Programming Massively Parallel Processors

Max von Buelow <max.von.buelow@gris.tu-darmstadt.de>

TECHNISCHE UNIVERSITÄT DARMSTADT

WS 2020/21
Sheet 2

Deadline: November 27th 2020, 23:55 CET
Submission: Code and PDF as separate moodle uploads

Introduction Your task is to implement a Gaussian filter to blur images. The implementation will consist of several CUDA kernels in order to get accustomed with the performance difference of the GPU's different memory types and caches.

Prerequisites, hints and modalities See sheet 1.

Task 2.1: CPU

One way to blur an image is by convolving it with a Gaussian filter kernel. Gaussian filter kernels are rotationally symmetric which allows separation of the 2D convolution operation into two 1D convolutions, horizontal and vertical. For a more thorough description of separable convolution please refer to the whitepaper of the `convolutionSeparable` CUDA sample which is also available online on `http://docs.nvidia.com` in the `CUDA Samples` section.

A CPU implementation of the two convolution operations is given in `conv_cpu.{cc,h}`. Additionally, `image.{cc,h}` contains a class which can load and save PPM images and should upload your images to the GPU. For generating a Gaussian filter kernel to use with the convolution you can use the supplied `filterkernel` class in `common.{cc,h}`.

Your task is to implement Gaussian blur by using the given code to load an image, generate a Gaussian filter kernel, apply Gaussian blur to the image and save the blurred image to a file named `out_cpu.ppm`. Your program should takes two command-line parameters. The first one is the file name of the image that should be blurred and the second is the filter kernel size. To use the CLI parameters, edit the corresponding line of the `job.sh.in` file and run CMake again.

Task 2.2: Global memory

Write a GPU implementation of the horizontal and vertical 1D convolution which works on a `ppm` image. For this task, you should only use global memory in your implementation. The `ppm` struct represents the pixels of an image in 4 bytes (1 byte per color channel, 4 color channels (RGBA)). The convolution should only be applied to the RGB color channels; you can ignore the alpha channel. Your implementation should be able to handle different filter kernel sizes. Finally, save the resulting image to a file named `out_gpu_gmem.ppm`.
Note: Be careful to handle out-of-bounds memory accesses at the borders of the image by clamping the accesses to the nearest valid pixel values.

Task 2.3: Shared memory

Copy your convolution CUDA kernels from task 2 and modify them to use shared memory for the pixel data. For an idea on how shared memory can help with convolution, see the whitepaper about the `convolutionSeparable` CUDA sample. The blurred image should be written to a file named `out_gpu_smem.ppm`.

Note: Do not implement everything from the whitepaper! A simple implementation which ignores the coalescing rules from old GPUs is completely sufficient.

Task 2.4: Constant memory

Copy your convolution CUDA kernels from task 2 and modify them to access the image in global memory and the filter kernel in constant memory. Because the amount of constant memory is determined at compile-time, you can restrict the maximum filter kernel size (e.g. 127). Save the blurred image to a file named `out_gpu_cmem.ppm`.

Task 2.5: L1/texture cache

Copy your convolution CUDA kernels from task 2 and modify them to access both pixel data and the filter kernel in global memory via the L1/texture cache by using the `__restrict__` keyword. For details, read chapter B.2.4 in the CUDA Programming Guide. Save the blurred image to a file named `out_gpu_tmem.ppm`.
Note: Usage of the L1/texture cache with the `__restrict__` keyword only has an effect on GPUs with a compute capability of 3.5 or higher. Additionally, you have to tell `nvcc` to compile the CUDA kernels for a specific compute capability by using the `-arch` option. (e.g. `-arch=sm_35`) Passing the option in CMake is done by setting the `CUDA_NVCC_FLAGS` variable: `set(CUDA_NVCC_FLAGS "-arch=sm_35")`

Task 2.6

Copy your convolution CUDA kernels from task 2 and modify them such that they combine all of the optimizations from the previous tasks: cache pixel data in shared memory by loading it from global memory via the L1/texture cache and access the filter kernel in constant memory. Save the blurred image to a file named `out_gpu_all.ppm`.

Task 2.7

Observe the runtime difference between each task's implementation by profiling your program using `nvprof` or the NVIDIA Visual Profiler. For details on both profilers refer to the documentation in the CUDA Toolkit or online at `http://docs.nvidia.com` under `Tools` and then `Profiler`. Copy the runtimes of the kernels to the answer section of the file `filtering.cc`. Use the 2048x2048 image.