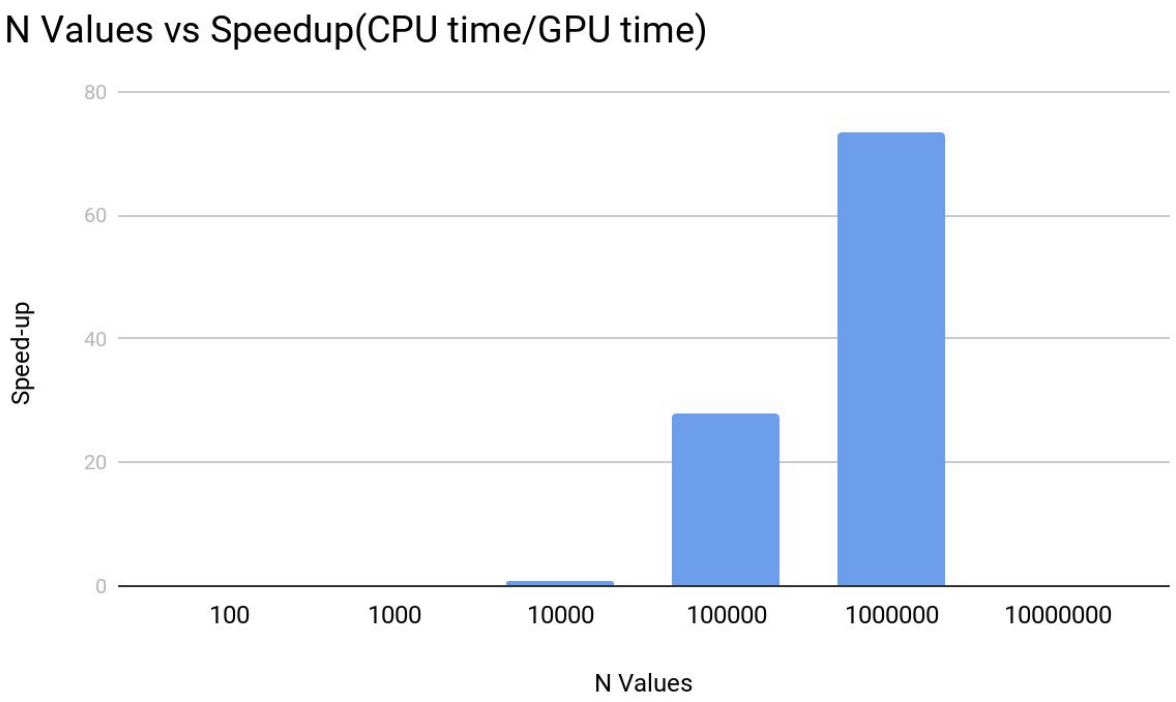

N	CPU/GPU
100	0.1242236025
1000	0.1193255512
10000	0.7582822086
100000	28.05297158
1000000	73.57955557
10000000	0

GPU N	AVG TIME
100	0.161
1000	0.1542
10000	0.163
100000	0.3096
1000000	11.2324
10000000	1035.16

CPU N	AVG TIME
100	0.02
1000	0.0184
10000	0.1236
100000	8.6852
1000000	826.475
10000000	

Explanation of Graph

The speedup of the lower N values 100 to 10,000 is less than 1 where the CPU is performing faster than the GPU. For these values, the GPU is spending more than 9/10 of its time on API calls to cudaMalloc, with nearly all of its time spent on API calls to cudaMalloc for N=100 (98%). Memory allocation overhead on GPU is higher with smaller values of N because the level of parallelism cannot make up for the time it takes to allocate memory on the device. For N=100, the GPU is also spending the most time out of all the N values on cudaMemCopy for GPU activities(~5% in total for DtoH and HtoD) so global memory access is affecting performance more for lower values of N and then decreases as N increases. However, once we reach values of N=100,000 and greater, the GPU starts shifting to spending more of its time on API calls to cudaMemCopy rather than memory allocation. At value N=100,000 API calls to cudaMalloc are 62% and decreases to 0.02% by N=10,000,000, and calls to cudaMemCopy are 36% and increase to 99.98% by N=10,000,000. The overhead shifts from memory allocation to global memory access, but this overhead is made up for the parallelization of larger data sizes, where performing the kernel computations becomes closer to and eventually reaches 100% of the time for GPU activities as N increases in size. This allows the GPU to have significant speedup compared with the CPU for these higher N values. Overhead from branch-divergence isn't noticeable from the profiling data, and I tried to optimize the kernel computations to decrease any divergence from my if statements. Also, CPU time to produce output for N=10,000,000 took more than 2 hours so speedup was not acquired for this N value, even though I used snappy3 to attempt to get this value.

nvprof outputs for all N values:

```
[vaa238@cuda5 ~]$ nvprof ./genprimes 100
==35812== NVPROF is profiling process 35812, command: ./genprimes 100
==35812== Profiling application: ./genprimes 100
==35812== Profiling result:
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:   94.78%  72.671us         1  72.671us  72.671us  72.671us  FindPrimes(int*, int)
                    2.84%  2.1760us         1  2.1760us  2.1760us  2.1760us  [CUDA memcpy DtoH]
                    2.38%  1.8240us         1  1.8240us  1.8240us  1.8240us  [CUDA memcpy HtoD]
      API calls:   98.42%  354.48ms         1  354.48ms  354.48ms  354.48ms  cudaMalloc
                    1.08%  3.8875ms       376  10.339us     282ns  392.86us  cuDeviceGetAttribute
                    0.24%  878.11us         4  219.53us  217.25us  221.51us  cuDeviceTotalMem
                    0.10%  352.14us         4  88.035us  80.467us  99.815us  cuDeviceGetName
                    0.09%  320.04us         1  320.04us  320.04us  320.04us  cudaFree
                    0.04%  134.78us         2  67.391us  46.574us  88.209us  cudaMemcpy
                    0.03%  95.361us         1  95.361us  95.361us  95.361us  cudaLaunch
                    0.00%  9.2650us         2  4.6320us     530ns  8.7350us  cudaSetupArgument
                    0.00%  8.6660us         8  1.0830us     402ns  4.4980us  cuDeviceGet
                    0.00%  4.0590us         3  1.3530us     367ns  2.9080us  cuDeviceGetCount
                    0.00%  3.8400us         1  3.8400us  3.8400us  3.8400us  cudaGetLastError
                    0.00%  2.9650us         1  2.9650us  2.9650us  2.9650us  cudaConfigureCall


[vaa238@cuda5 ~]$ nvprof ./genprimes 1000
==35896== NVPROF is profiling process 35896, command: ./genprimes 1000
==35896== Profiling application: ./genprimes 1000
==35896== Profiling result:
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:   99.14%  661.88us         1  661.88us  661.88us  661.88us  FindPrimes(int*, int)
                    0.53%  3.5520us         1  3.5520us  3.5520us  3.5520us  [CUDA memcpy DtoH]
                    0.33%  2.2080us         1  2.2080us  2.2080us  2.2080us  [CUDA memcpy HtoD]
      API calls:   95.93%  170.64ms         1  170.64ms  170.64ms  170.64ms  cudaMalloc
                    2.64%  4.6915ms       376  12.477us     304ns  811.92us  cuDeviceGetAttribute
                    0.58%  1.0378ms         4  259.46us  234.40us  283.95us  cuDeviceTotalMem
                    0.40%  706.72us         2  353.36us  29.088us  677.63us  cudaMemcpy
                    0.24%  429.02us         4  107.25us  90.067us  145.62us  cuDeviceGetName
                    0.16%  286.17us         1  286.17us  286.17us  286.17us  cudaFree
                    0.03%  59.891us         1  59.891us  59.891us  59.891us  cudaLaunch
                    0.01%  12.002us         8  1.5000us     420ns  6.2300us  cuDeviceGet
                    0.00%  6.3510us         2  3.1750us     405ns  5.9460us  cudaSetupArgument
                    0.00%  4.7910us         3  1.5970us     573ns  3.1200us  cuDeviceGetCount
                    0.00%  2.2520us         1  2.2520us  2.2520us  2.2520us  cudaConfigureCall
                    0.00%  2.2090us         1  2.2090us  2.2090us  2.2090us  cudaGetLastError
```

```
[vaa238@cuda5 ~]$ nvprof ./genprimes 10000
==36011== NVPROF is profiling process 36011, command: ./genprimes 10000
==36011== Profiling application: ./genprimes 10000
==36011== Profiling result:
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:   99.76%  6.4433ms         1  6.4433ms  6.4433ms  6.4433ms  FindPrimes(int*, int)
                    0.13%  8.1600us         1  8.1600us  8.1600us  8.1600us  [CUDA memcpy HtoD]
                    0.11%  7.2640us         1  7.2640us  7.2640us  7.2640us  [CUDA memcpy DtoH]
      API calls:   93.47%  185.44ms         1  185.44ms  185.44ms  185.44ms  cudaMalloc
                    3.28%  6.5170ms         2  3.2585ms  49.620us  6.4674ms  cudaMemcpy
                    2.23%  4.4217ms       376  11.759us     329ns  476.84us  cuDeviceGetAttribute
                    0.60%  1.1946ms         4  298.64us  263.37us  337.56us  cuDeviceTotalMem
                    0.26%  510.46us         4  127.61us  93.339us  191.01us  cuDeviceGetName
                    0.11%  212.28us         1  212.28us  212.28us  212.28us  cudaFree
                    0.04%  69.684us         1  69.684us  69.684us  69.684us  cudaLaunch
                    0.01%  11.089us         8  1.3860us     472ns  5.7660us  cuDeviceGet
                    0.00%  6.2730us         2  3.1360us     413ns  5.8600us  cudaSetupArgument
                    0.00%  4.8350us         3  1.6110us     391ns  3.2300us  cuDeviceGetCount
                    0.00%  2.3840us         1  2.3840us  2.3840us  2.3840us  cudaConfigureCall
                    0.00%  2.1480us         1  2.1480us  2.1480us  2.1480us  cudaGetLastError

[vaa238@cuda5 ~]$ nvprof ./genprimes 100000
==36050== NVPROF is profiling process 36050, command: ./genprimes 100000
==36050== Profiling application: ./genprimes 100000
==36050== Profiling result:
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:   99.92%  114.43ms         1  114.43ms  114.43ms  114.43ms  FindPrimes(int*, int)
                    0.04%  44.095us         1  44.095us  44.095us  44.095us  [CUDA memcpy DtoH]
                    0.04%  42.560us         1  42.560us  42.560us  42.560us  [CUDA memcpy HtoD]
      API calls:   62.20%  199.39ms         1  199.39ms  199.39ms  199.39ms  cudaMalloc
                   35.80%  114.76ms         2  57.379ms  196.26us  114.56ms  cudaMemcpy
                    1.42%  4.5575ms       376  12.121us     369ns  474.17us  cuDeviceGetAttribute
                    0.36%  1.1587ms         4  289.67us  279.14us  309.60us  cuDeviceTotalMem
                    0.12%  392.37us         4  98.092us  93.500us  104.98us  cuDeviceGetName
                    0.07%  210.99us         1  210.99us  210.99us  210.99us  cudaFree
                    0.02%  76.806us         1  76.806us  76.806us  76.806us  cudaLaunch
                    0.00%  12.639us         8  1.5790us     496ns  6.8630us  cuDeviceGet
                    0.00%  7.7880us         2  3.8940us     600ns  7.1880us  cudaSetupArgument
                    0.00%  5.8160us         3  1.9380us     493ns  4.2740us  cuDeviceGetCount
                    0.00%  2.7400us         1  2.7400us  2.7400us  2.7400us  cudaGetLastError
                    0.00%  2.6780us         1  2.6780us  2.6780us  2.6780us  cudaConfigureCall

[vaa238@cuda5 ~]$ nvprof ./genprimes 1000000
==36126== NVPROF is profiling process 36126, command: ./genprimes 1000000
==36126== Profiling application: ./genprimes 1000000
==36126== Profiling result:
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:   99.98%  8.85490s         1  8.85490s  8.85490s  8.85490s  FindPrimes(int*, int)
                    0.01%  1.0564ms         1  1.0564ms  1.0564ms  1.0564ms  [CUDA memcpy HtoD]
                    0.01%  952.43us         1  952.43us  952.43us  952.43us  [CUDA memcpy DtoH]
```

```
 API calls:  97.75%  8.85761s         2  4.42880s  1.3369ms  8.85627s  cudaMemcpy
             2.17%  196.75ms         1  196.75ms  196.75ms  196.75ms  cudaMalloc
             0.06%  5.3042ms       376  14.106us     343ns  1.1392ms  cuDeviceGetAttribute
             0.01%  1.1138ms         4  278.46us  274.80us  281.13us  cuDeviceTotalMem
             0.00%  391.80us         4  97.949us  92.837us  106.93us  cuDeviceGetName
             0.00%  308.23us         1  308.23us  308.23us  308.23us  cudaFree
             0.00%  65.323us         1  65.323us  65.323us  65.323us  cudaLaunch
             0.00%  10.993us         8  1.3740us     454ns  5.4250us  cuDeviceGet
             0.00%  6.2010us         2  3.1000us     414ns  5.7870us  cudaSetupArgument
             0.00%  4.8070us         3  1.6020us     372ns  3.2680us  cuDeviceGetCount
             0.00%  2.5900us         1  2.5900us  2.5900us  2.5900us  cudaConfigureCall
             0.00%  2.2320us         1  2.2320us  2.2320us  2.2320us  cudaGetLastError


[vaa238@cuda5 ~]$ nvprof ./genprimes 10000000
==36175== NVPROF is profiling process 36175, command: ./genprimes 10000000
==36175== Profiling application: ./genprimes 10000000
==36175== Profiling result:
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:  100.00%  937.265s         1  937.265s  937.265s  937.265s  FindPrimes(int*, int)
             0.00%  18.189ms         1  18.189ms  18.189ms  18.189ms  [CUDA memcpy DtoH]
             0.00%  12.471ms         1  12.471ms  12.471ms  12.471ms  [CUDA memcpy HtoD]
      API calls:  99.98%  937.297s         2  468.649s  12.785ms  937.284s  cudaMemcpy
             0.02%  159.08ms         1  159.08ms  159.08ms  159.08ms  cudaMalloc
             0.00%  4.7732ms       376  12.694us     272ns  519.14us  cuDeviceGetAttribute
             0.00%  1.3635ms         1  1.3635ms  1.3635ms  1.3635ms  cudaFree
             0.00%  1.2254ms         4  306.34us  282.91us  321.05us  cuDeviceTotalMem
             0.00%  391.52us         4  97.880us  81.785us  106.88us  cuDeviceGetName
             0.00%  88.130us         1  88.130us  88.130us  88.130us  cudaLaunch
             0.00%  12.520us         2  6.2600us     388ns  12.132us  cudaSetupArgument
             0.00%  10.272us         8  1.2840us     430ns  5.0230us  cuDeviceGet
             0.00%  4.2050us         3  1.4010us     374ns  2.6680us  cuDeviceGetCount
             0.00%  3.8130us         1  3.8130us  3.8130us  3.8130us  cudaConfigureCall
             0.00%  2.2910us         1  2.2910us  2.2910us  2.2910us  cudaGetLastError
```