Здание №2. Basic Image Convolution on NVIDIA GPUs using CUDA

Формулировка задания

В ходе выполнения первого задания была реализована базовая версия программы, которая производит загрузку изображений в двух режимах, применение фильтра одного из трех типов, замер времени работы программы в двух режимах (только ядер и ядер + время копирований).

Важно, чтобы замер времени для второго режима (много небольших картинок) так же производился в режиме wall_time (загрузки с диска + копирования на GPU + ядра суммарно для всех картинок из коллекции лошади/люди). Такой замер можно производить, к примеру, при помощи библиотеки chrono.

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

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

Дополнительные оптимизации для обработки набора из маленьких изображений:
- Выделение памяти (cudaMalloc) под обрабатываемые изображения 1 раз (а не каждый раз для каждого изображения заново);
- Обработка нескольких изображений за раз одним ядром или обработка нескольких изображений в конкурентном режиме при помощи CUDA-потоков;
- Одновременные копирования DtoH, HtoD и запуск ядер;
- 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__).

При этом особое внимание стоит уделить составлению отчета, где нужно исследовать эффекты от применяемых оптимизаций.

Входные данные

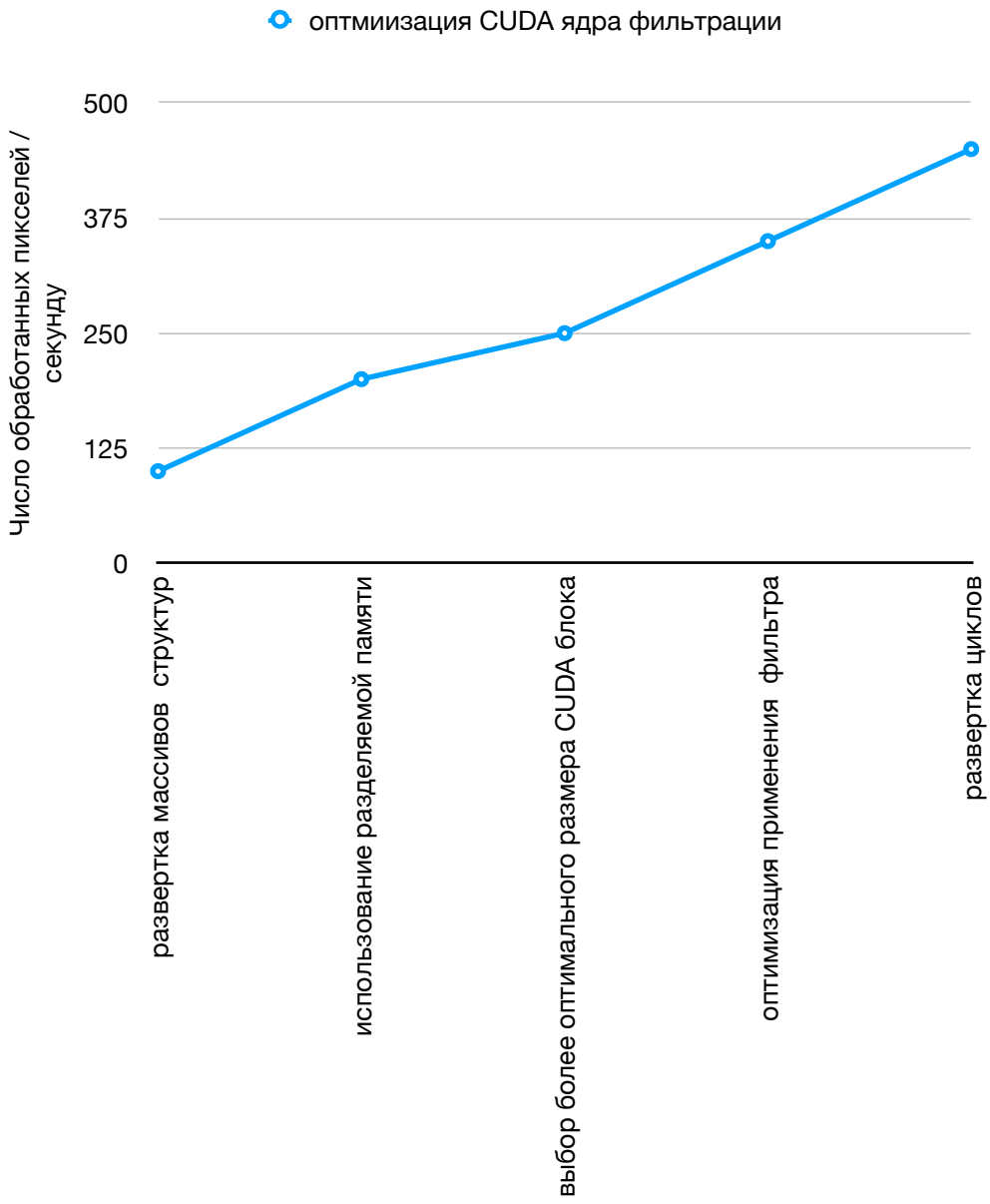
В качестве входных данных необходимо использовать данные двух видов:

1) одно изображение большого размера (например 2000x2000) — найдите любые большие изображения в интернете самостоятельно (чем больше — тем лучше)
2) много изображений маленького размера (300x300) — https://laurencemoroney.com/datasets.html

Требования к отчету

В отчете должны быть обязательно отражены следующие составляющие:

1. описание используемых оптимизаций и обоснование необходимости их применения для данной программы (за счет чего получаем ускорение).
2. замеры времени, демонстрирующие эффект от применения **КАЖДОЙ** оптимизации. Например: для изображения X время работы CUDA-ядра до развёртки массива структур в структуру массивов составляло Y секунд, после — Z секунд, полученное ускорение — Y/Z секунд.
3. График оптимизаций (пример на рисунке ниже). По оси X - последовательно применяемые оптимизации, по оси Y - производительность (число обрабатываемых пикселей в секунду) программы после применения данной оптимизации. Таких графика нужно сделать два, для времени работы CUDA ядра для одного большого изображения, а так же время работы всей программы для полной коллекции маленьких изображений.
4. Полную профилировку при помощи средства nvprof изначальной (задание 1) и финальной (конец выполнения задания 2) версии приложения для 1 большого и всех маленьких изображений: трасса работы, данные о профилировке ядра фильтрации.



Сроки выполнения задания

Выдача задания: 4 октября
Сдача задания: 15 ноября, 14:35 — 18:00.

Правила выставления зачета по курсу

Всего в семестре будет 4 задания. Каждое задание оценивается зачет/не зачет. Оценка за семестр складывается из сдачи заданий:

- в срок сданы все 4 задания — отлично
- в срок сданы 3 задания — хорошо
- в срок сданы 2 задания — удовлетворительно
- если сдано 1 или 0 заданий — не удовлетворительно

Сдача первого и второго заданий— обязательна.

Задания можно сдавать только в указанный срок (за исключением ситуации болезни/и др.).

Оценку «не удовлетворительно» можно исправить только на «удовлетворительно» путем сдачи всех заданий и дополнительной беседы с преподавателем по материалам курса.