CS 483 Project Report

Team_name: hawks
Names: Nikhil Paidipally, Daniel Niewierowski, Nicolas Nytko
Net_IDs: npaidi2, nnytko2, dniewi2
RAI_IDs: 5d97b1f588a5ec28f9cb94b0, 5d97b1f488a5ec28f9cb94ae, 5d97b1f388a5ec28f9cb94ac
Affiliation: uiuc

MILESTONE 1:

Kernels that collectively consume more than 90% of the program time:

- [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

CUDA calls that collectively consume more than 90% of the program time:

- cudaStreamCreateWithFlags
- cudaMemGetInfo
- cudaFree

Difference between kernels and API calls:

CUDA kernels are effectively like regular C functions that are executed an arbitrary number of times in parallel by an arbitrary number of different CUDA threads on the gpu/device.

CUDA API are effectively regular C functions that execute on the cpu/host and they are not executed in parallel like the CUDA kernels. Some examples of CUPA APIs include cudamalloc, cudaMemCpy, cudafree, etc.

Rai running MXNet on the CPU:

(CPU C Code)

```
Loading fashion-mnist data... done
Loading model... done
New Inference
```

```
Op Time: 11.330177
Op Time: 74.542234
Correctness: 0.7653 Model: ece408
100.71user 11.16system 1:29.81elapsed 124%CPU (0avgtext+0avgdata
6044220maxresident)k
0inpu
ts+0outputs (0major+2308767minor)pagefaults 0swaps
```

(CPU Python Code)

```
Loading fashion-mnist data... done
Loading model... done
New Inference
EvalMetric: {'accuracy': 0.8154}

17.90user 4.36system 0:09.78elapsed 227%CPU (0avgtext+0avgdata
6045072maxresident)k
0inputs+2824outputs (
0major+1603748minor)pagefaults 0swaps
```

Rai running MXNet on the GPU:

```
Loading fashion-mnist data... done
Loading model... done
New Inference
EvalMetric: {'accuracy': 0.8154}

4.96user 3.07system 0:04.63elapsed 173%CPU (0avgtext+0av
gdata 2996488maxresident)k
0inputs+1712outputs (0major+734016minor)pagefaults 0swaps
```