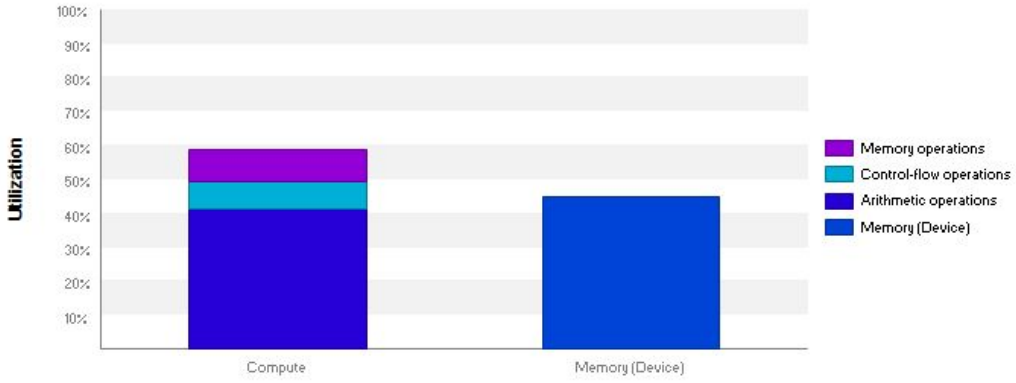
Team: Team
Bryan Lu (bryanlu2)
Omar Mbarki (ombarki2)
UIUC On-campus

Milestone 3

Timeline.nvprof

Kernel Performance Is Bound By Instruction And Memory Latency

This kernel exhibits low compute throughput and memory bandwidth utilization relative to the peak performance of "TITAN V". These utilization levels indicate that the performance of the kernel is most likely limited by the latency of arithmetic or memory operations. Achieved compute throughput and/or memory bandwidth below 60% of peak typically indicates latency issues.



It's no surprise that our very basic implementation of the kernel is bounded by memory latency. Using shared memory would solve the global memory bandwidth issue and would be much more efficient in terms of throughput.

Furthermore, we suffer 16.5% control divergence because of the if statement in line 33. There are ways to mitigate this issue, but we will focus our efforts into eliminating the memory bandwidth issue by using shared/constant memory and later down the line employing the unrolling technique to perform matrix multiplications instead of convolutions.

Milestone 2

Kernels that take up more than 90% of program time (m1.2.py):

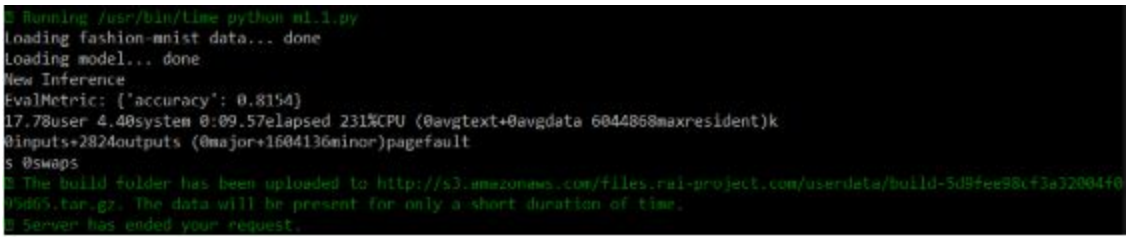
- CUDA memcpy HtoD
- volta_scudnn_128x64_relu_interior_nn_v1
- volta_gcgemm_64x32_nt
- fft2d_c2r_32x32
- volta_sgemm_128x128_tn
- op_generic_tensor_kernel
- fft2d_r2c_32x32

API calls that take up more than 90% of program time (m1.2.py):

- cudaStreamCreateWithFlags
- cudaMemGetInfo
- cudaFree

CUDA kernels are functions that are run by each individual thread that CUDA creates. The API calls are functions that run on the program host that manage the information needed by the CUDA threads to run (like memory allocations).

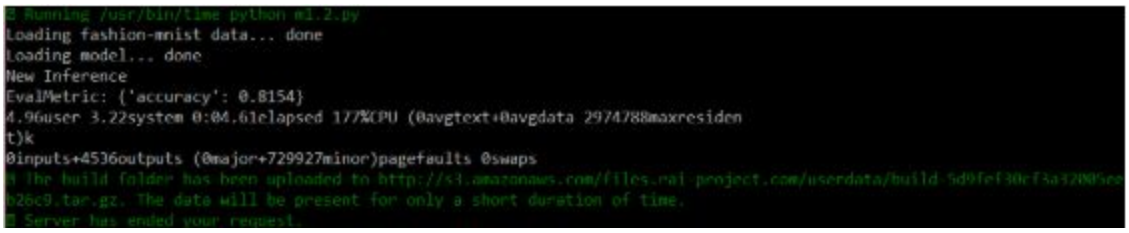
Output of rai running MXNet on CPU with program times:



CPU program run times:

- User: 17.78
- System: 4.40
- Elapsed: 9.57

Output of rai running MXNet on GPU with program times:



GPU program run times:

- User: 4.96
- System: 3.22
- Elapsed: 4.61

Output of CPU implementation:

```
* Running /usr/bin/time python m2.1.py
Loading fashion-mnist data... done
Loading model... done
New Inference
Op Time: 11.664715
Op Time: 60.728774
Correctness: 0.7653 Model: ece408
87.21user 10.91system 1:16.48elapsed 128%CPU
```

CPU implementation program times:

- User: 87.21
- System: 10.91
- Elapsed: 1:16.48

CPU implementation op times:

- 1: 11.664715
- 2: 60.728774