CSCI4130 – CUDA Programming Lab 4 Jacob Kilburn and Omar Naqib

1. Which CUDA machines have you tested?

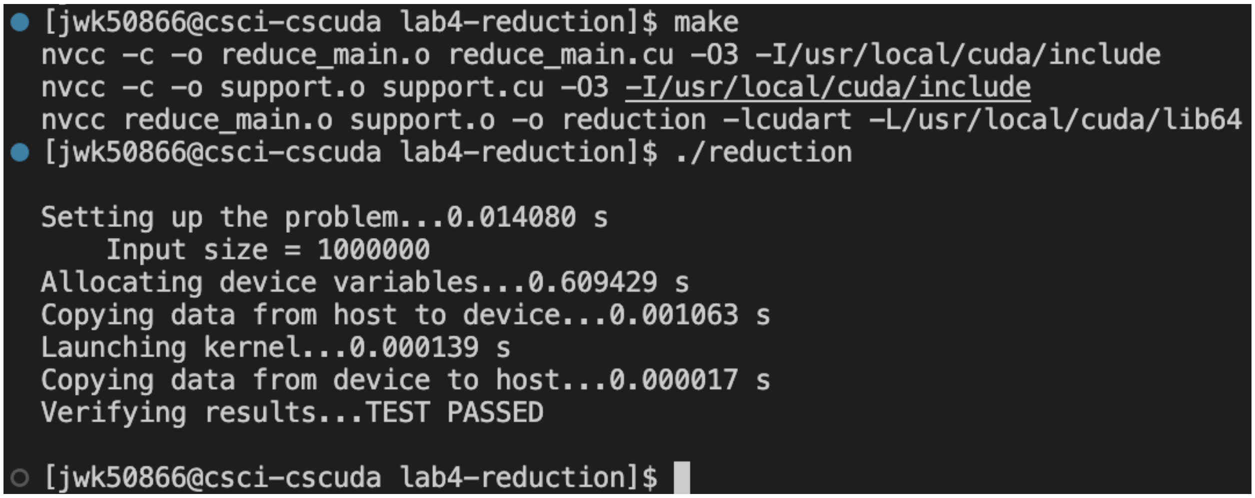
Our program works on cscuda and cuda3, we are both unable to log into cuda1 and cuda2.

2. Can your program compile properly?

Yes

3. Is your program working correctly and did it pass the test?

Yes



```
● [jwk50866@csci-cscuda lab4-reduction]$ make
  nvcc -c -o reduce_main.o reduce_main.cu -O3 -I/usr/local/cuda/include
  nvcc -c -o support.o support.cu -O3 -I/usr/local/cuda/include
  nvcc reduce_main.o support.o -o reduction -lcudart -L/usr/local/cuda/lib64
● [jwk50866@csci-cscuda lab4-reduction]$ ./reduction

  Setting up the problem...0.014080 s
      Input size = 1000000
  Allocating device variables...0.609429 s
  Copying data from host to device...0.001063 s
  Launching kernel...0.000139 s
  Copying data from device to host...0.000017 s
  Verifying results...TEST PASSED

○ [jwk50866@csci-cscuda lab4-reduction]$ 
```

4. How many times does a single thread block synchronize to reduce its portion of the array to a single value?

Since our code shifts count to the block size, the number of synchronizations can be represented by log base 2. In other words, since our block size is 512, the number of synchronizations is log$_2$(512) or 9 total synchronizations.

5. What is the minimum, maximum, and average number of "real" operations that a thread will perform? "Real" operations are those that directly contribute to the final reduction value.

Since threads with the highest index (block_size – 1) will only do one operation, the minimum is 1. Meanwhile, the threads with the lowest index will have to do an operation for each reduction which we already showed in the last questions to be 9, however; there will also need be a load instruction in our code making index 0 have to do 1 load operation and 9 addition operations, thus the maximum is 10. For the average, each of the 512 threads will have to do an initial load, then each will half a reduction step result in another (block_size – 1) operations. Thus, with a block size of 512, there will be 1023 operations over the 512 threads. Making the average be 1023/512 so approximately 2.

In summary:

Minimum = 1

Maximum = 10

Average = 2