Московский Государственный Университет имени М. В. Ломоносова
Факультет Вычислительной Математики и Кибернетики
Кафедра Суперкомпьютеров и Квантовой Информатики

Практикум CUDA

Отчет №2

Basic Image Convolution on NVIDIA GPUs using CUDA(with performance improvements)

Работу выполнил
Кислов Евгений

Москва 2021

Постановка задачи

Для выполнения второго задания необходимо реализовать следующие оптимизации разработанной на первом этапе программы:

Оптимизации для обработки больших и малых изображений:

1. Развертка массива, где хранится изображение из массива структур в структуру
2. массивов для улучшения шаблона доступа к глобальной памяти (Pixel * -> 3 массива
3. unsigned char* для хранения 3 компонент изображения);
4. Последовательный доступ к памяти от нитей варпа к массиву с изображением;
5. Использование разделяемой (shared) памяти для применения фильтра (по аналогии со stencil);
6. Использование 3х нитей для обработки r/g/b компонент;
7. Различные походы к передаче фильтра в матрицу (full unroll, константная память);
8. Развертка циклов, применяющих фильтров внутри каждой нити;
9. Подбор оптимальных значений размера CUDA блока;
10. Минимизация числа простаивающих нитей;

Дополнительные оптимизации для обработки набора из маленьких изображений:

1. Выделение памяти (cudaMalloc) под обрабатываемые изображения 1 раз (а не каждый раз для каждого изображения заново);
2. Обработка нескольких изображений за раз одним ядром или обработка нескольких
изображений в конкурентном режиме при помощи CUDA-потоков;
3. Одновременные копирования DtoH, HtoD и запуск ядер;
4. Параллельная работа с файлами обработка изображений на GPU для групп из N -
изображений: загружаем группу из N изображений с диска, пока их обрабатываем - грузим следующую. Сохранение на диск можно отключить (ifdef __NEED_TO_SAVE__).

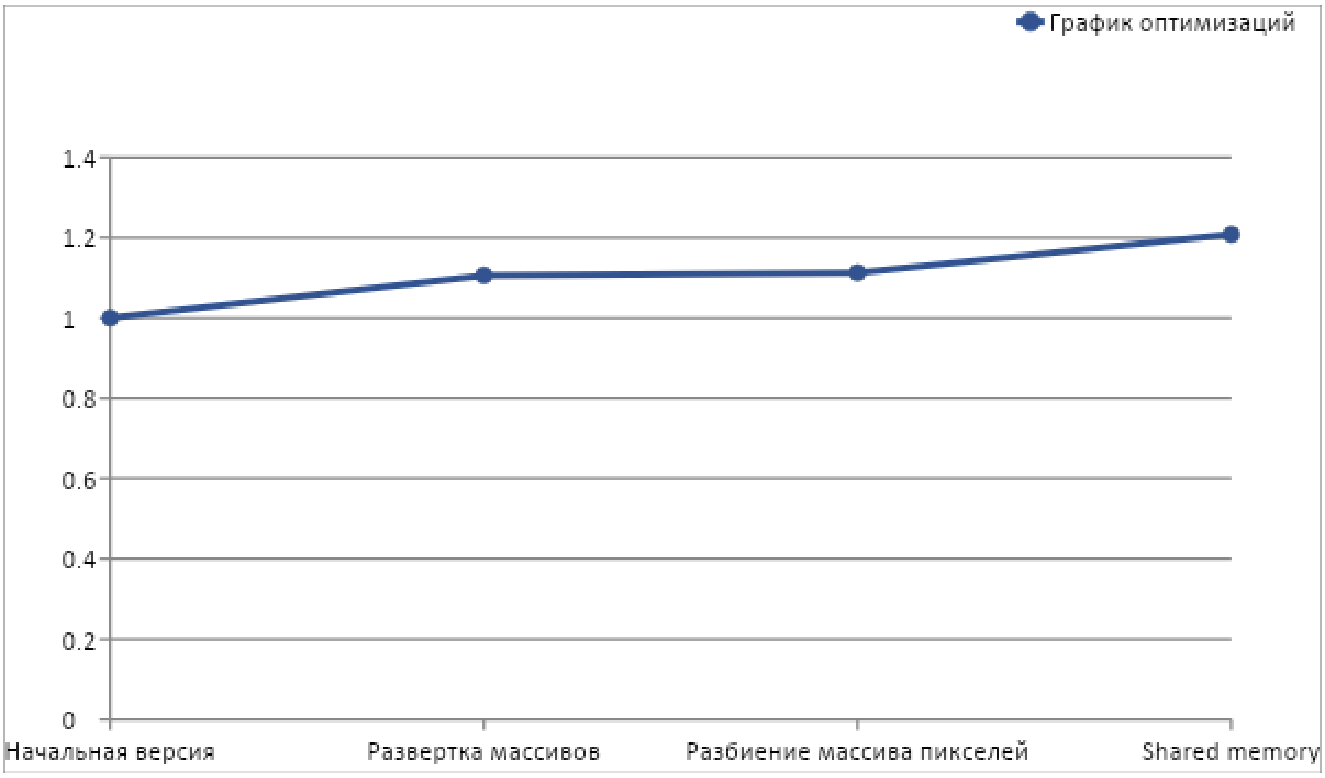
Результат

Из списка общих оптимизаций были выполнены все пункты, из списка оптимизаций для маленький изображений было выполнены пункты 2-4.

Ниже будут приведены замеры времени и графики для наиболее важных оптимизаций:

Вид оптимизации	Время вычисления на большой картинке(ms)	Время вычисления на маленьких картинках(ms)	Ускорение(в среднем по двум случаям)
Начальная версия	0.602816	0.26464	-

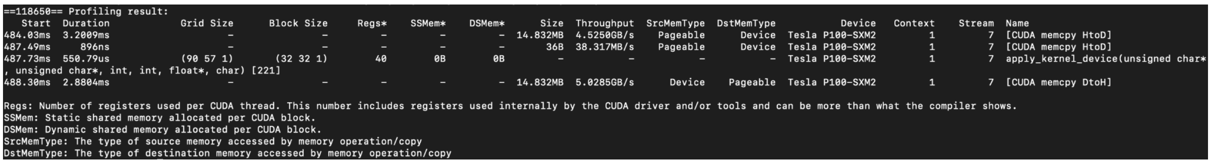
Развертка массивов	0.54472	0.216864	1.1
Разбиение массива пикселей на три	0.541745	0.213487	1.1
Shared memory	0.498944	0.19292	1.22



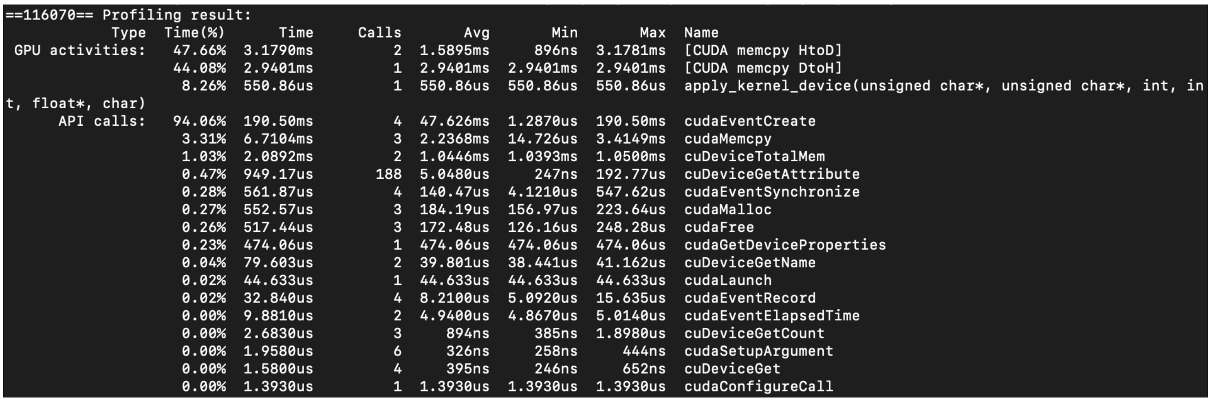
Графики оптимизаций:
После оптимизаций маленьких картинок время пересылки уменьшилось с 4.83 мс до 3.64 мс, дав ускорение в 1.32 раза.

Данные профилировщика nvprof в начальной версии:

Для больших

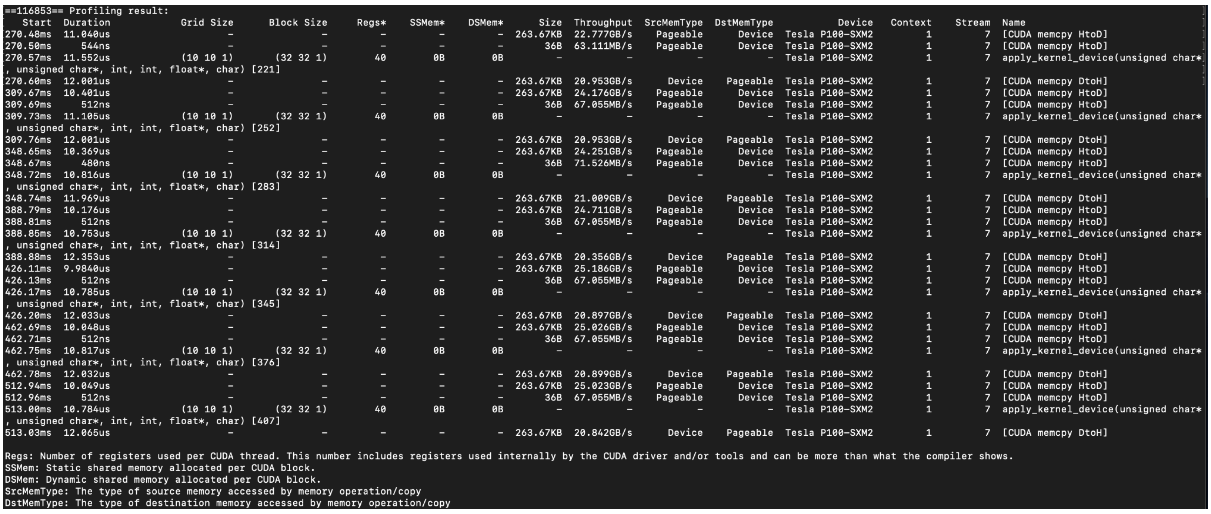


```
==118650== Profiling result:
   Start  Duration            Grid Size      Block Size     Regs*    SSMem*    DSMem*      Size  Throughput  SrcMemType  DstMemType           Device   Context    Stream  Name
484.03ms  3.2009ms                    -               -         -         -         -  14.832MB  4.5250GB/s    Pageable      Device  Tesla P100-SXM2         1         7  [CUDA memcpy HtoD]
487.49ms     896ns                    -               -         -         -         -       36B  38.317MB/s    Pageable      Device  Tesla P100-SXM2         1         7  [CUDA memcpy HtoD]
487.73ms  550.79us            (90 57 1)       (32 32 1)        40        0B        0B         -           -           -           -  Tesla P100-SXM2         1         7  apply_kernel_device(unsigned char*
, unsigned char*, int, int, float*, char) [221]
488.30ms  2.8804ms                    -               -         -         -         -  14.832MB  5.0285GB/s      Device    Pageable  Tesla P100-SXM2         1         7  [CUDA memcpy DtoH]

Regs: Number of registers used per CUDA thread. This number includes registers used internally by the CUDA driver and/or tools and can be more than what the compiler shows.
SSMem: Static shared memory allocated per CUDA block.
DSMem: Dynamic shared memory allocated per CUDA block.
SrcMemType: The type of source memory accessed by memory operation/copy
DstMemType: The type of destination memory accessed by memory operation/copy
```

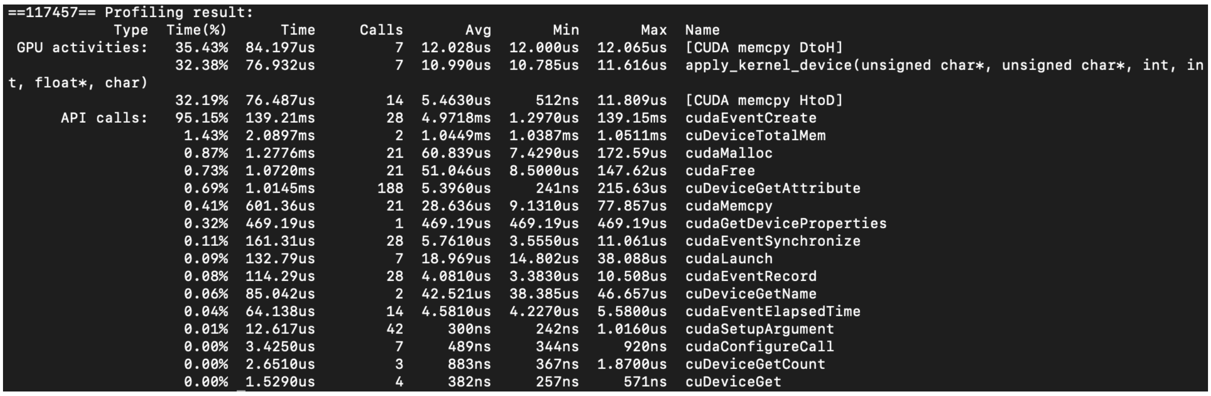


```
==116070== Profiling result:
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:   47.66%  3.1790ms         2  1.5895ms     896ns  3.1781ms  [CUDA memcpy HtoD]
                   44.08%  2.9401ms         1  2.9401ms  2.9401ms  2.9401ms  [CUDA memcpy DtoH]
                    8.26%  550.86us         1  550.86us  550.86us  550.86us  apply_kernel_device(unsigned char*, unsigned char*, int, in
t, float*, char)
      API calls:   94.06%  190.50ms         4  47.626ms  1.2870us  190.50ms  cudaEventCreate
                    3.31%  6.7104ms         3  2.2368ms  14.726us  3.4149ms  cudaMemcpy
                    1.03%  2.0892ms         2  1.0446ms  1.0393ms  1.0500ms  cuDeviceTotalMem
                    0.47%  949.17us       188  5.0480us     247ns  192.77us  cuDeviceGetAttribute
                    0.28%  561.87us         4  140.47us  4.1210us  547.62us  cudaEventSynchronize
                    0.27%  552.57us         3  184.19us  156.97us  223.64us  cudaMalloc
                    0.26%  517.44us         3  172.48us  126.16us  248.28us  cudaFree
                    0.23%  474.06us         1  474.06us  474.06us  474.06us  cudaGetDeviceProperties
                    0.04%  79.603us         2  39.801us  38.441us  41.162us  cuDeviceGetName
                    0.02%  44.633us         1  44.633us  44.633us  44.633us  cudaLaunch
                    0.02%  32.840us         4  8.2100us  5.0920us  15.635us  cudaEventRecord
                    0.00%  9.8810us         2  4.9400us  4.8670us  5.0140us  cudaEventElapsedTime
                    0.00%  2.6830us         3     894ns     385ns  1.8980us  cuDeviceGetCount
                    0.00%  1.9580us         6     326ns     258ns     444ns  cudaSetupArgument
                    0.00%  1.5800us         4     395ns     246ns     652ns  cuDeviceGet
                    0.00%  1.3930us         1  1.3930us  1.3930us  1.3930us  cudaConfigureCall
```

Для маленьких



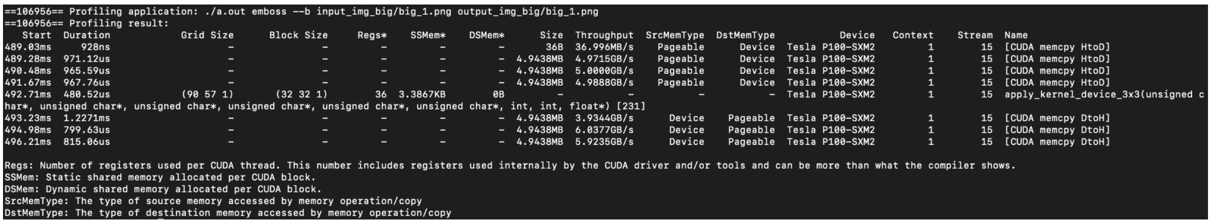
```
==116853== Profiling result:
   Start  Duration            Grid Size      Block Size     Regs*    SSMem*    DSMem*      Size  Throughput  SrcMemType  DstMemType           Device   Context    Stream  Name
270.48ms  11.040us                    -               -         -         -         -  263.67KB  22.777GB/s    Pageable      Device  Tesla P100-SXM2         1         7  [CUDA memcpy HtoD]
270.50ms     544ns                    -               -         -         -         -       36B  63.111MB/s    Pageable      Device  Tesla P100-SXM2         1         7  [CUDA memcpy HtoD]
270.57ms  11.552us           (10 10 1)       (32 32 1)        40        0B        0B         -           -           -           -  Tesla P100-SXM2         1         7  apply_kernel_device(unsigned char*
, unsigned char*, int, int, float*, char) [221]
270.60ms  12.001us                    -               -         -         -         -  263.67KB  20.953GB/s      Device    Pageable  Tesla P100-SXM2         1         7  [CUDA memcpy DtoH]
309.67ms  10.401us                    -               -         -         -         -  263.67KB  24.176GB/s    Pageable      Device  Tesla P100-SXM2         1         7  [CUDA memcpy HtoD]
309.69ms     512ns                    -               -         -         -         -       36B  67.055MB/s    Pageable      Device  Tesla P100-SXM2         1         7  [CUDA memcpy HtoD]
309.73ms  11.105us           (10 10 1)       (32 32 1)        40        0B        0B         -           -           -           -  Tesla P100-SXM2         1         7  apply_kernel_device(unsigned char*
, unsigned char*, int, int, float*, char) [252]
309.76ms  12.001us                    -               -         -         -         -  263.67KB  20.953GB/s      Device    Pageable  Tesla P100-SXM2         1         7  [CUDA memcpy DtoH]
348.65ms  10.369us                    -               -         -         -         -  263.67KB  24.251GB/s    Pageable      Device  Tesla P100-SXM2         1         7  [CUDA memcpy HtoD]
348.67ms     480ns                    -               -         -         -         -       36B  71.526MB/s    Pageable      Device  Tesla P100-SXM2         1         7  [CUDA memcpy HtoD]
348.72ms  10.816us           (10 10 1)       (32 32 1)        40        0B        0B         -           -           -           -  Tesla P100-SXM2         1         7  apply_kernel_device(unsigned char*
, unsigned char*, int, int, float*, char) [283]
348.74ms  11.969us                    -               -         -         -         -  263.67KB  21.009GB/s      Device    Pageable  Tesla P100-SXM2         1         7  [CUDA memcpy DtoH]
388.79ms  10.176us                    -               -         -         -         -  263.67KB  24.711GB/s    Pageable      Device  Tesla P100-SXM2         1         7  [CUDA memcpy HtoD]
388.81ms     512ns                    -               -         -         -         -       36B  67.055MB/s    Pageable      Device  Tesla P100-SXM2         1         7  [CUDA memcpy HtoD]
388.85ms  10.753us           (10 10 1)       (32 32 1)        40        0B        0B         -           -           -           -  Tesla P100-SXM2         1         7  apply_kernel_device(unsigned char*
, unsigned char*, int, int, float*, char) [314]
388.88ms  12.353us                    -               -         -         -         -  263.67KB  20.356GB/s      Device    Pageable  Tesla P100-SXM2         1         7  [CUDA memcpy DtoH]
426.11ms  9.9840us                    -               -         -         -         -  263.67KB  25.186GB/s    Pageable      Device  Tesla P100-SXM2         1         7  [CUDA memcpy HtoD]
426.13ms     512ns                    -               -         -         -         -       36B  67.055MB/s    Pageable      Device  Tesla P100-SXM2         1         7  [CUDA memcpy HtoD]
426.17ms  10.785us           (10 10 1)       (32 32 1)        40        0B        0B         -           -           -           -  Tesla P100-SXM2         1         7  apply_kernel_device(unsigned char*
, unsigned char*, int, int, float*, char) [345]
426.20ms  12.033us                    -               -         -         -         -  263.67KB  20.897GB/s      Device    Pageable  Tesla P100-SXM2         1         7  [CUDA memcpy DtoH]
462.69ms  10.048us                    -               -         -         -         -  263.67KB  25.026GB/s    Pageable      Device  Tesla P100-SXM2         1         7  [CUDA memcpy HtoD]
462.71ms     512ns                    -               -         -         -         -       36B  67.055MB/s    Pageable      Device  Tesla P100-SXM2         1         7  [CUDA memcpy HtoD]
462.75ms  10.817us           (10 10 1)       (32 32 1)        40        0B        0B         -           -           -           -  Tesla P100-SXM2         1         7  apply_kernel_device(unsigned char*
, unsigned char*, int, int, float*, char) [376]
462.78ms  12.032us                    -               -         -         -         -  263.67KB  20.899GB/s      Device    Pageable  Tesla P100-SXM2         1         7  [CUDA memcpy DtoH]
512.94ms  10.049us                    -               -         -         -         -  263.67KB  25.023GB/s    Pageable      Device  Tesla P100-SXM2         1         7  [CUDA memcpy HtoD]
512.96ms     512ns                    -               -         -         -         -       36B  67.055MB/s    Pageable      Device  Tesla P100-SXM2         1         7  [CUDA memcpy HtoD]
513.00ms  10.784us           (10 10 1)       (32 32 1)        40        0B        0B         -           -           -           -  Tesla P100-SXM2         1         7  apply_kernel_device(unsigned char*
, unsigned char*, int, int, float*, char) [407]
513.03ms  12.065us                    -               -         -         -         -  263.67KB  20.842GB/s      Device    Pageable  Tesla P100-SXM2         1         7  [CUDA memcpy DtoH]

Regs: Number of registers used per CUDA thread. This number includes registers used internally by the CUDA driver and/or tools and can be more than what the compiler shows.
SSMem: Static shared memory allocated per CUDA block.
DSMem: Dynamic shared memory allocated per CUDA block.
SrcMemType: The type of source memory accessed by memory operation/copy
DstMemType: The type of destination memory accessed by memory operation/copy
```



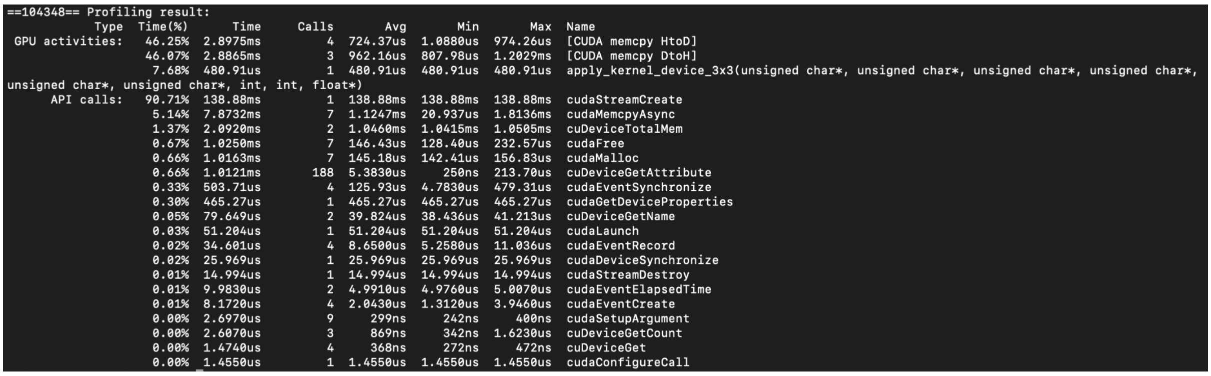
```
==117457== Profiling result:
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:   35.43%  84.197us         7  12.028us  12.000us  12.065us  [CUDA memcpy DtoH]
                   32.38%  76.932us         7  10.990us  10.785us  11.616us  apply_kernel_device(unsigned char*, unsigned char*, int, in
t, float*, char)
                   32.19%  76.487us        14  5.4630us     512ns  11.809us  [CUDA memcpy HtoD]
      API calls:   95.15%  139.21ms        28  4.9718ms  1.2970us  139.15ms  cudaEventCreate
                    1.43%  2.0897ms         2  1.0449ms  1.0387ms  1.0511ms  cuDeviceTotalMem
                    0.87%  1.2776ms        21  60.839us  7.4290us  172.59us  cudaMalloc
                    0.73%  1.0720ms        21  51.046us  8.5000us  147.62us  cudaFree
                    0.69%  1.0145ms       188  5.3960us     241ns  215.63us  cuDeviceGetAttribute
                    0.41%  601.36us        21  28.636us  9.1310us  77.857us  cudaMemcpy
                    0.32%  469.19us         1  469.19us  469.19us  469.19us  cudaGetDeviceProperties
                    0.11%  161.31us        28  5.7610us  3.5550us  11.061us  cudaEventSynchronize
                    0.09%  132.79us         7  18.969us  14.802us  38.088us  cudaLaunch
                    0.08%  114.29us        28  4.0810us  3.3830us  10.508us  cudaEventRecord
                    0.06%  85.042us         2  42.521us  38.385us  46.657us  cuDeviceGetName
                    0.04%  64.138us        14  4.5810us  4.2270us  5.5800us  cudaEventElapsedTime
                    0.01%  12.617us        42     300ns     242ns  1.0160us  cudaSetupArgument
                    0.00%  3.4250us         7     489ns     344ns     920ns  cudaConfigureCall
                    0.00%  2.6510us         3     883ns     367ns  1.8700us  cuDeviceGetCount
                    0.00%  1.5290us         4     382ns     257ns     571ns  cuDeviceGet
```

Данные профилировщика nvprof в оптимизорованной версии:

Для больших

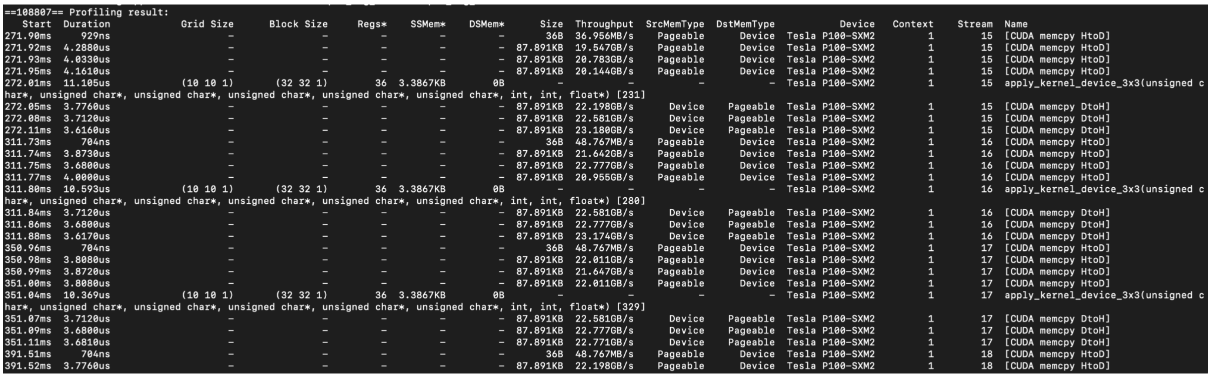


```
==106956== Profiling application: ./a.out emboss --b input_img_big/big_1.png output_img_big/big_1.png
==106956== Profiling result:
   Start  Duration            Grid Size      Block Size     Regs*    SSMem*    DSMem*      Size  Throughput  SrcMemType  DstMemType           Device   Context    Stream  Name
489.03ms     928ns                    -               -         -         -         -       36B  36.996MB/s    Pageable      Device  Tesla P100-SXM2         1        15  [CUDA memcpy HtoD]
489.28ms  971.12us                    -               -         -         -         -  4.9438MB  4.9715GB/s    Pageable      Device  Tesla P100-SXM2         1        15  [CUDA memcpy HtoD]
490.48ms  965.59us                    -               -         -         -         -  4.9438MB  5.0000GB/s    Pageable      Device  Tesla P100-SXM2         1        15  [CUDA memcpy HtoD]
491.67ms  967.76us                    -               -         -         -         -  4.9438MB  4.9888GB/s    Pageable      Device  Tesla P100-SXM2         1        15  [CUDA memcpy HtoD]
492.71ms  480.52us           (90 57 1)       (32 32 1)        36  3.3867KB        0B         -           -           -           -  Tesla P100-SXM2         1        15  apply_kernel_device_3x3(unsigned c
har*, unsigned char*, unsigned char*, unsigned char*, unsigned char*, unsigned char*, int, int, float*) [231]
493.23ms  1.2271ms                    -               -         -         -         -  4.9438MB  3.9344GB/s      Device    Pageable  Tesla P100-SXM2         1        15  [CUDA memcpy DtoH]
494.98ms  799.63us                    -               -         -         -         -  4.9438MB  6.0377GB/s      Device    Pageable  Tesla P100-SXM2         1        15  [CUDA memcpy DtoH]
496.21ms  815.06us                    -               -         -         -         -  4.9438MB  5.9235GB/s      Device    Pageable  Tesla P100-SXM2         1        15  [CUDA memcpy DtoH]

Regs: Number of registers used per CUDA thread. This number includes registers used internally by the CUDA driver and/or tools and can be more than what the compiler shows.
SSMem: Static shared memory allocated per CUDA block.
DSMem: Dynamic shared memory allocated per CUDA block.
SrcMemType: The type of source memory accessed by memory operation/copy
DstMemType: The type of destination memory accessed by memory operation/copy
```

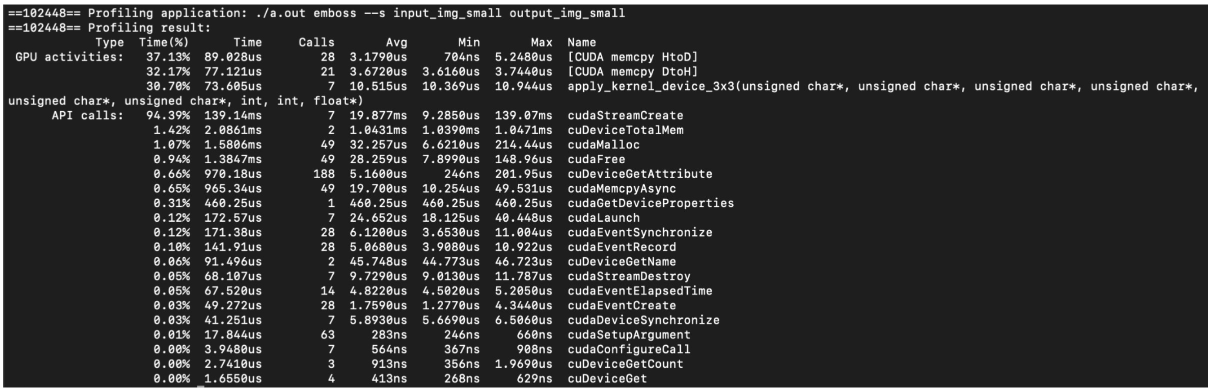


```
==104348== Profiling result:
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:   46.25%  2.8975ms         4  724.37us  1.0880us  974.26us  [CUDA memcpy HtoD]
                   46.07%  2.8865ms         3  962.16us  807.98us  1.2029ms  [CUDA memcpy DtoH]
                    7.68%  480.91us         1  480.91us  480.91us  480.91us  apply_kernel_device_3x3(unsigned char*, unsigned char*, unsigned char*, unsigned char*,
unsigned char*, unsigned char*, int, int, float*)
      API calls:   90.71%  138.88ms         1  138.88ms  138.88ms  138.88ms  cudaStreamCreate
                    5.14%  7.8732ms         7  1.1247ms  20.937us  1.8136ms  cudaMemcpyAsync
                    1.37%  2.0920ms         2  1.0460ms  1.0415ms  1.0505ms  cuDeviceTotalMem
                    0.67%  1.0250ms         7  146.43us  128.40us  232.57us  cudaFree
                    0.66%  1.0163ms         7  145.18us  142.41us  156.83us  cudaMalloc
                    0.66%  1.0121ms       188  5.3830us     250ns  213.70us  cuDeviceGetAttribute
                    0.33%  503.71us         4  125.93us  4.7830us  479.31us  cudaEventSynchronize
                    0.30%  465.27us         1  465.27us  465.27us  465.27us  cudaGetDeviceProperties
                    0.05%  79.649us         2  39.824us  38.436us  41.213us  cuDeviceGetName
                    0.03%  51.204us         1  51.204us  51.204us  51.204us  cudaLaunch
                    0.02%  34.601us         4  8.6500us  5.2580us  11.036us  cudaEventRecord
                    0.02%  25.969us         1  25.969us  25.969us  25.969us  cudaDeviceSynchronize
                    0.01%  14.994us         1  14.994us  14.994us  14.994us  cudaStreamDestroy
                    0.01%  9.9830us         2  4.9910us  4.9760us  5.0070us  cudaEventElapsedTime
                    0.01%  8.1720us         4  2.0430us  1.3120us  3.9460us  cudaEventCreate
                    0.00%  2.6970us         9     299ns     242ns     400ns  cudaSetupArgument
                    0.00%  2.6070us         3     869ns     342ns  1.6230us  cuDeviceGetCount
                    0.00%  1.4740us         4     368ns     272ns     472ns  cuDeviceGet
                    0.00%  1.4550us         1  1.4550us  1.4550us  1.4550us  cudaConfigureCall
```

Для маленьких



```
==108807== Profiling result:
   Start  Duration            Grid Size      Block Size     Regs*    SSMem*    DSMem*      Size  Throughput  SrcMemType  DstMemType           Device   Context    Stream  Name
271.90ms     929ns                    -               -         -         -         -       36B  36.956MB/s    Pageable      Device  Tesla P100-SXM2         1        15  [CUDA memcpy HtoD]
271.92ms  4.2880us                    -               -         -         -         -  87.891KB  19.547GB/s    Pageable      Device  Tesla P100-SXM2         1        15  [CUDA memcpy HtoD]
271.93ms  4.0330us                    -               -         -         -         -  87.891KB  20.783GB/s    Pageable      Device  Tesla P100-SXM2         1        15  [CUDA memcpy HtoD]
271.95ms  4.1610us                    -               -         -         -         -  87.891KB  20.144GB/s    Pageable      Device  Tesla P100-SXM2         1        15  [CUDA memcpy HtoD]
272.01ms  11.105us           (10 10 1)       (32 32 1)        36  3.3867KB        0B         -           -           -           -  Tesla P100-SXM2         1        15  apply_kernel_device_3x3(unsigned c
har*, unsigned char*, unsigned char*, unsigned char*, unsigned char*, unsigned char*, int, int, float*) [231]
272.05ms  3.7760us                    -               -         -         -         -  87.891KB  22.198GB/s      Device    Pageable  Tesla P100-SXM2         1        15  [CUDA memcpy DtoH]
272.08ms  3.7120us                    -               -         -         -         -  87.891KB  22.581GB/s      Device    Pageable  Tesla P100-SXM2         1        15  [CUDA memcpy DtoH]
272.11ms  3.6160us                    -               -         -         -         -  87.891KB  23.180GB/s      Device    Pageable  Tesla P100-SXM2         1        15  [CUDA memcpy DtoH]
311.73ms     704ns                    -               -         -         -         -       36B  48.767MB/s    Pageable      Device  Tesla P100-SXM2         1        16  [CUDA memcpy HtoD]
311.74ms  3.8730us                    -               -         -         -         -  87.891KB  21.642GB/s    Pageable      Device  Tesla P100-SXM2         1        16  [CUDA memcpy HtoD]
311.75ms  3.6800us                    -               -         -         -         -  87.891KB  22.777GB/s    Pageable      Device  Tesla P100-SXM2         1        16  [CUDA memcpy HtoD]
311.77ms  4.0000us                    -               -         -         -         -  87.891KB  20.955GB/s    Pageable      Device  Tesla P100-SXM2         1        16  [CUDA memcpy HtoD]
311.80ms  10.593us           (10 10 1)       (32 32 1)        36  3.3867KB        0B         -           -           -           -  Tesla P100-SXM2         1        16  apply_kernel_device_3x3(unsigned c
har*, unsigned char*, unsigned char*, unsigned char*, unsigned char*, unsigned char*, int, int, float*) [280]
311.84ms  3.7120us                    -               -         -         -         -  87.891KB  22.581GB/s      Device    Pageable  Tesla P100-SXM2         1        16  [CUDA memcpy DtoH]
311.86ms  3.6800us                    -               -         -         -         -  87.891KB  22.777GB/s      Device    Pageable  Tesla P100-SXM2         1        16  [CUDA memcpy DtoH]
311.88ms  3.6170us                    -               -         -         -         -  87.891KB  23.174GB/s      Device    Pageable  Tesla P100-SXM2         1        16  [CUDA memcpy DtoH]
350.96ms     704ns                    -               -         -         -         -       36B  48.767MB/s    Pageable      Device  Tesla P100-SXM2         1        17  [CUDA memcpy HtoD]
350.98ms  3.8080us                    -               -         -         -         -  87.891KB  22.011GB/s    Pageable      Device  Tesla P100-SXM2         1        17  [CUDA memcpy HtoD]
350.99ms  3.8720us                    -               -         -         -         -  87.891KB  21.647GB/s    Pageable      Device  Tesla P100-SXM2         1        17  [CUDA memcpy HtoD]
351.00ms  3.8080us                    -               -         -         -         -  87.891KB  22.011GB/s    Pageable      Device  Tesla P100-SXM2         1        17  [CUDA memcpy HtoD]
351.04ms  10.369us           (10 10 1)       (32 32 1)        36  3.3867KB        0B         -           -           -           -  Tesla P100-SXM2         1        17  apply_kernel_device_3x3(unsigned c
har*, unsigned char*, unsigned char*, unsigned char*, unsigned char*, unsigned char*, int, int, float*) [329]
351.07ms  3.7120us                    -               -         -         -         -  87.891KB  22.581GB/s      Device    Pageable  Tesla P100-SXM2         1        17  [CUDA memcpy DtoH]
351.09ms  3.6800us                    -               -         -         -         -  87.891KB  22.777GB/s      Device    Pageable  Tesla P100-SXM2         1        17  [CUDA memcpy DtoH]
351.11ms  3.6810us                    -               -         -         -         -  87.891KB  22.771GB/s      Device    Pageable  Tesla P100-SXM2         1        17  [CUDA memcpy DtoH]
391.51ms     704ns                    -               -         -         -         -       36B  48.767MB/s    Pageable      Device  Tesla P100-SXM2         1        18  [CUDA memcpy HtoD]
391.52ms  3.7760us                    -               -         -         -         -  87.891KB  22.198GB/s    Pageable      Device  Tesla P100-SXM2         1        18  [CUDA memcpy HtoD]
```



```
==102448== Profiling application: ./a.out emboss --s input_img_small output_img_small
==102448== Profiling result:
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:   37.13%  89.028us        28  3.1790us     704ns  5.2480us  [CUDA memcpy HtoD]
                   32.17%  77.121us        21  3.6720us  3.6160us  3.7440us  [CUDA memcpy DtoH]
                   30.70%  73.605us         7  10.515us  10.369us  10.944us  apply_kernel_device_3x3(unsigned char*, unsigned char*, unsigned char*, unsigned char*,
unsigned char*, unsigned char*, int, int, float*)
      API calls:   94.39%  139.14ms         7  19.877ms  9.2850us  139.07ms  cudaStreamCreate
                    1.42%  2.0861ms         2  1.0431ms  1.0390ms  1.0471ms  cuDeviceTotalMem
                    1.07%  1.5806ms        49  32.257us  6.6210us  214.44us  cudaMalloc
                    0.94%  1.3847ms        49  28.259us  7.8990us  148.96us  cudaFree
                    0.66%  970.18us       188  5.1600us     246ns  201.95us  cuDeviceGetAttribute
                    0.65%  965.34us        49  19.700us  10.254us  49.531us  cudaMemcpyAsync
                    0.31%  460.25us         1  460.25us  460.25us  460.25us  cudaGetDeviceProperties
                    0.12%  172.57us         7  24.652us  18.125us  40.448us  cudaLaunch
                    0.12%  171.38us        28  6.1200us  3.6530us  11.004us  cudaEventSynchronize
                    0.10%  141.91us        28  5.0680us  3.9080us  10.922us  cudaEventRecord
                    0.06%  91.496us         2  45.748us  44.773us  46.723us  cuDeviceGetName
                    0.05%  68.107us         7  9.7290us  9.0130us  11.787us  cudaStreamDestroy
                    0.05%  67.520us        14  4.8220us  4.5020us  5.2050us  cudaEventElapsedTime
                    0.03%  49.272us        28  1.7590us  1.2770us  4.3440us  cudaEventCreate
                    0.03%  41.251us         7  5.8930us  5.6690us  6.5060us  cudaDeviceSynchronize
                    0.01%  17.844us        63     283ns     246ns     660ns  cudaSetupArgument
                    0.00%  3.9480us         7     564ns     367ns     908ns  cudaConfigureCall
                    0.00%  2.7410us         3     913ns     356ns  1.9690us  cuDeviceGetCount
                    0.00%  1.6550us         4     413ns     268ns     629ns  cuDeviceGet
```

Выводы

- Развертка циклов и использование shared-памяти помогает значительно(вплоть до 20%) ускорить вычисления на GPU.
- В случае, когда в программе требуется вызывать несколько раз обработку на GPU, можно достичь ускорение(вплоть до 30%) при помощи CUDA-потоков.
- Транспорт по-прежнему является узким местом данной технологии.