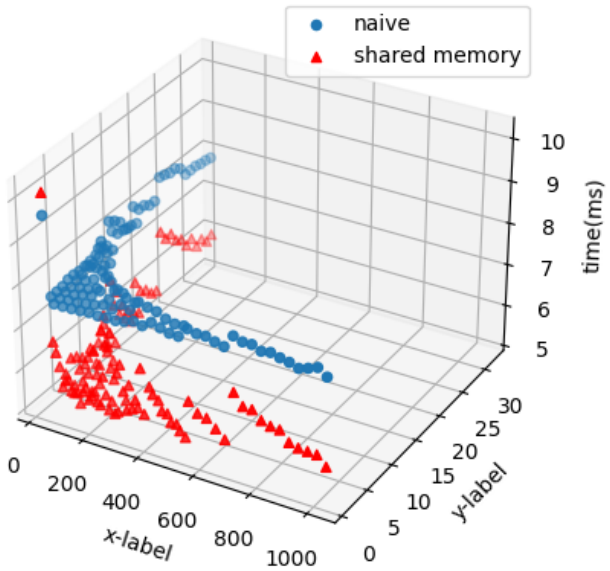
# CUDA 并行策略实验报告

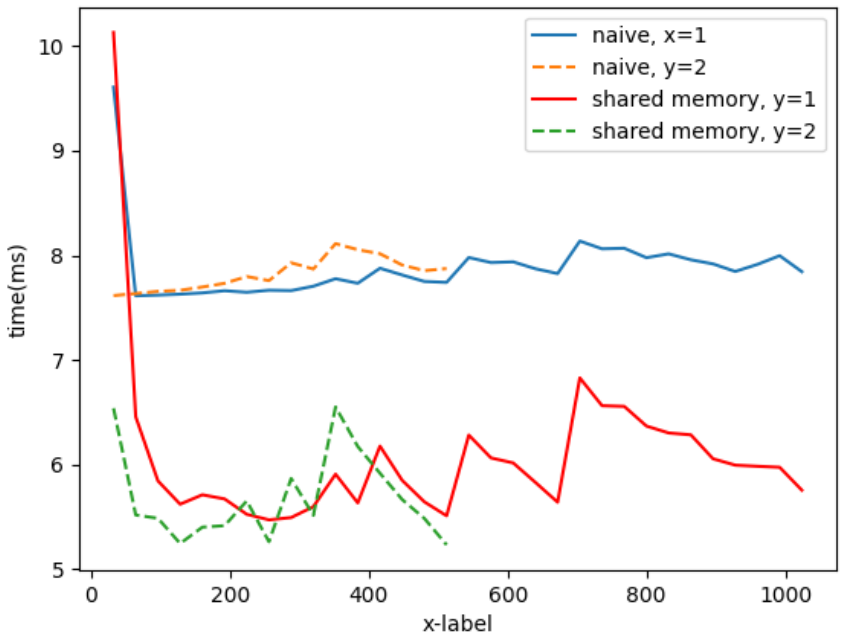
容逸朗　2020010869

## 测量结果

- 在不同的 thread block 大小组合 $(x, y)$ 以及 shared memory 的调用策略下，测量结果如下：




- 重点考虑 thread block size 为 $y = 1$ 和 $y = 2$ 的情况：




- 可以看到，当程序使用共享内存后，性能会随着线程块大小增长而出现较大的波动；

- 当 x 和 y 的乘积不为 2 的幂次时效率会降低；
- 更大的 thread block size 也不一定能带来性能的提升；
- 更完整的结果请参见附录。

# 分析

### 对于这个程序

1. 如何设置 thread block size 才可以达到最好的效果？为什么？

   线程大小为 $x = 128, y = 2$ 时效果最好：

   - 在使用共享内存的情况下用时 $5.24572ms$；
   - 不使用共享内存需要 $7.6652ms$；

   原因在于，要让程序有更好的运行条件，需要满足下列条件：

   - 较大的 thread block size（提升 SM 的利用率）
   - 同时 block size 不应过大（会导致 SM 利用率降低）
   - Thread block size 为 32 的倍数（一个 warp 是由连续 32 个线程组成的）

   因此设置 thread block size 为 $x = 128, y = 2$ 是一个比较好的选择。

2. Shared memory 总是带来优化吗？如果不是，为什么？

   在本次测试中，仅有线程大小 $x = 32, y = 1$ 时共享内存的效果较差，其余情况共享内存可以带来性能提升。一般而言，使用 shared memory 可以减少全局内存的访问次数，提升程序效率。但是当共享数据的使用次数较低的时候，从共享内存中放入 / 读出数据需要的两次操作可能会导致额外的操作时间。

3. Shared memory 在什么 thread block size 下有效果，什么时候没有？

   仅有线程大小 $x = 32, y = 1$ 时共享内存的效果较差，其余时间都比 naive 方法优秀。

4. 还有哪些可以优化的地方？

   还可以让访问的地址对齐，降低内存访问时长。

### 对于任意一个给定程序

1. 应该如何设置 thread block size?

   首先，由于同一个 block 中每个 warp （连续 32 个线程组成）同时只能执行一种指令，因此即使 warp 中有多余的硬件资源也不会被分配，所以 block size 应当为 32 的倍数；

   其次可以考虑 SM 处理的数据量，为了提升 SM 利用率，我们应当设置一个比 SM 最大并发线程数除以最大 block 数更大的值。

2. 应该如何决定 shared memory 的使用?

   当相邻进程间会使用相同的数据时，应当使用 shared memory。

# 附录

- 不同的线程块大小的运行结果如下：

| x | y | naive/ms | shared memory/ms |
|---|---|---|---|
| 32 | 1 | 9.60706 | 10.1293 |
| 32 | 2 | 7.6149 | 6.53871 |
| 32 | 3 | 7.6263 | 5.82725 |
| 32 | 4 | 7.64237 | 5.53721 |
| 32 | 5 | 7.70032 | 5.80913 |
| 32 | 6 | 7.70657 | 5.73397 |
| 32 | 7 | 7.68801 | 5.59666 |
| 32 | 8 | 7.70262 | 5.49618 |
| 32 | 9 | 7.7232 | 5.54566 |
| 32 | 10 | 7.75635 | 5.63876 |
| 32 | 11 | 7.85275 | 5.98604 |
| 32 | 12 | 7.77091 | 5.64942 |
| 32 | 13 | 8.01499 | 6.22619 |
| 32 | 14 | 7.85792 | 5.89034 |
| 32 | 15 | 7.83496 | 5.69075 |
| 32 | 16 | 7.78819 | 5.50345 |
| 32 | 17 | 8.20045 | 6.35119 |
| 32 | 18 | 8.03485 | 6.03389 |
| 32 | 19 | 8.02778 | 5.9566 |
| 32 | 20 | 7.93103 | 5.77127 |
| 32 | 21 | 7.93101 | 5.71968 |
| 32 | 22 | 8.41136 | 7.0748 |
| 32 | 23 | 8.37649 | 6.81543 |
| 32 | 24 | 8.31771 | 6.82143 |
| 32 | 25 | 8.32459 | 6.5662 |
| 32 | 26 | 8.22842 | 6.48532 |
| 32 | 27 | 7.9916 | 6.23027 |
| 32 | 28 | 7.99504 | 6.29557 |
| 32 | 29 | 8.00063 | 6.00291 |
| 32 | 30 | 7.95781 | 6.07726 |

| | | | |
|---|---|---|---|
| 32 | 31 | 7.94742 | 5.92051 |
| 32 | 32 | 7.91386 | 5.9907 |
| 64 | 1 | 7.61327 | 6.45651 |
| 64 | 2 | 7.63424 | 5.51795 |
| 64 | 3 | 7.68025 | 5.56853 |
| 64 | 4 | 7.68794 | 5.32385 |
| 64 | 5 | 7.74479 | 5.60601 |
| 64 | 6 | 7.77254 | 5.61526 |
| 64 | 7 | 7.88765 | 5.88987 |
| 64 | 8 | 7.81557 | 5.41075 |
| 64 | 9 | 8.08307 | 6.09641 |
| 64 | 10 | 7.9663 | 5.70267 |
| 64 | 11 | 8.51293 | 7.09545 |
| 64 | 12 | 8.41937 | 6.72138 |
| 64 | 13 | 8.28815 | 6.2865 |
| 64 | 14 | 8.02418 | 6.04203 |
| 64 | 15 | 8.00866 | 5.89214 |
| 64 | 16 | 7.95592 | 5.77647 |
| 96 | 1 | 7.61964 | 5.8441 |
| 96 | 2 | 7.65493 | 5.48678 |
| 96 | 3 | 7.67182 | 5.32165 |
| 96 | 4 | 7.73803 | 5.49045 |
| 96 | 5 | 7.79627 | 5.56333 |
| 96 | 6 | 8.01233 | 5.99817 |
| 96 | 7 | 7.88788 | 5.54743 |
| 96 | 8 | 8.40796 | 6.64258 |
| 96 | 9 | 8.02345 | 6.09385 |
| 96 | 10 | 7.99747 | 5.75479 |
| 128 | 1 | 7.62832 | 5.62131 |
| 128 | 2 | 7.6652 | 5.24572 |
| 128 | 3 | 7.74439 | 5.51139 |
| 128 | 4 | 7.7661 | 5.29974 |
| 128 | 5 | 7.88499 | 5.60486 |
| 128 | 6 | 8.3056 | 6.56304 |
| 128 | 7 | 8.00339 | 5.93109 |
| 128 | 8 | 7.94778 | 5.60017 |
| 160 | 1 | 7.64056 | 5.71112 |

| | | | |
|---|---|---|---|
| 160 | 2 | 7.69668 | 5.40177 |
| 160 | 3 | 7.7757 | 5.47996 |
| 160 | 4 | 7.84085 | 5.55609 |
| 160 | 5 | 8.14118 | 6.39086 |
| 160 | 6 | 7.92464 | 5.71249 |
| 192 | 1 | 7.66097 | 5.67184 |
| 192 | 2 | 7.7319 | 5.41613 |
| 192 | 3 | 7.95017 | 5.9103 |
| 192 | 4 | 8.15912 | 6.51511 |
| 192 | 5 | 7.95389 | 5.61786 |
| 224 | 1 | 7.64672 | 5.52477 |
| 224 | 2 | 7.79684 | 5.6526 |
| 224 | 3 | 7.81793 | 5.38081 |
| 224 | 4 | 7.91874 | 5.94606 |
| 256 | 1 | 7.66589 | 5.47176 |
| 256 | 2 | 7.75779 | 5.26234 |
| 256 | 3 | 8.10803 | 6.32625 |
| 256 | 4 | 7.91728 | 5.35204 |
| 288 | 1 | 7.66238 | 5.49364 |
| 288 | 2 | 7.92841 | 5.86761 |
| 288 | 3 | 7.97498 | 5.89774 |
| 320 | 1 | 7.7042 | 5.59788 |
| 320 | 2 | 7.86846 | 5.51409 |
| 320 | 3 | 7.92966 | 5.42423 |
| 352 | 1 | 7.77599 | 5.90815 |
| 352 | 2 | 8.11136 | 6.55069 |
| 384 | 1 | 7.73269 | 5.63347 |
| 384 | 2 | 8.05324 | 6.17251 |
| 416 | 1 | 7.87606 | 6.17638 |
| 416 | 2 | 8.01453 | 5.91698 |
| 448 | 1 | 7.81166 | 5.84902 |
| 448 | 2 | 7.90655 | 5.66598 |
| 480 | 1 | 7.74978 | 5.64367 |
| 480 | 2 | 7.8531 | 5.48487 |
| 512 | 1 | 7.74108 | 5.51029 |
| 512 | 2 | 7.87292 | 5.23275 |
| 544 | 1 | 7.97842 | 6.28164 |
| 576 | 1 | 7.93044 | 6.06352 |

| | | | |
|---|---|---|---|
| 608 | 1 | 7.93705 | 6.01558 |
| 640 | 1 | 7.87129 | 5.82964 |
| 672 | 1 | 7.82526 | 5.64036 |
| 704 | 1 | 8.13569 | 6.82813 |
| 736 | 1 | 8.06202 | 6.56258 |
| 768 | 1 | 8.06697 | 6.55604 |
| 800 | 1 | 7.97708 | 6.36806 |
| 832 | 1 | 8.01357 | 6.30199 |
| 864 | 1 | 7.95702 | 6.28423 |
| 896 | 1 | 7.91748 | 6.05529 |
| 928 | 1 | 7.84536 | 5.99437 |
| 960 | 1 | 7.91301 | 5.98361 |
| 992 | 1 | 7.99599 | 5.97423 |
| 1024 | 1 | 7.84309 | 5.75522 |