

**ECE 408 Final Project Milestone 1****DELIVERABLES:**

1. List of all kernels that collectively consume more than 90% of program time
  - a. CUDA memcpy HtoD (39.39%)
  - b. void cudnn::detail::implicit\_convolve\_sgemm (20.65%)
  - c. Volta\_cgemm\_64x32\_tn (12.11%)
  - d. Op\_generic\_tensor\_kernel (7.15%)
  - e. Fft2d\_c2r\_32x32 (5.74%)
  - f. Volta\_sgemm\_128x128\_tn (5.72%)
2. List of CUDA API that consume more than 90% of program time
  - a. cudaStreamCreateWithFlags (42.32%)
  - b. cudaMemGetInfo (33.58%)
  - c. cudaFree (21.37%)
3. Explanation of difference between kernels and API calls
  - a. CUDA kernels are essentially C functions defined by the user that are executed by threads on the GPU. CUDA API calls extend functionality through the runtime and Driver APIs which also hold the context. The context holds all of the management data to control and use the device (allocated memory, loaded modules that contain device code, mapping between CPU and GPU memory, etc). (<https://stackoverflow.com/questions/43244645/what-is-a-cuda-context>)
4. Output of RAI running on MXNet on the CPU (time m1.1.py)

```
EvalMetric: {'accuracy': 0.8236}
8.83user 3.76system 0:05.01elapsed 251%CPU (0avgtext+0avgdata
2470596maxresident)k
0inputs+2824outputs (0major+667706minor)pagefaults 0swaps
```

5. List Program Run time
  - a. 5.01 seconds
6. Output of RAI running on MXNet on the GPU

```
EvalMetric: {'accuracy': 0.8236}
4.28user 3.32system 0:04.32elapsed 176%CPU (0avgtext+0avgdata
2843476maxresident)k
8inputs+4552outputs (0major+660709minor)pagefaults 0swaps
```

7. List Program Run time
  - a. 4.32 seconds

## Example NVPROF output

==383== NVPROF is profiling process 383, command: python m1.2.py

Loading model... done

New Inference

EvalMetric: {'accuracy': 0.8236}

==383== Profiling application: python m1.2.py

==383== Profiling result:

	Type	Time(%)	Time	Calls	Avg	Min	Max	Name
<b>GPU activities:</b>	39.39%	16.127ms	20	806.35us	1.0880us	15.480ms	[CUDA memcpy HtoD]	
	20.65%	8.4531ms	1	8.4531ms	8.4531ms	8.4531ms	void	
cudnn::detail::implicit_convolve_sgemm<float, float, int=1024, int=5, int=5, int=3, int=3, int=3, int=1, bool=1, bool=0, bool=1>(int, int, int, float const *, int, float*,								
cudnn::detail::implicit_convolve_sgemm<float, float, int=1024, int=5, int=5, int=3, int=3, int=3, int=1, bool=1, bool=0, bool=1>*, kernel_conv_params, int, float, float, int, float, float, int, int)								
	12.11%	4.9587ms	1	4.9587ms	4.9587ms	4.9587ms		
volta_cgemm_64x32_tn								
	7.15%	2.9281ms	2	1.4641ms	24.864us	2.9033ms	void	
op_generic_tensor_kernel<int=2, float, float, float, int=256, cudnnGenericOp_t=7, cudnnNanPropagation_t=0, cudnnDimOrder_t=0, int=1>(cudnnTensorStruct, float*,								
cudnnTensorStruct, float const *, cudnnTensorStruct, float const *, float, float, float, float, dimArray, reducedDivisorArray)								
	5.74%	2.3506ms	1	2.3506ms	2.3506ms	2.3506ms	void	
fft2d_c2r_32x32<float, bool=0, bool=0, unsigned int=1, bool=0, bool=0>(float*, float2 const *, int, int, int, int, int, int, int, int, float, float, cudnn::reduced_divisor, bool, float*, float*, int2, int, int)								
	5.72%	2.3400ms	1	2.3400ms	2.3400ms	2.3400ms		
volta_sgemm_128x128_tn								
	4.60%	1.8821ms	1	1.8821ms	1.8821ms	1.8821ms	void	
cudnn::detail::pooling_fw_4d_kernel<float, float, cudnn::detail::maxpooling_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)								
	3.80%	1.5541ms	1	1.5541ms	1.5541ms	1.5541ms	void	
fft2d_r2c_32x32<float, bool=0, unsigned int=0, bool=0>(float2*, float const *, int, int, int, int, int, int, int, int, cudnn::reduced_divisor, bool, int2, int, int)								
	0.37%	152.42us	1	152.42us	152.42us	152.42us	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, unsigned int, mshadow::Shape<int=2>, int=2, int)								
	0.18%	75.072us	1	75.072us	75.072us	75.072us	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)
    0.07% 30.144us    13 2.3180us 1.2160us 7.5200us 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)
    0.06% 25.440us    1 25.440us 25.440us 25.440us volta_sgemm_32x128_tn
    0.06% 23.776us    2 11.888us 2.5920us 21.184us 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>, float, int=2, int=1>, float>>(mshadow::gpu, unsigned int, mshadow::Shape<int=2>,
int=2)
    0.04% 15.968us    1 15.968us 15.968us 15.968us void
fft2d_r2c_32x32<float, bool=0, unsigned int=1, bool=0>(float2*, float const *, int, int, int, int, int,
int, int, int, cudnn::reduced_divisor, bool, int2, int, int)
    0.02% 10.016us    9 1.1120us 992ns 1.5360us [CUDA memset]
0.02% 7.3280us    1 7.3280us 7.3280us 7.3280us [CUDA memcpy DtoH]
    0.01% 4.8000us    1 4.8000us 4.8000us 4.8000us 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<mshadow::gpu, int=3, float>, float, int=3, bool=1, int=2>,
float>>(mshadow::gpu, unsigned int, mshadow::Shape<int=2>, int=2)
    0.01% 3.2640us    1 3.2640us 3.2640us 3.2640us void flip_filter<float,
float>(float*, float const *, int, int, int, int)
    0.01% 2.6550us    1 2.6550us 2.6550us 2.6550us
compute_gemm_pointers(float2**, float2 const *, int, float2 const *, int, float2 const *, int, int)
API calls: 42.32% 3.00705s    22 136.68ms 12.851us 1.56016s
cudaStreamCreateWithFlags
    33.58% 2.38633s    24 99.430ms 104.11us 2.38104s cudaMemGetInfo
    21.37% 1.51843s    19 79.917ms 834ns 408.31ms cudaFree
    1.41% 99.952ms    912 109.60us 308ns 53.410ms cudaFuncSetAttribute
    0.46% 32.478ms    9 3.6086ms 33.713us 15.513ms cudaMemcpy2DAsync
    0.31% 21.688ms    29 747.87us 3.4170us 9.9361ms cudaStreamSynchronize
    0.18% 12.549ms    68 184.55us 5.7180us 2.8536ms cudaMalloc
    0.12% 8.4657ms    216 39.192us 889ns 5.9292ms
cudaEventCreateWithFlags
    0.10% 7.2073ms    6 1.2012ms 1.1090us 7.1385ms cudaEventCreate
    0.07% 4.7327ms    4 1.1832ms 424.50us 1.7514ms
cudaGetDeviceProperties

```

0.03%	2.4824ms	375	6.6190us	284ns	331.03us	cuDeviceGetAttribute
0.01%	789.41us	2	394.70us	51.193us	738.21us	cudaHostAlloc
0.01%	621.86us	30	20.728us	7.9970us	81.572us	cudaLaunchKernel
0.01%	610.11us	4	152.53us	94.017us	275.55us	cuDeviceTotalMem
0.01%	599.74us	4	149.94us	77.489us	246.88us	cudaStreamCreate
0.01%	469.55us	12	39.128us	5.9160us	88.270us	cudaMemcpy
0.01%	389.25us	9	43.250us	9.3750us	212.87us	cudaMemsetAsync
0.00%	323.78us	210	1.5410us	566ns	16.920us	cudaDeviceGetAttribute
0.00%	289.34us	4	72.334us	43.955us	103.40us	cuDeviceGetName
0.00%	172.31us	8	21.538us	13.755us	44.913us	
cudaStreamCreateWithPriority						
0.00%	155.99us	32	4.8740us	1.4400us	15.018us	cudaSetDevice
0.00%	106.65us	564	189ns	75ns	611ns	cudaGetLastError
0.00%	43.911us	18	2.4390us	599ns	4.7600us	cudaGetDevice
0.00%	23.685us	6	3.9470us	1.6840us	7.0150us	cudaEventRecord
0.00%	13.089us	1	13.089us	13.089us	13.089us	cudaBindTexture
0.00%	9.2010us	3	3.0670us	1.8970us	4.3720us	cudaStreamWaitEvent
0.00%	7.9230us	1	7.9230us	7.9230us	7.9230us	cuDeviceGetPCIBusId
0.00%	7.0690us	2	3.5340us	2.3280us	4.7410us	
cudaHostGetDevicePointer						
0.00%	6.1030us	6	1.0170us	401ns	2.3180us	cuDeviceGetCount
0.00%	6.0000us	2	3.0000us	1.5100us	4.4900us	
cudaDeviceGetStreamPriorityRange						
0.00%	5.2940us	18	294ns	121ns	673ns	cudaPeekAtLastError
0.00%	4.7520us	5	950ns	474ns	1.7100us	cuDeviceGet
0.00%	4.1730us	3	1.3910us	809ns	2.2560us	cuInit
0.00%	3.8930us	1	3.8930us	3.8930us	3	
.8930us cudaEventQuery						
0.00%	3.3850us	1	3.3850us	3.3850us	3.3850us	cudaUnbindTexture
0.00%	2.4530us	4	613ns	354ns	1.2000us	cuDeviceGetUuid
0.00%	1.9340us	3	644ns	330ns	1.1950us	cuDriverGetVersion
0.00%	1.7790us	4	444ns	262ns	777ns	cudaGetDeviceCount

## Milestone 2

### DELIVERABLES:

- List whole program run times
- List Op. times

Run #1: 100

```
*Running /usr/bin/time python m2.1.py 100
Loading fashion-mnist data... done
Loading model... done
New Inference
Op Time: 0.034078
Op Time: 0.074938
Correctness: 0.84 Model: ece408
```

Run #2: 1,000

```
*Running /usr/bin/time python m2.1.py 1000
Loading fashion-mnist data... done
Loading model... done
New Inference
Op Time: 0.243053
Op Time: 0.741502
Correctness: 0.852 Model: ece408
4.40user 2.85system 0:01.99elapsed 363%CPU (0avgtext+0avgdata 332360maxresident)k
0inputs+28240outputs (0major+110723minor)pagefaults 0swaps
```

Default: 10,000

```
Op Time: 2.437733
Op Time: 7.488936
Correctness: 0.8397 Model: ece408
15.27user 4.59system 0:11.51elapsed 172%CPU (0avgtext+0avgdata 1617608maxresident)k
```

## Milestone 3

Run #1: 100

```
*Running /usr/bin/time python m3.1.py 100
Loading fashion-mnist data... done
Loading model... done
New Inference
Op Time: 0.000075
Op Time: 0.000213
Correctness: 0.84 Model: ece408
4.16user 3.50system 0:04.18elapsed 183%CPU (0avgtext+0avgdata 2784952maxresident)k
8inputs+28000outputs (0major+624565minor)pagefaults 0swaps
```

Run #2: 1,000

```
*Running /usr/bin/time python m3.1.py 1000
Loading fashion-mnist data... done
Loading model... done
New Inference
Op Time: 0.000611
Op Time: 0.002004
Correctness: 0.852 Model: ece408
4.31user 3.21system 0:04.15elapsed 181%CPU (0avgtext+0avgdata 2776192maxresident)k
0inputs+4576outputs (0major+623696minor)pagefaults 0swaps
```

Default: 10,000

```
*Running /usr/bin/time python m3.1.py 10000
Loading fashion-mnist data... done
Loading model... done
New Inference
Op Time: 0.006043
Op Time: 0.021991
Correctness: 0.8397 Model: ece408
4.41user 3.48system 0:04.34elapsed 181%CPU (0avgtext+0avgdata 2844976maxresident)k
0inputs+4576outputs (0major+663183minor)pagefaults 0swaps
```

#### Milestone 4

##### Kernels in constant memory optimization:

In this optimization we essentially put all the kernels in constant memory instead of fetching them from global memory each time. This sped up execution by a small amount because the gpu no longer needed to wait for global memory each time and could instead get from the far faster constant memory. In the nvprof output we can see that a very large portion of the time is spent copying memory from the host to the device. This makes sense since there is a lot of data to copy like the input, output, and const kernel memory. Everything else is fairly insignificant in terms of time used. The computation looks to have taken less time than the copying, which says something about what we learned in class about how GPU's operation.

##### Nvprof output:

```
* Running nvprof python m4.1.py
Loading fashion-mnist data... done
==282== NVPROF is profiling process 282, command: python m4.1.py
Loading model... done
New Inference
Op Time: 0.005984
Op Time: 0.021129
Correctness: 0.8397 Model: ece408
==282== Profiling application: python m4.1.py
==282== Profiling result:
```

Type	Time(%)	Time	Calls	Avg	Min	Max	Name
------	---------	------	-------	-----	-----	-----	------

```

GPU activities: 52.19% 26.955ms      2 13.477ms 5.8940ms 21.061ms
mxnet::op::forward_kernel(float*, float const *, float const *, int, int, int, int, int, int)
    32.50% 16.785ms      20 839.25us 1.1200us 16.367ms [CUDA memcpy HtoD]
    4.83% 2.4969ms      2 1.2484ms 22.624us 2.4743ms
volta_sgemm_32x128_tn
    4.68% 2.4146ms      2 1.2073ms 732.92us 1.6817ms 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)
    3.15% 1.6250ms      2 812.49us 22.207us 1.6028ms void
op_generic_tensor_kernel<int=2, float, float, float, int=256, cudnnGenericOp_t=7,
cudnnNanPropagation_t=0, cudnnDimOrder_t=0, int=1>(cudnnTensorStruct, float*,
cudnnTensorStruct, float const *, cudnnTensorStruct, float const *, float, float, float, float,
dimArray, reducedDivisorArray)
    2.04% 1.0548ms      1 1.0548ms 1.0548ms 1.0548ms void
cudnn::detail::pooling_fw_4d_kernel<float, float, cudnn::detail::maxpooling_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)
    0.31% 158.30us      1 158.30us 158.30us 158.30us 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, unsigned int,
mshadow::Shape<int=2>, int=2, int)
    0.15% 75.103us      1 75.103us 75.103us 75.103us 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)
    0.05% 27.808us      13 2.1390us 1.1840us 6.4960us 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)
    0.05% 24.288us      2 12.144us 2.5920us 21.696us 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>, float, int=2, int=1>, float>>(mshadow::gpu, unsigned int, mshadow::Shape<int=2>,
int=2)

```

0.02%	11.295us	10	1.1290us	992ns	1.6320us	[CUDA memset]
0.01%	5.9840us	2	2.9920us	2.9120us	3.0720us	[CUDA memcpy DtoD]
0.01%	5.4720us	1	5.4720us	5.4720us	5.4720us	[CUDA memcpy DtoH]
0.01%	4.6720us	1	4.6720us	4.6720us	4.6720us	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<mshadow::gpu, int=3, float>, float, int=3, bool=1, int=2>,						
float>>(mshadow::gpu, unsigned int, mshadow::Shape<int=2>, int=2)						
API calls: 42.32% 3.00169s 22 136.44ms 14.044us 1.53483s						
cudaStreamCreateWithFlags						
34.35%	2.43651s	22	110.75ms	108.17us	2.43105s	cudaMemGetInfo
20.99%	1.48869s	18	82.705ms	875ns	396.71ms	cudaFree
0.71%	50.398ms	912	55.261us	307ns	36.773ms	cudaFuncSetAttribute
0.48%	34.330ms	9	3.8145ms	17.808us	16.514ms	cudaMemcpy2DAsync
0.41%	29.389ms	6	4.8981ms	3.7270us	21.061ms	cudaDeviceSynchronize
0.28%	19.739ms	216	91.383us	895ns	8.7924ms	
cudaEventCreateWithFlags						
0.17%	11.965ms	66	181.29us	5.4460us	1.7272ms	cudaMalloc
0.11%	7.9549ms	4	1.9887ms	470.49us	3.3376ms	
cudaGetDeviceProperties						
0.06%	4.4829ms	29	154.58us	2.9980us	2.1478ms	cudaStreamSynchronize
0.04%	2.5323ms	375	6.7520us	272ns	333.39us	cuDeviceGetAttribute
0.01%	787.82us	2	393.91us	84.316us	703.50us	cudaHostAlloc
0.01%	669.10us	4	167.27us	92.829us	276.86us	cuDeviceTotalMem
0.01%	624.88us	27	23.143us	8.5620us	64.383us	cudaLaunchKernel
0.01%	566.50us	4	141.62us	76.127us	220.07us	cudaStreamCreate
0.01%	470.95us	12	39.246us	8.0500us	86.436us	cudaMemcpy
0.00%	325.50us	10	32.550us	9.3450us	112.35us	cudaMemsetAsync
0.00%	293.17us	202	1.4510us	560ns	4.8010us	cudaDeviceGetAttribute
0.00%	274.81us	4	68.703us	45.592us	100.63us	cuDeviceGetName
0.00%	238.90us	8	29.862us	14.068us	71.035us	
cudaStreamCreateWithPriority						
0.00%	156.58us	29	5.3990us	1.0770us	16.726us	cudaSetDevice
0.00%	118.89us	557	213ns	79ns	771ns	cudaGetLastError
0.00%	86.434us	2	43.217us	37.954us	48.480us	cudaMemcpyToSymbol
0.00%	65.077us	4	16.269us	1.8870us	52.832us	cudaEventRecord
0.00%	43.422us	18	2.4120us	600ns	4.2970us	cudaGetDevice
0.00%	27.094us	6	4.5150us	1.3760us	11.713us	cudaEventCreate
0.00%	9.5680us	2	4.7840us	3.7110us	5.8570us	cudaEventQuery
0.00%	7.5360us	2	3.7680us	2.5570us	4.9790us	
cudaHostGetDevicePointer						
0.00%	6.4860us	20	324ns	110ns	657ns	cudaPeekAtLastError



0.00%	6.4510us	6	1.0750us	551ns	2.3380us	cuDeviceGetCount
0.00%	5.9450us	2	2.9720us	1.5960us	4.3490us	
cudaDeviceGetStreamPriorityRange						
0.00%	4.6490us	5	929ns	418ns	1.5200us	cuDeviceGet
0.00%	3.9440us	3	1.3140us	779ns	2.3110us	culnit
0.00%	3.7140us	1	3.7140us	3.7140us	3.7140us	cuDeviceGetPCIBusId
0.00%	2.4030us	4	600ns	322ns	1.2250us	cuDeviceGetUuid
0.00%	2.3840us	4	596ns	295ns	1.0180us	cudaGetDeviceCount
0.00%	1.7610us	3	587ns	317ns	1.1250us	cuDriverGetVersion

```
Op Time: 0.005866 Model: ece408 in ~/ece408_project on gitmaster x [23:38:48] Cuda
Op Time: 0.021038
Correctness: 0.8397 Model: ece408
4.29user 3.23system 0:04.36elapsed 172%CPU (@avgtext+@avgdata 283538@maxresident)k
@inputs+464@outputs (@major+661018@minor)pagefaults @swaps
```

### Shared memory convolution optimization:

In this optimization we used shared memory and tiled every image in order to try and get a speed up. However, it actually took longer and didn't save time perhaps because of the extra time it took to load into shared memory. Again in Nvprof output we can see that the most time is used copying all that data. However, it is significantly less than the copying for the constant memory optimization for some reason. For some reason though the algorithm took longer than the original, which maybe is because of extra floating point computation that had to be used in order to use shared memory and tiling.

### Nvprof output:

```
* Running nvprof python m4.1.py
Loading fashion-mnist data... done
==283== NVPROF is profiling process 283, command: python m4.1.py
Loading model... done
New Inference
Op Time: 0.006688
Op Time: 0.039945
Correctness: 0.8397 Model: ece408
==283== Profiling application: python m4.1.py
==283== Profiling result:
      Type Time(%)   Time    Calls    Avg     Min     Max  Name
GPU activities: 65.05% 46.549ms      2 23.274ms 6.6460ms 39.903ms
mxnet::op::forward_kernel(float*, float const *, float const *, int, int, int, int, int, int)
      23.89% 17.098ms      20 854.90us 1.0870us 16.571ms [CUDA memcpy HtoD]
      3.50% 2.5054ms       2 1.2527ms 21.408us 2.4839ms
volta_sgemm_32x128_tn
```

```

3.39% 2.4293ms      2 1.2147ms 738.84us 1.6905ms 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)
2.27% 1.6222ms      2 811.10us 21.952us 1.6002ms void
op_generic_tensor_kernel<int=2, float, float, float, int=256, cudnnGenericOp_t=7,
cudnnNanPropagation_t=0, cudnnDimOrder_t=0, int=1>(cudnnTensorStruct, float*,
cudnnTensorStruct, float const *, cudnnTensorStruct, float const *, float, float, float, float,
dimArray, reducedDivisorArray)
1.47% 1.0509ms      1 1.0509ms 1.0509ms 1.0509ms void
cudnn::detail::pooling_fw_4d_kernel<float, float, cudnn::detail::maxpooling_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)
0.22% 158.33us      1 158.33us 158.33us 158.33us 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, unsigned int,
mshadow::Shape<int=2>, int=2, int)
0.10% 75.103us      1 75.103us 75.103us 75.103us 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)
0.04% 27.551us      13 2.1190us 1.1520us 6.4310us 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)
0.03% 23.808us      2 11.904us 2.4960us 21.312us 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>, float, int=2, int=1>, float>>(mshadow::gpu, unsigned int, mshadow::Shape<int=2>,
int=2)
0.02% 11.775us      10 1.1770us 992ns 1.6320us [CUDA memset]
0.01% 7.9990us      1 7.9990us 7.9990us 7.9990us [CUDA memcpy DtoH]
0.01% 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<mshadow::gpu, int=3, float>, float, int=3, bool=1, int=2>,
float>>(mshadow::gpu, unsigned int, mshadow::Shape<int=2>, int=2)
  API calls: 43.63% 3.30590s    22 150.27ms 13.646us 1.65226s
cudaStreamCreateWithFlags
    33.69% 2.55286s    22 116.04ms 90.912us 2.54790s cudaMemGetInfo
    21.22% 1.60798s    18 89.332ms 934ns 436.06ms cudaFree
0.65% 49.010ms    6 8.1683ms 5.5220us 39.911ms cudaDeviceSynchronize
    0.46% 34.679ms    9 3.8532ms 35.671us 16.619ms cudaMemcpy2DAsync
    0.11% 8.0740ms    66 122.33us 6.4630us 1.0804ms cudaMalloc
    0.06% 4.9098ms    29 169.30us 3.0820us 2.2344ms cudaStreamSynchronize
    0.06% 4.7821ms    4 1.1955ms 594.51us 1.7482ms
cudaGetDeviceProperties
    0.03% 2.4967ms    375 6.6570us 287ns 336.92us cuDeviceGetAttribute
    0.01% 1.0539ms    216 4.8780us 903ns 164.36us
cudaEventCreateWithFlags
    0.01% 972.95us    912 1.0660us 308ns 28.289us cudaFuncSetAttribute
    0.01% 972.16us    10 97.215us 8.7890us 740.59us cudaMemsetAsync
    0.01% 753.49us    2 376.75us 30.463us 723.03us cudaHostAlloc
    0.01% 682.10us    4 170.53us 96.592us 280.71us cuDeviceTotalMem
    0.01% 589.57us    4 147.39us 96.650us 231.88us cudaStreamCreate
    0.01% 531.84us    27 19.697us 8.1210us 59.889us cudaLaunchKernel
    0.00% 310.86us    12 25.905us 9.2750us 65.413us cudaMemcpy
    0.00% 287.07us    4 71.766us 47.213us 106.39us cuDeviceGetName
    0.00% 188.34us    29 6.4940us 1.0100us 37.831us cudaSetDevice
    0.00% 172.83us    202 855ns 566ns 2.3720us cudaDeviceGetAttribute
    0.00% 146.59us    8 18.323us 9.6010us 53.927us
cudaStreamCreateWithPriority
    0.00% 66.765us    557 119ns 75ns 541ns cudaGetLastError
    0.00% 45.648us    6 7.6080us 1.3970us 34.856us cudaEventCreate
    0.00% 30.209us    18 1.6780us 610ns 3.9380us cudaGetDevice
    0.00% 13.473us    4 3.3680us 1.7230us 4.4740us cudaEventRecord
    0.00% 8.0900us    2 4.0450us 3.5980us 4.4920us cudaEventQuery
    0.00% 6.4420us    6 1.0730us 543ns 2.5680us cuDeviceGetCount
    0.00% 5.4950us    20 274ns 154ns 558ns cudaPeekAtLastError
    0.00% 5.1810us    2 2.5900us 2.3190us 2.8620us
cudaHostGetDevicePointer
    0.00% 4.5580us    5 911ns 508ns 1.6070us cuDeviceGet
    0.00% 4.4850us    1 4.4850us 4.4850us 4.4850us cuDeviceGetPCIBusId
    0.00% 4.4090us    3 1.4690us 930ns 2.3420us cuInit
    0.00% 3.4450us    2 1.7220us 1.6080us 1.8370us
cudaDeviceGetStreamPriorityRange
    0.00% 2.7260us    4 681ns 346ns 1.2350us cuDeviceGetUuid

```

0.00%	2.2490us	4	562ns	279ns	900ns	cudaGetDeviceCount
0.00%	1.9590us	3	653ns	367ns	1.1500us	cuDriverGetVersion

```
Op Time: 0.006586
Op Time: 0.039858
Correctness: 0.8397 Model: ece408
4.45user 3.26system 0:05.40elapsed:142%CPU(0avgtext+0avgdata 2840004maxresident)k
```

Report: Describe and analyze the optimizations

## Double Buffer Optimization

In this trial we implemented double buffering to reduce the number of syncthread calls (previously we had 2 - one to ensure data is loaded, the other to ensure data is consumed). In double buffering, pointers to shared memory alternate for each iteration, eliminating the inner loop syncthread calls and visibly reducing the operation time by a fraction of a millisecond. Unfortunately, there were some issues setting up NVVP so further analysis is not present but the matter will most likely be resolved by next checkpoint. For now, we have attached the NVPROF output.

```
* Running /usr/bin/time python m4.1.py
loading fashion-mnist data... done
loading model... done
New Inference
Op Time: 0.005792
Op Time: 0.019796
Correctness: 0.8397 Model: ece408
```

## Nvprof output:

```
* Running nvprof python m4.1.py
Loading fashion-mnist data... done
==283== NVPROF is profiling process 283, command: python m4.1.py
Loading model... done
New Inference
Op Time: 0.005886
Op Time: 0.019878
Correctness: 0.8397 Model: ece408
==283== Profiling application: python m4.1.py
==283== Profiling result:
   Type Time(%)   Time    Calls   Avg     Min     Max  Name
GPU activities: 51.64% 25.630ms    2 12.815ms 5.8105ms 19.820ms
mxnet::op::forward_kernel(float*, float const *, float const *, int, int, int, int, int, int)
   32.97% 16.367ms    20 818.35us 1.1200us 15.989ms [CUDA memcpy HtoD]
   4.77% 2.3661ms     2 1.1831ms 722.59us 1.6435ms 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)
    4.66% 2.3113ms      2 1.1557ms 21.472us 2.2898ms
volta_sgemm_32x128_tn
    3.25% 1.6121ms      2 806.04us 21.824us 1.5903ms void
op_generic_tensor_kernel<int=2, float, float, float, int=256, cudnnGenericOp_t=7,
cudnnNanPropagation_t=0, cudnnDimOrder_t=0, int=1>(cudnnTensorStruct, float*,
cudnnTensorStruct, float const *, cudnnTensorStruct, float const *, float, float, float, float,
dimArray, reducedDivisorArray)
    2.10% 1.0428ms      1 1.0428ms 1.0428ms 1.0428ms void
cudnn::detail::pooling_fw_4d_kernel<float, float, cudnn::detail::maxpooling_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)
    0.31% 152.83us      1 152.83us 152.83us 152.83us 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, unsigned int,
mshadow::Shape<int=2>, int=2, int)
    0.15% 75.264us      1 75.264us 75.264us 75.264us 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)
    0.06% 27.871us      13 2.1430us 1.1840us 6.4320us 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)
    0.05% 23.680us      2 11.840us 2.5600us 21.120us 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>, float, int=2, int=1>, float>>(mshadow::gpu, unsigned int, mshadow::Shape<int=2>,
int=2)
    0.02% 11.168us      10 1.1160us 992ns 1.6320us [CUDA memset]
    0.01% 5.7600us      1 5.7600us 5.7600us 5.7600us [CUDA memcpy DtoH]
    0.01% 5.1510us      2 2.5750us 2.4320us 2.7190us [CUDA memcpy DtoD]
    0.01% 5.0560us      1 5.0560us 5.0560us 5.0560us 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<mshadow::gpu, int=3, float>, float, int=3, bool=1, int=2>,
float>>(mshadow::gpu, unsigned int, mshadow::Shape<int=2>, int=2)
  API calls: 41.20% 3.35913s    22 152.69ms 98.856us 3.35334s cudaMemGetInfo
              37.94% 3.09287s    22 140.58ms 14.139us 1.56727s
cudaStreamCreateWithFlags
19.24% 1.56837s    18 87.132ms 839ns 430.35ms cudaFree
    0.41% 33.555ms    9 3.7283ms 18.372us 16.181ms cudaMemcpy2DAsync
    0.34% 28.018ms    6 4.6696ms 2.7620us 19.821ms cudaDeviceSynchronize
    0.32% 26.236ms   912 28.767us 312ns 7.5358ms cudaFuncSetAttribute
    0.22% 17.782ms    66 269.42us 5.8580us 5.9865ms cudaMalloc
    0.10% 7.9502ms   216 36.806us 895ns 5.4752ms
cudaEventCreateWithFlags
    0.06% 4.9316ms    4 1.2329ms 424.58us 1.8548ms
cudaGetDeviceProperties
    0.05% 4.3637ms    29 150.47us 2.5680us 2.0465ms cudaStreamSynchronize
    0.05% 4.2113ms   375 11.230us 286ns 1.8498ms cuDeviceGetAttribute
    0.01% 962.27us    8 120.28us 13.671us 738.47us
cudaStreamCreateWithPriority
    0.01% 748.26us    2 374.13us 48.475us 699.79us cudaHostAlloc
    0.01% 644.39us    4 161.10us 92.428us 273.52us cuDeviceTotalMem
    0.01% 621.00us   27 23.000us 8.8280us 57.041us cudaLaunchKernel
    0.01% 577.57us    4 144.39us 75.965us 229.16us cudaStreamCreate
    0.01% 430.78us   12 35.898us 5.9940us 88.814us cudaMemcpy
    0.00% 311.63us   10 31.162us 8.5100us 101.64us cudaMemcpyAsync
    0.00% 290.29us    4 72.571us 47.064us 104.55us cuDeviceGetName
    0.00% 289.23us   202 1.4310us 567ns 3.9070us cudaDeviceGetAttribute
    0.00% 152.57us   29 5.2610us 1.1310us 16.778us cudaSetDevice
    0.00% 114.50us  557 205ns 76ns 9.6110us cudaGetLastError
    0.00% 76.428us    4 19.107us 2.0780us 63.771us cudaEventRecord
    0.00% 73.877us    2 36.938us 33.464us 40.413us cudaMemcpyToSymbol
    0.00% 42.154us   18 2.3410us 594ns 4.1290us cudaGetDevice
    0.00% 33.572us    2 16.786us 4.9530us 28.619us
cudaHostGetDevicePointer
    0.00% 26.454us    6 4.4090us 1.3730us 8.9730us cudaEventCreate
    0.00% 6.6290us   20 331ns 124ns 651ns cudaPeekAtLastError
    0.00% 6.1630us    2 3.0810us 2.7760us 3.3870us cudaEventQuery
    0.00% 5.8730us    2 2.9360us 1.6680us 4.2050us
cudaDeviceGetStreamPriorityRange
    0.00% 5.4900us    6 915ns 331ns 1.9390us cuDeviceGetCount
    0.00% 4.5650us    1 4.5650us 4.5650us 4.5650us cuDeviceGetPCIBusId
    0.00% 4.4780us    5 895ns 467ns 1.6250us cuDeviceGet
    0.00% 4.4700us    3 1.4900us 881ns 2.4290us culnInit

```

0.00%	2.6610us	4	665ns	354ns	1.2080us	cuDeviceGetUuid
0.00%	2.3770us	4	594ns	189ns	1.3370us	cudaGetDeviceCount
0.00%	2.1340us	3	711ns	410ns	1.1370us	cuDriverGetVersion