Braden Dalit
May 19, 2023
CSC 656-01
Professor E. Wes Bethel

Coding Project 4
ANALYSIS QUESTIONS

1. **What is the MFLOP/s performance gain going from the CPU-only code to the final version of your CUDA code (the one with the cudaMemPrefetchAsync() call)? Show your work on how you compute this result.**
    a. Computing the % performance gain going from CPU only to Final CUDA code was figured out using this formula: (CPU only/ Final Cuda Code)*100 = (106598.9763/489.2929)*100 = 21,786.33 % increase.
2. **What is the memory bandwidth performance gain (or loss) going from the CPU-only code to the final version of your CUDA code (the one with the cudaMemPrefetchAsync() call)? Show your work on how you compute this result.**
    a. The CPU-only code has a % memory bandwidth utilization of 0.43% while The Final CUDA code has a 92.74% memory bandwidth utilization. Using the formula (Final Cuda % bandwidth / CPU only % bandwidth)*100 = (92.74/0.43)*100 = 21,567.44 % performance gain. If looking at the Memory Bandwidth Speeds (GB/s) CPU-only code has a rate of 3.8226 GB/s and the CUDA code has a rate of 832.8045 GB/s. Using a similar equation as above I get (832.8045/3.8226)*100 = 21,786.33 % increase. Which is identical to the MFLOP/s performance increase.
3. **For the final version of your CUDA code (the one with the cudaMemPrefetchAsync() call), what is the total number of concurrent threads being run? Show your work on how you arrive at this result.**
    a. Number Of Blocks Of Parallel Threads = (N + blocksize -1) / blocksize. Where block size = 256 and N = 2^26 and 256 Threads per Block
        i. Number Of Blocks Of Parallel Threads = (2^26 + 256 -1) / 256 = 262,145 blocks
        ii. If there are 256 Threads per Block then Number Of Parallel Threads = (N + blocksize - 1).
            1. Number of Parallel Threads = 2^26 +256 - 1 = 67,109,119 parallel threads.

PERFORMANCE TABLE

bdalit_656_Project4

File Edit View Insert Format Data Tools Extensions Help

Share

M24

	A	B	C	D	E
1	**Performance Table**	**kernel runtime (miliseconds)**	**MFLOP/s (MB)**	**My Bandwidth (GB/s)**	**% memory bandwidth utilized (GB/s)**
2	CPU vector addition	130.801	489.2929	3.8226	0.43%
3	CUDA vector addition: 1 thread, 1 thread block	6290.330287	10.1743	0.0795	0.01%
4	CUDA vector addition: 256 threads, 1 thread block	174.940367	365.8389	2.8581	0.32%
5	CUDA vector addition: 256 threads/block, many thread blocks	108.725287	588.6395	4.5987	0.51%
6	CUDA vector addition: 256 threads/block, many thread blocks, explicit data movement from host to GPU	0.600381	106598.9763	832.8045	92.74%
7		**kernel run time (seconds)**			
8		0.130801			
9		6.290330287			
10	**Formulas**	0.174940367			
11	MFLOP/s = #arithmetic operations / elapsed time in seconds / (1024*1024) # arithmetic ops = 2^26 = 67108864	0.108725287			
12		0.000600381			
13	**Memory bandwidth utilized** = bytes accessed / elapsed time in seconds (GB/s)				
14	**Total bytes accessed** = 2(read/write) * N * sizeof(float) = 2 * (2^26) * 4 bytes = 536870912				
15	**bytes to gb** = bytes/(1024*1024*1024) = 536870912/(1024*1024*1024) = 0.5				
16	**your rate** = bytes / time (second) | % of peak = your rate / theoretical rate				
17	**theoretical rate = Theoretical Bandwidth = 898GB/sec.** Stated by the nvidia cuda c++ Tesla V100 spec site: (0.877*10^9 *(4096/8)*2) / 10^9				
18					

MATHEMATICAL EQUATION REFERENCES

8.2.1. Theoretical Bandwidth Calculation

Theoretical bandwidth can be calculated using hardware specifications available in the product literature. For example, the NVIDIA Tesla V100 uses HBM2 (double data rate) RAM with a memory clock rate of 877 MHz and a 4096-bit-wide memory interface.

Using these data items, the peak theoretical memory bandwidth of the NVIDIA Tesla V100 is 898 GB/s:

$$(0.877 \times 10^9 \times (4096/8) \times 2) \div 10^9 = 898\text{GB/s}$$

In this calculation, the memory clock rate is converted in to Hz, multiplied by the interface width (divided by 8, to convert bits to bytes) and multiplied by 2 due to the double data rate. Finally, this product is divided by 10^9 to convert the result to GB/s.

Out of the Blocks

CUDA GPUs have many parallel processors grouped into Streaming Multiprocessors, or SMs. Each SM can run multiple concurrent thread blocks. As an example, a Tesla P100 GPU based on the Pascal GPU Architecture has 56 SMs, each capable of supporting up to 2048 active threads. To take full advantage of all these threads, I should launch the kernel with multiple thread blocks.

By now you may have guessed that the first parameter of the execution configuration specifies the number of thread blocks. Together, the blocks of parallel threads make up what is known as the *grid*. Since I have `N` elements to process, and 256 threads per block, I just need to calculate the number of blocks to get at least N threads. I simply divide `N` by the block size (being careful to round up in case `N` is not a multiple of `blockSize`).

```
int blockSize = 256;
int numBlocks = (N + blockSize - 1) / blockSize;
add<<<numBlocks, blockSize>>>(N, x, y);
```

gridDim.x = 4096

CODE RESULTS

CPU NODE ONLY

```
bdalit@perlmutter:login19:~/rsync/Downloads> nvcc vecadd_cpu.cpp -o vecadd_cpu
bdalit@perlmutter:login19:~/rsync/Downloads> ./vecadd_cpu
Elapsed time = 0.130801 seconds
Max error: 0
bdalit@perlmutter:login19:~/rsync/Downloads>
```

1 THREAD

```
gator@gatorbox: ~

Max error: 0
bdalit@perlmutter:login12:~/rsync/Downloads> nvcc vecadd_gpu_1t.cu -o vecadd_gpu_1t
bdalit@perlmutter:login12:~/rsync/Downloads> module load PrgEnv-nvidia

Lmod is automatically replacing "gcc/11.2.0" with "nvidia/22.7".


Lmod is automatically replacing "PrgEnv-gnu/8.3.3" with "PrgEnv-nvidia/8.3.3".


Due to MODULEPATH changes, the following have been reloaded:
  1) cray-mpich/8.1.25

bdalit@perlmutter:login12:~/rsync/Downloads> nvcc vecadd_gpu_1t.cu -o vecadd_gpu_1t
bdalit@perlmutter:login12:~/rsync/Downloads> salloc --nodes 1 --qos interactive --time 00:30:00 --constraint gpu --gpus 1 --account=m3930
salloc: Granted job allocation 8947750
salloc: Waiting for resource configuration
salloc: Nodes nid002409 are ready for job
bdalit@nid002409:~/rsync/Downloads> srun nsys nvprof ./vecadd_gpu_1t
WARNING: vecadd_gpu_1t and any of its children processes will be profiled.

Max error: 0
Generating '/tmp/nsys-report-e1fc.qdstrm'
[1/7] [========================100%] report1.nsys-rep
[2/7] [========================100%] report1.sqlite
SKIPPED: /global/u2/b/bdalit/rsync/Downloads/report1.sqlite does not contain NV Tools Extension (NVTX) data.
[3/7] Executing 'nvtxsum' stats report
[4/7] Executing 'cudaapisum' stats report

CUDA API Statistics:

 Time (%)  Total Time (ns)  Num Calls      Avg (ns)          Med (ns)         Min (ns)        Max (ns)       StdDev (ns)             Name
 --------  ---------------  ---------  ----------------  ----------------  -------------  -------------  ----------------  ----------------------
     92.3   76,235,793,506          2  38,117,896,753.0  38,117,896,753.0         65,186  76,235,728,320  53,906,754,370.3  cudaMallocManaged
      7.6    6,290,331,057          1   6,290,331,057.0   6,290,331,057.0  6,290,331,057   6,290,331,057               0.0  cudaDeviceSynchronize
      0.0       37,773,614          2      18,886,807.0      18,886,807.0     16,829,954      20,943,660       2,908,829.4  cudaFree
      0.0        1,535,567          1       1,535,567.0       1,535,567.0      1,535,567       1,535,567               0.0  cudaLaunchKernel
      0.0            1,202          1           1,202.0           1,202.0          1,202           1,202               0.0  cuModuleGetLoadingMode

[5/7] Executing 'gpukernsum' stats report

CUDA Kernel Statistics:

 Time (%)  Total Time (ns)  Instances     Avg (ns)         Med (ns)        Min (ns)       Max (ns)     StdDev (ns)            Name
 --------  ---------------  ---------  ---------------  ---------------  -------------  -------------  -----------  ------------------------
    100.0    6,290,330,287          1  6,290,330,287.0  6,290,330,287.0  6,290,330,287  6,290,330,287          0.0  add(int, float *, float *)

[6/7] Executing 'gpumemtimesum' stats report

CUDA Memory Operation Statistics (by time):

 Time (%)  Total Time (ns)  Count  Avg (ns)  Med (ns)  Min (ns)  Max (ns)  StdDev (ns)             Operation
 --------  ---------------  -----  --------  --------  --------  --------  -----------  --------------------------------
     67.2       26,075,683  3,072   8,488.2   3,711.0     2,014    41,280     11,230.6  [CUDA Unified Memory memcpy HtoD]
     32.8       12,724,077  1,536   8,283.9   2,895.5     1,535    42,016     11,444.0  [CUDA Unified Memory memcpy DtoH]

[7/7] Executing 'gpumemsizesum' stats report

CUDA Memory Operation Statistics (by size):

 Total (MB)  Count  Avg (MB)  Med (MB)  Min (MB)  Max (MB)  StdDev (MB)             Operation
 ----------  -----  --------  --------  --------  --------  -----------  --------------------------------
    536.871  3,072     0.175     0.033     0.004     1.044        0.301  [CUDA Unified Memory memcpy HtoD]
    268.435  1,536     0.175     0.033     0.004     1.044        0.301  [CUDA Unified Memory memcpy DtoH]

Generated:
    /global/u2/b/bdalit/rsync/Downloads/report1.nsys-rep
    /global/u2/b/bdalit/rsync/Downloads/report1.sqlite
bdalit@nid002409:~/rsync/Downloads>
```

256 THREAD

```
bdalit@nid002409:~/rsync/Downloads> nvcc vecadd_gpu_256t.cu -o vecadd_gpu_256t
bdalit@nid002409:~/rsync/Downloads> srun nsys nvprof ./vecadd_gpu_256t
WARNING: vecadd_gpu_256t and any of its children processes will be profiled.

Max error: 0
Generating '/tmp/nsys-report-184a.qdstrm'
[1/7] [========================100%] report2.nsys-rep
[2/7] [========================100%] report2.sqlite
SKIPPED: /global/u2/b/bdalit/rsync/Downloads/report2.sqlite does not contain NV Tools Extension (NVTX) data.
[3/7] Executing 'nvtxsum' stats report
[4/7] Executing 'cudaapisum' stats report

CUDA API Statistics:

 Time (%)  Total Time (ns)  Num Calls    Avg (ns)       Med (ns)      Min (ns)     Max (ns)     StdDev (ns)           Name
 --------  ---------------  ---------  -------------  -------------  -----------  -----------  -------------  ----------------------
     52.1      236,222,854          2  118,111,427.0  118,111,427.0       38,705  236,184,149  166,980,044.8  cudaMallocManaged
     38.6      174,938,680          1  174,938,680.0  174,938,680.0  174,938,680  174,938,680            0.0  cudaDeviceSynchronize
      9.3       41,997,057          2   20,998,528.5   20,998,528.5   19,546,019   22,451,038    2,054,158.6  cudaFree
      0.0           61,309          1       61,309.0       61,309.0       61,309       61,309            0.0  cudaLaunchKernel
      0.0              962          1          962.0          962.0          962          962            0.0  cuModuleGetLoadingMode

[5/7] Executing 'gpukernsum' stats report

CUDA Kernel Statistics:

 Time (%)  Total Time (ns)  Instances    Avg (ns)       Med (ns)      Min (ns)     Max (ns)    StdDev (ns)           Name
 --------  ---------------  ---------  -------------  -------------  -----------  -----------  -----------  -------------------------
    100.0      174,940,367          1  174,940,367.0  174,940,367.0  174,940,367  174,940,367          0.0  add(int, float *, float *)

[6/7] Executing 'gpumemtimesum' stats report

CUDA Memory Operation Statistics (by time):

 Time (%)  Total Time (ns)  Count  Avg (ns)  Med (ns)  Min (ns)  Max (ns)  StdDev (ns)              Operation
 --------  ---------------  -----  --------  --------  --------  --------  -----------  --------------------------------
     67.2       26,103,545  3,072   8,497.2   3,503.5     2,015    43,584     11,267.2  [CUDA Unified Memory memcpy HtoD]
     32.8       12,716,363  1,536   8,278.9   2,895.5     1,534    41,952     11,439.5  [CUDA Unified Memory memcpy DtoH]

[7/7] Executing 'gpumemsizesum' stats report

CUDA Memory Operation Statistics (by size):

 Total (MB)  Count  Avg (MB)  Med (MB)  Min (MB)  Max (MB)  StdDev (MB)              Operation
 ----------  -----  --------  --------  --------  --------  -----------  --------------------------------
    536.871  3,072     0.175     0.033     0.004     1.044        0.301  [CUDA Unified Memory memcpy HtoD]
    268.435  1,536     0.175     0.033     0.004     1.044        0.301  [CUDA Unified Memory memcpy DtoH]

Generated:
    /global/u2/b/bdalit/rsync/Downloads/report2.nsys-rep
    /global/u2/b/bdalit/rsync/Downloads/report2.sqlite
bdalit@nid002409:~/rsync/Downloads>
```

256 THREAD BLOCK

```
bdalit@nid002409:~/rsync/Downloads> nvcc vecadd_gpu_256t_manyblocks.cu -o vecadd_gpu_256t_manyblocks
bdalit@nid002409:~/rsync/Downloads> srun nsys nvprof ./vecadd_gpu_256t_manyblocks
WARNING: vecadd_gpu_256t_manyblocks and any of its children processes will be profiled.

Max error: 0
Generating '/tmp/nsys-report-6497.qdstrm'
[1/7] [========================100%] report3.nsys-rep
[2/7] [========================100%] report3.sqlite
SKIPPED: /global/u2/b/bdalit/rsync/Downloads/report3.sqlite does not contain NV Tools Extension (NVTX) data.
[3/7] Executing 'nvtxsum' stats report
[4/7] Executing 'cudaapisum' stats report

CUDA API Statistics:

 Time (%)  Total Time (ns)  Num Calls    Avg (ns)       Med (ns)      Min (ns)     Max (ns)     StdDev (ns)            Name
 --------  ---------------  ---------  -------------  -------------  -----------  -----------  -------------  ----------------------
     60.0      232,556,464          2  116,278,232.0  116,278,232.0       35,880  232,520,584  164,391,510.7  cudaMallocManaged
     28.0      108,720,689          1  108,720,689.0  108,720,689.0  108,720,689  108,720,689            0.0  cudaDeviceSynchronize
     11.9       46,297,712          2   23,148,856.0   23,148,856.0   22,200,273   24,097,439    1,341,498.9  cudaFree
      0.0           59,556          1       59,556.0       59,556.0       59,556       59,556            0.0  cudaLaunchKernel
      0.0            1,133          1        1,133.0        1,133.0        1,133        1,133            0.0  cuModuleGetLoadingMode

[5/7] Executing 'gpukernsum' stats report

CUDA Kernel Statistics:

 Time (%)  Total Time (ns)  Instances    Avg (ns)       Med (ns)      Min (ns)     Max (ns)    StdDev (ns)            Name
 --------  ---------------  ---------  -------------  -------------  -----------  -----------  -----------  -------------------------
    100.0      108,725,287          1  108,725,287.0  108,725,287.0  108,725,287  108,725,287          0.0  add(int, float *, float *)

[6/7] Executing 'gpumemtimesum' stats report

CUDA Memory Operation Statistics (by time):

 Time (%)  Total Time (ns)  Count  Avg (ns)  Med (ns)  Min (ns)  Max (ns)  StdDev (ns)              Operation
 --------  ---------------  -----  --------  --------  --------  --------  -----------  --------------------------------
     65.7       24,277,578  8,191   2,963.9   2,207.0     2,014    40,960      3,866.6  [CUDA Unified Memory memcpy HtoD]
     34.3       12,700,908  1,530   8,301.2   2,831.5     1,535    42,047     11,459.1  [CUDA Unified Memory memcpy DtoH]

[7/7] Executing 'gpumemsizesum' stats report

CUDA Memory Operation Statistics (by size):

 Total (MB)  Count  Avg (MB)  Med (MB)  Min (MB)  Max (MB)  StdDev (MB)              Operation
 ----------  -----  --------  --------  --------  --------  -----------  --------------------------------
    268.173  1,530     0.175     0.033     0.004     1.044        0.301  [CUDA Unified Memory memcpy DtoH]
    226.918  8,191     0.028     0.008     0.004     1.044        0.103  [CUDA Unified Memory memcpy HtoD]

Generated:
    /global/u2/b/bdalit/rsync/Downloads/report3.nsys-rep
    /global/u2/b/bdalit/rsync/Downloads/report3.sqlite
bdalit@nid002409:~/rsync/Downloads>
```

256 THREAD BLOCK PREFETCH

```
bdalit@nid002409:~/rsync/Downloads> nvcc vecadd_gpu_256t_manyblocks_prefetch.cu -o vecadd_gpu_256t_manyblocks_prefetch
bdalit@nid002409:~/rsync/Downloads> srun nsys nvprof ./vecadd_gpu_256t_manyblocks_prefetch
WARNING: vecadd_gpu_256t_manyblocks_prefetch and any of its children processes will be profiled.

Max error: 0
Generating '/tmp/nsys-report-2a12.qdstrm'
[1/7] [========================100%] report4.nsys-rep
[2/7] [========================100%] report4.sqlite
SKIPPED: /global/u2/b/bdalit/rsync/Downloads/report4.sqlite does not contain NV Tools Extension (NVTX) data.
[3/7] Executing 'nvtxsum' stats report
[4/7] Executing 'cudaapisum' stats report

CUDA API Statistics:

 Time (%)  Total Time (ns)  Num Calls    Avg (ns)       Med (ns)      Min (ns)    Max (ns)     StdDev (ns)           Name
 --------  ---------------  ---------  -------------  -------------  ----------  -----------  -------------  ----------------------
     71.5      206,426,554          2  103,213,277.0  103,213,277.0      33,004  206,393,550  145,918,941.4  cudaMallocManaged
     14.7       42,451,215          2   21,225,607.5   21,225,607.5  19,599,803   22,851,412    2,299,234.8  cudaFree
      7.5       21,604,807          1   21,604,807.0   21,604,807.0  21,604,807   21,604,807            0.0  cudaDeviceSynchronize
      6.3       18,157,880          2    9,078,940.0    9,078,940.0     277,698   17,880,182   12,446,835.8  cudaMemPrefetchAsync
      0.0           47,682          1       47,682.0       47,682.0      47,682       47,682            0.0  cudaLaunchKernel
      0.0            1,273          1        1,273.0        1,273.0       1,273        1,273            0.0  cuModuleGetLoadingMode

[5/7] Executing 'gpukernsum' stats report

CUDA Kernel Statistics:

 Time (%)  Total Time (ns)  Instances  Avg (ns)   Med (ns)   Min (ns)  Max (ns)  StdDev (ns)           Name
 --------  ---------------  ---------  ---------  ---------  --------  --------  -----------  -------------------------
    100.0          600,381          1  600,381.0  600,381.0   600,381   600,381          0.0  add(int, float *, float *)

[6/7] Executing 'gpumemtimesum' stats report

CUDA Memory Operation Statistics (by time):

 Time (%)  Total Time (ns)  Count  Avg (ns)  Med (ns)  Min (ns)  Max (ns)  StdDev (ns)             Operation
 --------  ---------------  -----  --------  --------  --------  --------  -----------  --------------------------------
     61.9       20,660,815    256  80,706.3  80,640.0    80,575    85,408        370.6  [CUDA Unified Memory memcpy HtoD]
     38.1       12,732,003  1,536   8,289.1   2,831.5     1,535    42,016     11,442.8  [CUDA Unified Memory memcpy DtoH]

[7/7] Executing 'gpumemsizesum' stats report

CUDA Memory Operation Statistics (by size):

 Total (MB)  Count  Avg (MB)  Med (MB)  Min (MB)  Max (MB)  StdDev (MB)             Operation
 ----------  -----  --------  --------  --------  --------  -----------  --------------------------------
    536.871    256     2.097     2.097     2.097     2.097        0.000  [CUDA Unified Memory memcpy HtoD]
    268.435  1,536     0.175     0.033     0.004     1.044        0.301  [CUDA Unified Memory memcpy DtoH]

Generated:
    /global/u2/b/bdalit/rsync/Downloads/report4.nsys-rep
    /global/u2/b/bdalit/rsync/Downloads/report4.sqlite
bdalit@nid002409:~/rsync/Downloads>
```