

GPU Assignment: Image processing

Stefano Sandonà

Vrije Universiteit Amsterdam, Holland

1 GPUs: NVIDIA GTX480

The aim of this assignment was to learn how to use many-core accelerators, GPUs in this particular case, to parallelize data-intensive code. All the implementations were written for the **NVIDIA GTX480**, using CUDA, a parallel computing platform and programming model invented by NVIDIA. Programming with CUDA, there is a straightforward mapping onto hardware, for this reason it is necessary to study the available HW before start developing an application. The architecture of the given accelerator is shown in Figure 1, its main characteristics and limits are shown in Table 1.

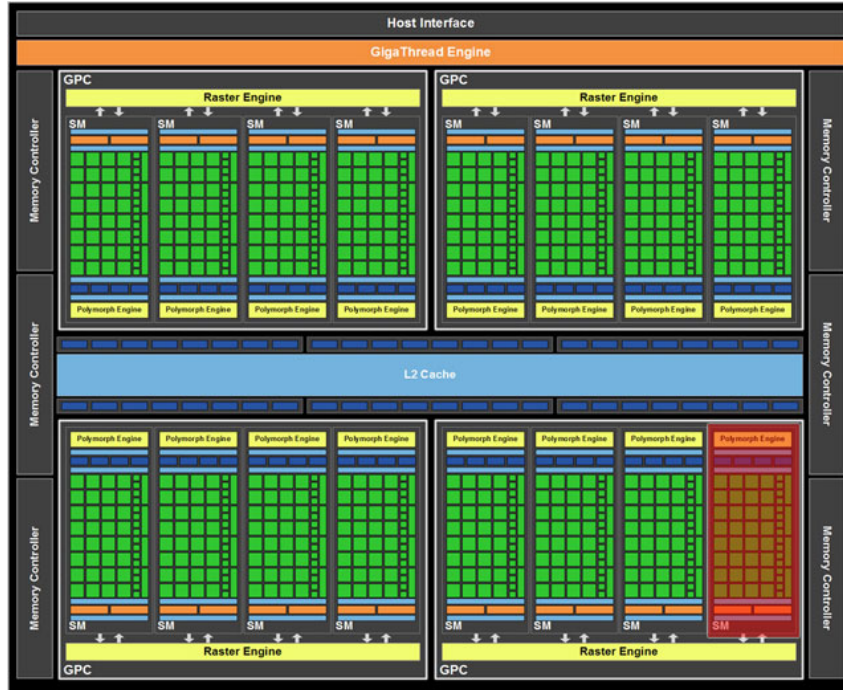


Figure 1: NVIDIA GTX480 Architecture

2 CImg

The image processing library used in this project was CImg, a small, modern and open-source toolkit developed for C++. CImg implements the RGB color model, an additive color model in which red, green, and blue light are added together in various ways to reproduce a broad array of colors. Each colored image of size $N \times M$ is composed by three parts (R,G,B) of the same size, so that $N \times M \times 3$ values are necessary to define an image. The Figure 2 shows an example of image composition.

Microarchitecture	Fermi
Compute capability (version)	2.0
Maximum dimensionality of grid of thread blocks	3
Maximum x-dimension of a grid of thread blocks	65535
Maximum y-, or z-dimension of a grid of thread blocks	65535
Maximum dimensionality of thread block	3
Maximum x- or y-dimension of a block	1024
Maximum number of threads per block	1024
Cores per SM (warp size)	32
SM	15
Cores	480 (32 * 15)
Maximum number of resident blocks per multiprocessor	8
Maximum number of resident warps per multiprocessor	48
Maximum number of resident threads per multiprocessor	1536 (48 * 32)
Number of 32-bit registers per multiprocessor	32K
Maximum amount of shared memory per multiprocessor	48K
Theoretical Throughput	1345 GFLOPS
Theoretical Bandwidth	177.4 GB/s

Table 1: NVIDIA GTX480 Specifications

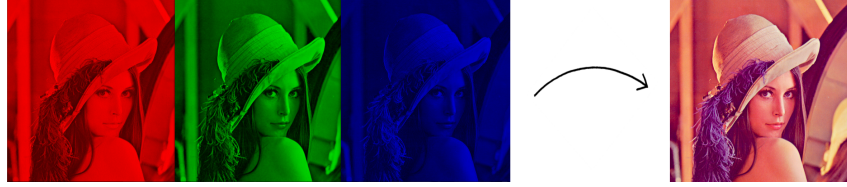


Figure 2: RGB model

3 The processing flow

Using CUDA there are two parts of the code: the device code, or GPU code, or the Kernel, that is a sequential program, written for one thread and executed for all and the HOST code, or CPU code, that is used to instantiate the grid, run the kernel, manage the memory. Figure 3 shows the processing flow of a CUDA application. In the particular case of image processing, everything starts from the CPU, that stores the image from a file into a local buffer, allocates IN and OUT buffers on the GPU (*cudaMalloc*) and copies the image into the GPU's IN buffer (*cudaMemcpy*). After that, the CPU launches the GPU kernel with a defined grid configuration (*kernel_function* «*gridDim*, *blockDim*»(*params*)), that is executed by the GPU following the SIMT (Single Instruction, Multiple Threads) NVIDIA model. The threads are executed in parallel in each core, and they read the assigned part of IN data and generates the assigned part of OUT data. At the end, the results are copied out back to the CPU (*cudaMemcpy*) and the image is written to a file by the CPU.

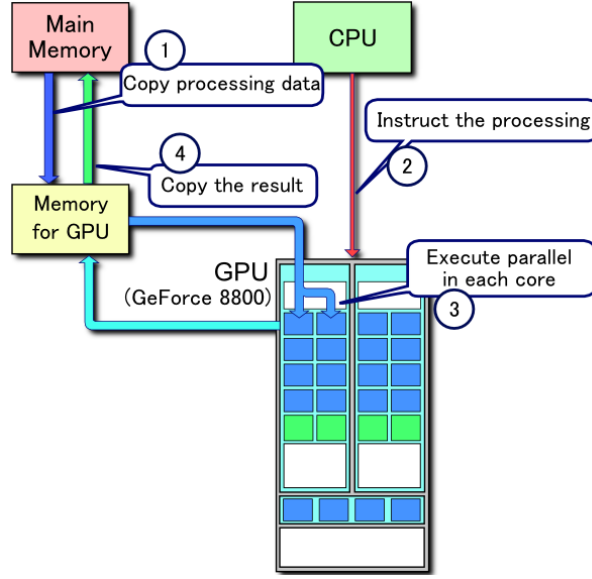


Figure 3: CUDA processing flow

4 CUDA grid configuration

In CUDA, as mentioned before, there is a strict mapping with the hardware, so that a hardware virtualization model is fixed with the concepts of thread, block and grid. Each **thread** executes the kernel code, running on one CUDA core. The threads are logically grouped into **thread blocks**, so that the threads of the same block will run on the same multiprocessor. The thread blocks are logically organized in a **Grid**, that represent the entire dataset. The blocks and the grid can be of 1D, 2D or 3D. The most important thing programming with GPUs, to make use of all their power, is to make them as busy as possible. Switching between concurrent warps has no overhead because registers and share memory are partitioned and not stored/restored, so that if one warp has to wait for example for a memory access and another warp is ready, there is a switch to hide the latency. The thread scheduling is really quick, and this allow the blocks to be swapped in and out really quickly. For this reason an high number of warps is needed and instantiating a grid, is good to have a number of blocks much bigger than the number of available multiprocessor and a block size that can be higher than the amount of cuda cores available per SM. Another interesting aspect to take into consideration, is the block size. The block is divided into warps, so that Threads 0..31 are part of Warp 0, Threads 32..63 are part of Warp 1 and so on. Considering this, it is good to have blocks with a size multiple of 32, so that no useless threads are launched. For what concerning the grid configuration, an image is a 2D structure and for this reason it is intuitive to set up also a 2D grid. By the way, setting up a 2D grid, there is not only one way to follow. For this particular project three possible block configurations were considered, 1D (1x256), 2D (8x32), 2D (16x32) and two possible grid configurations: dynamic and fixed. In the rest part of this document, for dynamic grid configurations, **M1** indicates a dynamic grid of 1D blocks, in which the grid has a height of $image_height$, a width of $(\lceil image_width / 256 \rceil)$ and the threads on the right border if the image width is not a multiple of $block_width$ are idle (Figure 4a); **M2** indicates a dynamic grid of 1D blocks, in which the size of the grid is the same as M1, but the threads are consecutive and only the threads on the last block(s) are idle if the image's number of pixel is not a multiple of the size of the block (Figure 4b); **M3** indicates a dynamic grid of 2D blocks, in which the grid has a width of $\lceil image_width / 32 \rceil$, a height of $\lceil image_height / 8 \rceil$ and the overflow is the same as M1 (Figure 5a); **M4** indicates a dynamic grid of 2D blocks, in which the size of the grid is the same as M3, but the overflow is the same as M2 (Figure 5b). The advantages using a fixed grid configuration are that it is possible to support any problem size even if it exceeds the largest grid size the CUDA device supports and that it is possible to limit the number of blocks, for example launching a number of blocks that is a multiple of the number of multiprocessors on the device, to balance utilization.

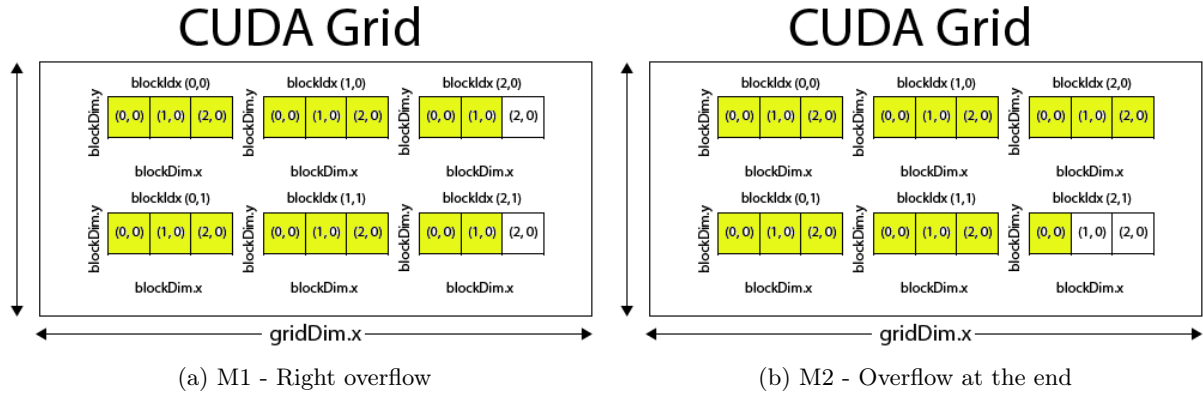


Figure 4: 1D blocks, Kernel configurations, image of 16 pixels

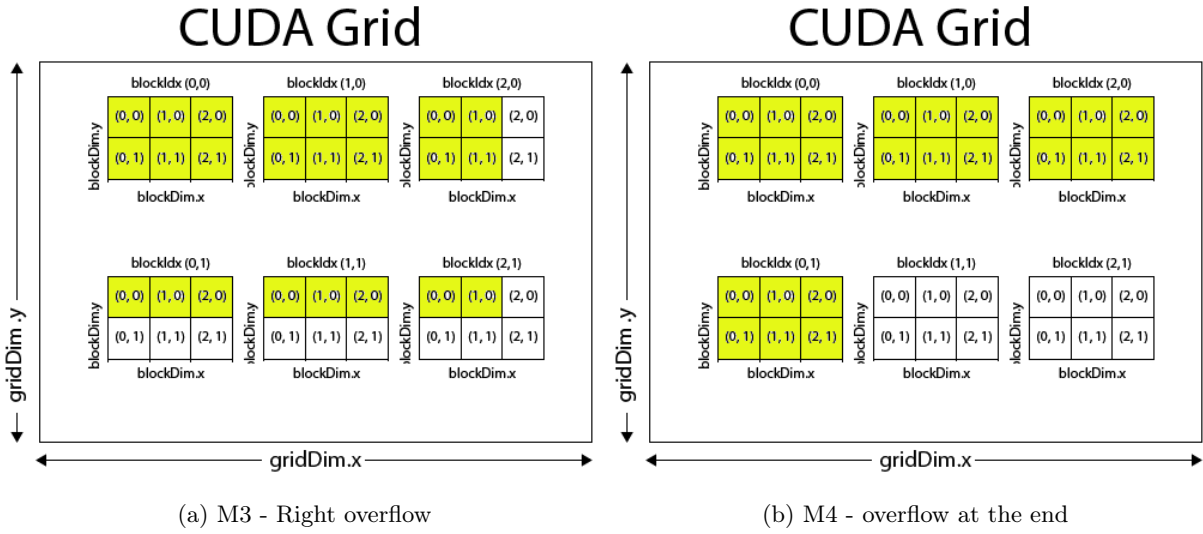


Figure 5: 2D blocks possible kernel configurations, image of 24 pixels

5 Coalesced memory access

One of the main bottlenecks with the GPUs are the global memory accesses that are expensive, for this reason it is better to maximize the use of bytes that travel from the DRAM to the Streaming Multiprocessor. CUDA uses a SIMT approach, in which all threads of a warp execute the same instruction, so that global memory accesses are effectuated "per warp". The threads in a warp (32) provide 32 addresses and the hardware converts these addresses into memory transactions. The memory is divided into regions of 128 bytes, so that bytes 0...127 are part of Region 0, bytes 128...255 are part of Region 1 and so on. The memory is accessed per region, that means if one thread wants the byte 0, the entire Region 0 is loaded. A kernel is correctly designed, if consecutive threads access consecutive memory addresses, so that all the requested addresses fall on the same region and only one transaction is performed. Behaving in this "coalesced way" instead of having one access per thread, these accesses are grouped and the total memory overhead is reduced. Developing all the three algorithms this aspect was taken into consideration.

6 Algorithm 1: Grayscale Conversion and Darkening

From an RGB image, the output of this algorithm is a darker grayscale image. The gray value of a pixel is generated by weighting the three values ($0.3 \cdot R$, $0.59 \cdot G$, $0.11 \cdot B$) and then summing them together. To darken the obtained grayscale image, the final pixel value is multiplied by a constant (0.6). The Figure 6 shows an example of the result. The sequential algorithm, simply go through the entire image and computes for each pixel the corresponding value (Listing 1).

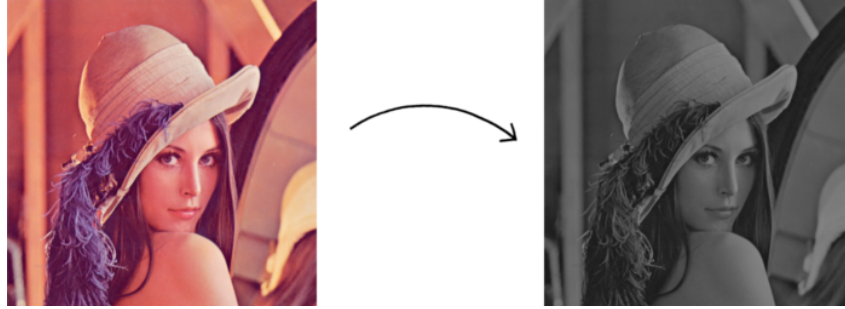


Figure 6: Grayscale Conversion and Darkening

```
//H=image_height, W=image_width
for ( int y = 0; y < H; y++ ) {
  for ( int x = 0; x < W; x++ ) {
    float grayPix = 0.0f;
    float r = static_cast< float >(inputImage[(y * W) + x]);
    float g = static_cast< float >(inputImage[(W * H) + (y * W) + x]);
    float b = static_cast< float >(inputImage[(2 * W * H) + (y * W) + x]);
    grayPix = ((0.3f * r) + (0.59f * g) + (0.11f * b));
    grayPix = (grayPix * 0.6f) + 0.5f;
    darkGrayImage[(y * W) + x] = static_cast< unsigned char >(grayPix);
  }
}
```

Listing 1: Darker Sequential code

6.1 Parallelization

6.2 One pixel per thread

After allocating into the GPU global memory some space to contain the input image ($3 * image_width * image_height * sizeof(unsigned char)$), copying the input image there and allocating some memory to content the output image ($image_width * image_height * sizeof(unsigned char)$), the kernel is ready to be launched. The GPU code is the same as the code content inside the loop of the sequential version (Listing 1), but instead of using the indexes of the loop to access the image pixels, it uses the index associated with the thread. In this first method, only one pixel is computed per thread. To exploit the coalesced memory access, two consecutive threads computes/accesses the values of two consecutive pixels. The 4 different grid configurations (explained in Section 4) were tested for this program, in order to establish the best. In Figure 7 are reported the obtained results. As shown methods M1 and M2 obtained the best results, so with unidimensional blocks the algorithm achieved better speedups. This is probably due to the fewest amount of operations to calculate the index of the pixel associated with each thread, because in this particular task there is no clear advantage on using 1D blocks instead of 2D blocks.

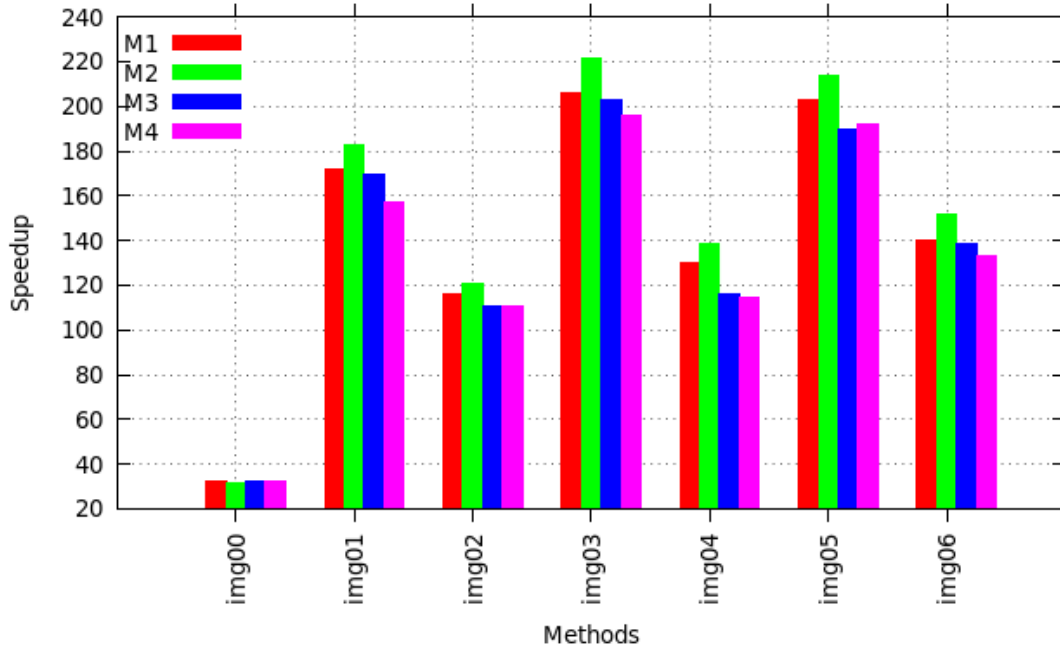


Figure 7: Speedups Comparison

6.3 Optimization - More pixels per thread

To exploit all the power of GPUs, as said, they have to be as busy as possible, because the warp scheduling is very fast and can hide the latency. Anyway, reducing the number of thread blocks by increasing the work per thread revealed to achieve better results. Lots of experiments were performed to establish the right number of pixels to compute by each thread, all taking into account the coalesced memory access. To maintain the coalescing, one thread computes the associated first pixel index with the formula shown in Listing 2 and then for the following pixels, it adds each time to this the total amount of pixels contained in the grid($gridDim.x * blockDim.x * gridDim.y * blockDim.y$). The final code, is the one shown in Listing 3.

```
unsigned int pixel=((blockIdx.y * gridDim.x + blockIdx.x)
                  * blockDim.x) + threadIdx.x
```

Listing 2: Corresponding first pixel

```
for(i = ((blockIdx.y * gridDim.x + blockIdx.x) * blockDim.x) + threadIdx.x;
    i < width * height;
    i += (gridDim.x * blockDim.x) * (gridDim.y * blockDim.y)) {
    float grayPix = 0.0f;
    float r = static_cast<float>(inputImage[i]);
    float g = static_cast<float>(inputImage[(width * height) + i]);
    float b = static_cast<float>(inputImage[(2 * width * height) + i]);
    grayPix = ((0.3f * r) + (0.59f * g) + (0.11f * b));
    grayPix = (grayPix * 0.6f) + 0.5f;
    outputDarkGrayImage[i] = static_cast<unsigned char>(grayPix);
}
```

Listing 3: Darker Parallel Code

Figures 8 and 9 represent the speedups achieved for different thread loads, considering both a dynamic and a fixed grid. Figure 10 compares the best results for the two configurations. As it is possible to see, the first approach achieved better results. With a number of 10 pixels per thread, the program obtained the best results, so this is a good thread load to help the scheduler. This load was sufficient to both hide the latency and reduce the warp scheduling overhead. Figure 11 shows the comparison between the best results obtained with a single pixel per thread and the results obtained with the new version. It is clear

that the new version is an optimization, in fact for each image the speedup increased. This solution is the one adopted. In Table 2 and 3 are shown the detailed execution times and speedups achieved. Table 4 shows the total time considering also the memory overhead. Table 5 exposes the Achieved Throughput and Bandwidth. For the Throughput, for each pixel there are 3 additions and 4 multiplications. For the Bandwidth, for each pixel there are 3 reads and 1 write of 1 byte. Table 6 exposes the differences obtained comparing the output of the sequential algorithm with the output of the parallel algorithm. These differences could derive from a difference in floating point precision between the CPU and GPU, or from different ordering of floating point operations. In Section 9 is reported a detailed analysis using the NVIDIA Virtual Profiler.

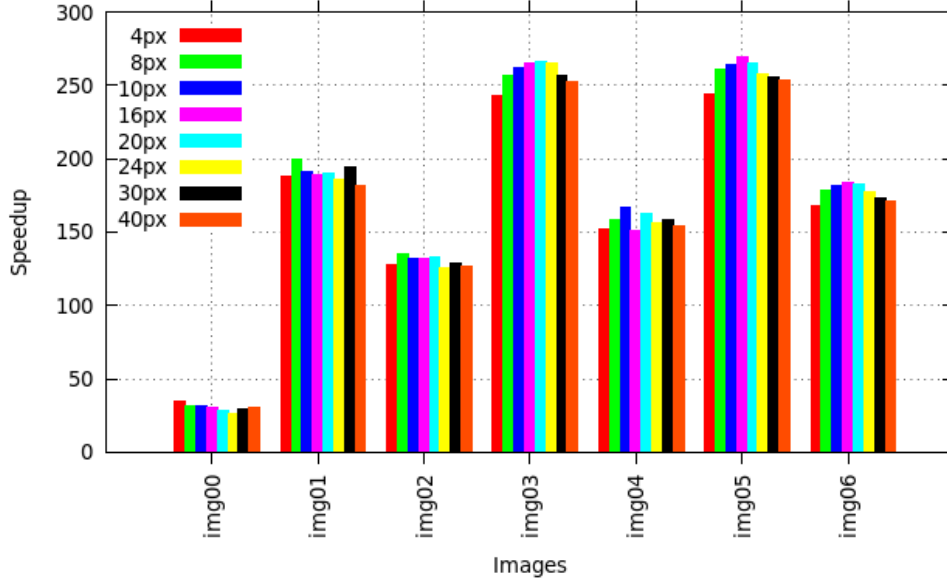


Figure 8: Speedup with different number of pixels per thread

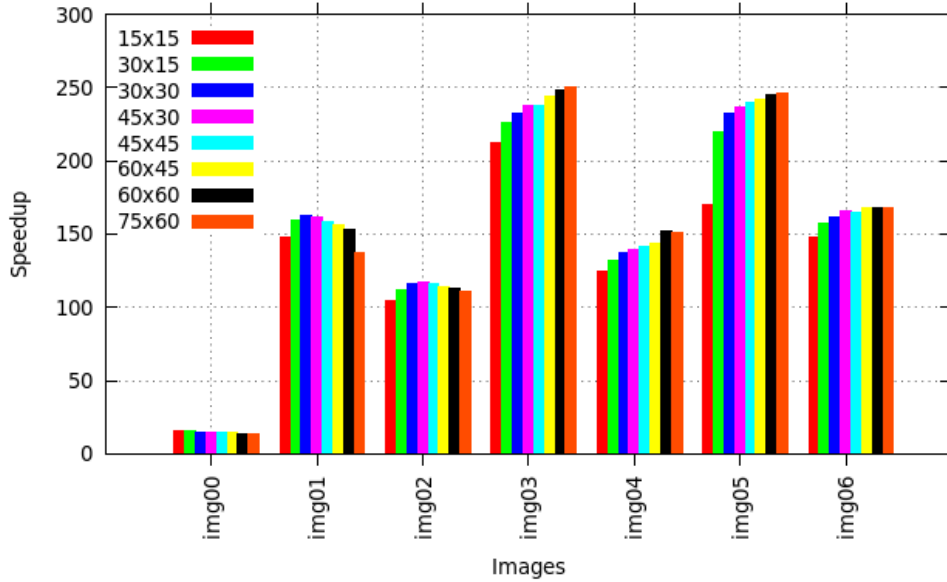


Figure 9: Speedup with fixed grid

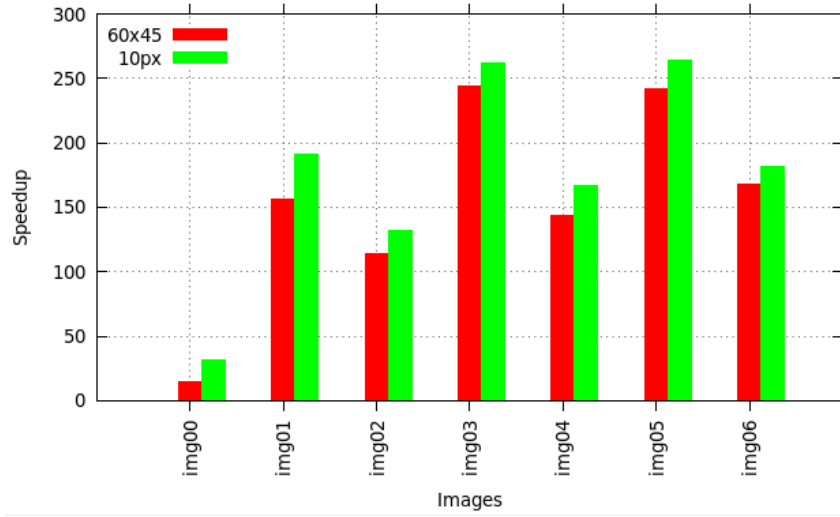


Figure 10: Solutions Comparison

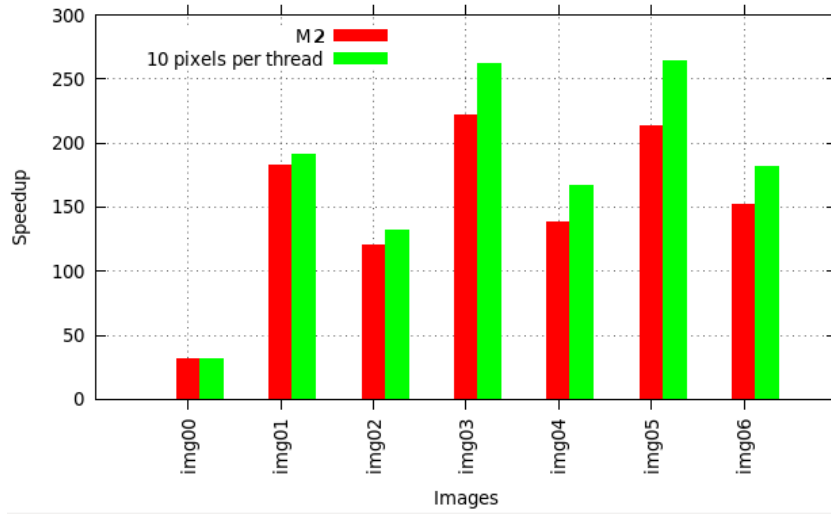


Figure 11: Solutions comparison

	Sequential	M2	10 pixels per thread
img00	0.002416	0.000076	0,000076
img01	0.036925	0.000202	0,000193
img02	0.033726	0.000279	0,000255
img03	0.340821	0.001537	0,001299
img04	0.151429	0.001091	0,000907
img05	0.445976	0.002090	0,001692
img06	0.588814	0.003880	0,003239

Table 2: Execution Times

	M2	10 pixels per thread
img00	31.79	31.79
img01	182.80	191.32
img02	120.88	132.26
img03	221.74	262.37
img04	138.80	166.96
img05	213.39	263.58
img06	151.76	181.79

Table 3: Speedups

Images	Execution Time (s)	Memory Time(s)	Total(s)
Img00	0,000076	0,000547	0,077502
Img01	0,000193	0,004562	0,081866
Img02	0,000255	0,006141	0,072205
Img03	0,001299	0,031491	0,108873
Img04	0,000907	0,025354	0,103170
Img05	0,001692	0,049163	0,125342
Img06	0,003239	0,079937	0,158472

Table 4: Measured times

	Ach. Throughput (GFLOPS/s)	Ach. Bandwidth (GB/s)	Compute utilization (%)	Bandwidth utilization (%)
Img00	24,14	13,80	1,80	7,78
Img01	95,08	54,33	7,07	30,63
Img02	101,20	57,83	7,52	32,60
Img03	136,02	77,72	10,11	43,81
Img04	129,48	73,99	9,63	41,71
Img05	137,11	78,35	10,19	44,16
Img06	141,63	80,93	10,53	45,62

Table 5: Achieved Throughput and Bandwidth

	img00	img01	img02	img03	img04	img05	img06
Pixels above threshold	111	231	135	735	1503	9	4206

Table 6: Differences between the Sequential and the Parallel output

7 Algorithm 2: Histogram Computation

From an RGB image, the output of this algorithm is a grayscale image with the relative histogram of 256 possible values of gray. The histogram measures how often a value of gray is used in an image. The sequential algorithm, simply go through the entire image, computing for each pixel the corresponding gray value and incrementing the corresponding counter (Listing 4). The Figure 12 shows an example of the result.

```
//H=image_height , W=image_width
for ( int y = 0; y < H; y++ ) {
  for ( int x = 0; x < W; x++ ) {
    float grayPix = 0.0f;
    float r = static_cast< float >(inputImage[(y * W) + x]);
    float g = static_cast< float >(inputImage[(W * H) + (y * W) + x]);
    float b = static_cast< float >(inputImage[(2 * W * H) + (y * W) + x]);
    grayPix = ((0.3f * r) + (0.59f * g) + (0.11f * b)) + 0.5f;
    grayImage[(y * W) + x] = static_cast< unsigned char >(grayPix);
    histogram[static_cast< unsigned int >(grayPix)] += 1;
  }
}
```

Listing 4: Sequential code

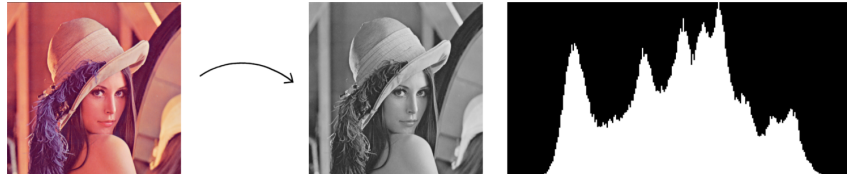


Figure 12: Histogram Computation

7.1 Parallelization

7.2 One pixel per thread

After copying the input image ($3 * image_width * image_height * sizeof(unsigned\ char)$) and the initial empty histogram ($256 * sizeof(unsigned\ int)$) into the GPU global memory and allocating some memory to content the output image ($image_width * image_height * sizeof(unsigned\ char)$), the kernel is ready to be launched. As for the previous algorithm, the GPU code is almost the same as the code content inside the loop of the sequential version (Listing 4), but instead of using the indexes of the loop to access the image pixels, it uses the index associated with the thread. Also in this case, to exploit the coalesced memory access, two consecutive threads computes/accesses the values of two consecutive pixels. The interesting aspect of computing histograms in parallel, is that it is possible that two threads read at the same time the same value of gray for a pixel, so that at the same time they try to increment the corresponding histogram bin. For a correct execution, to avoid wrong updates, the increment has to be an atomic operation, so that only one thread at the time will modify the bin value. The logic behind an atomic operation is that each thread "locks" the variable, modify it, and "unlocks" it, so that the other threads that try to modify the same variable has to wait that this will be "unlocked". The worst case, in the histogram scenario, is a monochrome image (only one color), in which all the threads try to modify the same bin, which implies a long waiting queue of threads. The general idea to have high performance, is to reduce the number of conflicts at the minimum. Also in this algorithm, the speedup is not related to the shape of the block, in fact the performance of the algorithm using the atomics depends on the image. Dealing with an image with horizontal stripes of different colors of 1 pixel, for example, with a 1D block (1x256) there will be all the 256 threads that try to increment the same histogram bin. In the same scenario, using a 2D block (8x32) there will be only 32 threads that try to increment the same histogram value. If the image instead, is composed of vertical stripes of 32 pixels wide, it happens the opposite, the 1D approach will obtain more performance. For these reasons, for the initial tests, to remove the fortuity, the given images were substituted with images of the same sizes but composed by

only one color (monochrome). Different implementations were tested. The first approach consisted of using only the histogram array stored in the global memory and modifying it with atomic add operations (Figure 13). The second approach consisted of using an additional histogram array stored in shared memory, performing the atomic add operations there and at the end, assigning a certain bin to a thread and reducing the shared histograms in the global one (Figure 14). Doing this, not all the launched threads will compete for the same bin, but only the threads in the same block during the computation and the threads to which is assigned the same bin during the reduction phase. Figure 15 shows the obtained results for the two approaches. From this image it is clear that the second approach drastically increased the speedups and that the block configuration (1D or 2D) doesn't affect the performance.

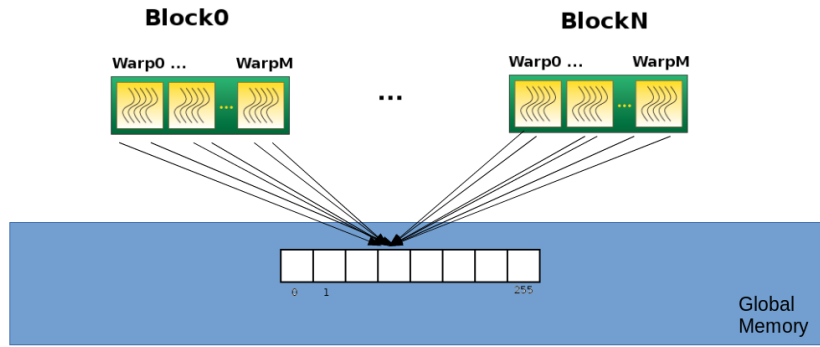


Figure 13: Global histogram - Concurrency

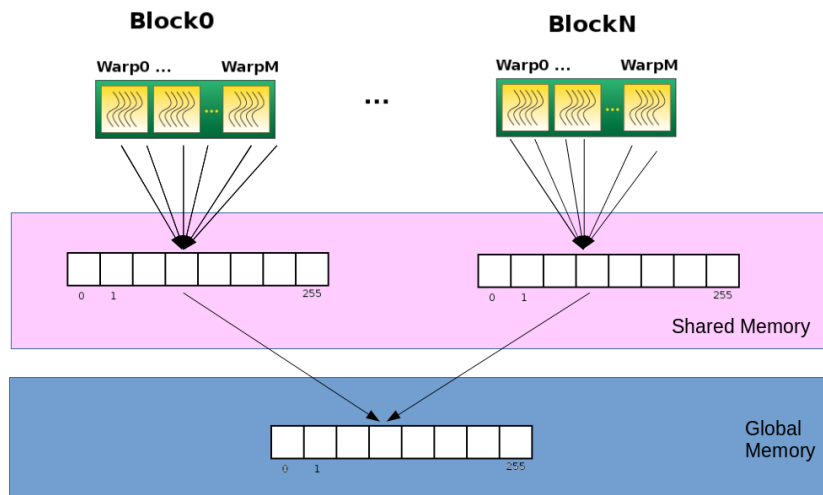


Figure 14: Shared histogram - Concurrency

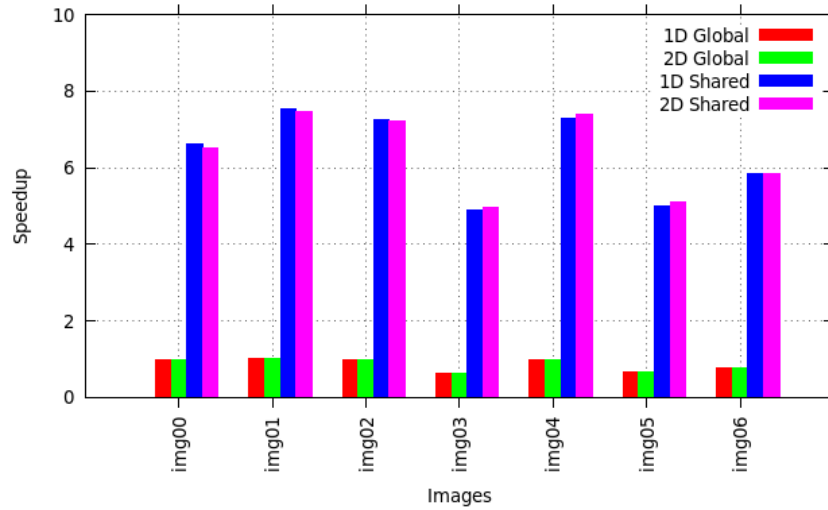


Figure 15: Global Atomics vs Shared Atomics (Monochrome images)

An additional "privatization" of the histogram could be performed. Instead of having one histogram per block, it is possible to assign one histogram per warp (Figure 16), so that the concurrency at the end is only among the 32 warp threads during the computation and between the threads to which is assigned the same bin during the reducing phase. Figure 17 shows the performance of this last method compared to the previous. The number of concurrent histogram modifications was reduced, so that better speedups were achieved.

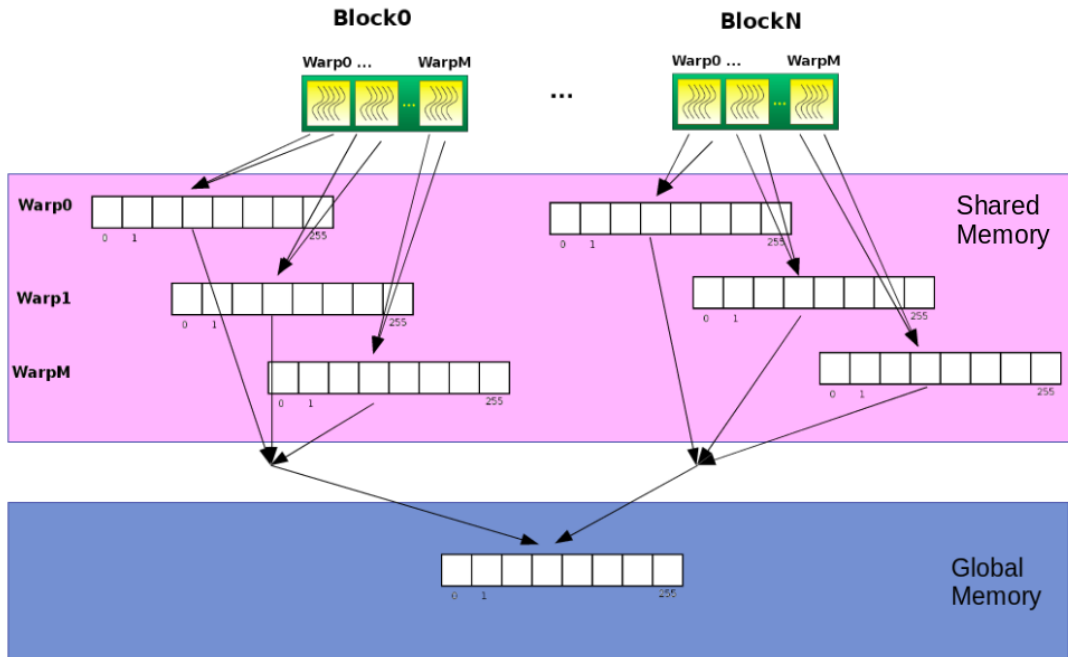


Figure 16: Shared Warps histogram - Concurrency

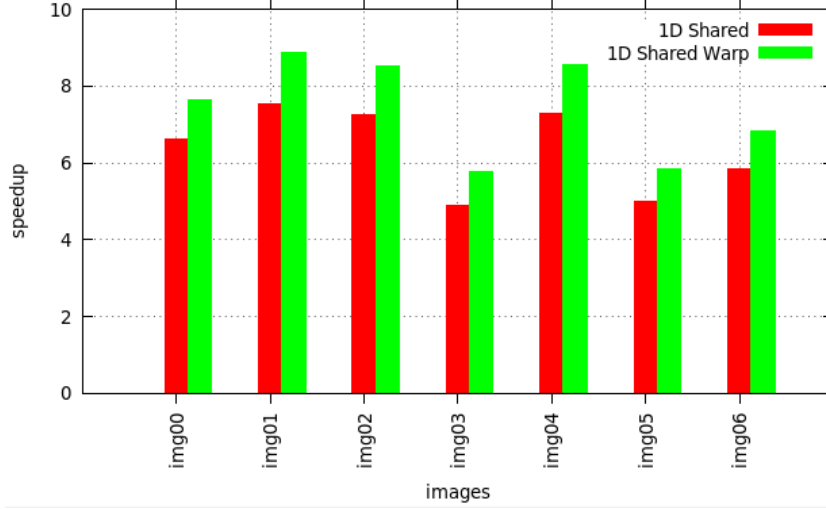


Figure 17: Shared Atomics with Warps vs Shared Atomics (Monochrome images)

After a testing phase with monochrome images, the tests were performed with the real given images. Figure 18 represents the obtained results. As it is possible to see, in this case the warp solution is much worse than the shared version, that means for standard images, the time spent to reduce the single warp histograms is bigger than the time spent to compute the atomic operations.

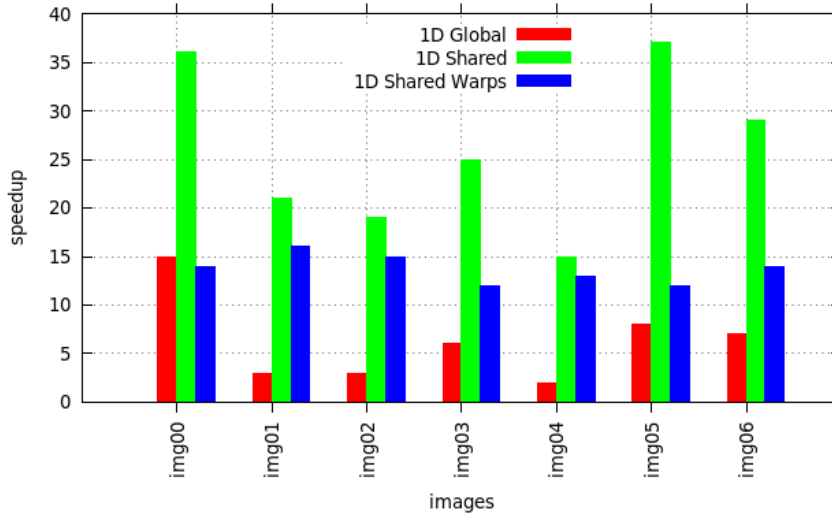


Figure 18: Global Atomics vs Shared Atomics vs Shared Atomics with Warps

7.3 Optimization - More pixels per thread

As for the previous algorithm, some additional tests were performed assigning the computation of more pixels per thread. The last two approaches (Shared Atomics and Shared Atomics with Warps) were improved to test this solution and both on the real images obtained better results. This is due to the reduction of the scheduling load and to the decrease of histograms reduction operations. Different work loads were assigned to each thread and measured. A dynamic grid and fixed grid configurations were tested and Figures 19, 20, 21 and 22 show the obtained results. Figure 23 compares the best obtained configurations for the two approaches. As shown, the second approach (Shared Atomics with Warps), differently than the one pixel per thread version, obtains for almost all the images the best results. This result could surprise, but means that the decrement on the total number of reduction operations is critical for this approach. For the small images (e.g. Img00) the time spent for the final reduction remains high in comparison with the total computation due to the low amount of pixels to compute. The final adopted solution is the Shared Atomics with Warps with with a fixed grid, because obtained better results for

almost all the images and it is a more scalable solution. Tables 7 and 8 show the detailed obtained results. Table 9 shows the total time of the adopted solution, considering also the memory overhead. Table 10 exposes the Achieved Throughput and Bandwidth. For the Throughput, for each pixel there are 3 additions, 3 multiplications and 1 shared memory atomic add operation and for each block there are 256×8 additions (warps histogram reduction) and 256 global memory atomic add operations. For the Bandwidth, for each pixel there are 3 reads and 1 write (of 1 byte) and for each block there are 256 writes (of 4 bytes) for the atomic add operations. Tables 11 and 12 show the differences obtained comparing the outputs of the sequential algorithm with the outputs of the parallel algorithm.

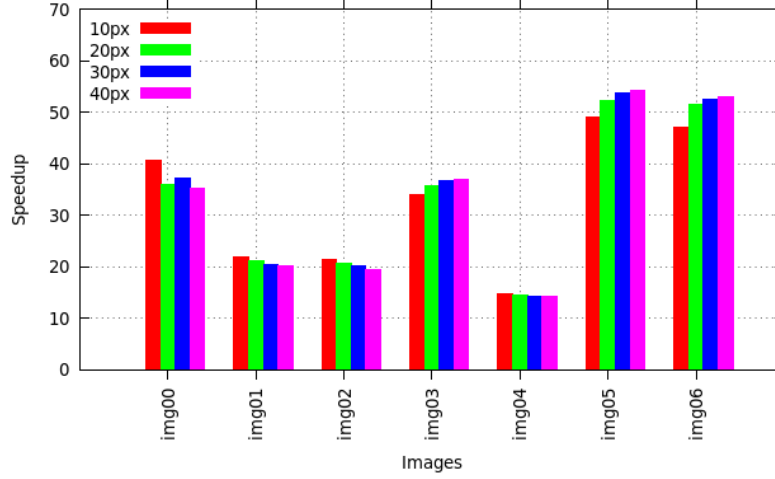


Figure 19: Shared Atomics, 1D blocks - Different number of assigned pixels per thread

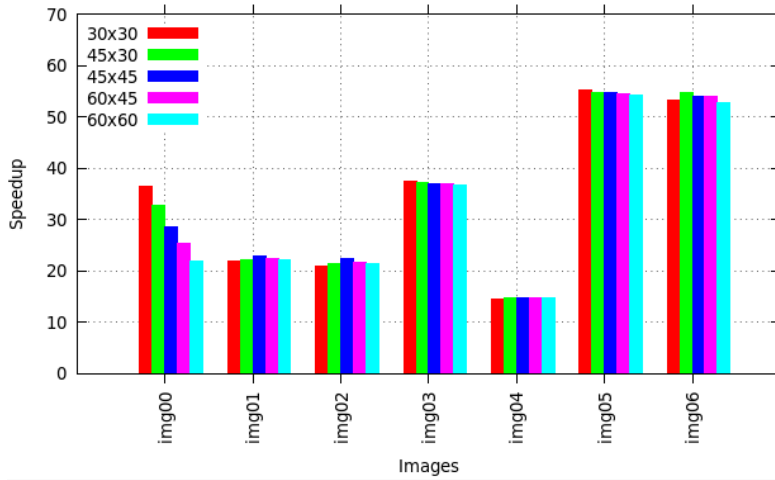


Figure 20: Shared Atomics, 1D blocks - Fixed grid

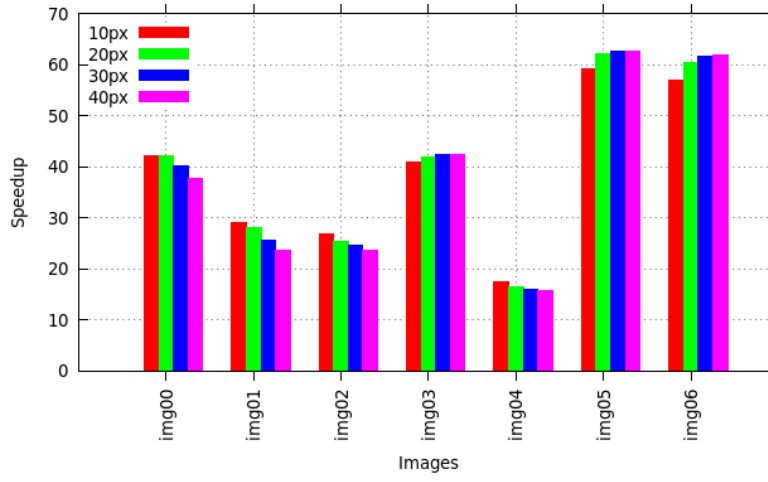


Figure 21: Shared Atomics with Warps, 1D blocks - Different number of assigned pixels per thread

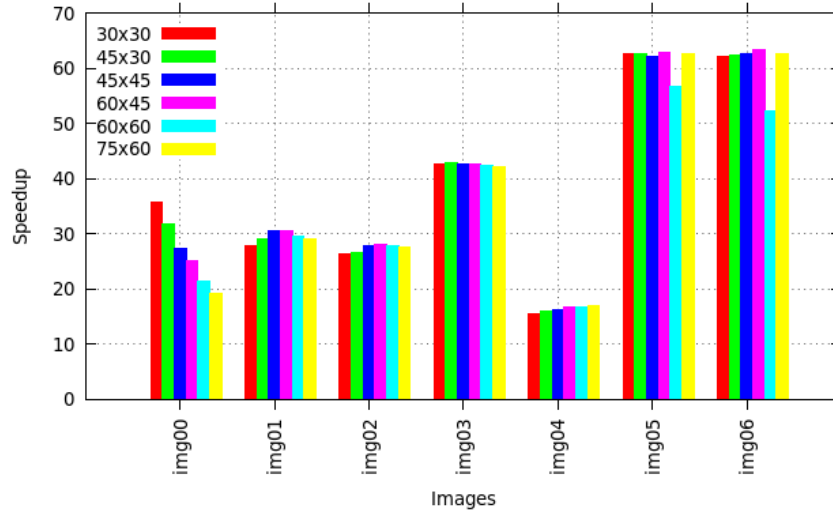


Figure 22: Shared Atomics with Warps, 1D blocks - Fixed grid

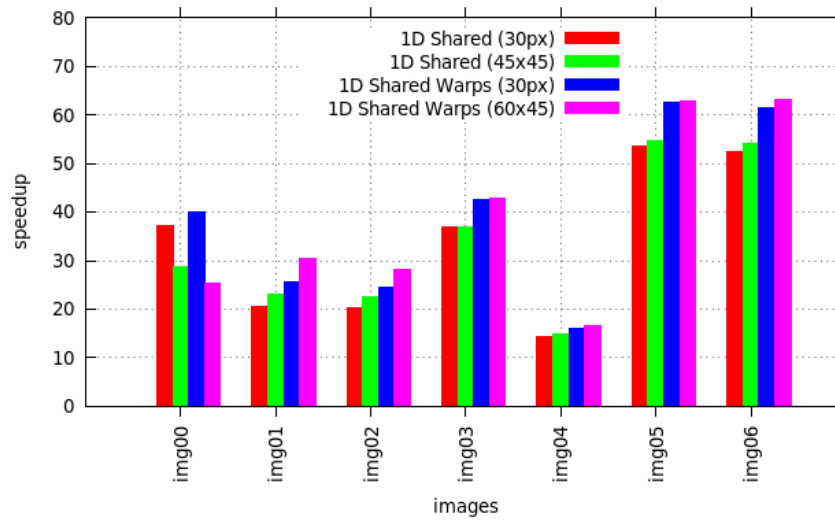


Figure 23: Best solutions comparison

	Sequential	30px	60x45
Img00	0,003208	0,000080	0,000127
Img01	0,032379	0,001264	0,001062
Img02	0,043784	0,001778	0,001553
Img03	0,200681	0,004724	0,004698
Img04	0,197973	0,012325	0,011871
Img05	0,267590	0,004278	0,004260
Img06	0,616924	0,010026	0,009756

Table 7: Execution Times

	30px	60x45
Img00	40.10	25.26
Img01	25.62	30.49
Img02	24.63	28.19
Img03	42.48	42.72
Img04	16.06	16.68
Img05	62.55	62.81
Img06	61.53	63.24

Table 8: Speedups

Images	Execution Time (s)	Memory Time(s)	Total(s)
Img00	0,000127	0.000645	0.070765
Img01	0,001062	0.003966	0.070738
Img02	0,001553	0.006266	0.073200
Img03	0,004698	0.037846	0.118953
Img04	0,011871	0.019135	0.096683
Img05	0,004260	0.036651	0.105514
Img06	0,009756	0.097466	0.182628

Table 9: Measured times

Images	Ach. Throughput (GFLOPS/s)	Ach. Bandwidth (GB/s)	Compute utilization (%)	Bandwidth utilization (%)
Img00	63,43	30,03	4,72	16,93
Img01	23,14	12,48	1,72	7,03
Img02	20,62	11,28	1,53	6,36
Img03	38,93	22,08	2,89	12,45
Img04	10,42	5,89	0,77	3,32
Img05	55,92	31,77	4,16	17,91
Img06	47,66	27,15	3,54	15,31

Table 10: Achieved Throughput and Bandwidth

	img00	img01	img02	img03	img04	img05	img06
Pixels above threshold	324	12	0	24	0	12	48

Table 11: Differences between the Sequential and the Parallel output (histogram)

Images	Img00	Img01	Img02	Img03	Img04	Img05	Img06
Pixels above threshold	777	5322	849	3450	8004	1431	28956

Table 12: Differences between the Sequential and the Parallel output (gray image)

8 Algorithm 3: Smoothing

Smoothing is the process of removing noise from an image by the means of statistical analysis. To remove the noise, each point is replaced by a weighted average of its neighbours. In this way small-scale structures are removed from the image. In this case a two-dimensional 5-points triangular smooth filter was used. The Figure 24 shows an example of the result.

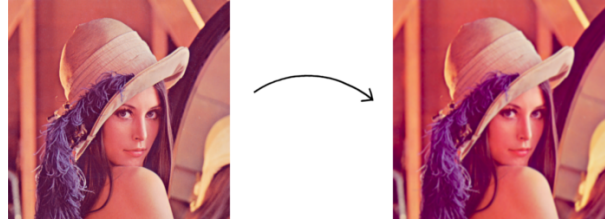


Figure 24: Smoothing

8.1 Parallelization

8.2 One pixel per thread

After copying the input image into the GPU global memory ($3 * image_width * image_height * sizeof(unsigned char)$) and allocating some memory to content the output image ($3 * image_width * image_height * sizeof(unsigned char)$), the kernel is ready to be launched. As for the previous algorithms, the GPU code is almost the same as the code content inside the loop of the sequential version, but instead of using the indexes of the loop to access the image pixels, it uses the index associated with the thread. Also in this case, to exploit the coalesced memory access, two consecutive threads computes/accesses the values of two consecutive pixels. This particular algorithm (filter) deals with square areas of the image (5*5 pixels), so that using 2D blocks, the threads can efficiently share memory and prevent a lot of global memory accesses. Figure 25 shows two threads (Blue and Red) that apply the filter to the assigned pixel (A and B). To apply the filter to the assigned pixel, a thread has to read the 25 neighbor pixels. With a proper access, while one thread reads a neighbor (for example the top left as shown), the following thread in the same warp, in the same moment reads the following pixel, so that the global memory is accessed in a coalasced way. Another advantage, using 2D blocks, is that the algorithm is more readable, in fact it is very similar to the sequential one. The first consideration is that a filter is a structure of data that doesn't change during the computation, that is read lots of time and that is small. For this reason it is a perfect candidate to be stored in the constant memory. Another interesting aspect, is that the same pixel is read multiple times by different threads reading the matrix of neighbor to compute the filter. For this reason two different approaches were tested at the begin. The first approach is more simple, very similar to the sequential version and uses only global memory accesses; the second approach, first copies in the shared memory the image portion owned by a block and than compute the filter accessing the shared memory. Figure 26 shows the obtained results. As it is possible to understand, using bigger blocks (16x32 instead of 8x32) the performance increased. The explanation is that this algorithm is based on data locality so that using bigger blocks the memory is better shared. Another interesting aspect, is that the approach that doesn't use the Shared Memory obtained better results. This could surprise but means the global memory accesses are so well done that the time spent to load the image in shared memory and to access the shared memory is bigger. For this reason the first approach is adopted.

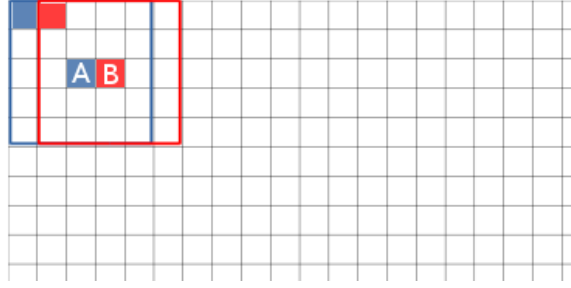


Figure 25: Pixels access

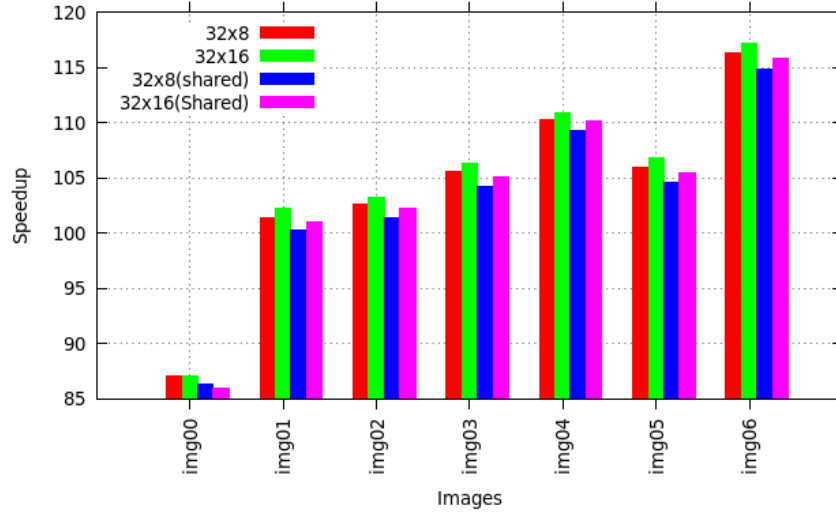


Figure 26: Image Portion in Shared Memory VS Global Memory Accesses (different block sizes)

8.3 Optimization - More pixels per thread

As for the previous algorithms, some additional tests were performed assigning the computation of more pixels per thread. Different thread loads were tested, considering both a dynamic and a fixed grid. Figures 27 and 28 show the obtained results. With a dynamic approach, varying the number of pixels computed per thread doesn't change a lot the final speedups, except for small images. With a fixed height grid the results are almost the same for every configuration. Figure 29 shows the comparison of the best dynamic and fixed grid configurations. The results are not so different but the fixed (height) grid approach achieved a little more speedups. For this reason and because it is a more scalable solution, it is the one chosen. Tables 13 and 14 show the detailed obtained results. Table 15 shows the total time of the adopted solution, considering also the memory overhead. Table 16 exposes the Achieved Throughput and Bandwidth. For the Throughput, for each pixel there are 225 additions, 75 multiplications and 3 divisions. For the Bandwidth, for each pixel there are 75 reads and 3 writes (of 1 byte). Table 17 shows the differences obtained comparing the output of the sequential algorithm with output of the parallel algorithm.

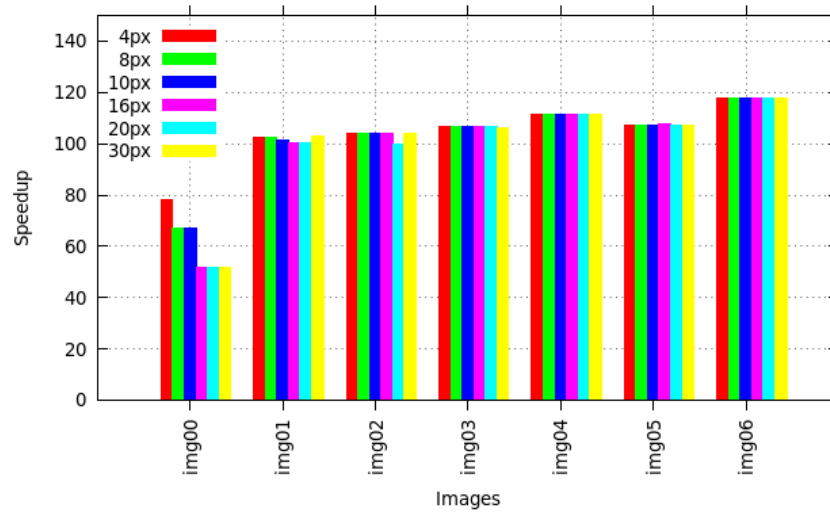


Figure 27: More pixels per thread

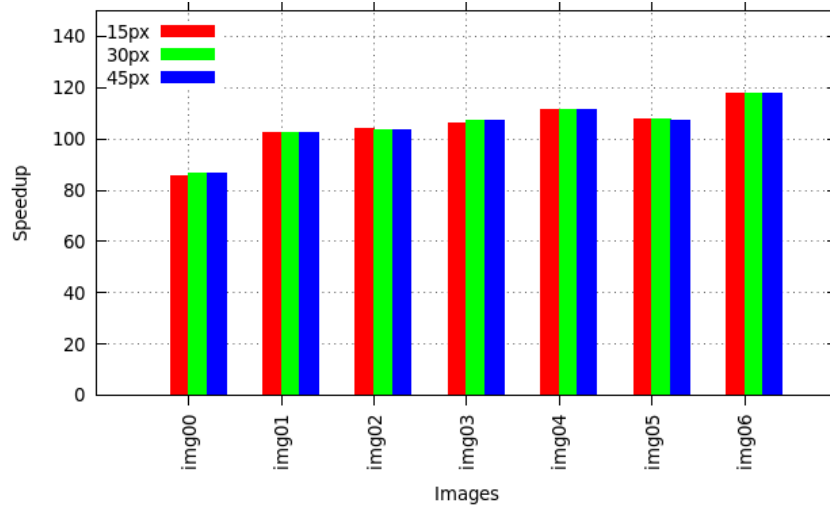


Figure 28: Fixed Height and Dynamic width

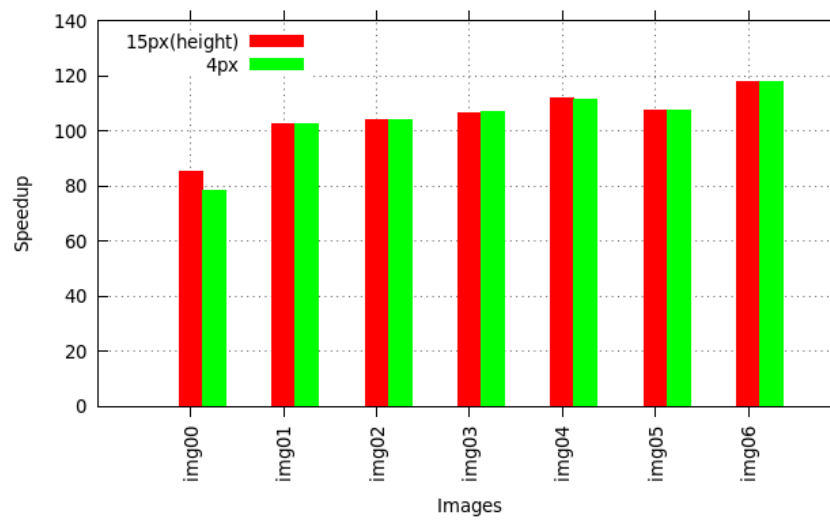


Figure 29: Solutions comparison

Images	Sequential	15px (height)	4px
Img00	0,067991	0,000797	0,000868
Img01	0,672104	0,006553	0,006552
Img02	0,947251	0,009093	0,009114
Img03	6,496388	0,061039	0,060738
Img04	4,695295	0,042040	0,042152
Img05	8,563947	0,079630	0,079795
Img06	18,571981	0,157342	0,157640

Table 13: Execution Times

Images	15px (height)	4px
Img00	85,31	78,33
Img01	102,56	102,58
Img02	104,17	103,93
Img03	106,43	106,96
Img04	111,69	111,39
Img05	107,55	107,32
Img06	118,04	117,81

Table 14: Speedups

Images	Execution Time (s)	Memory Time(s)	Total(s)
Img00	0,000797	0.000977	0.071893
Img01	0,006553	0.006250	0.090151
Img02	0,009093	0.009142	0.094102
Img03	0,061039	0.057633	0.194930
Img04	0,042040	0.038645	0.156782
Img05	0,079630	0.061299	0.209716
Img06	0,157342	0.147982	0.381236

Table 15: Measured times

Images	Ach. Throughput (GFLOPS/s)	Ach. Bandwidth (GB/s)	Compute utilization (%)	Bandwidth utilization (%)
Img00	99,66	25,66	7,41	14,46
Img01	121,21	31,20	9,01	17,59
Img02	122,84	31,62	9,13	17,83
Img03	125,30	32,25	9,32	18,18
Img04	120,92	31,13	8,99	17,55
Img05	126,11	32,46	9,38	18,30
Img06	126,21	32,49	9,38	18,31

Table 16: Achieved Throughput and Bandwidth

	img00	img01	img02	img03	img04	img05	img06
Pixels above threshold	0	0	0	0	0	0	0

Table 17: Differences between the Sequential and the Parallel output

9 Algorithms analysis with the Visual Profiler

After collecting the profile of the applications using **nvprof** (`prun -v -np 1 -native -l gpu=GTX480' nvprof -analysis-metrics -output-profile file_name CUDAapp image`), the output files were evaluated using the **Nvidia Visual Profiler** to better understand the applications bottlenecks. To have information about the lines of code that cause divergence execution or some uncoalesced issues the option `-lineinfo` was included for the compiling phase.

9.1 Darker

The choosen configuration, uses unidimensional *Blocks* of (1x256) threads, 12 *Registers* and 0 bytes of *Shared Memory*. Table 18 shows some interesting measurements. For all the images the profiler found no issues for *Divergent Execution* (threads that follow different if branches) and a *Warp Execution Efficiency* (ratio of the average active threads per warp to the maximum number of threads per warp supported on a multiprocessor) of 100%. This last value comes from the fact that is used a block of 256 threads, that is a multiple of 32 (warp size), so that no useless threads are launched and from the fact that there is no thread divergent execution, so that the threads within a warp can execute in a SIMT way, avoiding inactive threads within the warp. All the images, obtained an occupancy over 91% except the first that obtained 80.2%. This is probably correlated to its small size, but in any case the occupancy is not limiting the performance of the application. For all the images exepct the 5th, the Profiler found no issues related to Global Memory Access Pattern. The 5th image, is the only one of the given set that has an amount of pixels that is not a multiple of 128. The memory in the device is allocated with a 128-byte line granularity, so probably, the addresses fall within 2 cache lines when try to read the G and B values and 2 transaction insteads of one are performed. Figures 30a and 30b show an example in which 32 consecutive words of 4 bytes are requested by the threads in a warp. In the first case, the accesses are perfectly aligned, so that only 128 bytes move when there is a miss (100% of utilization). In the second case, the accesses are not aligned and 256 bytes move (50% utilization). The 2 lines of code in the Listing 5 are the source of the uncoalesced access for *image05*. To check the differences, a compilation with the option `-Xptxas -dlcm=cg` was tried. With this option, the search of data go directly to L2, and on L2 miss go to DRAM. In addition, the transactions are not performed by 128 byte lines but by segments of 32 byte, that for unaligned accesses means a better bus utilization. Figures 30c and 30d show what happens in the same scenario of aligned and unaligned accesses explained before. In the first case, 128 bytes move on a miss (100% utilization), while in the second case at most 160 bytes move (at least 80% of utilization). By the way, using this method, worst results are obtained for all the images. In fact, this last method is very good when the accesses are very scattered and this is not the case.

Images	Warp Execution Efficiency (%)	Occupancy (%)	Branch Divergence (%)
Img00	100	80.2	No issues
Img01	100	91.9	No issues
Img02	100	92.3	No issues
Img03	100	93	No issues
Img04	100	92.5	No issues
Img05	100	92.3	No issues
Img06	100	92.8	No issues

Table 18: Visual Profiler Results

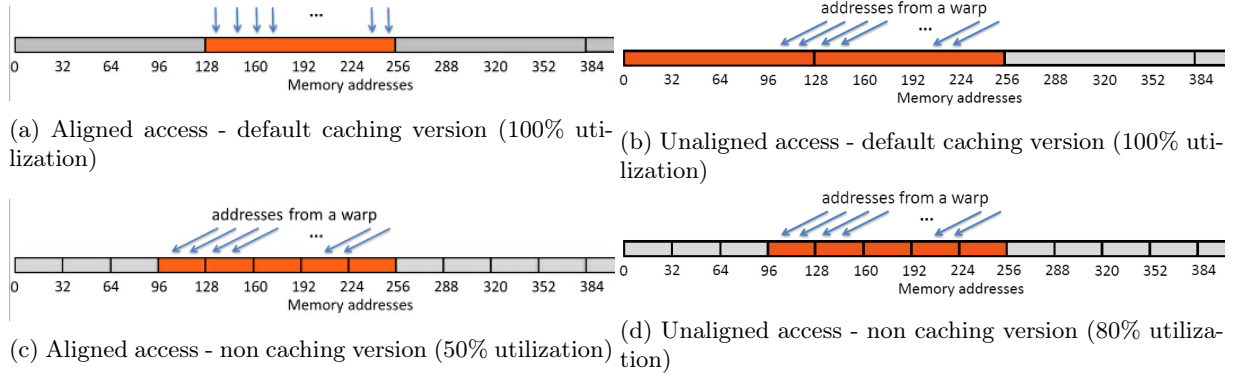


Figure 30: Possible access alignments

```

float g = static_cast< float >(inputImage[(width * height) +
                                           (y * width) + x]);
float b = static_cast< float >(inputImage[(2 * width * height) +
                                           (y * width) + x]);

```

Listing 5: Unaligned accesses

9.2 Histogram

The chosen configuration, uses unidimensional *Blocks* of (1x256) threads, 14 *Registers* and 8192 bytes of *Shared Memory* (8 histograms of 256 unsigned int). Table 19 shows some interesting measurements. For all the images the profiler found issues for *Divergent Execution* and this is due to atomic operations on the shared memory. On the *GTX 480*, in fact, *_shared_memory atomics* are implemented not as a single machine instruction, but by a sequence of machine (SASS) instructions that form a loop. This loop is essentially contending for a lock. When the lock is acquired by a particular thread, that thread will then complete the requested memory operation atomically on the identified shared memory cell, and then release the lock. The process of looping to acquire the lock necessarily involves branch divergence. This problem there would not be using global atomics, that are generally implemented as a single atomic SASS instruction, but as shown before with global atomics worst results were obtained. The profiler, in addition, shows a *Warp Execution Efficiency* that is variable from 50 to 90%. This last value comes directly from the fact that there is thread divergent execution, so that some threads within a warp are inactive during the execution. All the images, obtained an occupancy over 91.5% so the occupancy is not limiting the performance of the application. As for the previous algorithm and for the same reason, for all the images except the 5th, the Profiler found no issues related to Global Memory Access Pattern. The 2 lines of code that lead to uncoalesced accesses for *image05* are the same as the ones for the previous algorithm (Listing 5). For what concerning *bank conflicts*, no optimization can be done, because the accesses to the histograms in shared memory depends on the image.

Images	Warp Execution Efficiency (%)	Occupancy (%)	Branch Divergence (%)
Img00	90.7	91.5	74.6
Img01	60.9	98.8	93.1
Img02	52.3	99	91.2
Img03	55.2	99.4	90
Img04	72.5	99.6	90
Img05	64.2	99.4	82.2
Img06	59.5	99.5	86.1

Table 19: Visual Profiler Results

9.3 Smooth

The chosen configuration, uses bidimensional *Blocks* of (16x32) threads, 21 *Registers* and 0 bytes of *Shared Memory*. Table 20 shows some interesting measurements. The results seem to be worst than the previous algorithms and there is an explanation. Using 21 registers for each thread, means 10752 registers for each block and this register usage is likely preventing the kernel from fully utilizing the GPU because the *GTX 480* provides "just" up to 32768 registers for each block. By the way, as suggested by the Profiler, the occupancy "is large enough that an increment may not improve performance". The number of registers for the algorithm can't be reduced and testing the same program with smaller blocks the occupancy increased but the overall program obtained worst results as shown before. The explanation is that, if it is possible to achieve coalesced accesses to global memory, then a lower occupancy will be sufficient to hide the latency to global memory. For all the images the profiles outlined some issues for global memory access patterns. The source of uncoalescing is the first line of code of Listing 6. The problem is that even if are used blocks of 32 threads wide, to apply the filter each thread has to read the 25 pixels matrix of neighbors and, due to the fact that the memory in the device is allocated with a 128-byte line granularity, doing this for certain blocks the addresses to read fall within 2 cache lines and 2 transaction instead of one are performed. For *image5* in addition to the first line of Listing 6, also the second is a source of uncoalescing, for the same reason of the previous algorithms. Also in this case, to check the differences, a compilation with the option *-Xptxas -dlcm=cg* was tried, but also for this program, worst results are obtained for all the images, so the accesses are not sufficiently scattered to achieve performance. For what concerning the *branch divergence*, is strange that is reported only for some images, because following the code structure, reading the matrix of neighbors, there should be problems of divergence in all the blocks in the first and last column of the grid. The results of the *Warps Execution Efficiency* derived from the branch divergence.

Images	Warp Execution Efficiency (%)	Occupancy (%)	Branch Divergence (%)
Img00	100	66.1	No issues
Img01	99.9	66.3	3.9
Img02	99.4	66.3	13.8
Img03	100	66.6	No issues
Img04	99.6	66.4	7.4
Img05	99.9	66.6	No issues
Img06	100	66.5	No issues

Table 20: Visual Profiler Results

```
smoothPix += static_cast< float >(inputImage[(z * width * height) +
                                           (fy * width) + fx]) *
                                           filter[filterItem];
smoothImage[(z * width * height) + (i * width) + j] =
    static_cast< unsigned char >(smoothPix + 0.5f);
```

Listing 6: Unaligned accesses

References

- [1] John Cheng, Max Grossman, and Ty McKercher. *Professional Cuda C Programming*. John Wiley & Sons, 2014.
- [2] *Parallel Forall*. 2016. URL: <https://devblogs.nvidia.com/parallelforall/>.
- [3] *NVIDIA Developer Forum*. 2016. URL: <https://devtalk.nvidia.com/>.
- [4] *NVIDIA Developer*. 2016. URL: <https://developer.nvidia.com/>.
- [5] *NVIDIA Documentation*. 2016. URL: <http://docs.nvidia.com/>.