CS610 PROGRAMMING FOR PERFORMANCE
Assignment 4

November 5, 2024

SAHIL BASIA
241110061

Note

- All the problems I have tested on GPU3 and GPU0. Their was a lot of change in output I got from both the machines. Like around 10x to 20x difference in speedup.
- I have included the results of problems using GPU3 as asked. nvprof was not working on GPU3 due to compute compatibility was greater than 8.0. So I used nvprof on GPU0 and attached result.txt files for nvprof results.
- I have used nvidia-smi command to show the characteristics of GPU3 and GPU0. The screenshots are attached in end.

Ans: Problem - 1

In this problem, the results were totally dependent on the GPU used. Not only this, but the results also varied a lot after repeated execution. For UVM and pinned memory part I used block side 2 as it gave best results.

Command Used

```
nvcc -std=c++17 -arch=sm_61 -lineinfo -src-in-ptx -ccbin
/bin/g++-10 p1.cu -o p1.out

./p1.out
```

Results/Evaluation

Stencil result

Time taken by CPU stencil execution is: 32.141 ms

Part 1 result
Time taken by kernel1 execution is: 2.25344 ms

Part 2 result
Time taken by kernel2_1 with block side = 1 execution is: 1.96576 ms

Time taken by kernel2_2 with block side = 2 execution is: 1.85533 ms

Time taken by kernel2_4 with block side = 4 execution is: 1.81837 ms

Time taken by kernel2_8 with block side = 8 execution is: 1.82557 ms

Part 3 result
Time taken by kernel2_part3 with block side = 2 execution is: 1.88944 ms

Part 4 result
Time taken by kernel2_part4 execution is = 1.04368 ms

Part 5 result
Time taken by kernel2_part5 execution is = 5.96486 ms

AVG SPEEDUPS

Speedup of kernel1 over stencil = 14.263070
Speedup of kernel2_1 over stencil = 15.197538

```
Speedup of kernel2_2 over stencil = 16.278326
Speedup of kernel2_4 over stencil = 16.323035
Speedup of kernel2_8 over stencil = 16.561773
Speedup of kernel2_part3 over stencil = 16.320117
Speedup of kernel2_part4 pinned memory over stencil = 28.839018
Speedup of kernel2_part5 unified memory over stencil = 8.061749
```

Ans: **Problem - 2**

Command Used

```
nvcc -std=c++17 -arch=sm_61 -lineinfo -src-in-ptx -ccbin
/bin/g++-10  p2.cu -o p2.out

./p2.out or ./p2.out $((2**24))$ 512
here $((2**24))$ this is the value of N in the code and
512 is the number of threads per block. I tested with
different versions so I used this approach to test.
```

Results/Evaluation

```
Time taken by Thrust implementation: 839 ms
Time taken by CUDA implementation: 45.6233 ms
No differences found between base and test versions

CUDA speedup over Thrust: 18.3897

Last value in CUDA output: 16777215
Last value in Thrust output: 16777215
```

Ans: **Problem - 3**

Part - 1

Command Used

```
nvcc -std=c++17 -arch=sm_61 -lineinfo -src-in-ptx -ccbin
/bin/g++-10  pr3_1.cu -o pr3_1.out

./pr3_1.out
```

Results/Evaluation

A new result file will be created

Part - 2

Command Used

```
nvcc -std=c++17 -arch=sm_61 -lineinfo -src-in-ptx -ccbin
/bin/g++-10  pr3_2.cu -o pr3_2.out

./pr3_2.out
```

Results/Evaluation

A new result file will be created

Part - 3

Command Used

```
nvcc -std=c++17 -arch=sm_61 -lineinfo -src-in-ptx -ccbin
/bin/g++-10  pr3_3.cu -o pr3_3.out

./pr3_3.out
```

Results/Evaluation

A new result file will be created

Part - 4

Command Used

```
nvcc -std=c++17 -arch=sm_61 -lineinfo -src-in-ptx -ccbin
/bin/g++-10  pr3_4.cu -o pr3_4.out

./pr3_4.out
```

Results/Evaluation

A new result file will be created

Ans: Problem - 4

In this, I introduced branchless programming in the kernel to optimize the code further. The results were astonishing but varied from GPU to GPU. In this report, all results are based on GPU3. Rest Shared memory concept is also used.

Command Used

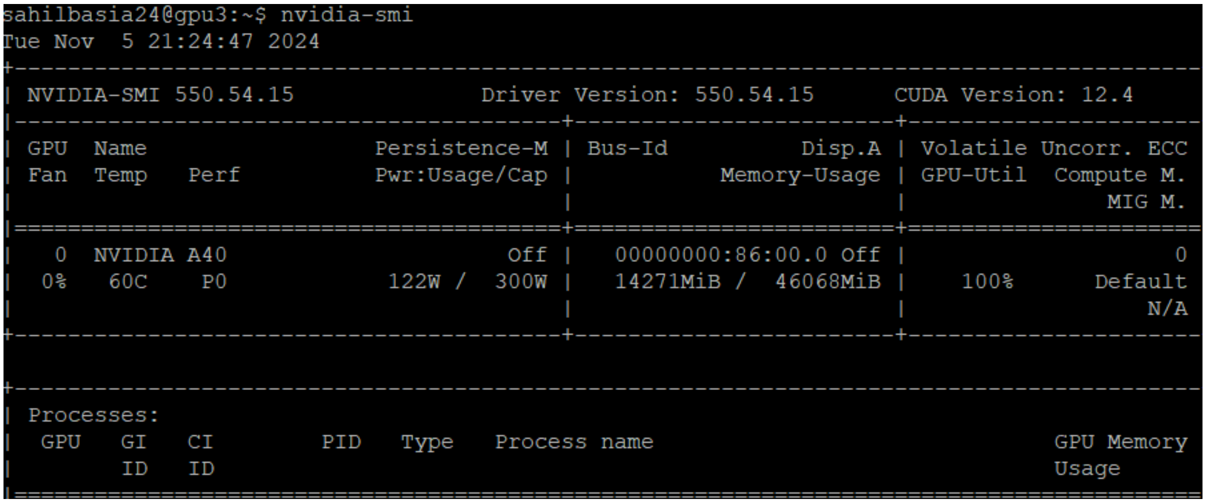
```
nvcc -std=c++17 -arch=sm_61 -lineinfo -src-in-ptx -ccbin
/bin/g++-10  p4.cu -o p4.out

./p4.out
```

Results/Evaluation

```
GPU Execution time for 2D convolution (normal): 0.040288 ms
GPU Execution time for 2D convolution (optimized): 0.029728 ms
No differences found between base and optimized versions
GPU Execution time for 2D convolution (shared memory): 0.016128
ms
No differences found between base and optimized versions
GPU Execution time for 3D convolution (normal): 0.97872 ms
GPU Execution time for 3D convolution (optimized): 0.95008 ms
No differences found between base and shared_mem versions
GPU Execution time for 3D convolution (shared memory): 0.872864
ms
No differences found between base and shared_mem versions
```

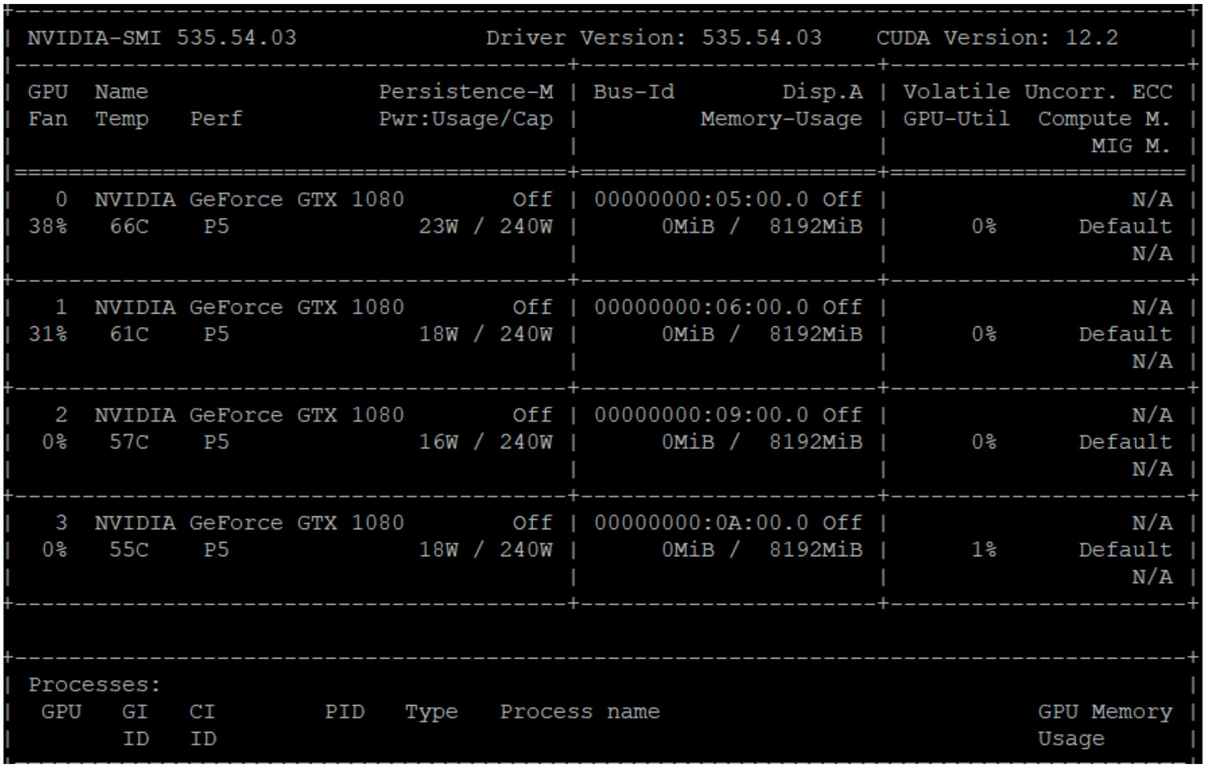
AVG SPEEDUPS

```
Speedup of 2D optimized over 2D normal: 1.355221
Speedup of 2D shared memory over 2D normal: 2.498016
Speedup of 3D optimized over 3D normal: 1.030145
Speedup of 3D shared memory over 3D normal: 1.121274
```



```
sahilbasia24@gpu3:~$ nvidia-smi
Tue Nov  5 21:24:47 2024
+-----------------------------------------------------------------------------------------
| NVIDIA-SMI 550.54.15              Driver Version: 550.54.15      CUDA Version: 12.4
|-----------------------------------------+------------------------+----------------------
| GPU  Name                 Persistence-M | Bus-Id          Disp.A | Volatile Uncorr. ECC
| Fan  Temp   Perf          Pwr:Usage/Cap |           Memory-Usage | GPU-Util  Compute M.
|                                         |                        |               MIG M.
|=========================================+========================+======================
|   0  NVIDIA A40                     Off |   00000000:86:00.0 Off |                    0
|  0%   60C    P0            122W /  300W |   14271MiB /  46068MiB |    100%      Default
|                                         |                        |                  N/A
+-----------------------------------------+------------------------+----------------------

+-----------------------------------------------------------------------------------------
| Processes:
|  GPU   GI   CI        PID   Type   Process name                              GPU Memory
|        ID   ID                                                               Usage
|=========================================================================================
```

Figure 1: GPU3_nvidia_smi



```
+---------------------------------------------------------------------------------------+
| NVIDIA-SMI 535.54.03              Driver Version: 535.54.03    CUDA Version: 12.2     |
|-----------------------------------------+----------------------+----------------------+
| GPU  Name                 Persistence-M | Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp   Perf          Pwr:Usage/Cap |         Memory-Usage | GPU-Util  Compute M. |
|                                         |                      |               MIG M. |
|=========================================+======================+======================|
|   0  NVIDIA GeForce GTX 1080        Off | 00000000:05:00.0 Off |                  N/A |
| 38%   66C    P5              23W / 240W |      0MiB /  8192MiB |      0%      Default |
|                                         |                      |                  N/A |
+-----------------------------------------+----------------------+----------------------+
|   1  NVIDIA GeForce GTX 1080        Off | 00000000:06:00.0 Off |                  N/A |
| 31%   61C    P5              18W / 240W |      0MiB /  8192MiB |      0%      Default |
|                                         |                      |                  N/A |
+-----------------------------------------+----------------------+----------------------+
|   2  NVIDIA GeForce GTX 1080        Off | 00000000:09:00.0 Off |                  N/A |
|  0%   57C    P5              16W / 240W |      0MiB /  8192MiB |      0%      Default |
|                                         |                      |                  N/A |
+-----------------------------------------+----------------------+----------------------+
|   3  NVIDIA GeForce GTX 1080        Off | 00000000:0A:00.0 Off |                  N/A |
|  0%   55C    P5              18W / 240W |      0MiB /  8192MiB |      1%      Default |
|                                         |                      |                  N/A |
+-----------------------------------------+----------------------+----------------------+

+---------------------------------------------------------------------------------------+
| Processes:                                                                            |
|  GPU   GI   CI        PID   Type   Process name                            GPU Memory |
|        ID   ID                                                             Usage      |
```

Figure 2: GPU0_nvidia_smi

Problem -1

```
==24411== Profiling application: ./p1.out
==24411== Profiling result:
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:   45.26%  7.3191ms         7  1.0456ms  637.77us  1.2915ms  [CUDA memcpy DtoH]
                   29.93%  4.8398ms         3  1.6133ms  147.10us  4.5452ms  kernel2_part3(float const *, float*)
                   12.15%  1.9654ms         2  982.71us  698.82us  1.2666ms  [CUDA memcpy HtoD]
                    2.78%  449.48us         1  449.48us  449.48us  449.48us  kernel2_1(float const *, float*)
                    2.68%  433.41us         1  433.41us  433.41us  433.41us  kernel2_4(float const *, float*)
                    2.67%  431.52us         1  431.52us  431.52us  431.52us  kernel2_2(float const *, float*)
                    2.44%  394.21us         1  394.21us  394.21us  394.21us  kernel1(float const *, float*)
                    2.10%  339.33us         1  339.33us  339.33us  339.33us  kernel2_8(float const *, float*)
      API calls:   78.21%  197.26ms         8  24.658ms  84.527us  196.33ms  cudaMalloc
                    8.05%  20.311ms         2  10.156ms  27.778us  20.284ms  cudaMallocManaged
                    5.04%  12.711ms         9  1.4124ms  734.56us  1.9544ms  cudaMemcpy
                    3.58%  9.0218ms        10  902.18us  183.99us  4.1156ms  cudaFree
                    2.01%  5.0683ms         2  2.5342ms  2.4839ms  2.5844ms  cudaHostAlloc
                    1.82%  4.5912ms         8  573.90us  5.1370us  4.5479ms  cudaEventSynchronize
                    0.85%  2.1488ms         2  1.0744ms  1.0321ms  1.1167ms  cudaFreeHost
                    0.30%  751.60us       404  1.8600us     149ns  89.570us  cuDeviceGetAttribute
                    0.07%  177.28us         8  22.160us  20.381us  23.732us  cudaLaunchKernel
                    0.04%  105.96us        16  6.6220us  2.0950us  13.786us  cudaEventRecord
                    0.01%  19.550us         4  4.8870us  3.1390us  9.2820us  cuDeviceGetName
                    0.01%  12.750us         8  1.5930us  1.3600us  2.0150us  cudaEventElapsedTime
                    0.00%  8.7840us         4  2.1960us     892ns  5.1780us  cuDeviceGetPCIBusId
                    0.00%  8.2880us         2  4.1440us     559ns  7.7290us  cudaEventCreate
                    0.00%  3.7700us         2  1.8850us     668ns  3.1020us  cudaEventDestroy
                    0.00%  1.5260us         8     190ns     143ns     409ns  cuDeviceGet
                    0.00%  1.3060us         4     326ns     206ns     564ns  cuDeviceTotalMem
                    0.00%  1.0390us         3     346ns     170ns     666ns  cuDeviceGetCount
                    0.00%     793ns         4     198ns     160ns     272ns  cuDeviceGetUuid

==24411== Unified Memory profiling result:
Device "NVIDIA GeForce GTX 1080 (0)"
   Count  Avg Size  Min Size  Max Size  Total Size  Total Time  Name
     295  55.538KB  4.0000KB  0.9766MB  16.00000MB  1.731594ms  Host To Device
      48  170.67KB  4.0000KB  0.9961MB  8.000000MB  678.3730us  Device To Host
      39         -         -         -           -  4.792965ms  Gpu page fault groups
Total CPU Page faults: 72
```

Problem -2

```
==24784== Profiling application: ./p2.out
==24784== Profiling result:
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:   54.68%  33.579ms         3  11.193ms  18.368us  33.538ms  cuda_sum(unsigned int*, unsigned int
                   16.58%  10.181ms         1  10.181ms  10.181ms  10.181ms  [CUDA memcpy DtoH]
                   16.54%  10.158ms         1  10.158ms  10.158ms  10.158ms  [CUDA memcpy HtoD]
                    4.97%  3.0550ms         2  1.5275ms  6.7520us  3.0482ms  add_block_sums(unsigned int*, unsigne
                    4.88%  2.9953ms         1  2.9953ms  2.9953ms  2.9953ms  void thrust::cuda_cub::core::_kernel_
                    2.33%  1.4337ms         1  1.4337ms  1.4337ms  1.4337ms  void thrust::cuda_cub::core::_kernel_
                    0.01%  6.6880us         1  6.6880us  6.6880us  6.6880us  void thrust::cuda_cub::core::_kernel_
      API calls:   66.56%  195.31ms         9  21.701ms  2.8390us  194.66ms  cudaMalloc
                   12.52%  36.733ms         5  7.3466ms  9.4430us  33.623ms  cudaDeviceSynchronize
                    6.95%  20.408ms         2  10.204ms  10.145ms  10.264ms  cudaMemcpyAsync
                    6.91%  20.276ms         2  10.138ms  34.964us  20.241ms  cudaMallocManaged
                    5.14%  15.079ms        11  1.3708ms  2.3350us  6.0849ms  cudaFree
                    1.61%  4.7178ms         4  1.1794ms  3.5320us  2.9999ms  cudaStreamSynchronize
                    0.24%  707.94us       404  1.7520us     116ns  93.134us  cuDeviceGetAttribute
                    0.04%  112.59us         8  14.073us  4.4880us  29.930us  cudaLaunchKernel
                    0.01%  19.610us         4  4.9020us  2.9870us  9.7190us  cuDeviceGetName
                    0.01%  16.323us         2  8.1610us  5.6160us  10.707us  cudaEventRecord
                    0.00%  11.706us         1  11.706us  11.706us  11.706us  cudaFuncGetAttributes
                    0.00%  10.541us         2  5.2700us     671ns  9.8700us  cudaEventCreate
                    0.00%  9.5650us         4  2.3910us     654ns  6.3700us  cuDeviceGetPCIBusId
                    0.00%  6.1890us         1  6.1890us  6.1890us  6.1890us  cudaEventSynchronize
                    0.00%  4.5650us        37     123ns      96ns     261ns  cudaGetLastError
                    0.00%  4.4520us         2  2.2260us     709ns  3.7430us  cudaEventDestroy
                    0.00%  4.2950us         9     477ns     250ns  1.6520us  cudaGetDevice
                    0.00%  2.8530us         5     570ns     245ns  1.4830us  cudaDeviceGetAttribute
                    0.00%  1.9930us         1  1.9930us  1.9930us  1.9930us  cudaEventElapsedTime
                    0.00%  1.4020us         8     175ns     121ns     420ns  cuDeviceGet
                    0.00%  1.1400us         4     285ns     239ns     410ns  cuDeviceTotalMem
                    0.00%     880ns         6     146ns     109ns     217ns  cudaPeekAtLastError
                    0.00%     812ns         3     270ns     155ns     401ns  cuDeviceGetCount
                    0.00%     667ns         4     166ns     130ns     246ns  cuDeviceGetUuid
                    0.00%     233ns         1     233ns     233ns     233ns  cudaGetDeviceCount

==24784== Unified Memory profiling result:
Device "NVIDIA GeForce GTX 1080 (0)"
   Count  Avg Size  Min Size  Max Size  Total Size  Total Time  Name
     856  76.561KB  4.0000KB  0.9922MB  64.00000MB  11.26996ms  Host To Device
     384  170.67KB  4.0000KB  0.9961MB  64.00000MB  10.37151ms  Device To Host
     339         -         -         -           -  30.72517ms  Gpu page fault groups
Total CPU Page faults: 384
```

Problem -3

```
==26003== Profiling application: ./pr3_1.out
==26003== Profiling result:
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:   55.60%  110.261s     25045  4.4025ms  4.3396ms  20.985ms  [CUDA memcpy HtoH]
                   40.26%  79.8428s     25045  3.1880ms  3.1818ms  6.2691ms  [CUDA memset]
                    4.14%  8.20832s     25045  327.74us  299.91us  1.2651ms  computeKernel(double*, double*, doubl
                    0.00%  6.8490us         4  1.7120us  1.4080us  2.4330us  [CUDA memcpy HtoD]
      API calls:   57.70%  110.488s     25049  4.4109ms  3.9190us  20.995ms  cudaMemcpy
                   42.04%  80.5008s     25045  3.2142ms  3.2076ms  6.4350ms  cudaMemset
                    0.14%  277.49ms     25045  11.079us  8.8250us  46.481us  cudaLaunchKernel
                    0.11%  208.86ms         2  104.43ms     901ns  208.86ms  cudaEventCreate
                    0.01%  11.829ms         1  11.829ms  11.829ms  11.829ms  cudaHostAlloc
                    0.00%  4.9040ms         1  4.9040ms  4.9040ms  4.9040ms  cudaFreeHost
                    0.00%  801.28us       404  1.9830us     136ns  102.28us  cuDeviceGetAttribute
                    0.00%  250.88us         4  62.719us  2.9760us  218.40us  cudaFree
                    0.00%  112.63us         4  28.156us  2.9660us  100.02us  cudaMalloc
                    0.00%  21.578us         4  5.3940us  3.4640us  10.019us  cuDeviceGetName
                    0.00%  16.308us         2  8.1540us  5.9830us  10.325us  cudaEventRecord
                    0.00%  10.893us         4  2.7230us     706ns  8.0450us  cuDeviceGetPCIBusId
                    0.00%  5.3800us         1  5.3800us  5.3800us  5.3800us  cudaEventSynchronize
                    0.00%  1.8430us         1  1.8430us  1.8430us  1.8430us  cudaEventElapsedTime
                    0.00%  1.8380us         2     919ns     513ns  1.3250us  cudaEventDestroy
                    0.00%  1.6810us         8     210ns     131ns     474ns  cuDeviceGet
                    0.00%  1.2200us         4     305ns     193ns     490ns  cuDeviceTotalMem
                    0.00%     868ns         4     217ns     174ns     317ns  cuDeviceGetUuid
                    0.00%     817ns         3     272ns     163ns     434ns  cuDeviceGetCount
```

Problem -4

```
==25081== Profiling application: ./p4.out
==25081== Profiling result:
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:   38.69%  262.15us         6  43.691us  2.4320us  85.249us  [CUDA memcpy DtoH]
                   20.75%  140.55us         1  140.55us  140.55us  140.55us  kernel3D_optimized(float const *, flo
                   19.56%  132.51us         2  66.256us  4.0640us  128.45us  [CUDA memcpy HtoD]
                   18.45%  124.99us         1  124.99us  124.99us  124.99us  kernel3D_linear(float const *, float*
                    1.29%  8.7360us         1  8.7360us  8.7360us  8.7360us  kernel2D_linear(float const *, float*
                    1.26%  8.5440us         1  8.5440us  8.5440us  8.5440us  kernel2D_optimized(float const *, flo
      API calls:   94.59%  279.47ms         2  139.74ms  1.0270us  279.47ms  cudaEventCreate
                    2.78%  8.2219ms         8  1.0277ms  14.488us  2.4989ms  cudaMemcpy
                    2.24%  6.6236ms         4  1.6559ms  6.5340us  6.5215ms  cudaFree
                    0.26%  762.91us       404  1.8880us     131ns  89.633us  cuDeviceGetAttribute
                    0.07%  211.37us         4  52.843us  3.0060us  105.38us  cudaMalloc
                    0.02%  47.479us         6  7.9130us  1.1830us  18.948us  cudaLaunchKernel
                    0.01%  30.305us        12  2.5250us  1.6350us  7.0830us  cudaEventRecord
                    0.01%  30.038us         6  5.0060us  4.7710us  5.1390us  cudaEventSynchronize
                    0.01%  20.766us         4  5.1910us  3.6670us  8.9720us  cuDeviceGetName
                    0.01%  14.809us         4  3.7020us     761ns  9.1630us  cuDeviceGetPCIBusId
                    0.00%  6.7570us         6  1.1260us     846ns  1.6790us  cudaEventElapsedTime
                    0.00%  1.8040us         2     902ns     410ns  1.3940us  cudaEventDestroy
                    0.00%  1.6290us         8     203ns     136ns     439ns  cuDeviceGet
                    0.00%  1.1910us         4     297ns     205ns     516ns  cuDeviceTotalMem
                    0.00%     898ns         4     224ns     167ns     309ns  cuDeviceGetUuid
                    0.00%     809ns         3     269ns     170ns     453ns  cuDeviceGetCount
```