Name: Yiming Li
NetID: yiming22
Section: AL1

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").

```
<output here>
Loading fashion-mnist data...Done
Loading model...Done
Conv-GPU==
Layer Time: 68.8763 ms
Op Time: 1.63944 ms
Conv-GPU==
Layer Time: 51.8532 ms
Op Time: 6.2645 ms

Test Accuracy: 0.886


real    0m9.660s
user    0m9.269s
sys     0m0.328s
```

*The build folder has been uploaded to http://s3.amazonaws.com/files.rai-project.com/userdata/build-6184c17df5b88942145451fb.tar.gz. The data will be present for only a short duration of time.

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	0.175098 ms	0.632573 ms	0m1.190s	0.86
1000	1.63944 ms	6.2645 ms	0m9.660s	0.886
10000	16.0852 ms	62.7071 ms	1m37.766s	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).

<answer here>
conv_forward_kernel (100.0% Time)

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).

<answer here>
cudaMemcpy (76.6% Time)
cudaMalloc (15.9% Time)

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

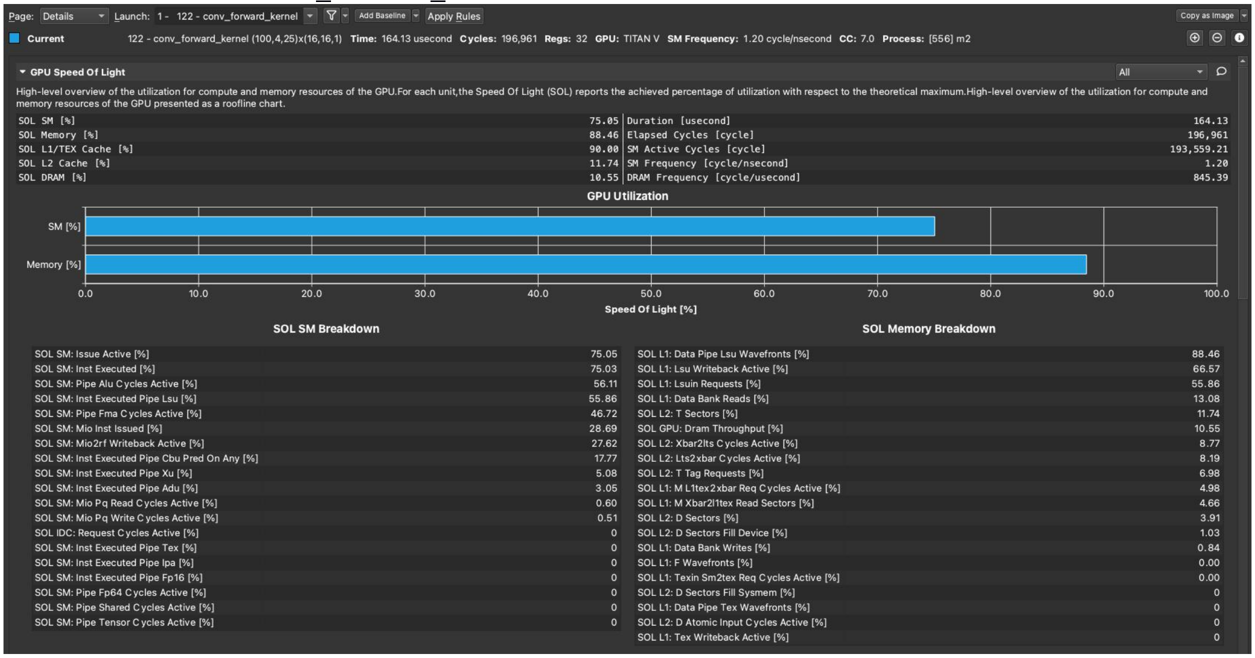
<answer here>
Kernels are C++ functions that are executed N times in parallel by N different CUDA threads when called. A kernel is defined using the __global__ declaration specifier.
Example is the kernel "conv_forward_kernel()" in this milestone2 code.

However, API calls are calls made by the code into the CUDA driver or runtime libraries. It is only executed once like regular C++ functions.
Examples are "cudaMalloc()", "cudaMemcpy()", etc.

6. Show a screenshot of the GPU SOL utilization

<nsight output here>

Launch: 1 -   122 - conv_forward_kernel



Current 122 - conv_forward_kernel (100,4,25)x(16,16,1) **Time:** 164.13 usecond **Cycles:** 196,961 **Regs:** 32 **GPU:** TITAN V **SM Frequency:** 1.20 cycle/nsecond **CC:** 7.0 **Process:** [556] m2

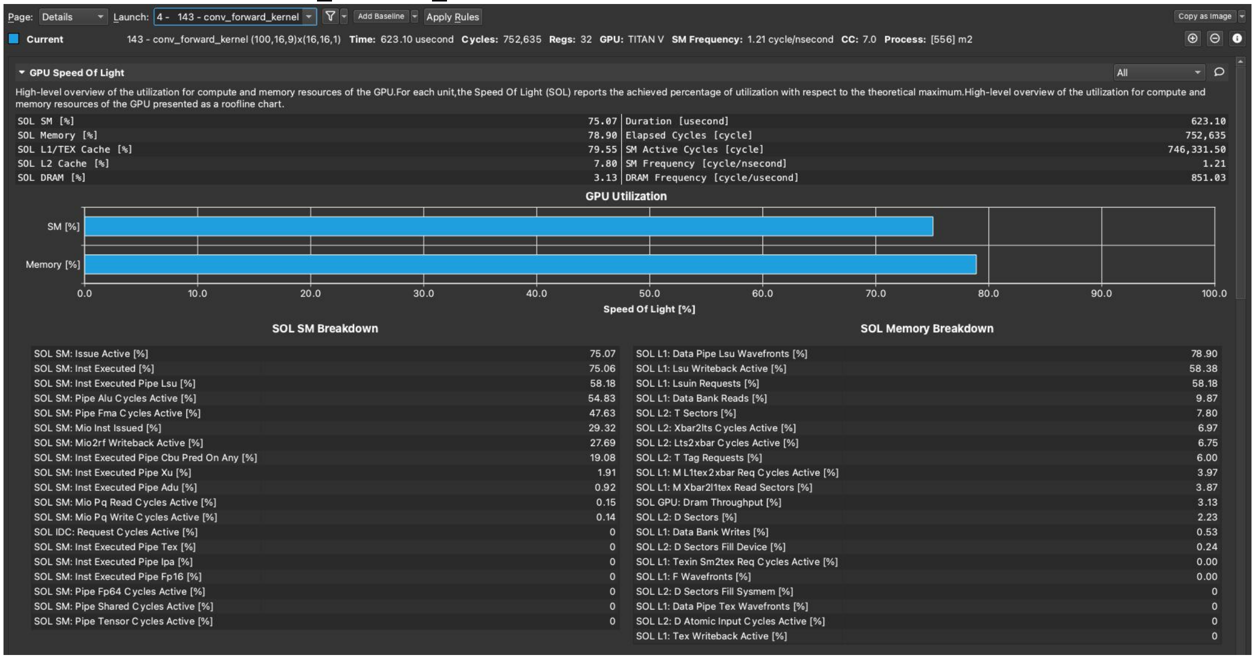
GPU Speed Of Light

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 [%]	75.05	Duration [usecond]	164.13
SOL Memory [%]	88.46	Elapsed Cycles [cycle]	196,961
SOL L1/TEX Cache [%]	90.00	SM Active Cycles [cycle]	193,559.21
SOL L2 Cache [%]	11.74	SM Frequency [cycle/nsecond]	1.20
SOL DRAM [%]	10.55	DRAM Frequency [cycle/usecond]	845.39

SOL SM Breakdown		SOL Memory Breakdown	
SOL SM: Issue Active [%]	75.05	SOL L1: Data Pipe Lsu Wavefronts [%]	88.46
SOL SM: Inst Executed [%]	75.03	SOL L1: Lsu Writeback Active [%]	66.57
SOL SM: Pipe Alu Cycles Active [%]	56.11	SOL L1: Lsuin Requests [%]	55.86
SOL SM: Inst Executed Pipe Lsu [%]	55.86	SOL L1: Data Bank Reads [%]	13.08
SOL SM: Pipe Fma Cycles Active [%]	46.72	SOL L2: T Sectors [%]	11.74
SOL SM: Mio Inst Issued [%]	28.69	SOL GPU: Dram Throughput [%]	10.55
SOL SM: Mio2rf Writeback Active [%]	27.62	SOL L2: Xbar2lts Cycles Active [%]	8.77
SOL SM: Inst Executed Pipe Cbu Pred On Any [%]	17.77	SOL L2: Lts2xbar Cycles Active [%]	8.19
SOL SM: Inst Executed Pipe Xu [%]	5.08	SOL L2: T Tag Requests [%]	6.98
SOL SM: Inst Executed Pipe Adu [%]	3.05	SOL L1: M L1tex2xbar Req Cycles Active [%]	4.98
SOL SM: Mio Pq Read Cycles Active [%]	0.60	SOL L1: M Xbar2l1tex Read Sectors [%]	4.66
SOL SM: Mio Pq Write Cycles Active [%]	0.51	SOL L2: D Sectors [%]	3.91
SOL IDC: Request Cycles Active [%]	0	SOL L2: D Sectors Fill Device [%]	1.03
SOL SM: Inst Executed Pipe Tex [%]	0	SOL L1: Data Bank Writes [%]	0.84
SOL SM: Inst Executed Pipe Ipa [%]	0	SOL L1: F Wavefronts [%]	0.00
SOL SM: Inst Executed Pipe Fp16 [%]	0	SOL L1: Texin Sm2tex Req Cycles Active [%]	0.00
SOL SM: Pipe Fp64 Cycles Active [%]	0	SOL L2: D Sectors Fill Sysmem [%]	0
SOL SM: Pipe Shared Cycles Active [%]	0	SOL L1: Data Pipe Tex Wavefronts [%]	0
SOL SM: Pipe Tensor Cycles Active [%]	0	SOL L2: D Atomic Input Cycles Active [%]	0
		SOL L1: Tex Writeback Active [%]	0

Launch: 4 -   143 - conv_forward_kernel



Current 143 - conv_forward_kernel (100,16,9)x(16,16,1) **Time:** 623.10 usecond **Cycles:** 752,635 **Regs:** 32 **GPU:** TITAN V **SM Frequency:** 1.21 cycle/nsecond **CC:** 7.0 **Process:** [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 [%]	75.07	Duration [usecond]	623.10
SOL Memory [%]	78.90	Elapsed Cycles [cycle]	752,635
SOL L1/TEX Cache [%]	79.55	SM Active Cycles [cycle]	746,331.50
SOL L2 Cache [%]	7.80	SM Frequency [cycle/nsecond]	1.21
SOL DRAM [%]	3.13	DRAM Frequency [cycle/usecond]	851.03

SOL SM Breakdown		SOL Memory Breakdown	
SOL SM: Issue Active [%]	75.07	SOL L1: Data Pipe Lsu Wavefronts [%]	78.90
SOL SM: Inst Executed [%]	75.06	SOL L1: Lsu Writeback Active [%]	58.38
SOL SM: Inst Executed Pipe Lsu [%]	58.18	SOL L1: Lsuin Requests [%]	58.18
SOL SM: Pipe Alu Cycles Active [%]	54.83	SOL L1: Data Bank Reads [%]	9.87
SOL SM: Pipe Fma Cycles Active [%]	47.63	SOL L2: T Sectors [%]	7.80
SOL SM: Mio Inst Issued [%]	29.32	SOL L2: Xbar2lts Cycles Active [%]	6.97
SOL SM: Mio2rf Writeback Active [%]	27.69	SOL L2: Lts2xbar Cycles Active [%]	6.75
SOL SM: Inst Executed Pipe Cbu Pred On Any [%]	19.08	SOL L2: T Tag Requests [%]	6.00
SOL SM: Inst Executed Pipe Xu [%]	1.91	SOL L1: M L1tex2xbar Req Cycles Active [%]	3.97
SOL SM: Inst Executed Pipe Adu [%]	0.92	SOL L1: M Xbar2l1tex Read Sectors [%]	3.87
SOL SM: Mio Pq Read Cycles Active [%]	0.15	SOL GPU: Dram Throughput [%]	3.13
SOL SM: Mio Pq Write Cycles Active [%]	0.14	SOL L2: D Sectors [%]	2.23
SOL IDC: Request Cycles Active [%]	0	SOL L1: Data Bank Writes [%]	0.53
SOL SM: Inst Executed Pipe Tex [%]	0	SOL L2: D Sectors Fill Device [%]	0.24
SOL SM: Inst Executed Pipe Ipa [%]	0	SOL L1: Texin Sm2tex Req Cycles Active [%]	0.00
SOL SM: Inst Executed Pipe Fp16 [%]	0	SOL L1: F Wavefronts [%]	0.00
SOL SM: Pipe Fp64 Cycles Active [%]	0	SOL L2: D Sectors Fill Sysmem [%]	0
SOL SM: Pipe Shared Cycles Active [%]	0	SOL L1: Data Pipe Tex Wavefronts [%]	0
SOL SM: Pipe Tensor Cycles Active [%]	0	SOL L2: D Atomic Input Cycles Active [%]	0
		SOL L1: Tex Writeback Active [%]	0