

Patrick Hanrahan
Benjamin Hobbs
Assignment 5
GPU

Part 1: Blur Filter

For the blur filter we averaged 5x5 blocks of an input image using CUDA GPU threading commands. The average of the 5x5 block was placed in the center of the equivalent 5x5 output block. To insure we did not write to an out of bounds address, we checked boundary conditions with an “if” statement.

```
nvidia@tegra-ubuntu:~/Desktop/Assignment_5/code/filter$ ./lab5
FPS = 29.3334
FPS = 41.7086
FPS = 61.9868
FPS = 41.7087
```

Part II: Sobel Filter

USING OPEN CV:

```
nvidia@tegra-ubuntu:~/Desktop/Assignment_5/code/sobel$ ./lab5 1024 1024 0
Using OpenCV
FPS = 15.7864
FPS = 16.1703
FPS = 17.3225
FPS = 17.2385
```

USING CPU:

```
nvidia@tegra-ubuntu:~/Desktop/Assignment_5/code/sobel$ ./lab5 1024 1024 1
Using CPU
FPS = 18.5704
FPS = 18.0074
FPS = 18.2068
FPS = 19.8103
```

USING GPU:

```
nvidia@tegra-ubuntu:~/Desktop/Assignment_5/code/sobel$ ./lab5 1024 1024
Using GPU
FPS = 34.9697
FPS = 34.5561
FPS = 22.9989
FPS = 24.9088
```

```
nvidia@tegra-ubuntu:~/Desktop/Assignment_5/code/sobel$ ./lab5 4096 4096
Using GPU
FPS = 17.0605
FPS = 21.3192
FPS = 21.2824
FPS = 21.3917
FPS = 21.5907
```

```
nvidia@tegra-ubuntu:~/Desktop/Assignment_5/code/sobel$ ./lab5 512 512
Using GPU
FPS = 39.7147
FPS = 57.5451
FPS = 71.7398
FPS = 64.8412
```

```
nvidia@tegra-ubuntu:~/Desktop/Assignment_5/code/sobel$ ./lab5 512 1024
Using GPU
FPS = 37.2168
FPS = 33.358
FPS = 34.6448
FPS = 36.1099
```

```
nvidia@tegra-ubuntu:~/Desktop/Assignment_5/code/sobel$ ./lab5 1024 512
Using GPU
FPS = 30.0601
FPS = 42.2698
FPS = 47.3036
FPS = 46.5513
```

Part III

```
nvidia@tegra-ubuntu:~/Desktop/Assignment_5/code/matrix$ ./mm 16 16
Time CPU = 0.09ms, Time GPU = 0.11ms, Speedup = 0.81x, RMSE = 0.00000
nvidia@tegra-ubuntu:~/Desktop/Assignment_5/code/matrix$ ./mm 1024 1024
Time CPU = 1535.44ms, Time GPU = 27.69ms, Speedup = 55.45x, RMSE = 0.00011
nvidia@tegra-ubuntu:~/Desktop/Assignment_5/code/matrix$ ./mm 16 1024
Time CPU = 20.82ms, Time GPU = 0.53ms, Speedup = 39.63x, RMSE = 0.00000
nvidia@tegra-ubuntu:~/Desktop/Assignment_5/code/matrix$ ./mm 1024 16
Time CPU = 0.69ms, Time GPU = 0.19ms, Speedup = 3.60x, RMSE = 0.00011
nvidia@tegra-ubuntu:~/Desktop/Assignment_5/code/matrix$ ./mm 64 64
Time CPU = 0.59ms, Time GPU = 0.13ms, Speedup = 4.65x, RMSE = 0.00000
nvidia@tegra-ubuntu:~/Desktop/Assignment_5/code/matrix$ ./mm 128 128
```

Time CPU = 3.43ms, Time GPU = 0.17ms, Speedup = 20.66x, RMSE = 0.00001
 nvidia@tegra-ubuntu:~/Desktop/Assignment_5/code/matrix\$./mm 128 64
 Time CPU = 0.87ms, Time GPU = 0.07ms, Speedup = 12.42x, RMSE = 0.00001
 nvidia@tegra-ubuntu:~/Desktop/Assignment_5/code/matrix\$./mm 64 128
 Time CPU = 1.42ms, Time GPU = 0.15ms, Speedup = 9.44x, RMSE = 0.00000

nvidia@tegra-ubuntu:~/Desktop/Assignment_5/code/matrix

```

1 flop_hp_efficiency
1 flop_sp_efficiency
1 flop_dp_efficiency
1 l2_utilization
1 dram_utilization
1 half_precision_fpu_utilization
1 nvprof -m all ./mm 512 512
==7379== NVPROF is profiling process 7379, command: ./mm 512 512
==7379== Some kernel(s) will be replayed on device 0 in order to collect all events/metrics.
==7379== Replaying kernel "block_m_kernel(float const *, float const *, float*, int, int)" (done)
Time CPU = 185.41ms, Time GPU = 1373.96ms, Speedup = 0.13x, RMSE = 0.00004
==7379== Profiling application: ./mm 512 512
==7379== Profiling result:
==7379== Metric result:
Invocations
Device "NVIDIA Tegra X2 (0)"
Kernel: block_m_kernel(float const *, float const *, float*, int, int)
1 inst_per_warp Instructions per warp 1.7990e+03 1.7990e+03 1.7990e+03
1 branch_efficiency Branch Efficiency 100.00% 100.00% 100.00%
1 warp_execution_efficiency Warp Execution Efficiency 100.00% 100.00% 100.00%
1 warp_nonpred_execution_efficiency Warp Non-Predicated Execution Efficiency 99.89% 99.89% 99.89%
1 inst_replay_overhead Instruction Replay Overhead 0.000026 0.000026 0.000026
1 shared_load_transactions_per_request Shared Memory Load Transactions Per Request 1.200000 1.200000 1.200000
1 shared_store_transactions_per_request Shared Memory Store Transactions Per Request 1.000000 1.000000 1.000000
1 local_load_transactions_per_request Local Memory Load Transactions Per Request 0.000000 0.000000 0.000000
1 local_store_transactions_per_request Local Memory Store Transactions Per Request 0.000000 0.000000 0.000000
1 gld_transactions_per_request Global Load Transactions Per Request 0.000000 0.000000 0.000000
1 gst_transactions_per_request Global Store Transactions Per Request 4.000000 4.000000 4.000000
1 shared_store_transactions Shared Store Transactions 524288 524288 524288
1 shared_load_transactions Shared Load Transactions 6291456 6291456 6291456
1 local_load_transactions Local Load Transactions 0 0 0
1 local_store_transactions Local Store Transactions 0 0 0
1 gld_transactions Global Load Transactions 4194304 4194304 4194304
1 gst_transactions Global Store Transactions 32768 32768 32768
1 sysmem_read_transactions System Memory Read Transactions 0 0 0
1 sysmem_write_transactions System Memory Write Transactions 0 0 0
1 l2_read_transactions L2 Read Transactions 2097265 2097265 2097265
1 l2_write_transactions L2 Write Transactions 32817 32817 32817
1 global_hit_rate Global Hit Rate 50.00% 50.00% 50.00%
1 local_hit_rate Local Hit Rate 0.00% 0.00% 0.00%
1 gld_requested_throughput Requested Global Load Throughput 18.600GB/s 18.600GB/s 18.599GB/s
1 gst_requested_throughput Requested Global Store Throughput 297.600MB/s 297.600MB/s 297.55MB/s
1 gld_throughput Global Load Throughput 18.600GB/s 18.600GB/s 18.599GB/s
1 gst_throughput Global Store Throughput 297.600MB/s 297.600MB/s 297.55MB/s
1 local_memory_overhead Local Memory Overhead 0.00% 0.00% 0.00%
1 tex_cache_hit_rate Unified Cache Hit Rate 50.38% 50.38% 50.38%
1 tex_cache_throughput Unified Cache Throughput 18.600GB/s 18.600GB/s 18.599GB/s
1 l2_tex_read_throughput L2 Throughput (Texture Reads) 18.600GB/s 18.600GB/s 18.599GB/s
1 l2_tex_write_throughput L2 Throughput (Texture Writes) 148.800MB/s 148.800MB/s 148.77MB/s
1 l2_read_throughput L2 Throughput (Reads) 18.601GB/s 18.601GB/s 18.600GB/s
1 l2_write_throughput L2 Throughput (Writes) 298.05MB/s 298.05MB/s 297.55MB/s
1 sysmem_read_throughput System Memory Read Throughput 0.000000/s 0.000000/s 0.000000/s
1 sysmem_write_throughput System Memory Write Throughput 0.000000/s 0.000000/s 0.000000/s
1 local_load_throughput Local Memory Load Throughput 0.000000/s 0.000000/s 0.000000/s
1 local_store_throughput Local Memory Store Throughput 0.000000/s 0.000000/s 0.000000/s
1 shared_load_throughput Shared Memory Load Throughput 223.20GB/s 223.20GB/s 223.20GB/s
1 shared_store_throughput Shared Memory Store Throughput 18.600GB/s 18.600GB/s 18.599GB/s
1 gld_efficiency Global Memory Load Efficiency 100.00% 100.00% 100.00%
1 gst_efficiency Global Memory Store Efficiency 100.00% 100.00% 100.00%
1 tex_cache_transactions Unified Cache Transactions 2097152 2097152 2097152
1 flop_count_dp Floating Point Operations(Double Precision) 0 0 0
1 flop_count_dp_add Floating Point Operations(Double Precision Add) 0 0 0
1 flop_count_dp_fma Floating Point Operations(Double Precision FMA) 0 0 0
1 flop_count_dp_mul Floating Point Operations(Double Precision Mul) 0 0 0

```

Metric Description	Min	Max	Avg
Instructions per warp	1.7990e+03	1.7990e+03	1.7990e+03
Branch Efficiency	100.00%	100.00%	100.00%
Warp Execution Efficiency	100.00%	100.00%	100.00%

We investigated the difference in warp Execution Efficiency with different block sizes. After changing the block size to 7, one can see the efficiency was reduced from 100 percent to 76.5 percent (below).

Metric Description		Min	Max	Avg
nt, int)	Instructions per warp	2.9470e+03	2.9470e+03	2.9470e+03
	Branch Efficiency	100.00%	100.00%	100.00%
	Warp Execution Efficiency	76.56%	76.56%	76.56%
	Warp Non-Predicated Execution Efficiency	76.51%	76.51%	76.51%
	Instruction Replay Overhead	0.000012	0.000012	0.000012