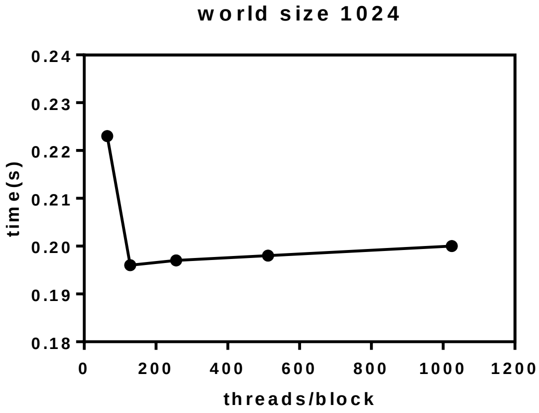
Parallel Performance Analysis and Report

1.

For the world size 1024*1024, the plot of execution time versus thread blocksize is shown below.



The optimal number of threads per block is 128.

The results from an **nvprof** run is:

For 64 threads per block:

```
This is the Game of Life running in parallel on a GPU.
==97052== NVPROF is profiling process 97052, command: ./gol 3 1024 1024 64 0
==97052== Profiling application: ./gol 3 1024 1024 64 0
==97052== Profiling result:
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:  100.00%  33.793ms      1024  33.000us  31.647us  1.3774ms  gol_kernel(unsigned char const *, unsigned int,
unsigned int, unsigned char*)
      API calls:   85.05%  196.96ms         2  98.479ms  13.789us  196.95ms  cudaMallocManaged
                   10.34%  23.945ms         1  23.945ms  23.945ms  23.945ms  cudaDeviceSynchronize
                    3.84%  8.9015ms      1024  8.6920us  7.9860us  33.161us  cudaLaunchKernel
                    0.37%  853.32us         1  853.32us  853.32us  853.32us  cuDeviceTotalMem
                    0.28%  638.27us        97  6.5800us     226ns  240.93us  cuDeviceGetAttribute
                    0.10%  231.97us         2  115.98us  33.362us  198.61us  cudaFree
                    0.02%  52.349us         1  52.349us  52.349us  52.349us  cuDeviceGetName
                    0.00%  4.2500us         1  4.2500us  4.2500us  4.2500us  cuDeviceGetPCIBusId
                    0.00%  2.4670us         3     822ns     464ns  1.4530us  cuDeviceGetCount
                    0.00%  1.0310us         2     515ns     348ns     683ns  cuDeviceGet
                    0.00%     375ns         1     375ns     375ns     375ns  cuDeviceGetUuid
```

The time for running the gol_kernel function is 33.793ms comparing to 223ms total runtime.

The cudaMallocManaged function takes up the highest proportion of the total runtime, which is 85.05%.

The cudaDeviceSynchronize function takes up 10.34%.

The cudaLaunchKernel function takes up 3.84%.

For 128 threads per block:

```
This is the Game of Life running in parallel on a GPU.
==97130== NVPROF is profiling process 97130, command: ./gol 3 1024 1024 128 0
==97130== Profiling application: ./gol 3 1024 1024 128 0
==97130== Profiling result:
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:  100.00%  28.292ms      1024  27.628us  26.080us  1.5042ms  gol_kernel(unsigned char const *, unsigned int, 
unsigned int, unsigned char*)
      API calls:   88.71%  226.30ms         2  113.15ms  20.315us  226.28ms  cudaMallocManaged
                    6.25%  15.936ms      1024  15.562us  14.542us  57.211us  cudaLaunchKernel
                    4.25%  10.842ms         1  10.842ms  10.842ms  10.842ms  cudaDeviceSynchronize
                    0.36%  929.65us         1  929.65us  929.65us  929.65us  cuDeviceTotalMem
                    0.31%  798.55us        97  8.2320us     289ns  308.34us  cuDeviceGetAttribute
                    0.09%  221.95us         2  110.98us  34.992us  186.96us  cudaFree
                    0.02%  62.001us         1  62.001us  62.001us  62.001us  cuDeviceGetName
                    0.00%  3.2560us         1  3.2560us  3.2560us  3.2560us  cuDeviceGetPCIBusId
                    0.00%  2.6560us         3     885ns     529ns  1.5340us  cuDeviceGetCount
                    0.00%  1.0990us         2     549ns     350ns     749ns  cuDeviceGet
                    0.00%     502ns         1     502ns     502ns     502ns  cuDeviceGetUuid
```

The time for running the gol_kernel function is 28.292ms.

The cudaMallocManaged function takes up the highest proportion of the total runtime, which is 88.71%.

The cudaDeviceSynchronize function takes up 4.25%.

For 256 threads per block:

```
This is the Game of Life running in parallel on a GPU.
==97367== NVPROF is profiling process 97367, command: ./gol 3 1024 1024 256 0
==97367== Profiling application: ./gol 3 1024 1024 256 0
==97367== Profiling result:
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:  100.00%  28.089ms      1024  27.431us  26.080us  1.3238ms  gol_kernel(unsigned char const *, unsigned int, 
unsigned int, unsigned char*)
      API calls:   87.08%  199.98ms         2  99.990ms  12.017us  199.97ms  cudaMallocManaged
                    8.41%  19.323ms         1  19.323ms  19.323ms  19.323ms  cudaDeviceSynchronize
                    3.43%  7.8701ms      1024  7.6850us  7.0810us  29.645us  cudaLaunchKernel
                    0.58%  1.3342ms         1  1.3342ms  1.3342ms  1.3342ms  cuDeviceTotalMem
                    0.38%  865.75us        97  8.9250us     355ns  331.56us  cuDeviceGetAttribute
                    0.09%  205.29us         2  102.65us  33.151us  172.14us  cudaFree
                    0.03%  70.001us         1  70.001us  70.001us  70.001us  cuDeviceGetName
                    0.00%  3.6350us         1  3.6350us  3.6350us  3.6350us  cuDeviceGetPCIBusId
                    0.00%  3.2570us         3  1.0850us     623ns  1.8470us  cuDeviceGetCount
                    0.00%  1.4820us         2     741ns     564ns     918ns  cuDeviceGet
                    0.00%     692ns         1     692ns     692ns     692ns  cuDeviceGetUuid
```

The time for running the gol_kernel function is 28.089ms.

The cudaMallocManaged function takes up the highest proportion of the total runtime.

And also we can find that the cudaDeviceSynchronize function takes up larger proportion compared to the smaller threads number (128 threads).

For 512 threads per block:

```
This is the Game of Life running in parallel on a GPU.
==97478== NVPROF is profiling process 97478, command: ./gol 3 1024 1024 512 0
==97478== Profiling application: ./gol 3 1024 1024 512 0
==97478== Profiling result:
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:  100.00%  28.088ms      1024  27.429us  25.983us  1.4054ms  gol_kernel(unsigned char const *, unsigned int,
unsigned int, unsigned char*)
      API calls:   87.15%  199.40ms         2  99.699ms  17.242us  199.38ms  cudaMallocManaged
                    7.01%  16.045ms         1  16.045ms  16.045ms  16.045ms  cudaDeviceSynchronize
                    4.76%  10.890ms      1024  10.634us  9.8320us  36.366us  cudaLaunchKernel
                    0.59%  1.3408ms         1  1.3408ms  1.3408ms  1.3408ms  cuDeviceTotalMem
                    0.38%  858.08us        97  8.8460us     359ns  329.86us  cuDeviceGetAttribute
                    0.08%  188.52us         2  94.257us  29.212us  159.30us  cudaFree
                    0.03%  70.783us         1  70.783us  70.783us  70.783us  cuDeviceGetName
                    0.00%  3.4080us         1  3.4080us  3.4080us  3.4080us  cuDeviceGetPCIBusId
                    0.00%  3.2870us         3  1.0950us     650ns  1.8340us  cuDeviceGetCount
                    0.00%  1.4440us         2     722ns     484ns     960ns  cuDeviceGet
                    0.00%     691ns         1     691ns     691ns     691ns  cuDeviceGetUuid
```

The time for running the gol_kernel function is 28.088ms.

The cudaMallocManaged function takes up the highest proportion of the total runtime.

The cudaDeviceSynchronize function takes up larger proportion compared to the smaller threads number (128 threads).

For 1024 threads per block:

```
This is the Game of Life running in parallel on a GPU.
==97661== NVPROF is profiling process 97661, command: ./gol 3 1024 1024 1024 0
==97661== Profiling application: ./gol 3 1024 1024 1024 0
==97661== Profiling result:
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:  100.00%  28.880ms      1024  28.203us  26.816us  1.2341ms  gol_kernel(unsigned char const *, unsigned int,
unsigned int, unsigned char*)
      API calls:   86.73%  198.66ms         2  99.331ms  13.556us  198.65ms  cudaMallocManaged
                    8.18%  18.744ms         1  18.744ms  18.744ms  18.744ms  cudaDeviceSynchronize
                    3.98%  9.1232ms      1024  8.9090us  8.1660us  32.823us  cudaLaunchKernel
                    0.60%  1.3639ms         1  1.3639ms  1.3639ms  1.3639ms  cuDeviceTotalMem
                    0.37%  848.90us        97  8.7510us     353ns  325.82us  cuDeviceGetAttribute
                    0.10%  230.96us         2  115.48us  33.672us  197.29us  cudaFree
                    0.03%  69.772us         1  69.772us  69.772us  69.772us  cuDeviceGetName
                    0.00%  3.3470us         3  1.1150us     641ns  1.9040us  cuDeviceGetCount
                    0.00%  2.9060us         1  2.9060us  2.9060us  2.9060us  cuDeviceGetPCIBusId
                    0.00%  1.4580us         2     729ns     505ns     953ns  cuDeviceGet
                    0.00%     653ns         1     653ns     653ns     653ns  cuDeviceGetUuid
```

The time for running the gol_kernel function is 28.880ms.

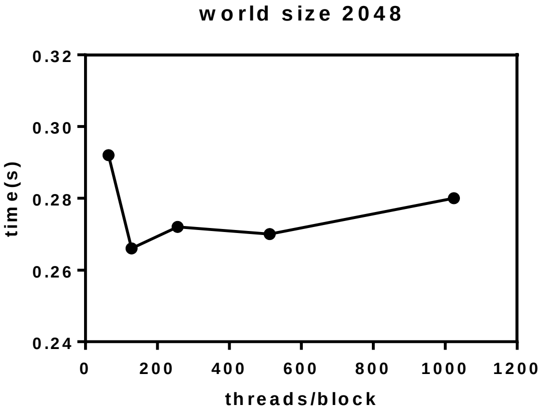
The cudaMallocManaged function takes up the highest proportion of the total runtime, which is 86.73%.

The cudaDeviceSynchronize function takes up larger proportion compared to the smaller threads number (128 threads).

The number of threads is 128 yielding the fastest execution time. When the number of threads goes higher, the time spent in the synchronization is increased, slowing down the updating rate.

2.

For the world size 2048*2048, the plot of execution time versus thread blocksize is shown below.



The optimal number of threads per block is 128.

The output from an **nvprof** run is:

For 64 threads per block:

```
This is the Game of Life running in parallel on a GPU.
==98017== NVPROF is profiling process 98017, command: ./gol 3 2048 1024 64 0
==98017== Profiling application: ./gol 3 2048 1024 64 0
==98017== Profiling result:
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:  100.00%  114.14ms      1024  111.47us  97.343us  3.6803ms  gol_kernel(unsigned char const *, unsigned int, 
unsigned int, unsigned char*)
      API calls:   63.23%  198.90ms         2  99.452ms  50.534us  198.85ms  cudaMallocManaged
                   32.30%  101.62ms         1  101.62ms  101.62ms  101.62ms  cudaDeviceSynchronize
                    3.61%  11.371ms      1024  11.104us  10.549us  41.283us  cudaLaunchKernel
                    0.43%  1.3378ms         1  1.3378ms  1.3378ms  1.3378ms  cuDeviceTotalMem
                    0.27%  857.99us        97  8.8450us     355ns  330.53us  cuDeviceGetAttribute
                    0.12%  385.85us         2  192.93us  184.14us  201.71us  cudaFree
                    0.02%  73.073us         1  73.073us  73.073us  73.073us  cuDeviceGetName
                    0.00%  3.7720us         1  3.7720us  3.7720us  3.7720us  cuDeviceGetPCIBusId
                    0.00%  3.4460us         3  1.1480us     836ns  1.7620us  cuDeviceGetCount
                    0.00%  1.3360us         2     668ns     489ns     847ns  cuDeviceGet
                    0.00%     605ns         1     605ns     605ns     605ns  cuDeviceGetUuid
```

The time for running the gol_kernel function is 114.14ms.

The cudaMallocManaged function takes up the highest proportion of the total runtime, which is 63.23%.

The cudaDeviceSynchronize function takes up 32.30% of the total runtime.

For 128 threads per block:

```
This is the Game of Life running in parallel on a GPU.
==98084== NVPROF is profiling process 98084, command: ./gol 3 2048 1024 128 0
==98084== Profiling application: ./gol 3 2048 1024 128 0
==98084== Profiling result:
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:  100.00%  97.028ms      1024  94.753us  90.784us  3.9262ms  gol_kernel(unsigned char const *, unsigned int,
unsigned int, unsigned char*)
      API calls:   67.02%  200.35ms         2  100.18ms  48.887us  200.31ms  cudaMallocManaged
                   28.41%  84.935ms         1  84.935ms  84.935ms  84.935ms  cudaDeviceSynchronize
                    3.67%  10.982ms      1024  10.724us  10.173us  40.981us  cudaLaunchKernel
                    0.46%  1.3666ms         1  1.3666ms  1.3666ms  1.3666ms  cuDeviceTotalMem
                    0.29%  871.63us        97  8.9850us     357ns  334.47us  cuDeviceGetAttribute
                    0.12%  362.43us         2  181.21us  179.67us  182.76us  cudaFree
                    0.02%  71.282us         1  71.282us  71.282us  71.282us  cuDeviceGetName
                    0.00%  3.9320us         3  1.3100us     666ns  2.3670us  cuDeviceGetCount
                    0.00%  3.6930us         1  3.6930us  3.6930us  3.6930us  cuDeviceGetPCIBusId
                    0.00%  1.4550us         2     727ns     500ns     955ns  cuDeviceGet
                    0.00%     662ns         1     662ns     662ns     662ns  cuDeviceGetUuid

==98084== Unified Memory profiling result:
Device "Tesla V100-SXM2-16GB (0)"
   Count  Avg Size  Min Size  Max Size  Total Size  Total Time  Name
      39  210.05KB  64.000KB  960.00KB  8.000000MB  363.8400us  Host To Device
      21         -         -         -           -  3.832832ms  Gpu page fault groups
Total CPU Page faults: 24
======== Error: Application returned non-zero code 1
```

For 256 threads per block:

```
This is the Game of Life running in parallel on a GPU.
==98268== NVPROF is profiling process 98268, command: ./gol 3 2048 1024 256 0
==98268== Profiling application: ./gol 3 2048 1024 256 0
==98268== Profiling result:
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:  100.00%  97.081ms      1024  94.805us  90.719us  4.0251ms  gol_kernel(unsigned char const *, unsigned int,
unsigned int, unsigned char*)
      API calls:   66.77%  198.42ms         2  99.210ms  40.409us  198.38ms  cudaMallocManaged
                   29.40%  87.359ms         1  87.359ms  87.359ms  87.359ms  cudaDeviceSynchronize
                    2.94%  8.7348ms      1024  8.5300us  7.9960us  35.460us  cudaLaunchKernel
                    0.45%  1.3427ms         1  1.3427ms  1.3427ms  1.3427ms  cuDeviceTotalMem
                    0.29%  860.18us        97  8.8670us     358ns  331.04us  cuDeviceGetAttribute
                    0.13%  387.91us         2  193.96us  181.45us  206.47us  cudaFree
                    0.02%  71.005us         1  71.005us  71.005us  71.005us  cuDeviceGetName
                    0.00%  4.2910us         1  4.2910us  4.2910us  4.2910us  cuDeviceGetPCIBusId
                    0.00%  3.5470us         3  1.1820us     664ns  2.0390us  cuDeviceGetCount
                    0.00%  1.5380us         2     769ns     518ns  1.0200us  cuDeviceGet
                    0.00%     589ns         1     589ns     589ns     589ns  cuDeviceGetUuid

==98268== Unified Memory profiling result:
Device "Tesla V100-SXM2-16GB (0)"
   Count  Avg Size  Min Size  Max Size  Total Size  Total Time  Name
      39  210.05KB  64.000KB  960.00KB  8.000000MB  358.8480us  Host To Device
      21         -         -         -           -  3.942560ms  Gpu page fault groups
Total CPU Page faults: 24
======== Error: Application returned non-zero code 1
```

For 512 threads per block:

```
This is the Game of Life running in parallel on a GPU.
==98332== NVPROF is profiling process 98332, command: ./gol 3 2048 1024 512 0
==98332== Profiling application: ./gol 3 2048 1024 512 0
==98332== Profiling result:
```

```
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:  100.00%  98.726ms      1024  96.412us  92.256us  4.0143ms  gol_kernel(unsigned char const *, unsigned int,
unsigned int, unsigned char*)
      API calls:   66.10%  194.47ms         2  97.235ms  38.106us  194.43ms  cudaMallocManaged
                   30.66%  90.207ms         1  90.207ms  90.207ms  90.207ms  cudaDeviceSynchronize
                    2.58%  7.6008ms      1024  7.4220us  6.9200us  32.407us  cudaLaunchKernel
                    0.29%  865.00us         1  865.00us  865.00us  865.00us  cuDeviceTotalMem
                    0.21%  622.86us        97  6.4210us     228ns  239.94us  cuDeviceGetAttribute
                    0.13%  390.65us         2  195.32us  184.31us  206.33us  cudaFree
                    0.02%  50.356us         1  50.356us  50.356us  50.356us  cuDeviceGetName
                    0.00%  3.1190us         1  3.1190us  3.1190us  3.1190us  cuDeviceGetPCIBusId
                    0.00%  2.3060us         3     768ns     427ns  1.2720us  cuDeviceGetCount
                    0.00%     838ns         2     419ns     308ns     530ns  cuDeviceGet
                    0.00%     469ns         1     469ns     469ns     469ns  cuDeviceGetUuid

==98332== Unified Memory profiling result:
Device "Tesla V100-SXM2-16GB (0)"
   Count  Avg Size  Min Size  Max Size  Total Size  Total Time  Name
      40  204.80KB  64.000KB  960.00KB  8.000000MB  364.1600us  Host To Device
      21         -         -         -           -  3.921632ms  Gpu page fault groups
Total CPU Page faults: 24
======== Error: Application returned non-zero code 1
```

For 1024 threads per block:

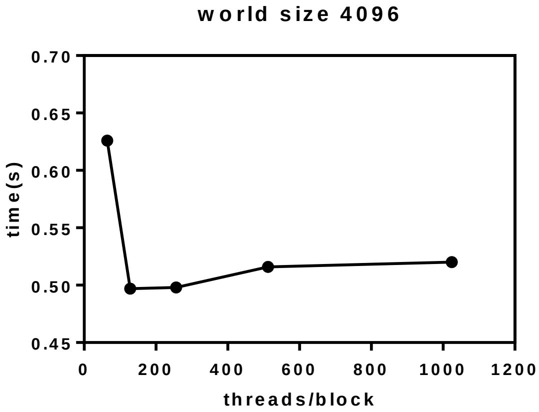
```
This is the Game of Life running in parallel on a GPU.
==98415== NVPROF is profiling process 98415, command: ./gol 3 2048 1024 1024 0
==98415== Profiling application: ./gol 3 2048 1024 1024 0
==98415== Profiling result:
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:  100.00%  102.24ms      1024  99.848us  95.808us  3.8674ms  gol_kernel(unsigned char const *, unsigned int,
unsigned int, unsigned char*)
      API calls:   65.77%  197.93ms         2  98.963ms  48.247us  197.88ms  cudaMallocManaged
                   29.92%  90.052ms         1  90.052ms  90.052ms  90.052ms  cudaDeviceSynchronize
                    3.67%  11.035ms      1024  10.776us  10.179us  42.725us  cudaLaunchKernel
                    0.29%  870.41us         1  870.41us  870.41us  870.41us  cuDeviceTotalMem
                    0.21%  630.39us        97  6.4980us     226ns  244.18us  cuDeviceGetAttribute
                    0.12%  361.32us         2  180.66us  179.70us  181.61us  cudaFree
                    0.02%  51.660us         1  51.660us  51.660us  51.660us  cuDeviceGetName
                    0.00%  3.1820us         1  3.1820us  3.1820us  3.1820us  cuDeviceGetPCIBusId
                    0.00%  2.4720us         3     824ns     406ns  1.4490us  cuDeviceGetCount
                    0.00%     913ns         2     456ns     350ns     563ns  cuDeviceGet
                    0.00%     362ns         1     362ns     362ns     362ns  cuDeviceGetUuid

==98415== Unified Memory profiling result:
Device "Tesla V100-SXM2-16GB (0)"
   Count  Avg Size  Min Size  Max Size  Total Size  Total Time  Name
      36  227.56KB  64.000KB  896.00KB  8.000000MB  357.8880us  Host To Device
      19         -         -         -           -  3.778976ms  Gpu page fault groups
Total CPU Page faults: 24
======== Error: Application returned non-zero code 1
```

According to nvprof analysis result, the time spent in the CUDA gol kernel function is 114.14ms, 97.028ms, 97.081ms, 98.726ms, and 102.24ms for each block size individually in 2048*2048 world.

3.

For the world size 4096*4096, the plot of execution time versus thread blocksize is shown below.



The optimal number of threads per block is 128.

The output from an **nvprof** run is:

For 64 threads per block:

```
This is the Game of Life running in parallel on a GPU.
==99150== NVPROF is profiling process 99150, command: ./gol 3 4096 1024 64 0
==99150== Profiling application: ./gol 3 4096 1024 64 0
==99150== Profiling result:
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:  100.00%  417.25ms      1024  407.47us  374.88us  15.155ms  gol_kernel(unsigned char const *, unsigned int,
unsigned int, unsigned char*)
      API calls:   64.79%  401.26ms         1  401.26ms  401.26ms  401.26ms  cudaDeviceSynchronize
                   32.27%  199.87ms         2  99.936ms  53.778us  199.82ms  cudaMallocManaged
                    2.43%  15.029ms      1024  14.676us  8.0680us  5.3667ms  cudaLaunchKernel
                    0.22%  1.3354ms         1  1.3354ms  1.3354ms  1.3354ms  cuDeviceTotalMem
                    0.14%  860.62us        97  8.8720us     359ns  332.06us  cuDeviceGetAttribute
                    0.14%  850.72us         2  425.36us  416.85us  433.87us  cudaFree
                    0.01%  71.974us         1  71.974us  71.974us  71.974us  cuDeviceGetName
                    0.00%  3.4040us         1  3.4040us  3.4040us  3.4040us  cuDeviceGetPCIBusId
                    0.00%  3.3910us         3  1.1300us     764ns  1.7400us  cuDeviceGetCount
                    0.00%  1.4120us         2     706ns     553ns     859ns  cuDeviceGet
                    0.00%     614ns         1     614ns     614ns     614ns  cuDeviceGetUuid

==99150== Unified Memory profiling result:
Device "Tesla V100-SXM2-16GB (0)"
   Count  Avg Size  Min Size  Max Size  Total Size  Total Time  Name
     155  211.41KB  64.000KB  960.00KB  32.00000MB  1.400256ms  Host To Device
      83         -         -         -           -  14.76621ms  Gpu page fault groups
Total CPU Page faults: 96
======== Error: Application returned non-zero code 1
```

For 128 threads per block:

```
This is the Game of Life running in parallel on a GPU.
==99219== NVPROF is profiling process 99219, command: ./gol 3 4096 1024 128 0
==99219== Profiling application: ./gol 3 4096 1024 128 0
==99219== Profiling result:
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:  100.00%  346.88ms      1024  338.75us  298.72us  15.671ms  gol_kernel(unsigned char const *, unsigned int,
unsigned int, unsigned char*)
      API calls:   59.27%  330.42ms         1  330.42ms  330.42ms  330.42ms  cudaDeviceSynchronize
                   37.54%  209.28ms         2  104.64ms  60.940us  209.22ms  cudaMallocManaged
                    2.75%  15.350ms      1024  14.989us  9.6520us  4.5532ms  cudaLaunchKernel
                    0.16%  876.45us         2  438.22us  412.70us  463.75us  cudaFree
                    0.15%  857.29us         1  857.29us  857.29us  857.29us  cuDeviceTotalMem
                    0.11%  624.98us        97  6.4430us     230ns  241.83us  cuDeviceGetAttribute
                    0.01%  51.485us         1  51.485us  51.485us  51.485us  cuDeviceGetName
                    0.00%  4.3890us         1  4.3890us  4.3890us  4.3890us  cuDeviceGetPCIBusId
                    0.00%  2.2620us         3     754ns     472ns  1.2520us  cuDeviceGetCount
                    0.00%     931ns         2     465ns     337ns     594ns  cuDeviceGet
                    0.00%     406ns         1     406ns     406ns     406ns  cuDeviceGetUuid

==99219== Unified Memory profiling result:
Device "Tesla V100-SXM2-16GB (0)"
   Count  Avg Size  Min Size  Max Size  Total Size  Total Time  Name
     153  214.17KB  64.000KB  960.00KB  32.00000MB  1.306784ms  Host To Device
      83         -         -         -           -  15.33757ms  Gpu page fault groups
Total CPU Page faults: 96
======== Error: Application returned non-zero code 1
```

For 256 threads per block:

```
This is the Game of Life running in parallel on a GPU.
==99387== NVPROF is profiling process 99387, command: ./gol 3 4096 1024 256 0
==99387== Profiling application: ./gol 3 4096 1024 256 0
==99387== Profiling result:
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:  100.00%  320.50ms      1024  312.99us  298.37us  14.984ms  gol_kernel(unsigned char const *, unsigned int,
unsigned int, unsigned char*)
      API calls:   58.93%  304.76ms         1  304.76ms  304.76ms  304.76ms  cudaDeviceSynchronize
                   37.63%  194.62ms         2  97.308ms  57.239us  194.56ms  cudaMallocManaged
                    2.84%  14.668ms      1024  14.324us  7.4760us  4.3999ms  cudaLaunchKernel
                    0.26%  1.3365ms         1  1.3365ms  1.3365ms  1.3365ms  cuDeviceTotalMem
                    0.17%  864.10us        97  8.9080us     353ns  334.17us  cuDeviceGetAttribute
                    0.16%  847.17us         2  423.58us  418.02us  429.15us  cudaFree
                    0.01%  71.378us         1  71.378us  71.378us  71.378us  cuDeviceGetName
                    0.00%  3.2850us         3  1.0950us     664ns  1.7620us  cuDeviceGetCount
                    0.00%  3.2560us         1  3.2560us  3.2560us  3.2560us  cuDeviceGetPCIBusId
                    0.00%  1.4300us         2     715ns     549ns     881ns  cuDeviceGet
                    0.00%     631ns         1     631ns     631ns     631ns  cuDeviceGetUuid

==99387== Unified Memory profiling result:
Device "Tesla V100-SXM2-16GB (0)"
   Count  Avg Size  Min Size  Max Size  Total Size  Total Time  Name
     145  225.99KB  64.000KB  960.00KB  32.00000MB  1.331808ms  Host To Device
      81         -         -         -           -  14.70131ms  Gpu page fault groups
Total CPU Page faults: 96
======== Error: Application returned non-zero code 1
```

For 512 threads per block:

```
This is the Game of Life running in parallel on a GPU.
==99437== NVPROF is profiling process 99437, command: ./gol 3 4096 1024 512 0
==99437== Profiling application: ./gol 3 4096 1024 512 0
==99437== Profiling result:
```

```
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:  100.00%  326.75ms      1024  319.09us  303.26us  16.157ms  gol_kernel(unsigned char const *, unsigned int,
unsigned int, unsigned char*)
      API calls:   59.26%  309.85ms         1  309.85ms  309.85ms  309.85ms  cudaDeviceSynchronize
                   37.11%  194.01ms         2  97.007ms  55.078us  193.96ms  cudaMallocManaged
                    3.03%  15.842ms      1024  15.470us  7.6950us  5.9601ms  cudaLaunchKernel
                    0.26%  1.3404ms         1  1.3404ms  1.3404ms  1.3404ms  cuDeviceTotalMem
                    0.17%  867.51us        97  8.9430us     355ns  333.16us  cuDeviceGetAttribute
                    0.16%  860.13us         2  430.07us  420.43us  439.70us  cudaFree
                    0.01%  71.142us         1  71.142us  71.142us  71.142us  cuDeviceGetName
                    0.00%  3.8710us         1  3.8710us  3.8710us  3.8710us  cuDeviceGetPCIBusId
                    0.00%  3.2300us         3  1.0760us     693ns  1.6990us  cuDeviceGetCount
                    0.00%  1.4370us         2     718ns     445ns     992ns  cuDeviceGet
                    0.00%     668ns         1     668ns     668ns     668ns  cuDeviceGetUuid

==99437== Unified Memory profiling result:
Device "Tesla V100-SXM2-16GB (0)"
   Count  Avg Size  Min Size  Max Size  Total Size  Total Time  Name
     146  224.44KB  64.000KB  960.00KB  32.00000MB  1.327136ms  Host To Device
      80         -         -         -           -  15.86672ms  Gpu page fault groups
Total CPU Page faults: 96
======== Error: Application returned non-zero code 1
```

For 1024 threads per block:

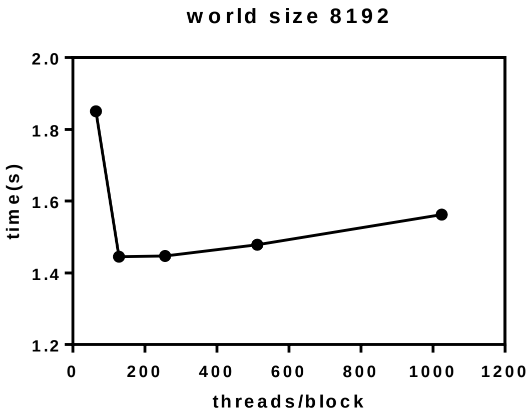
```
This is the Game of Life running in parallel on a GPU.
==99494== NVPROF is profiling process 99494, command: ./gol 3 4096 1024 1024 0
==99494== Profiling application: ./gol 3 4096 1024 1024 0
==99494== Profiling result:
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:  100.00%  376.71ms      1024  367.88us  333.73us  16.240ms  gol_kernel(unsigned char const *, unsigned int,
unsigned int, unsigned char*)
      API calls:   62.20%  359.69ms         1  359.69ms  359.69ms  359.69ms  cudaDeviceSynchronize
                   34.51%  199.59ms         2  99.795ms  68.523us  199.52ms  cudaMallocManaged
                    2.74%  15.866ms      1024  15.493us  10.560us  4.1027ms  cudaLaunchKernel
                    0.23%  1.3340ms         1  1.3340ms  1.3340ms  1.3340ms  cuDeviceTotalMem
                    0.15%  872.13us        97  8.9910us     359ns  335.68us  cuDeviceGetAttribute
                    0.15%  864.94us         2  432.47us  420.95us  443.99us  cudaFree
                    0.01%  70.488us         1  70.488us  70.488us  70.488us  cuDeviceGetName
                    0.00%  3.4610us         1  3.4610us  3.4610us  3.4610us  cuDeviceGetPCIBusId
                    0.00%  3.3790us         3  1.1260us     701ns  1.7910us  cuDeviceGetCount
                    0.00%  1.3710us         2     685ns     502ns     869ns  cuDeviceGet
                    0.00%     670ns         1     670ns     670ns     670ns  cuDeviceGetUuid

==99494== Unified Memory profiling result:
Device "Tesla V100-SXM2-16GB (0)"
   Count  Avg Size  Min Size  Max Size  Total Size  Total Time  Name
     143  229.15KB  64.000KB  960.00KB  32.00000MB  1.387968ms  Host To Device
      76         -         -         -           -  15.85123ms  Gpu page fault groups
Total CPU Page faults: 96
======== Error: Application returned non-zero code 1
```

According to nvprof, the time spent in the CUDA gol kernel function is 417.25ms, 346.88ms, 320.50ms, 326.75ms, and 376.71ms for each block size individually in 4096*4096 world.

4.

For the world size 8192*8192, the plot of execution time versus thread blocksize is shown below.



When block size is 128 threads, it yields the fastest execution time, which is 1.445s.

The output from an **nvprof** run is:

For 64 threads per block:

```
This is the Game of Life running in parallel on a GPU.
==99716== NVPROF is profiling process 99716, command: ./gol 3 8192 1024 64 0
==99716== Profiling application: ./gol 3 8192 1024 64 0
==99716== Profiling result:
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:  100.00%  1.58800s      1024  1.5508ms  1.4842ms  50.948ms  gol_kernel(unsigned char const *, unsigned int,
unsigned int, unsigned char*)
      API calls:   85.61%  1.53489s         1  1.53489s  1.53489s  1.53489s  cudaDeviceSynchronize
                   11.24%  201.47ms         2  100.73ms  73.814us  201.40ms  cudaMallocManaged
                    2.91%  52.183ms      1024  50.959us  7.3050us  42.504ms  cudaLaunchKernel
                    0.15%  2.6854ms         2  1.3427ms  1.2766ms  1.4088ms  cudaFree
                    0.05%  918.71us         1  918.71us  918.71us  918.71us  cuDeviceTotalMem
                    0.04%  654.80us        97  6.7500us     244ns  253.60us  cuDeviceGetAttribute
                    0.00%  50.441us         1  50.441us  50.441us  50.441us  cuDeviceGetName
                    0.00%  3.5060us         1  3.5060us  3.5060us  3.5060us  cuDeviceGetPCIBusId
                    0.00%  2.4950us         3     831ns     442ns  1.5020us  cuDeviceGetCount
                    0.00%     965ns         2     482ns     354ns     611ns  cuDeviceGet
                    0.00%     401ns         1     401ns     401ns     401ns  cuDeviceGetUuid

==99716== Unified Memory profiling result:
Device "Tesla V100-SXM2-16GB (0)"
   Count  Avg Size  Min Size  Max Size  Total Size  Total Time  Name
     608  215.58KB  64.000KB  960.00KB  128.0000MB  5.308448ms  Host To Device
     322         -         -         -           -  49.45325ms  Gpu page fault groups
Total CPU Page faults: 384
======== Error: Application returned non-zero code 1
```

For 128 threads per block:

```
This is the Game of Life running in parallel on a GPU.
==99770== NVPROF is profiling process 99770, command: ./gol 3 8192 1024 128 0
==99770== Profiling application: ./gol 3 8192 1024 128 0
==99770== Profiling result:
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:  100.00%  1.25950s      1024  1.2300ms  1.1797ms  51.277ms  gol_kernel(unsigned char const *, unsigned int,
unsigned int, unsigned char*)
      API calls:   82.67%  1.20661s         1  1.20661s  1.20661s  1.20661s  cudaDeviceSynchronize
                   13.42%  195.89ms         2  97.946ms  79.597us  195.81ms  cudaMallocManaged
                    3.56%  51.988ms      1024  50.769us  7.0870us  43.053ms  cudaLaunchKernel
                    0.19%  2.7268ms         2  1.3634ms  1.3028ms  1.4239ms  cudaFree
                    0.09%  1.3406ms         1  1.3406ms  1.3406ms  1.3406ms  cuDeviceTotalMem
                    0.06%  858.16us        97  8.8470us     353ns  333.92us  cuDeviceGetAttribute
                    0.00%  69.847us         1  69.847us  69.847us  69.847us  cuDeviceGetName
                    0.00%  4.0720us         3  1.3570us     670ns  2.5740us  cuDeviceGetCount
                    0.00%  3.8030us         1  3.8030us  3.8030us  3.8030us  cuDeviceGetPCIBusId
                    0.00%  1.4450us         2     722ns     484ns     961ns  cuDeviceGet
                    0.00%     519ns         1     519ns     519ns     519ns  cuDeviceGetUuid

==99770== Unified Memory profiling result:
Device "Tesla V100-SXM2-16GB (0)"
   Count  Avg Size  Min Size  Max Size  Total Size  Total Time  Name
     609  215.23KB  64.000KB  960.00KB  128.0000MB  5.113856ms  Host To Device
     329         -         -         -           -  50.17453ms  Gpu page fault groups
Total CPU Page faults: 384
======== Error: Application returned non-zero code 1
```

For 256 threads per block:

```
This is the Game of Life running in parallel on a GPU.
==99822== NVPROF is profiling process 99822, command: ./gol 3 8192 1024 256 0
==99822== Profiling application: ./gol 3 8192 1024 256 0
==99822== Profiling result:
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:  100.00%  1.26453s      1024  1.2349ms  1.1805ms  54.625ms  gol_kernel(unsigned char const *, unsigned int,
unsigned int, unsigned char*)
      API calls:   82.82%  1.20826s         1  1.20826s  1.20826s  1.20826s  cudaDeviceSynchronize
                   13.10%  191.12ms         2  95.560ms  80.977us  191.04ms  cudaMallocManaged
                    3.79%  55.353ms      1024  54.055us  7.0710us  46.439ms  cudaLaunchKernel
                    0.18%  2.6300ms         2  1.3150ms  1.2636ms  1.3665ms  cudaFree
                    0.06%  854.87us         1  854.87us  854.87us  854.87us  cuDeviceTotalMem
                    0.04%  640.78us        97  6.6060us     228ns  250.11us  cuDeviceGetAttribute
                    0.00%  50.643us         1  50.643us  50.643us  50.643us  cuDeviceGetName
                    0.00%  3.2750us         1  3.2750us  3.2750us  3.2750us  cuDeviceGetPCIBusId
                    0.00%  2.2620us         3     754ns     416ns  1.2680us  cuDeviceGetCount
                    0.00%     885ns         2     442ns     305ns     580ns  cuDeviceGet
                    0.00%     429ns         1     429ns     429ns     429ns  cuDeviceGetUuid

==99822== Unified Memory profiling result:
Device "Tesla V100-SXM2-16GB (0)"
   Count  Avg Size  Min Size  Max Size  Total Size  Total Time  Name
     604  217.01KB  64.000KB  960.00KB  128.0000MB  5.151520ms  Host To Device
     324         -         -         -           -  53.46205ms  Gpu page fault groups
Total CPU Page faults: 384
======== Error: Application returned non-zero code 1
```

For 512 threads per block:

```
This is the Game of Life running in parallel on a GPU.
==99895== NVPROF is profiling process 99895, command: ./gol 3 8192 1024 512 0
==99895== Profiling application: ./gol 3 8192 1024 512 0
==99895== Profiling result:
```

```
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:  100.00%  1.30616s      1024  1.2755ms  1.1998ms  56.502ms  gol_kernel(unsigned char const *, unsigned int,
unsigned int, unsigned char*)
      API calls:   82.81%  1.24783s         1  1.24783s  1.24783s  1.24783s  cudaDeviceSynchronize
                   13.09%  197.29ms         2  98.646ms  83.208us  197.21ms  cudaMallocManaged
                    3.81%  57.421ms      1024  56.075us  7.2710us  48.032ms  cudaLaunchKernel
                    0.18%  2.6778ms         2  1.3389ms  1.2776ms  1.4003ms  cudaFree
                    0.06%  855.15us         1  855.15us  855.15us  855.15us  cuDeviceTotalMem
                    0.04%  655.31us        97  6.7550us     226ns  249.32us  cuDeviceGetAttribute
                    0.00%  51.002us         1  51.002us  51.002us  51.002us  cuDeviceGetName
                    0.00%  4.8240us         1  4.8240us  4.8240us  4.8240us  cuDeviceGetPCIBusId
                    0.00%  2.5480us         3     849ns     439ns  1.5600us  cuDeviceGetCount
                    0.00%  1.0930us         2     546ns     322ns     771ns  cuDeviceGet
                    0.00%     461ns         1     461ns     461ns     461ns  cuDeviceGetUuid

==99895== Unified Memory profiling result:
Device "Tesla V100-SXM2-16GB (0)"
   Count  Avg Size  Min Size  Max Size  Total Size  Total Time  Name
     608  215.58KB  64.000KB  960.00KB  128.0000MB  5.388480ms  Host To Device
     323         -         -         -           -  55.12368ms  Gpu page fault groups
Total CPU Page faults: 384
======== Error: Application returned non-zero code 1
```

For 1024 threads per block:

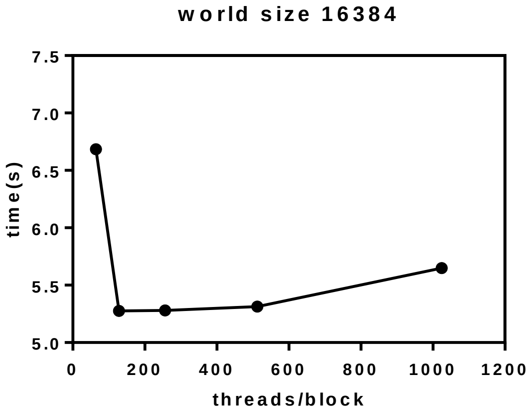
```
This is the Game of Life running in parallel on a GPU.
==100063== NVPROF is profiling process 100063, command: ./gol 3 8192 1024 1024 0
==100063== Profiling application: ./gol 3 8192 1024 1024 0
==100063== Profiling result:
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:  100.00%  1.36159s      1024  1.3297ms  1.2445ms  54.477ms  gol_kernel(unsigned char const *, unsigned int,
unsigned int, unsigned char*)
      API calls:   83.60%  1.30520s         1  1.30520s  1.30520s  1.30520s  cudaDeviceSynchronize
                   12.54%  195.80ms         2  97.902ms  69.614us  195.73ms  cudaMallocManaged
                    3.55%  55.360ms      1024  54.062us  8.2480us  44.886ms  cudaLaunchKernel
                    0.21%  3.2615ms         2  1.6307ms  1.5502ms  1.7113ms  cudaFree
                    0.06%  895.54us         1  895.54us  895.54us  895.54us  cuDeviceTotalMem
                    0.04%  663.58us        97  6.8410us     246ns  258.79us  cuDeviceGetAttribute
                    0.00%  51.094us         1  51.094us  51.094us  51.094us  cuDeviceGetName
                    0.00%  3.2300us         1  3.2300us  3.2300us  3.2300us  cuDeviceGetPCIBusId
                    0.00%  2.4810us         3     827ns     405ns  1.4980us  cuDeviceGetCount
                    0.00%     906ns         2     453ns     342ns     564ns  cuDeviceGet
                    0.00%     406ns         1     406ns     406ns     406ns  cuDeviceGetUuid

==100063== Unified Memory profiling result:
Device "Tesla V100-SXM2-16GB (0)"
   Count  Avg Size  Min Size  Max Size  Total Size  Total Time  Name
     565  231.99KB  64.000KB  960.00KB  128.0000MB  5.272384ms  Host To Device
     300         -         -         -           -  53.03814ms  Gpu page fault groups
Total CPU Page faults: 384
======== Error: Application returned non-zero code 1
```

According to nvprof, the time spent in the CUDA gol kernel function is 1.588s, 1.259s, 1.264s, 1.306s and 1.361s for each block size individually in 8192*8192 world.

5.

For the world size 16384*16384, the plot of execution time versus thread blocksize is shown below.



When block size is 128 threads, it yields the fastest execution time, which is 5.275s.

The output from an **nvprof** run is:

For 64 threads per block:

```
This is the Game of Life running in parallel on a GPU.
==100158== NVPROF is profiling process 100158, command: ./gol 3 16384 1024 64 0
==100158== Profiling application: ./gol 3 16384 1024 64 0
==100158== Profiling result:
           Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:  100.00%  6.25640s      1024  6.1098ms  5.9235ms  189.73ms  gol_kernel(unsigned char const *, unsigned int,
unsigned int, unsigned char*)
      API calls:   93.69%  6.05997s         1  6.05997s  6.05997s  6.05997s  cudaDeviceSynchronize
                    3.07%  198.85ms         2  99.423ms  84.063us  198.76ms  cudaMallocManaged
                    3.02%  195.54ms      1024  190.96us  6.9100us  181.50ms  cudaLaunchKernel
                    0.18%  11.807ms         2  5.9035ms  5.6786ms  6.1284ms  cudaFree
                    0.02%  1.3299ms         1  1.3299ms  1.3299ms  1.3299ms  cuDeviceTotalMem
                    0.01%  858.44us        97  8.8490us     352ns  334.76us  cuDeviceGetAttribute
                    0.00%  71.013us         1  71.013us  71.013us  71.013us  cuDeviceGetName
                    0.00%  3.6920us         3  1.2300us     629ns  2.1230us  cuDeviceGetCount
                    0.00%  3.6460us         1  3.6460us  3.6460us  3.6460us  cuDeviceGetPCIBusId
                    0.00%  1.4050us         2     702ns     501ns     904ns  cuDeviceGet
                    0.00%     586ns         1     586ns     586ns     586ns  cuDeviceGetUuid

==100158== Unified Memory profiling result:
Device "Tesla V100-SXM2-16GB (0)"
   Count  Avg Size  Min Size  Max Size  Total Size  Total Time  Name
    2484  211.07KB  64.000KB  960.00KB  512.0000MB  19.79568ms  Host To Device
    1304         -         -         -           -  183.8939ms  Gpu page fault groups
Total CPU Page faults: 1536
======== Error: Application returned non-zero code 1
```

For 128 threads per block:

```
This is the Game of Life running in parallel on a GPU.
==100245== NVPROF is profiling process 100245, command: ./gol 3 16384 1024 128 0
==100245== Profiling application: ./gol 3 16384 1024 128 0
==100245== Profiling result:
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:  100.00%  5.01916s      1024  4.9015ms  4.7089ms  185.04ms  gol_kernel(unsigned char const *, unsigned int,
unsigned int, unsigned char*)
      API calls:   92.38%  4.82866s         1  4.82866s  4.82866s  4.82866s  cudaDeviceSynchronize
                    3.77%  197.28ms         2  98.640ms  86.241us  197.19ms  cudaMallocManaged
                    3.63%  189.57ms      1024  185.13us  7.0160us  176.83ms  cudaLaunchKernel
                    0.19%  9.9018ms         2  4.9509ms  4.8164ms  5.0854ms  cudaFree
                    0.02%  854.53us         1  854.53us  854.53us  854.53us  cuDeviceTotalMem
                    0.01%  627.17us        97  6.4650us     224ns  244.77us  cuDeviceGetAttribute
                    0.00%  51.168us         1  51.168us  51.168us  51.168us  cuDeviceGetName
                    0.00%  4.3390us         1  4.3390us  4.3390us  4.3390us  cuDeviceGetPCIBusId
                    0.00%  2.3910us         3     797ns     475ns  1.4320us  cuDeviceGetCount
                    0.00%  1.0000us         2     500ns     367ns     633ns  cuDeviceGet
                    0.00%     394ns         1     394ns     394ns     394ns  cuDeviceGetUuid

==100245== Unified Memory profiling result:
Device "Tesla V100-SXM2-16GB (0)"
   Count  Avg Size  Min Size  Max Size  Total Size  Total Time  Name
    2477  211.66KB  64.000KB  960.00KB  512.0000MB  19.85034ms  Host To Device
    1299         -         -         -           -  180.0971ms  Gpu page fault groups
Total CPU Page faults: 1536
======== Error: Application returned non-zero code 1
```

For 256 threads per block:

```
This is the Game of Life running in parallel on a GPU.
==100421== NVPROF is profiling process 100421, command: ./gol 3 16384 1024 256 0
==100421== Profiling application: ./gol 3 16384 1024 256 0
==100421== Profiling result:
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:  100.00%  5.03297s      1024  4.9150ms  4.7110ms  197.15ms  gol_kernel(unsigned char const *, unsigned int,
unsigned int, unsigned char*)
      API calls:   92.18%  4.83033s         1  4.83033s  4.83033s  4.83033s  cudaDeviceSynchronize
                    3.85%  201.71ms      1024  196.98us  6.9170us  189.08ms  cudaLaunchKernel
                    3.75%  196.51ms         2  98.256ms  84.587us  196.43ms  cudaMallocManaged
                    0.19%  9.8976ms         2  4.9488ms  4.7935ms  5.1041ms  cudaFree
                    0.02%  854.35us         1  854.35us  854.35us  854.35us  cuDeviceTotalMem
                    0.01%  638.29us        97  6.5800us     229ns  252.37us  cuDeviceGetAttribute
                    0.00%  49.981us         1  49.981us  49.981us  49.981us  cuDeviceGetName
                    0.00%  3.0630us         1  3.0630us  3.0630us  3.0630us  cuDeviceGetPCIBusId
                    0.00%  2.7020us         3     900ns     442ns  1.7070us  cuDeviceGetCount
                    0.00%     955ns         2     477ns     340ns     615ns  cuDeviceGet
                    0.00%     375ns         1     375ns     375ns     375ns  cuDeviceGetUuid

==100421== Unified Memory profiling result:
Device "Tesla V100-SXM2-16GB (0)"
   Count  Avg Size  Min Size  Max Size  Total Size  Total Time  Name
    2453  213.73KB  64.000KB  960.00KB  512.0000MB  19.88413ms  Host To Device
    1295         -         -         -           -  191.9648ms  Gpu page fault groups
Total CPU Page faults: 1536
======== Error: Application returned non-zero code 1
```

For 512 threads per block:

```
This is the Game of Life running in parallel on a GPU.
==100480== NVPROF is profiling process 100480, command: ./gol 3 16384 1024 512 0
==100480== Profiling application: ./gol 3 16384 1024 512 0
==100480== Profiling result:
```

```
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:  100.00%  5.11348s      1024  4.9936ms  4.7829ms  198.83ms  gol_kernel(unsigned char const *, unsigned int,
unsigned int, unsigned char*)
      API calls:   92.19%  4.90859s         1  4.90859s  4.90859s  4.90859s  cudaDeviceSynchronize
                    3.83%  203.94ms      1024  199.16us  6.9410us  190.66ms  cudaLaunchKernel
                    3.75%  199.42ms         2  99.709ms  79.800us  199.34ms  cudaMallocManaged
                    0.19%  9.9373ms         2  4.9686ms  4.8195ms  5.1178ms  cudaFree
                    0.03%  1.3429ms         1  1.3429ms  1.3429ms  1.3429ms  cuDeviceTotalMem
                    0.02%  871.88us        97  8.9880us     359ns  342.63us  cuDeviceGetAttribute
                    0.00%  73.187us         1  73.187us  73.187us  73.187us  cuDeviceGetName
                    0.00%  3.3640us         3  1.1210us     700ns  1.8770us  cuDeviceGetCount
                    0.00%  3.3590us         1  3.3590us  3.3590us  3.3590us  cuDeviceGetPCIBusId
                    0.00%  1.4410us         2     720ns     574ns     867ns  cuDeviceGet
                    0.00%     693ns         1     693ns     693ns     693ns  cuDeviceGetUuid

==100480== Unified Memory profiling result:
Device "Tesla V100-SXM2-16GB (0)"
   Count  Avg Size  Min Size  Max Size  Total Size  Total Time  Name
    2398  218.64KB  64.000KB  960.00KB  512.0000MB  19.90192ms  Host To Device
    1243         -         -         -           -  193.5486ms  Gpu page fault groups
Total CPU Page faults: 1536
======== Error: Application returned non-zero code 1
```

For 1024 threads per block:

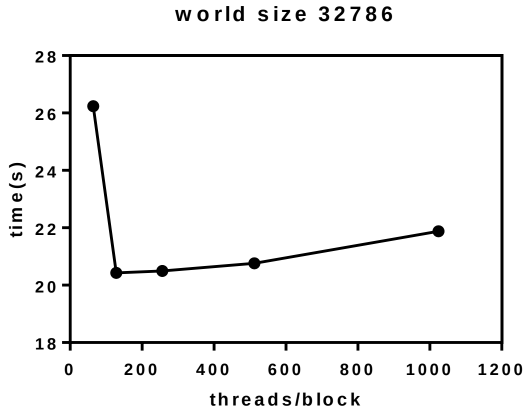
```
This is the Game of Life running in parallel on a GPU.
==100538== NVPROF is profiling process 100538, command: ./gol 3 16384 1024 1024 0
==100538== Profiling application: ./gol 3 16384 1024 1024 0
==100538== Profiling result:
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:  100.00%  5.28385s      1024  5.1600ms  4.9633ms  194.78ms  gol_kernel(unsigned char const *, unsigned int,
unsigned int, unsigned char*)
      API calls:   92.63%  5.08366s         1  5.08366s  5.08366s  5.08366s  cudaDeviceSynchronize
                    3.63%  199.28ms      1024  194.61us  6.9860us  186.59ms  cudaLaunchKernel
                    3.52%  193.25ms         2  96.623ms  88.491us  193.16ms  cudaMallocManaged
                    0.18%  9.8242ms         2  4.9121ms  4.6990ms  5.1251ms  cudaFree
                    0.02%  1.3358ms         1  1.3358ms  1.3358ms  1.3358ms  cuDeviceTotalMem
                    0.02%  858.32us        97  8.8480us     358ns  330.84us  cuDeviceGetAttribute
                    0.00%  70.808us         1  70.808us  70.808us  70.808us  cuDeviceGetName
                    0.00%  3.6950us         1  3.6950us  3.6950us  3.6950us  cuDeviceGetPCIBusId
                    0.00%  3.1950us         3  1.0650us     608ns  1.7010us  cuDeviceGetCount
                    0.00%  1.5000us         2     750ns     496ns  1.0040us  cuDeviceGet
                    0.00%     513ns         1     513ns     513ns     513ns  cuDeviceGetUuid

==100538== Unified Memory profiling result:
Device "Tesla V100-SXM2-16GB (0)"
   Count  Avg Size  Min Size  Max Size  Total Size  Total Time  Name
    2259  232.09KB  64.000KB  960.00KB  512.0000MB  19.02544ms  Host To Device
    1173         -         -         -           -  189.8370ms  Gpu page fault groups
Total CPU Page faults: 1536
======== Error: Application returned non-zero code 1
```

According to nvprof, the time spent in the CUDA gol kernel function is 6.256s, 5.019s, 5.033s, 5.113s and 5.284s for each block size individually in 16384*16384 world.

6.

For the world size 32786*32786, the plot of execution time versus thread blocksize is shown below.



When block size is 128 threads, it yields the fastest execution time, which is 20.426s.

The output from an **nvprof** run is:

For 64 threads per block:

```
This is the Game of Life running in parallel on a GPU.
==100757== NVPROF is profiling process 100757, command: ./gol 3 32786 1024 64 0
==100757== Profiling application: ./gol 3 32786 1024 64 0
==100757== Profiling result:
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:  100.00%  24.9784s      1024  24.393ms  23.701ms  728.54ms  gol_kernel(unsigned char const *, unsigned int,
unsigned int, unsigned char*)
      API calls:   96.06%  24.2257s         1  24.2257s  24.2257s  24.2257s  cudaDeviceSynchronize
                    2.98%  751.79ms      1024  734.17us  6.9860us  720.37ms  cudaLaunchKernel
                    0.80%  201.71ms         2  100.86ms  130.98us  201.58ms  cudaMallocManaged
                    0.15%  38.414ms         2  19.207ms  18.523ms  19.891ms  cudaFree
                    0.00%  851.68us         1  851.68us  851.68us  851.68us  cuDeviceTotalMem
                    0.00%  627.30us        97  6.4670us     224ns  245.35us  cuDeviceGetAttribute
                    0.00%  50.776us         1  50.776us  50.776us  50.776us  cuDeviceGetName
                    0.00%  4.0510us         1  4.0510us  4.0510us  4.0510us  cuDeviceGetPCIBusId
                    0.00%  2.8230us         3     941ns     401ns  1.9030us  cuDeviceGetCount
                    0.00%  1.0070us         2     503ns     346ns     661ns  cuDeviceGet
                    0.00%     445ns         1     445ns     445ns     445ns  cuDeviceGetUuid

==100757== Unified Memory profiling result:
Device "Tesla V100-SXM2-16GB (0)"
   Count  Avg Size  Min Size  Max Size  Total Size  Total Time  Name
   10091  208.06KB  64.000KB  960.00KB  2.002319GB  77.05830ms  Host To Device
    5095         -         -         -           -  706.7547ms  Gpu page fault groups
Total CPU Page faults: 6156
======== Error: Application returned non-zero code 1
```

For 128 threads per block:

```
This is the Game of Life running in parallel on a GPU.
==100881== NVPROF is profiling process 100881, command: ./gol 3 32786 1024 128 0
==100881== Profiling application: ./gol 3 32786 1024 128 0
==100881== Profiling result:
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:  100.00%  20.0098s      1024  19.541ms  18.826ms  701.80ms  gol_kernel(unsigned char const *, unsigned int,
unsigned int, unsigned char*)
      API calls:   95.25%  19.2887s         1  19.2887s  19.2887s  19.2887s  cudaDeviceSynchronize
                    3.56%  720.16ms      1024  703.28us  7.0600us  693.53ms  cudaLaunchKernel
                    0.99%  200.64ms         2  100.32ms  115.06us  200.53ms  cudaMallocManaged
                    0.19%  38.400ms         2  19.200ms  18.612ms  19.788ms  cudaFree
                    0.01%  1.3513ms         1  1.3513ms  1.3513ms  1.3513ms  cuDeviceTotalMem
                    0.00%  871.92us        97  8.9880us     361ns  337.99us  cuDeviceGetAttribute
                    0.00%  73.218us         1  73.218us  73.218us  73.218us  cuDeviceGetName
                    0.00%  4.0710us         3  1.3570us     778ns  2.3850us  cuDeviceGetCount
                    0.00%  3.5600us         1  3.5600us  3.5600us  3.5600us  cuDeviceGetPCIBusId
                    0.00%  1.4940us         2     747ns     547ns     947ns  cuDeviceGet
                    0.00%     637ns         1     637ns     637ns     637ns  cuDeviceGetUuid

==100881== Unified Memory profiling result:
Device "Tesla V100-SXM2-16GB (0)"
   Count  Avg Size  Min Size  Max Size  Total Size  Total Time  Name
   10033  209.27KB  64.000KB  960.00KB  2.002319GB  77.21370ms  Host To Device
    5112         -         -         -           -  683.4538ms  Gpu page fault groups
Total CPU Page faults: 6156
======== Error: Application returned non-zero code 1
```

For 256 threads per block:

```
This is the Game of Life running in parallel on a GPU.
==101518== NVPROF is profiling process 101518, command: ./gol 3 32786 1024 256 0
==101518== Profiling application: ./gol 3 32786 1024 256 0
==101518== Profiling result:
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:  100.00%  20.0525s      1024  19.582ms  18.844ms  726.02ms  gol_kernel(unsigned char const *, unsigned int,
unsigned int, unsigned char*)
      API calls:   95.15%  19.3072s         1  19.3072s  19.3072s  19.3072s  cudaDeviceSynchronize
                    3.67%  744.39ms      1024  726.94us  7.0800us  717.74ms  cudaLaunchKernel
                    0.98%  198.16ms         2  99.078ms  113.19us  198.04ms  cudaMallocManaged
                    0.19%  38.504ms         2  19.252ms  18.624ms  19.880ms  cudaFree
                    0.01%  1.3371ms         1  1.3371ms  1.3371ms  1.3371ms  cuDeviceTotalMem
                    0.00%  877.04us        97  9.0410us     358ns  339.75us  cuDeviceGetAttribute
                    0.00%  71.177us         1  71.177us  71.177us  71.177us  cuDeviceGetName
                    0.00%  5.1780us         1  5.1780us  5.1780us  5.1780us  cuDeviceGetPCIBusId
                    0.00%  3.8990us         3  1.2990us     629ns  2.4370us  cuDeviceGetCount
                    0.00%  1.5410us         2     770ns     484ns  1.0570us  cuDeviceGet
                    0.00%     626ns         1     626ns     626ns     626ns  cuDeviceGetUuid

==101518== Unified Memory profiling result:
Device "Tesla V100-SXM2-16GB (0)"
   Count  Avg Size  Min Size  Max Size  Total Size  Total Time  Name
    9739  215.59KB  64.000KB  960.00KB  2.002319GB  76.62653ms  Host To Device
    5084         -         -         -           -  707.0591ms  Gpu page fault groups
Total CPU Page faults: 6156
======== Error: Application returned non-zero code 1
```

For 512 threads per block:

```
This is the Game of Life running in parallel on a GPU.
==101593== NVPROF is profiling process 101593, command: ./gol 3 32786 1024 512 0
==101593== Profiling application: ./gol 3 32786 1024 512 0
==101593== Profiling result:
```

```
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:  100.00%  20.3965s      1024  19.918ms  19.144ms  759.87ms  gol_kernel(unsigned char const *, unsigned int,
unsigned int, unsigned char*)
      API calls:   95.05%  19.6171s         1  19.6171s  19.6171s  19.6171s  cudaDeviceSynchronize
                    3.77%  778.54ms      1024  760.30us  7.0250us  751.64ms  cudaLaunchKernel
                    0.94%  195.01ms         2  97.504ms  116.05us  194.89ms  cudaMallocManaged
                    0.22%  45.496ms         2  22.748ms  22.003ms  23.493ms  cudaFree
                    0.00%  853.64us         1  853.64us  853.64us  853.64us  cuDeviceTotalMem
                    0.00%  628.95us        97  6.4830us     223ns  246.41us  cuDeviceGetAttribute
                    0.00%  50.579us         1  50.579us  50.579us  50.579us  cuDeviceGetName
                    0.00%  4.0270us         1  4.0270us  4.0270us  4.0270us  cuDeviceGetPCIBusId
                    0.00%  2.4240us         3     808ns     413ns  1.3960us  cuDeviceGetCount
                    0.00%  1.4920us         2     746ns     331ns  1.1610us  cuDeviceGet
                    0.00%     377ns         1     377ns     377ns     377ns  cuDeviceGetUuid

==101593== Unified Memory profiling result:
Device "Tesla V100-SXM2-16GB (0)"
   Count  Avg Size  Min Size  Max Size  Total Size  Total Time  Name
    9661  217.33KB  64.000KB  960.00KB  2.002319GB  76.43962ms  Host To Device
    4945         -         -         -           -  740.7441ms  Gpu page fault groups
Total CPU Page faults: 6156
======== Error: Application returned non-zero code 1
```

For 1024 threads per block:

```
This is the Game of Life running in parallel on a GPU.
==101828== NVPROF is profiling process 101828, command: ./gol 3 32786 1024 1024 0
==101828== Profiling application: ./gol 3 32786 1024 1024 0
==101828== Profiling result:
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:  100.00%  21.0946s      1024  20.600ms  19.850ms  739.08ms  gol_kernel(unsigned char const *, unsigned int,
unsigned int, unsigned char*)
      API calls:   95.19%  20.3352s         1  20.3352s  20.3352s  20.3352s  cudaDeviceSynchronize
                    3.55%  758.47ms      1024  740.70us  6.9700us  730.96ms  cudaLaunchKernel
                    1.07%  228.42ms         2  114.21ms  169.99us  228.25ms  cudaMallocManaged
                    0.18%  38.669ms         2  19.335ms  18.626ms  20.043ms  cudaFree
                    0.00%  931.04us         1  931.04us  931.04us  931.04us  cuDeviceTotalMem
                    0.00%  783.34us        97  8.0750us     289ns  305.77us  cuDeviceGetAttribute
                    0.00%  67.437us         1  67.437us  67.437us  67.437us  cuDeviceGetName
                    0.00%  3.3000us         1  3.3000us  3.3000us  3.3000us  cuDeviceGetPCIBusId
                    0.00%  2.7820us         3     927ns     540ns  1.5470us  cuDeviceGetCount
                    0.00%  1.0230us         2     511ns     459ns     564ns  cuDeviceGet
                    0.00%     486ns         1     486ns     486ns     486ns  cuDeviceGetUuid

==101828== Unified Memory profiling result:
Device "Tesla V100-SXM2-16GB (0)"
   Count  Avg Size  Min Size  Max Size  Total Size  Total Time  Name
    8891  236.15KB  64.000KB  960.00KB  2.002319GB  74.78941ms  Host To Device
    4608         -         -         -           -  718.5453ms  Gpu page fault groups
Total CPU Page faults: 6156
======== Error: Application returned non-zero code 1
```

According to nvprof, the time spent in the CUDA gol kernel function is 24.978s, 20.010s, 20.052s, 20.396s and 21.095s for each block size individually in 32786*32786 world.

From these analysis results, we can find that the largest number of threads won't yield the fastest execution time. At the beginning, more threads have better parallel performance. But when the threads number goes up to a certain amount, increasing the number of threads would increase the time for synchronizing all threads within the blocks, resulting in longer execution time. The optimal number of threads per block in my experiment is 128. The results in the nvprof command can also support the reasoning.

The table of the "cell updates per second" rate for each thread blocksize / worldsize configuration is shown below:

"Cells Updates per Second" Rate

threads	1024	2048	4096	8192	16384	32786
64	4814985758	14708792110	27443880486	37145663101	41124761661	41951365161
128	5478274612	16146493594	34567141215	47556731305	52109555819	53888177769
256	5450466112	15790320941	34497729285	47490999818	52050351627	53701513348
512	5422938505	15907286281	33294320124	46494909835	51736854309	53018636824
1024	5368709120	15339168914	33038209969	43994543365	48668184657	50300229361

The cell with the yellow highlight is the one yielding the fastest "cell updates per second" rate. So the configuration with 128 threads per block and 32786*32786 world size has the best performance.

In order to run the jobs easier, we can add the thread number per block as the 4th argument and add the boolean value to control the output is on or off as the 5th argument. So the argument count "argc" should be 6. If argc is not equal to 6, an error is shown and the program is terminated. The argument vector "argv" is an array of pointers to character strings that contain the arguments, one per string. For example, the command line arguments "./gol 4 64 2 2 0" will execute the GOL program using pattern 4 for a world size of 64*64, performing two iterations, assigning 2 threads per block and turning off the output.