Profiling Matrix Multiplication using *gprof, Valgrind, nvprof*

Naveen Himthani (120010001)

Observations:

N = 1000

A) openMP

1) Cache Hit/Miss Rate (using valgrind)

```
==240326==
==240326== Events    : Ir Dr Dw I1mr D1mr D1mw ILmr DLmr DLmw
==240326== Collected : 49058836175 20020018233 2006257541 1491
1190205704 203467 1404 2503 190395
==240326==
==240326== I   refs:      49,058,836,175
==240326== I1  misses:             1,491
==240326== LLi misses:             1,404
==240326== I1  miss rate:           0.0%
==240326== LLi miss rate:           0.0%
==240326==
==240326== D   refs:      22,026,275,774  (20,020,018,233 rd + 2,006,257,541 wr)
==240326== D1  misses:     1,190,409,171  ( 1,190,205,704 rd +       203,467 wr)
==240326== LLd misses:           192,898  (         2,503 rd +       190,395 wr)
==240326== D1  miss rate:            5.4% (           5.9%   +         0.0%  )
==240326== LLd miss rate:            0.0% (           0.0%   +         0.0%  )
==240326==
==240326== LL refs:        1,190,410,662  ( 1,190,207,195 rd +       203,467 wr)
==240326== LL misses:            194,302  (         3,907 rd +       190,395 wr)
==240326== LL miss rate:             0.0% (           0.0%   +         0.0%  )
```

2) Flat profile (using gprof)

```
Each sample counts as 0.01 seconds.
  %   cumulative   self              self     total
 time   seconds   seconds    calls  ms/call  ms/call  name
 100.20      8.67     8.67       20   433.38   433.38  matrix_multiply
  0.12      8.68     0.01       14     0.72     0.72  init_pattern_matrices_omp
  0.00      8.68     0.00        3     0.00     0.00  create_matrix
  0.00      8.68     0.00        3     0.00     0.00  destroy_matrix
```

3) Call graph

```
granularity: each sample hit covers 2 byte(s) for 0.12% of 8.68 seconds

index % time    self  children    called     name
                                             <spontaneous>
[1]    100.0    0.00    8.68                 main [1]
                8.67    0.00      20/20        matrix_multiply [2]
                0.01    0.00      14/14        init_pattern_matrices_omp [3]
                0.00    0.00       3/3         create_matrix [4]
                0.00    0.00       3/3         destroy_matrix [5]
-----------------------------------------------
                8.67    0.00      20/20        main [1]
[2]     99.9    8.67    0.00      20           matrix_multiply [2]
-----------------------------------------------
                0.01    0.00      14/14        main [1]
[3]      0.1    0.01    0.00      14           init_pattern_matrices_omp [3]
-----------------------------------------------
                0.00    0.00       3/3         main [1]
[4]      0.0    0.00    0.00       3           create_matrix [4]
-----------------------------------------------
                0.00    0.00       3/3         main [1]
[5]      0.0    0.00    0.00       3           destroy_matrix [5]
-----------------------------------------------
```

B) MPI (np = 4)

1) Cache Hit/misses

```
==241786==
==241786== Events    : Ir Dr Dw I1mr D1mr D1mw ILmr DLmr DLmw
==241786== Collected : 67595720 15670579 12294533 83245 160287 39417 8753
11341 14879
==241786==
==241786== I   refs:      67,595,720
==241786== I1  misses:        83,245
==241786== LLi misses:         8,753
==241786== I1  miss rate:       0.12%
==241786== LLi miss rate:       0.1%
==241786==
==241786== D   refs:      27,965,112  (15,670,579 rd + 12,294,533 wr)
==241786== D1  misses:       199,704  (   160,287 rd +     39,417 wr)
==241786== LLd misses:        26,220  (    11,341 rd +     14,879 wr)
==241786== D1  miss rate:        0.7% (       1.0%   +       0.3%  )
==241786== LLd miss rate:        0.0% (       0.0%   +       0.1%  )
==241786==
==241786== LL refs:          282,949  (   243,532 rd +     39,417 wr)
==241786== LL misses:         34,973  (    20,094 rd +     14,879 wr)
==241786== LL miss rate:         0.0% (       0.0%   +       0.1%  )
```

2) Flat Profile

Each sample counts as 0.01 seconds.

% time	cumulative seconds	self seconds	calls	self s/call	total s/call	name
97.48	1.39	1.39	1	1.39	1.39	matrix_multiply

3) Call Graph

granularity: each sample hit covers 2 byte(s) for 0.72% of 1.39 seconds

```
index % time    self  children    called     name
                1.39    0.00       1/1          main [2]
[1]    100.0    1.39    0.00       1          matrix_multiply [1]
-----------------------------------------------
                                              <spontaneous>
[2]    100.0    0.00    1.39                 main [2]
                1.39    0.00       1/1          matrix_multiply [1]
-----------------------------------------------
```

C) CUDA (Algorithm - 1 in the code)

1) Profile (using nvprof)

```
==242922== Profiling application: ./a.out
==242922== Profiling result:
Time(%)      Time     Calls       Avg       Min       Max  Name
 99.82%  1.65479s      1000  1.6548ms  1.6444ms  1.6731ms  matmul_one(float*, float*, float*, int)
  0.12%  1.9696ms         2  984.79us  983.94us  985.64us  [CUDA memcpy HtoD]
  0.06%  933.57us         1  933.57us  933.57us  933.57us  [CUDA memcpy DtoH]

==242922== API calls:
Time(%)      Time     Calls       Avg       Min       Max  Name
 84.80%  1.65452s         3  551.51ms  1.3727ms  1.65174s  cudaMemcpy
 14.55%  283.87ms         3  94.623ms  138.45us  283.57ms  cudaMalloc
  0.50%  9.7111ms      1000  9.7110us  8.3990us  73.152us  cudaLaunch
  0.05%  1.0359ms      4000     258ns     164ns  8.4080us  cudaSetupArgument
  0.05%  920.28us       166  5.5430us     302ns  205.60us  cuDeviceGetAttribute
  0.02%  447.79us      1000     447ns     346ns  7.7670us  cudaConfigureCall
  0.02%  432.99us         3  144.33us  102.51us  221.97us  cudaFree
  0.01%  130.42us         2  65.211us  56.086us  74.336us  cuDeviceTotalMem
  0.01%  115.33us         2  57.663us  45.661us  69.665us  cuDeviceGetName
  0.00%  4.9870us         2  2.4930us     592ns  4.3950us  cuDeviceGetCount
  0.00%  3.9270us         4     981ns     329ns  2.1860us  cuDeviceGet
```

2) Cache hit/misses

```
==242939==
==242939== Events    : Ir Dr Dw I1mr D1mr D1mw ILmr DLmr DLmw
==242939== Collected : 89806721 30323057 16174554 496114 822350 543024
7402 66721 337520
==242939==
==242939== I   refs:      89,806,721
==242939== I1  misses:       496,114
==242939== LLi misses:         7,402
==242939== I1  miss rate:       0.55%
==242939== LLi miss rate:       0.0%
==242939==
==242939== D   refs:      46,497,611  (30,323,057 rd + 16,174,554 wr)
==242939== D1  misses:     1,365,374  (   822,350 rd +    543,024 wr)
==242939== LLd misses:       404,241  (    66,721 rd +    337,520 wr)
==242939== D1  miss rate:        2.9% (       2.7%   +        3.3%  )
==242939== LLd miss rate:        0.8% (       0.2%   +        2.0%  )
==242939==
==242939== LL refs:        1,861,488  ( 1,318,464 rd +    543,024 wr)
==242939== LL misses:        411,643  (    74,123 rd +    337,520 wr)
==242939== LL miss rate:         0.3% (       0.0%   +        2.0%  )
```

3) Call Graph (using gprof)

granularity: each sample hit covers 2 byte(s) for 100.00% of 0.01 seconds

```
index % time    self  children    called     name
                                             <spontaneous>
[1]    100.0    0.01    0.00                 main [1]
                0.00    0.00       1/1          matmul_caller_one(float*, float*, float*, float*,
float*, float*, int) [217]
-----------------------------------------------
                0.00    0.00    2000/2000       matmul_caller_one(float*, float*, float*,
float*, float*, float*, int) [217]
[213]    0.0    0.00    0.00    2000         dim3::dim3(unsigned int, unsigned int,
unsigned int) [213]
-----------------------------------------------
                0.00    0.00    1000/1000
__device_stub__Z10matmul_onePfS_S_i(float*, float*, float*, int) [216]
[214]    0.0    0.00    0.00    1000         cudaError cudaLaunch<char>(char*) [214]
-----------------------------------------------
                0.00    0.00    1000/1000       matmul_caller_one(float*, float*, float*,
float*, float*, float*, int) [217]
[215]    0.0    0.00    0.00    1000         matmul_one(float*, float*, float*, int) [215]
                0.00    0.00    1000/1000
__device_stub__Z10matmul_onePfS_S_i(float*, float*, float*, int) [216]
-----------------------------------------------
                0.00    0.00    1000/1000       matmul_one(float*, float*, float*, int) [215]
[216]    0.0    0.00    0.00    1000
__device_stub__Z10matmul_onePfS_S_i(float*, float*, float*, int) [216]
                0.00    0.00    1000/1000       cudaError cudaLaunch<char>(char*) [214]
-----------------------------------------------
                0.00    0.00       1/1          main [1]
```

```
[217]    0.0    0.00    0.00       1        matmul_caller_one(float*, float*, float*, float*,
float*, float*, int) [217]
             0.00    0.00    2000/2000       dim3::dim3(unsigned int, unsigned int,
unsigned int) [213]
             0.00    0.00    1000/1000       matmul_one(float*, float*, float*, int) [215]
             0.00    0.00       1/1          std::ceil(float) [222]
-----------------------------------------------
             0.00    0.00       1/1          __cudaUnregisterBinaryUtil() [256]
[218]    0.0    0.00    0.00       1        ____nv_dummy_param_ref(void*) [218]
-----------------------------------------------
             0.00    0.00       1/1
__sti____cudaRegisterAll_46_tmpxft_0003b494_00000000_9_matmul_cuda_cpp1_
ii_main() [221]
[219]    0.0    0.00    0.00       1        __nv_cudaEntityRegisterCallback(void**) [219]
             0.00    0.00       1/1
__nv_save_fatbinhandle_for_managed_rt(void**) [220]
-----------------------------------------------
             0.00    0.00       1/1          __nv_cudaEntityRegisterCallback(void**) [219]
[220]    0.0    0.00    0.00       1
__nv_save_fatbinhandle_for_managed_rt(void**) [220]
-----------------------------------------------
             0.00    0.00       1/1          __libc_csu_init [1048]
[221]    0.0    0.00    0.00       1
__sti____cudaRegisterAll_46_tmpxft_0003b494_00000000_9_matmul_cuda_cpp1_
ii_main() [221]
             0.00    0.00       1/1          __nv_cudaEntityRegisterCallback(void**) [219]
-----------------------------------------------
             0.00    0.00       1/1          matmul_caller_one(float*, float*, float*, float*,
float*, float*, int) [217]
[222]    0.0    0.00    0.00       1        std::ceil(float) [222]
```

Performance Analysis:

Already submitted in previous reports in the form of time taken for various matrix dimensions