HW4 – Report

1. **Name:** 徐嘉駿, **Institute:** 資應所, **Student ID:** 107065528

2. Implement

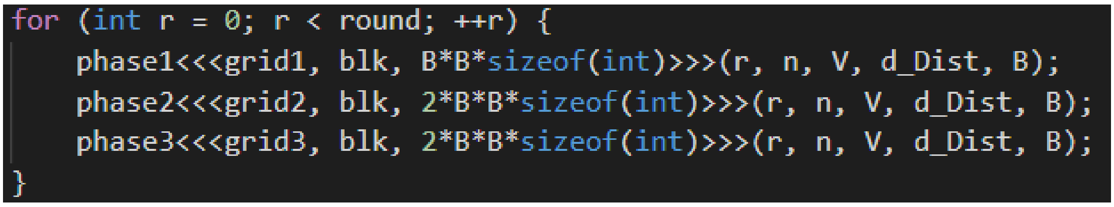
- **Divide Data:** 如同這次作業 spec 的切割方法，將 data 切成 blocks，以下為 configuration:

  (1) **Blocking Factor:** 32

  (2) **Blocks:** (data 量/32)^2

  (3) **Threads:** 32*32

- **Implementation:** 如同這次作業 spec 的實作方法，將過程分為好幾 rounds，每 round 分為 3 個 phases:

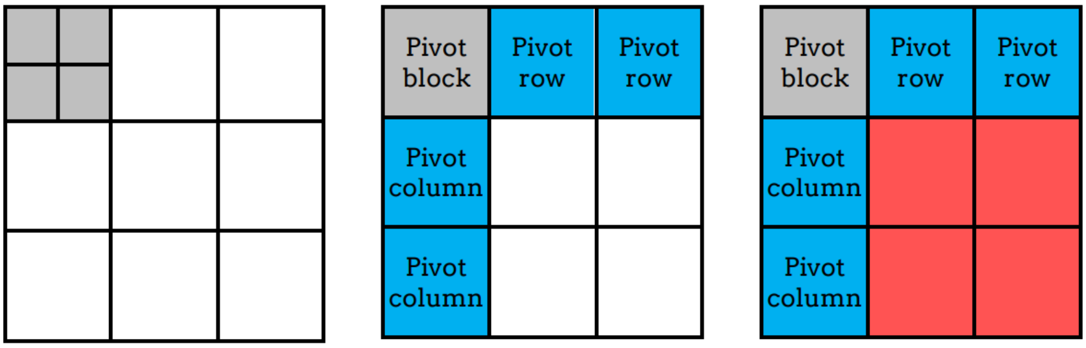


```
for (int r = 0; r < round; ++r) {
    phase1<<<grid1, blk, B*B*sizeof(int)>>>(r, n, V, d_Dist, B);
    phase2<<<grid2, blk, 2*B*B*sizeof(int)>>>(r, n, V, d_Dist, B);
    phase3<<<grid3, blk, 2*B*B*sizeof(int)>>>(r, n, V, d_Dist, B);
}
```

(1) **Phase 1:** 運算 pivot block

(2) **Phase 2:** 運算 pivot row blocks & pivot column blocks

(3) **Phase 3:** 運算剩下的 blocks



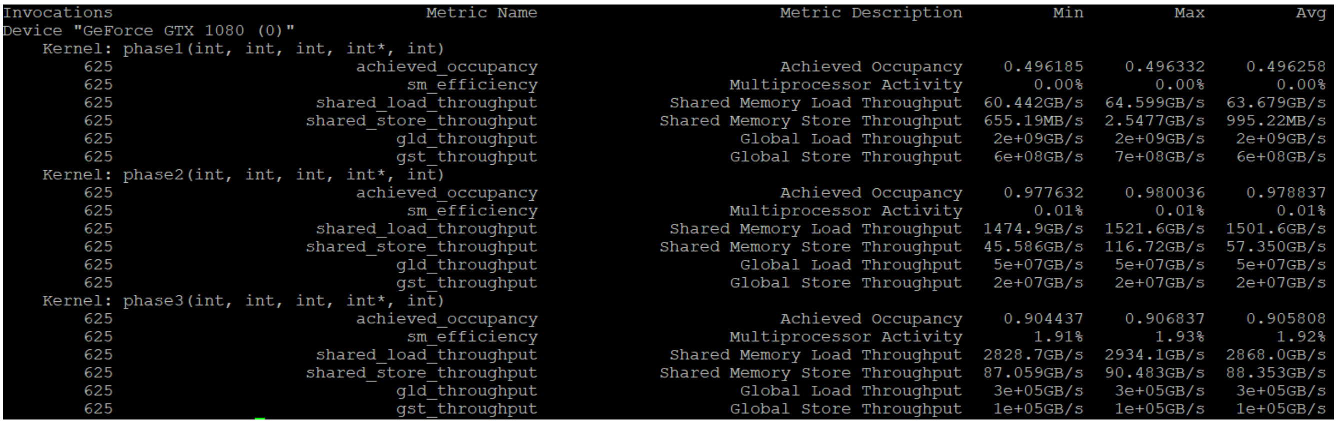
(a) Phase 1 (b) Phase 2 (c) Phase 3

3. Profiling Results

使用 p20k1 做 measurement

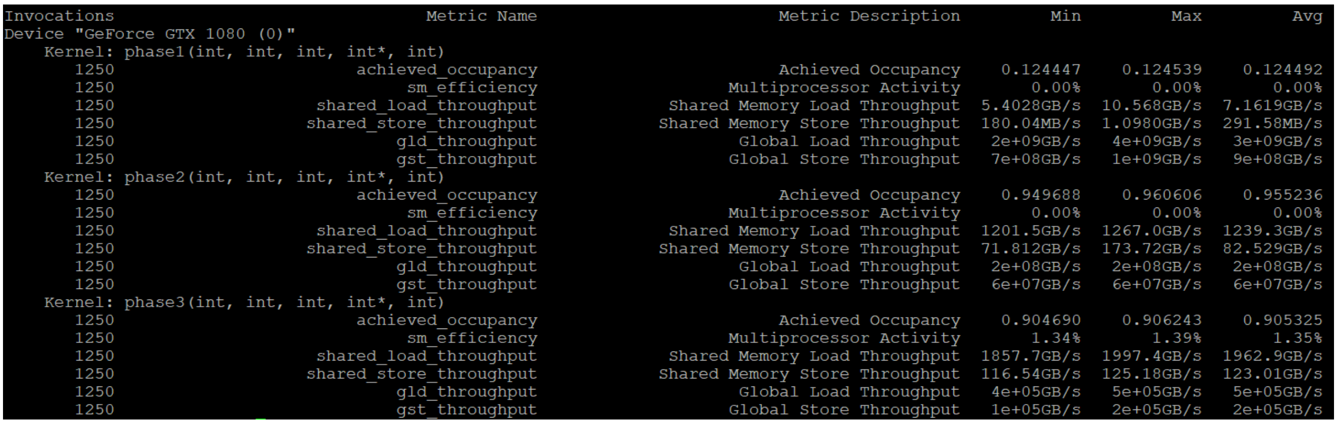
以下為 Occupancy, sm efficiency, shared memory load/store throughput, global load/store throughput 測量之結果:

- **Block dimension: (32, 32)**



```
Invocations                               Metric Name                        Metric Description         Min         Max         Avg
Device "GeForce GTX 1080 (0)"
    Kernel: phase1(int, int, int, int*, int)
        625                        achieved_occupancy                        Achieved Occupancy    0.496185    0.496332    0.496258
        625                             sm_efficiency                   Multiprocessor Activity       0.00%       0.00%       0.00%
        625                    shared_load_throughput             Shared Memory Load Throughput  60.442GB/s  64.599GB/s  63.679GB/s
        625                   shared_store_throughput            Shared Memory Store Throughput  655.19MB/s  2.5477GB/s  995.22MB/s
        625                            gld_throughput                    Global Load Throughput   2e+09GB/s   2e+09GB/s   2e+09GB/s
        625                            gst_throughput                   Global Store Throughput   6e+08GB/s   7e+08GB/s   6e+08GB/s
    Kernel: phase2(int, int, int, int*, int)
        625                        achieved_occupancy                        Achieved Occupancy    0.977632    0.980036    0.978837
        625                             sm_efficiency                   Multiprocessor Activity       0.01%       0.01%       0.01%
        625                    shared_load_throughput             Shared Memory Load Throughput  1474.9GB/s  1521.6GB/s  1501.6GB/s
        625                   shared_store_throughput            Shared Memory Store Throughput  45.586GB/s  116.72GB/s  57.350GB/s
        625                            gld_throughput                    Global Load Throughput   5e+07GB/s   5e+07GB/s   5e+07GB/s
        625                            gst_throughput                   Global Store Throughput   2e+07GB/s   2e+07GB/s   2e+07GB/s
    Kernel: phase3(int, int, int, int*, int)
        625                        achieved_occupancy                        Achieved Occupancy    0.904437    0.906837    0.905808
        625                             sm_efficiency                   Multiprocessor Activity       1.91%       1.93%       1.92%
        625                    shared_load_throughput             Shared Memory Load Throughput  2828.7GB/s  2934.1GB/s  2868.0GB/s
        625                   shared_store_throughput            Shared Memory Store Throughput  87.059GB/s  90.483GB/s  88.353GB/s
        625                            gld_throughput                    Global Load Throughput   3e+05GB/s   3e+05GB/s   3e+05GB/s
        625                            gst_throughput                   Global Store Throughput   1e+05GB/s   1e+05GB/s   1e+05GB/s
```

- **Block dimension: (16, 16)**



```
Invocations                               Metric Name                        Metric Description         Min         Max         Avg
Device "GeForce GTX 1080 (0)"
    Kernel: phase1(int, int, int, int*, int)
       1250                        achieved_occupancy                        Achieved Occupancy    0.124447    0.124539    0.124492
       1250                             sm_efficiency                   Multiprocessor Activity       0.00%       0.00%       0.00%
       1250                    shared_load_throughput             Shared Memory Load Throughput  5.4028GB/s  10.568GB/s  7.1619GB/s
       1250                   shared_store_throughput            Shared Memory Store Throughput  180.04MB/s  1.0980GB/s  291.58MB/s
       1250                            gld_throughput                    Global Load Throughput   2e+09GB/s   4e+09GB/s   3e+09GB/s
       1250                            gst_throughput                   Global Store Throughput   7e+08GB/s   1e+09GB/s   9e+08GB/s
    Kernel: phase2(int, int, int, int*, int)
       1250                        achieved_occupancy                        Achieved Occupancy    0.949688    0.960606    0.955236
       1250                             sm_efficiency                   Multiprocessor Activity       0.00%       0.00%       0.00%
       1250                    shared_load_throughput             Shared Memory Load Throughput  1201.5GB/s  1267.0GB/s  1239.3GB/s
       1250                   shared_store_throughput            Shared Memory Store Throughput  71.812GB/s  173.72GB/s  82.529GB/s
       1250                            gld_throughput                    Global Load Throughput   2e+08GB/s   2e+08GB/s   2e+08GB/s
       1250                            gst_throughput                   Global Store Throughput   6e+07GB/s   6e+07GB/s   6e+07GB/s
    Kernel: phase3(int, int, int, int*, int)
       1250                        achieved_occupancy                        Achieved Occupancy    0.904690    0.906243    0.905325
       1250                             sm_efficiency                   Multiprocessor Activity       1.34%       1.39%       1.35%
       1250                    shared_load_throughput             Shared Memory Load Throughput  1857.7GB/s  1997.4GB/s  1962.9GB/s
       1250                   shared_store_throughput            Shared Memory Store Throughput  116.54GB/s  125.18GB/s  123.01GB/s
       1250                            gld_throughput                    Global Load Throughput   4e+05GB/s   5e+05GB/s   5e+05GB/s
       1250                            gst_throughput                   Global Store Throughput   1e+05GB/s   2e+05GB/s   2e+05GB/s
```

結論: 在 occupancy 中，block dimension 16x16 明顯比 block dimension 32x32 來的小很多，代表同時 active 的 warp 數比較少。由於 GeForce GTX 1080 每個 block 一次最多只能 launch 1024 個 threads，所以 blocking factor 最多只能設置 32，此為 bottleneck 所在。

4. Experiment & Analysis

- **Time Distribution:**

  使用以下 test cases 做 measurement:

  p11k1: vector ->11000, edge->505586

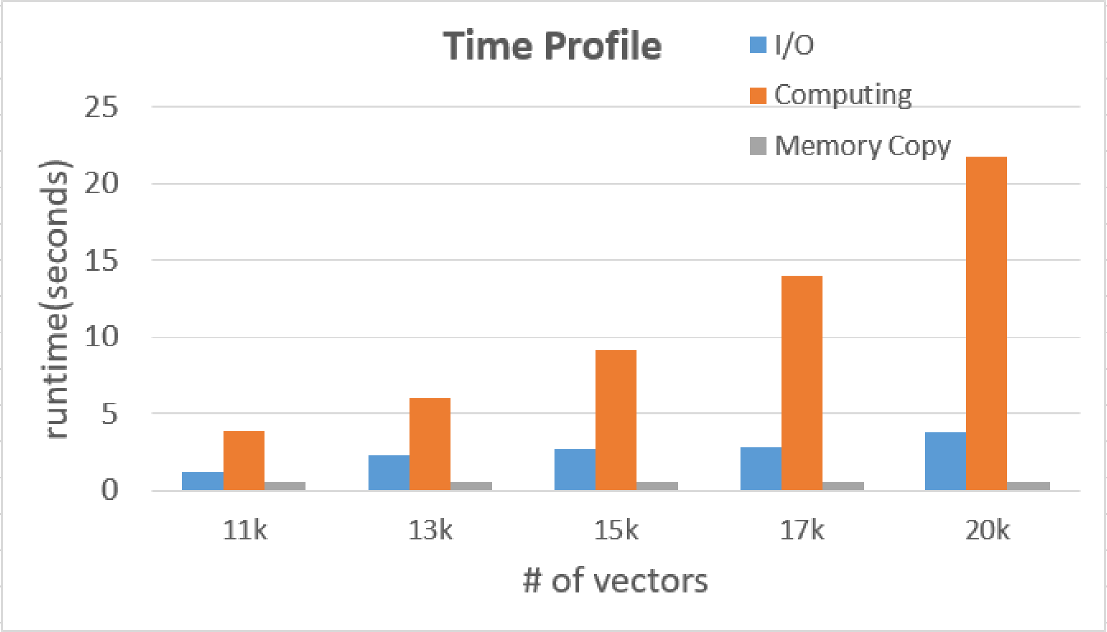
  p13k1: vector ->13000, edge->1829967

  p15k1: vector ->15000, edge->5591272

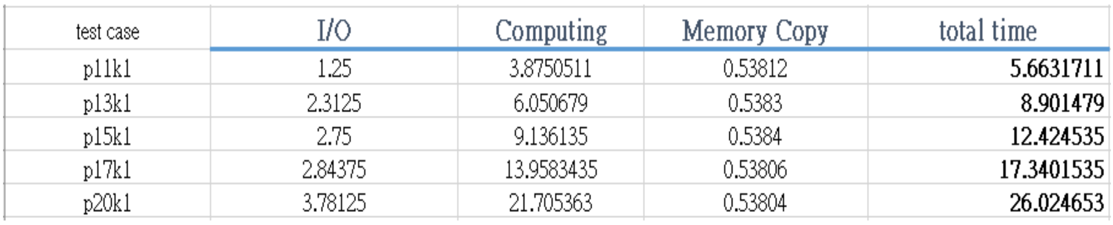
  p17k1: vector ->17000, edge->4326829

  p20k1: vector ->20000, edge->264275

  (1) Time Distribution



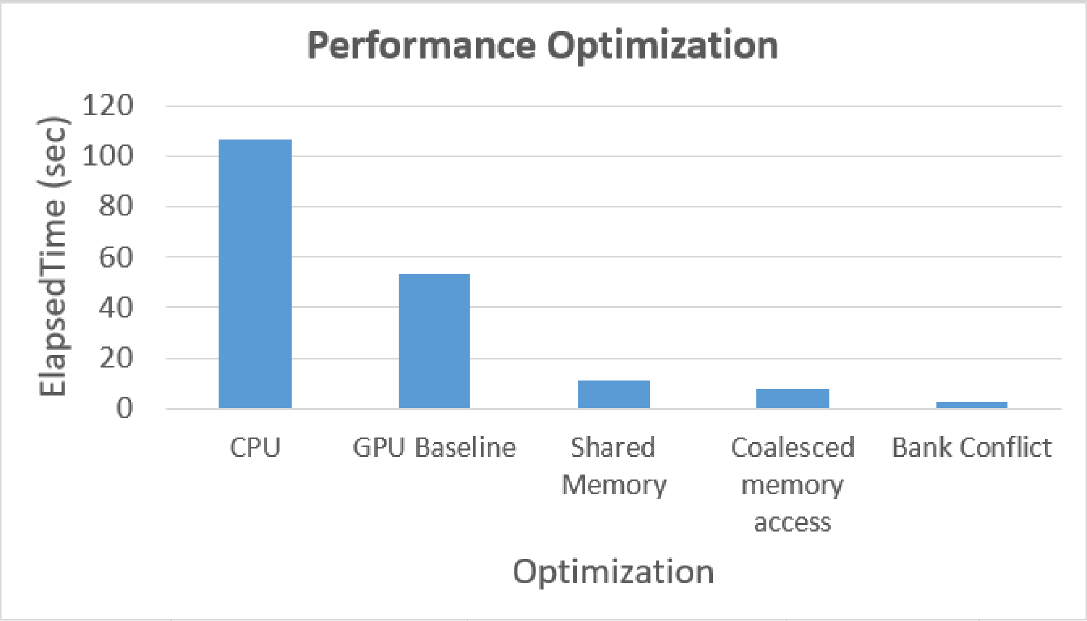
數值:



test case	I/O	Computing	Memory Copy	total time
p11k1	1.25	3.8750511	0.53812	**5.6631711**
p13k1	2.3125	6.050679	0.5383	**8.901479**
p15k1	2.75	9.136135	0.5384	**12.424535**
p17k1	2.84375	13.9583435	0.53806	**17.3401535**
p20k1	3.78125	21.705363	0.53804	**26.024653**

(2) Optimization

使用 c21.1 做 measurement



5. Conclusion

這次的作業我寫了好幾個版本，從一開始只用 global memory 做存取、copy 至 shared memory 到最後解決 bank conflict 的問題，一步一步慢慢加速，也從中了解到記憶體存取方法及位置對於 GPU 計算的重要性。