Names: Brian Tang (bt3), Joseph Adamo (jdadamo2), Kyle Jew (kjew2)

Team Name: thread_beast

Affiliation: On-Campus students

List of kernels consuming more than 90% of program time:

None, highest is [CUDA memcpy HtoD] at 30.04% of time.

List of CUDA API calls consuming more than 90% of program time:

None, highest is cudaStreamCreateWithFlags at 41.19% of time.

Kernels and API calls difference:

Kernels are programmer-defined C functions and when launched, are executed N times in parallel by N different threads. However, CUDA API calls are pre-defined extensions to the C language and meant for easing the experience for programmers to set up programs for execution by the device.

Output of rai running MXNET on CPU:

```
“Loading fashion-mnist data... done

Loading model... done

New Inference

EvalMetric: {'accuracy': 0.8154}

18.26user 4.46system 0:09.56elapsed 237%CPU (0avgtext+0avgdata 6047060maxresident)k

0inputs+2824outputs (0major+1601873minor)pagefaults 0swaps”
```

Program run time: 9.56 seconds

Output of rai running MXNET on GPU:

```
“Loading fashion-mnist data... done

Loading model... done

New Inference

EvalMetric: {'accuracy': 0.8154}

4.97user 3.25system 0:04.59elapsed 179%CPU (0avgtext+0avgdata 2968484maxresident)k

0inputs+4536outputs (0major+733238minor)pagefaults 0swaps”
```

Program run time: 4.59 seconds

Whole Program Execution Time: 1 minute 16.47 seconds

Op Times:

Python m2.1.py:

```
        Op Time: 12.307846
        Op Time: 59.309954
        Correctness: 0.7653
```

At 100 images:

```
        Op Time: 1.082469
        Op Time: 5.923644
        Correctness: 0.767
```

At 1000 images:

```
        Op Time: 0.108870
        Op Time: 0.590093
        Correctness: 0.76
```

At 10000 images:

```
        Op Time: 10.855807
        Op Time: 60.478481
        Correctness: 0.7653
```

Milestone 3:

Correctness and Timing at 100 images:

```
Op Time: 0.000282
Op Time: 0.000924
Correctness: 0.76 Model: ece408
4.84user 2.65system 0:06.72elapsed 111%CPU (0avgtext+0avgdata 2783704maxresi
dent)k
0inputs+4560outputs (0major+636682minor)pagefaults 0swaps
```

At 1000 images:

Op Time: 0.002764

Op Time: 0.009408

Correctness: 0.767 Model: ece408

4.82user 2.77system 0:04.38elapsed 173%CPU (0avgtex

t+0avgdata 2811072maxresident)k

0inputs+4560outputs (0major+641440minor)pagefaults 0swaps

At 10000 images:

Op Time: 0.027439

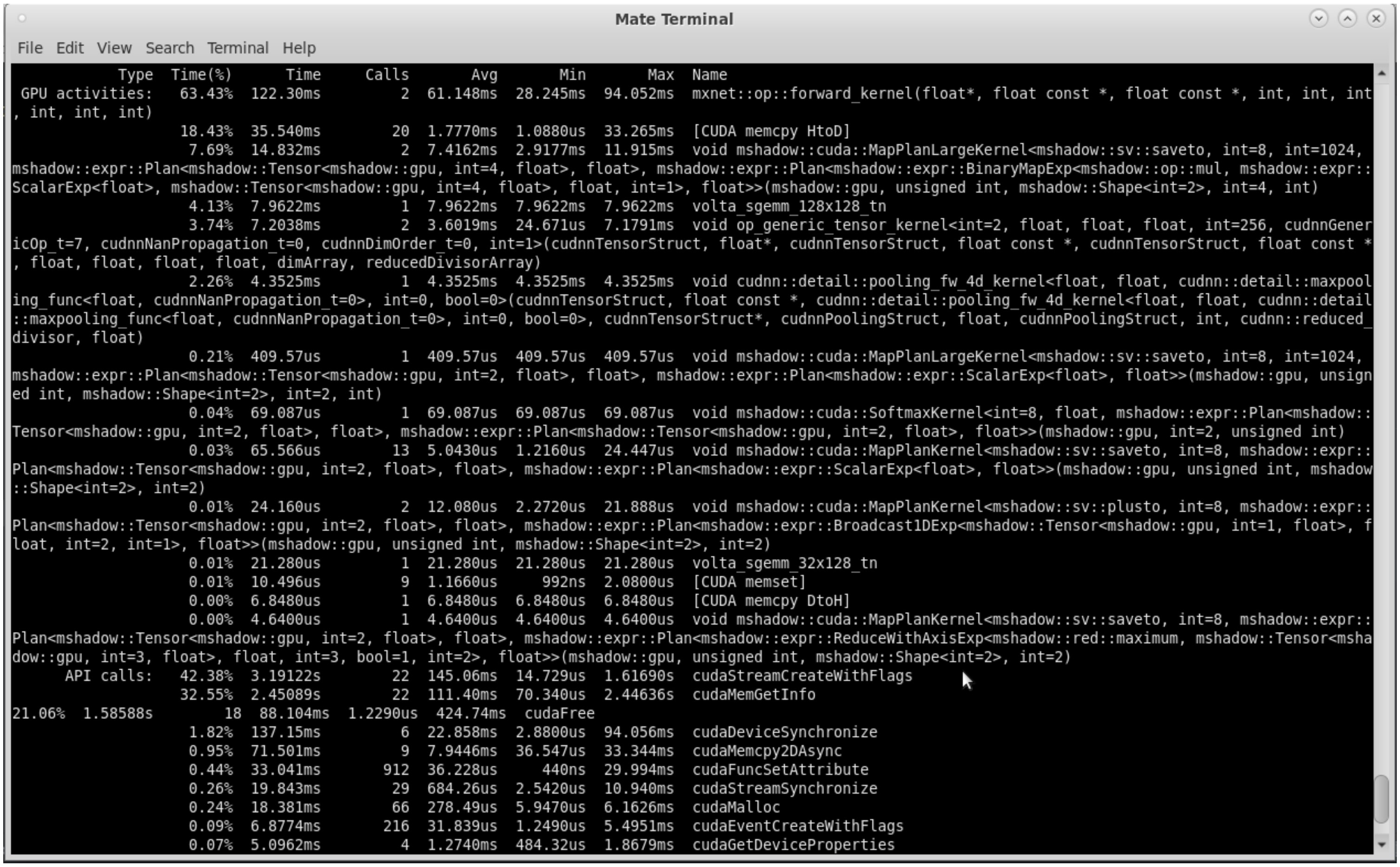
Op Time: 0.093477

Correctness: 0.7653 Model: ece408

5.19user 3.16system 0:04.87elapsed 171%CPU (0avgtext+0avgdata 2981280maxresident)k

0inputs+4560outputs (0major+734975minor)pagefaults 0swaps

NVPROF Execution:



```
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:   63.43%  122.30ms         2  61.148ms  28.245ms  94.052ms  mxnet::op::forward_kernel(float*, float const *, float const *, int, int, int
, int, int, int)
                   18.43%  35.540ms        20  1.7770ms  1.0880us  33.265ms  [CUDA memcpy HtoD]
                    7.69%  14.832ms         2  7.4162ms  2.9177ms  11.915ms  void mshadow::cuda::MapPlanLargeKernel<mshadow::sv::saveto, int=8, int=1024,
mshadow::expr::Plan<mshadow::Tensor<mshadow::gpu, int=4, float>, float>, mshadow::expr::Plan<mshadow::expr::BinaryMapExp<mshadow::op::mul, mshadow::expr::
ScalarExp<float>, mshadow::Tensor<mshadow::gpu, int=4, float>, float, int=1>, float>>(mshadow::gpu, unsigned int, mshadow::Shape<int=2>, int=4, int)
                    4.13%  7.9622ms         1  7.9622ms  7.9622ms  7.9622ms  volta_sgemm_128x128_tn
                    3.74%  7.2038ms         2  3.6019ms  24.671us  7.1791ms  void op_generic_tensor_kernel<int=2, float, float, float, int=256, cudnnGener
icOp_t=7, cudnnNanPropagation_t=0, cudnnDimOrder_t=0, int=1>(cudnnTensorStruct, float*, cudnnTensorStruct, float const *, cudnnTensorStruct, float const *
, float, float, float, float, dimArray, reducedDivisorArray)
                    2.26%  4.3525ms         1  4.3525ms  4.3525ms  4.3525ms  void cudnn::detail::pooling_fw_4d_kernel<float, float, cudnn::detail::maxpool
ing_func<float, cudnnNanPropagation_t=0>, int=0, bool=0>(cudnnTensorStruct, float const *, cudnn::detail::pooling_fw_4d_kernel<float, float, cudnn::detail
::maxpooling_func<float, cudnnNanPropagation_t=0>, int=0, bool=0>, cudnnTensorStruct*, cudnnPoolingStruct, float, cudnnPoolingStruct, int, cudnn::reduced_
divisor, float)
                    0.21%  409.57us         1  409.57us  409.57us  409.57us  void mshadow::cuda::MapPlanLargeKernel<mshadow::sv::saveto, int=8, int=1024,
mshadow::expr::Plan<mshadow::Tensor<mshadow::gpu, int=2, float>, float>, mshadow::expr::Plan<mshadow::expr::ScalarExp<float>, float>>(mshadow::gpu, unsign
ed int, mshadow::Shape<int=2>, int=2, int)
                    0.04%  69.087us         1  69.087us  69.087us  69.087us  void mshadow::cuda::SoftmaxKernel<int=8, float, mshadow::expr::Plan<mshadow::
Tensor<mshadow::gpu, int=2, float>, float>, mshadow::expr::Plan<mshadow::Tensor<mshadow::gpu, int=2, float>, float>>(mshadow::gpu, int=2, unsigned int)
                    0.03%  65.566us        13  5.0430us  1.2160us  24.447us  void mshadow::cuda::MapPlanKernel<mshadow::sv::saveto, int=8, mshadow::expr::
Plan<mshadow::Tensor<mshadow::gpu, int=2, float>, float>, mshadow::expr::Plan<mshadow::expr::ScalarExp<float>, float>>(mshadow::gpu, unsigned int, mshadow
::Shape<int=2>, int=2)
                    0.01%  24.160us         2  12.080us  2.2720us  21.888us  void mshadow::cuda::MapPlanKernel<mshadow::sv::plusto, int=8, mshadow::expr::
Plan<mshadow::Tensor<mshadow::gpu, int=2, float>, float>, mshadow::expr::Plan<mshadow::expr::Broadcast1DExp<mshadow::Tensor<mshadow::gpu, int=1, float>, f
loat, int=2, int=1>, float>>(mshadow::gpu, unsigned int, mshadow::Shape<int=2>, int=2)
                    0.01%  21.280us         1  21.280us  21.280us  21.280us  volta_sgemm_32x128_tn
                    0.01%  10.496us         9  1.1660us     992ns  2.0800us  [CUDA memset]
                    0.00%  6.8480us         1  6.8480us  6.8480us  6.8480us  [CUDA memcpy DtoH]
                    0.00%  4.6400us         1  4.6400us  4.6400us  4.6400us  void mshadow::cuda::MapPlanKernel<mshadow::sv::saveto, int=8, mshadow::expr::
Plan<mshadow::Tensor<mshadow::gpu, int=2, float>, float>, mshadow::expr::Plan<mshadow::expr::ReduceWithAxisExp<mshadow::red::maximum, mshadow::Tensor<msha
dow::gpu, int=3, float>, float, int=3, bool=1, int=2>, float>>(mshadow::gpu, unsigned int, mshadow::Shape<int=2>, int=2)
      API calls:   42.38%  3.19122s        22  145.06ms  14.729us  1.61690s  cudaStreamCreateWithFlags
                   32.55%  2.45089s        22  111.40ms  70.340us  2.44636s  cudaMemGetInfo
21.06%  1.58588s        18  88.104ms  1.2290us  424.74ms  cudaFree
                    1.82%  137.15ms         6  22.858ms  2.8800us  94.056ms  cudaDeviceSynchronize
                    0.95%  71.501ms         9  7.9446ms  36.547us  33.344ms  cudaMemcpy2DAsync
                    0.44%  33.041ms       912  36.228us     440ns  29.994ms  cudaFuncSetAttribute
                    0.26%  19.843ms        29  684.26us  2.5420us  10.940ms  cudaStreamSynchronize
                    0.24%  18.381ms        66  278.49us  5.9470us  6.1626ms  cudaMalloc
                    0.09%  6.8774ms       216  31.839us  1.2490us  5.4951ms  cudaEventCreateWithFlags
                    0.07%  5.0962ms         4  1.2740ms  484.32us  1.8679ms  cudaGetDeviceProperties
```

Many issues with trying to install NVVP. We had the disk space failure problem and then referred to the Instructor's answer in the Piazza Post @352. The steps detailed by Ayush were not successful for us, as we were being denied access to install the runfile from CUDA download page. "Access Denied. The username you have entered cannot authenticate with Duo Security. Please contact system administrator".

As it seems there are no Office Hours until Monday earliest, please excuse us our allow a late submission for this Nvidia profiling portion. We have been successful in performing everything else required in this milestone but have run into logistic problems with NVVP (it seems many other groups have the same problems).