 (-25.64%)
SOL L1: Data Bank Reads [%]	9.72 (-27.95%)
SOL L2: D Sectors Fill Device [%]	9.50 (-14.80%)
SOL L2: T Sectors [%]	8.69 (-45.43%)
SOL L2: Xbar2lts Cycles Active [%]	8.18 (-35.29%)