Министерство образования и науки Российской Федерации

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО»

Кафедра математической
кибернетики и компьютерных наук

ВЫЧИСЛЕНИЯ НА ВИДЕОКАРТАХ

КУРСОВАЯ РАБОТА

студента 3 курса 351 группы
направления 09.03.04 — Программная инженерия
факультета КНиИТ
Григорьева Алексея Александровича

Научный руководитель
доцент _____ М. С. Семенов

Заведующий кафедрой
к. ф.-м. н. _____ С. В. Миронов

Саратов 2019

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Будет добавлено позже. Немного об истории, сравнение характеристик процессора и видеокарты

Цели курсовой работы:

- ознакомиться с теорией, необходимой для написания эффективных алгоритмов, исполняющихся на видеокарте, с использованием OpenCL;
- понять общие свойства архитектуры видеокарты и тем самым научиться оптимизировать алгоритмы;
- получить практический опыт разработки программ на видеокартах с помощью OpenCL;

1 Краткая теория

Составление эффективных алгоритмов вычисления на видеокарте в значительной степени отличается от привычных алгоритмов, исполняющихся на процессоре. При составлении программного кода необходимо учитывать как и общие особенности видеокарт, так и, возможно, характеристики конкретного устройства, для которого программируется алгоритм.

В данном разделе будет рассмотрена типовая модель видеокарты и основные понятия OpenCL, с которыми будем оперировать в данной работе.

1.1 Типовая модель видеокарты

Рассмотрим следующую архитектуру вычислительного устройства, используемого в видеокартах Nvidia 1.

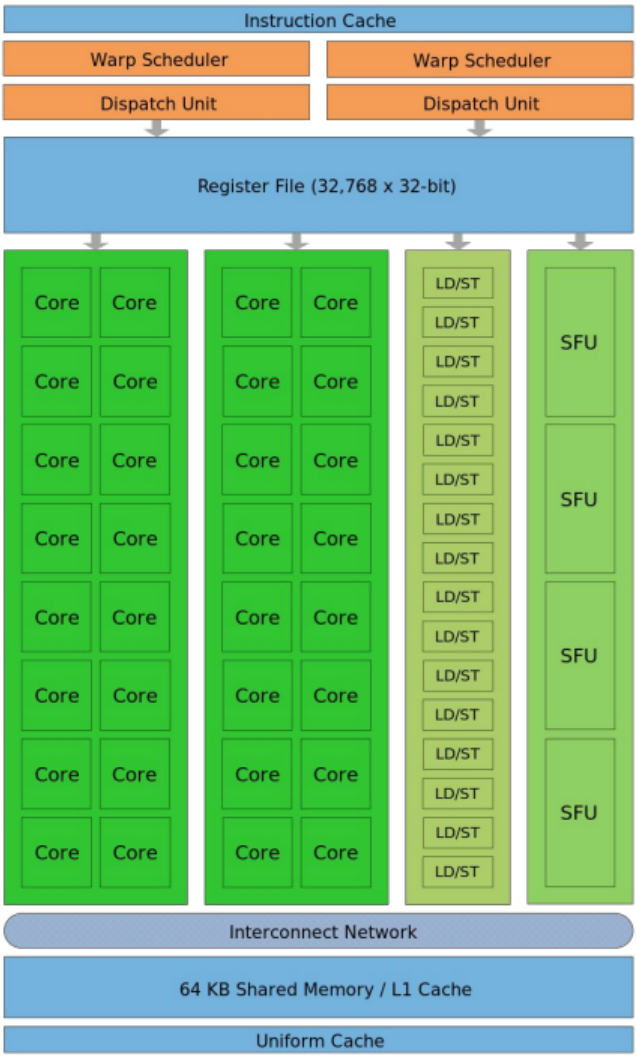

Рисунок 1 – Архитектура потокого мультипроцессора Fermi.

Вычислительное устройство в архитектуре Nvidia имеет 32 **ядра** (CUDA

cores), каждое из которых в состоянии работы является **потоком**. В отличии от процессора, ядра выполняют более узкий набор задач, что позволяет с меньшими затратами увеличить их количество в устройстве [1]. Для управления ими существует **warp scheduler**, выполняющий роль указателя на инструкции соответствуя архитектуре SIMD. Данные для вычислений потоки берут из **локальной памяти** (shared memory), общей для всех ядер. Достигается это с использованием **устройств загрузки и хранения** (load-store units), соответственно способных загружать, а также сохранять данные в локальную память. Между всеми 32 ядрами вычислительного устройства динамически распределяются **регистры**, самая быстрая память, доступная им. У мультипроцессора в наличии намного больше регистров, чем могло быть нужно для выполнения программы. Это сделано для сокрытия времени на загрузку памяти и быстрого переключения контекста, подробнее - в разделе 2.1.

Количество таких устройств в видеокарте определяется следующим образом:

Количество ядер в видеокарте / 32, в случае Nvidia

Количество ядер в видеокарте / 64, в случае AMD

В терминологии Nvidia, поток из всех (32) активных ядер вычислительного устройства образует **warp**, Например, видеокарта Nvidia Geforce GTX 1050 Ti имеет 768 ядер CUDA, и, соответственно, 24 warp.

1.2 Основные понятия OpenCL

OpenCL – открытый для свободного пользования программный интерфейс для создания параллельных приложений, использующих многоядерные структуры как и центрального процессора (CPU), так и графического (GPU). Использование API необходимо для обеспечения совместимости программы с различными устройствами [2].

При построении задач, определяется **рабочее пространство** (NDRange), представляющее собой все возможные в рамках задачи значения индексов потоков. Размер рабочего пространства определяется программистом на этапе инициализации OpenCL программы. Рабочее пространство может представлять:

- одномерный массив длиной N элементов;
- двумерную сетку размерности NxM;

— трехмерное пространство размерностью NxMxP.

Код, выполняющийся параллельно на ядрах процессора, называется **kernel**. Копия kernel выполняется для каждого индекса рабочего пространства и называется **work-item** с глобальным ID, соответствующим некоторому ID рабочего пространства. Kernel для всех work-item в рабочем пространстве имеют одинаковый код и входные параметры, но может иметь различный путь выполнения программы соответственно своему глобальному индексу - индекс в рабочем пространстве, полученному с использованием функции `get_global_id()`. Kernel в отличии от остальной программы полностью выполняется на видеокарте [3].

Группа work-item называется **work-group**, и за каждой группой закреплен собственный warp (см. предыдущий раздел), в рамках которого work-item могут синхронизироваться. Для каждой рабочей группы существует ее индекс в рабочем пространстве, и каждый work-item может узнать свой индекс внутри рабочей группы. Нетрудно заметить следующее соотношение:

$$\text{global ID} = \text{group ID} * \text{размер группы} + \text{local ID}$$

Размер рабочей группы аналогично рабочему пространству определяется программистом.

Каждое ядро, выполняя заданный kernel, является work-item в некоторой рабочей группой, на которые разделено рабочее пространство NDRange.

Рассмотрим на примере следующей схемы 2 другие виды сущностей, с которыми будет взаимодействие в OpenCL.

— Платформа — драйвер, модель взаимодействия OpenCL и устройства. Распространены платформы от следующих производителей: Nvidia, Intel, AMD.
— Программа — хостовая часть, организующая подготовку к вычислениям и набор kernel-подпрограмм.
— Kernel — программа, исполняющаяся на видеокарте в каждом ядре.
— Контекст — окружение, в котором исполняется kernel.
— Объект памяти — создаваемый в контексте объект.
— Буфер — произвольный массив данных.

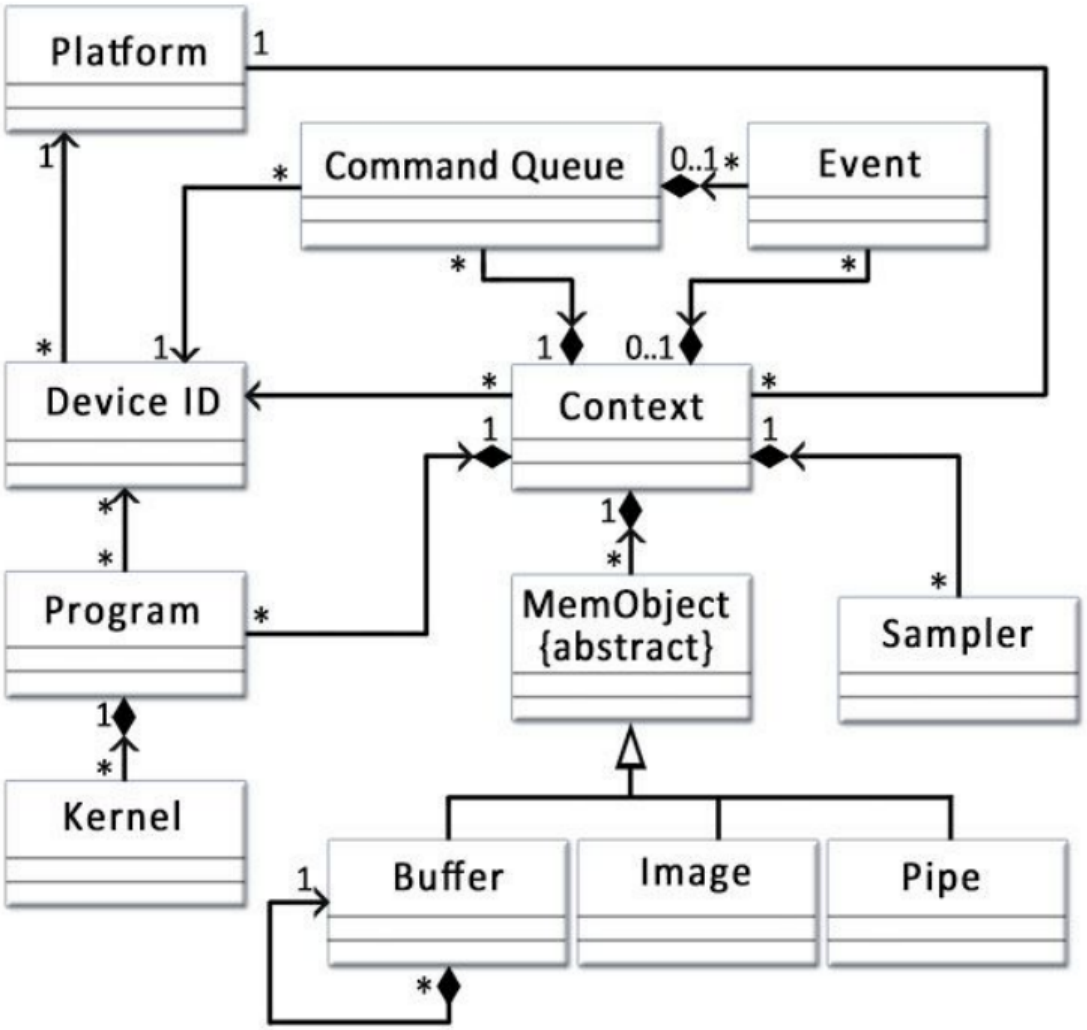

Рисунок 2 – Основные сущности в OpenCL.

2 Алгоритмы на видеокарте

В данном разделе будет рассмотрена анализ и практическая реализация алгоритмов на видеокарте, включая:

— описание общих требований к алгоритмам на основе доступа к памяти и параллельного исполнения;
— настройка среды разработки Microsoft Visual Studio 2017 под выполнение параллельных программ с использованием OpenCL;
— написание программ для задач, использующих входные данные разных размерностей.

2.1 Требования к алгоритмам

Любой алгоритм можно вычислить на видеокарте, но эффективность в сравнении с реализацией на центральном процессоре зависит от корректного построения алгоритма для видеокарты.

Массовый параллелизм заключается в том что задачу можно разбить на рабочие группы так, что не будет требоваться постоянная синхронизация между work-item из разных рабочих групп.

Следует вспомнить, что все потоки в warp выполняют одинаковые инструкции в любой момент времени. Какая инструкция будет выполняться определяется с помощью warp scheduler, единого для всех потоков в warp. Рассмотрим следующий фрагмент кода:

```
if (predicate) {
    value = x[i];
}
else {
    value = y[i];
}
```

Учитывая сказанное выше, все потоки при срабатывания if-части должны выполнить внутреннюю часть, однако это не совсем так, и если у потока предикат — `False`, он будет спрятан от выполнения внутренней части, аналогично и с `else`-частью. Однако несмотря на то что результат выполнения конструкции `if-else` будет верным, часть потоков будет простаивать, ожидая выполнение маскированных для них частей.

Данная ситуация называется code divergence, и она может стать причиной низкой производительности программы. Этого можно избежать, если организовать код таким образом чтобы для всех потоков предикат возвращал одинаковое значение, тогда конструкция не соответствующая ему будет пропущена указателем на инструкции. Если это невозможно, то для эффективного выполнения алгоритма рекомендуется отказаться от многочисленных сложных ветвлений, так как сложность выполнения фрагмента алгоритма будет вычисляться как сумма if- и else- частей вместо максимума как в последовательных программах.

При выборе размера рабочих групп стоит учитывать особенности алгоритма, однако, есть некоторые общие правила, которых необходимо придерживаться.

1. Размер рабочей группы не должен быть меньше warp.
2. Размер рабочей группы должен быть кратен 32 (64 если используется AMD).

В противном случае, некоторые потоки будут простаивать, ожидая пока остальные завершат свою работу

Как известно, операции с памятью являются одними из самых долгих по времени выполнения, в связи с этим было решено сделать разбиение задач на рабочие группы, в результате у видеокарт появился аналог имеющегося у процессоров hyper-threading. Он заключается в использовании каждым вычислительным устройством регистров для переключения контекста при задержке, созданной обращением к памяти (latency) [4].

Другими словами, warp может быстро сохранить состояние выполнения в данной рабочей группе и пока выполняется долгая операция обращения к памяти, вычислительное устройство может переключиться на другой warp в рабочей группе, и если второй warp хочет выполнить операцию обращения к памяти, то происходит возвращение к первому warp если доступ к памяти завершился, либо активируется третий warp и так далее. Следствие — высокая вычислительная мощность и большая пропускная способность видеокарты [5].

Количество одновременно активных warp в рабочей группе определяется как минимум из:

— количества регистров / количество используемых в warp регистров;
— количества локальной памяти / количество используемой локальной па-

мяти;

— максимально допустимого количества warp (~10).

В соответствии с этим существует величина **occupancy**, определяемая соотношением

среднее кол-во активных warp / максимальное кол-во активных warp

Не всегда высокий occupancy означает что программа имеет высокую производительность. Например, если доступ к памяти в программе очень быстрый, только из регистров, то необходимости в сокрытии задержки и переключения контекста нет, и occupancy будет низким.

Однако, низкий occupancy и высокая задержка при обращении к памяти может означать что программа написана не достаточно эффективно, и ей необходимы улучшения, если это возможно.

Чем больше на одном вычислителе warp — тем реже все warp оказываются в состоянии «ждем запрос памяти» и тем реже вычислитель будет простаивать, т.к. тем чаще у него находится рабочая группа в которой можно что-то посчитать [4].

Если потоки из одного warp делают запрос к памяти, то эти запросы склеются в столько запросов, сколькими кеш-линиями покрываются запрошенные данные.

Другими словами, если потоки запрашивают данные, которые в памяти лежат подряд, то достигнутая пропускная способность будет максимальная так как запросы «склеются». Размер кеш-линии обычно от 32 до 128 байт.

Если приложение использует OpenCL 1.x, то размеры NDRange должны нацело (без остатка) делиться на размеры рабочих групп. Там, где данные образуют NDRange с другим размером, необходимо самостоятельно изменить их чтобы выполнялось это условие, например, добавлением нулей или средних значений, которые не будут значимо влиять на результат вычислений [6].

В OpenCL 2.0 появилась новая возможность, в которой устранена данная проблема. Речь идет о так называемых неоднородных рабочих группах: выполняемый модуль OpenCL 2.0 может разделить NDRange на рабочие группы неоднородного размера по любому измерению. Если разработчик укажет размер рабочей группы, на который размер NDRange не делится нацело, выполняемый модуль разделит NDRange таким образом, чтобы создать как мож-

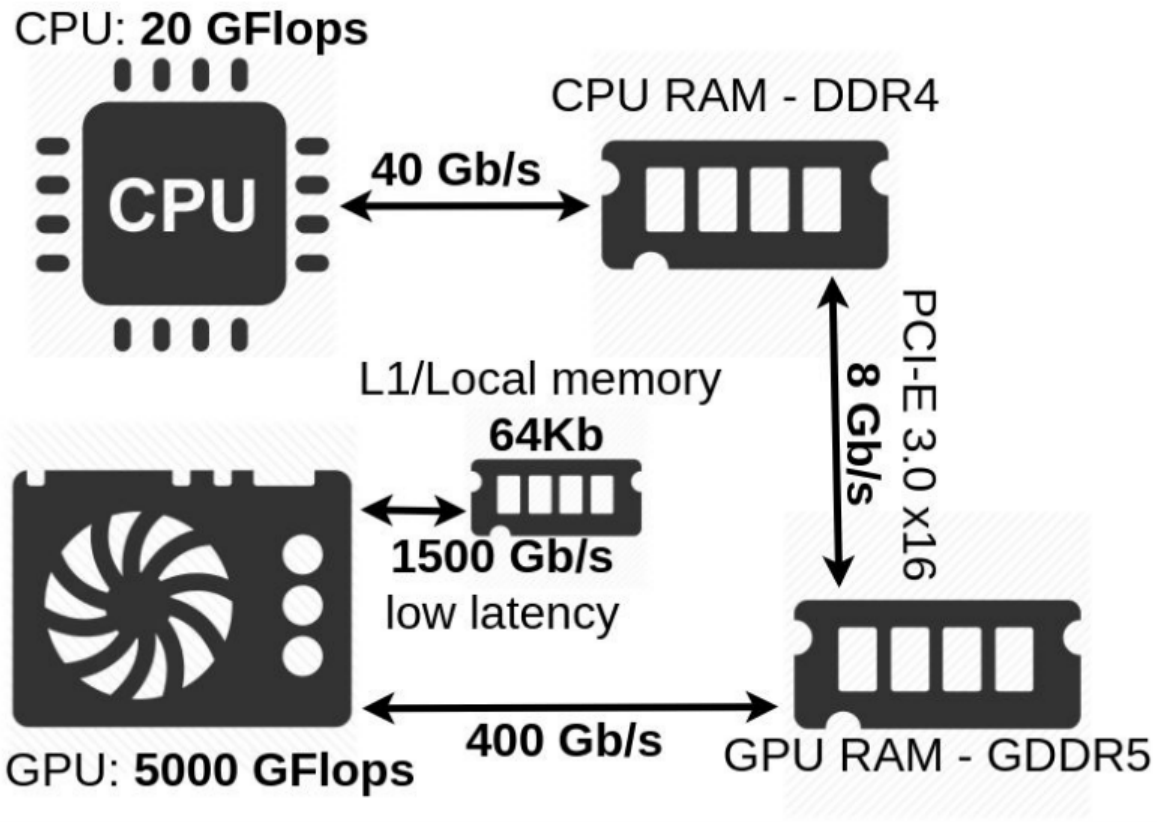

Рисунок 3 – Доступные ресурсы — память.

но больше рабочих групп с указанным размером, а остальные рабочие группы будут иметь другой размер. Например, для NDRange размером 1918x1078 рабочих элементов при размере рабочей группы 16x16 элементов среда выполнения OpenCL 2.0 разделит NDRange, как показано на приведенном ниже рисунке 4.

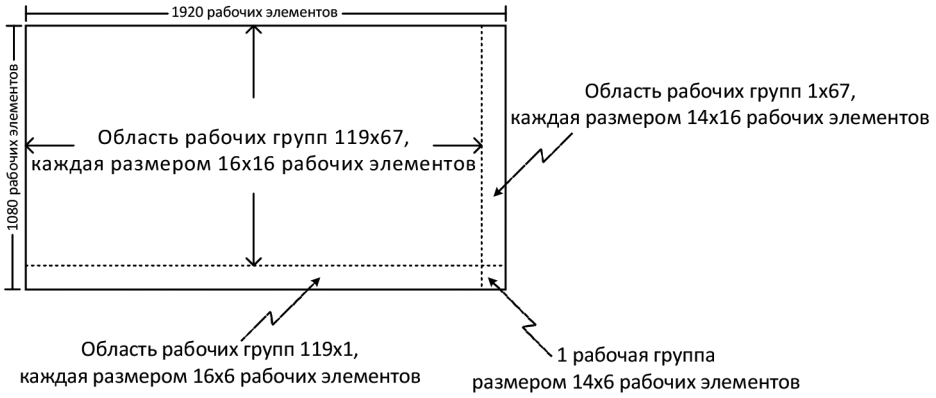

Рисунок 4 – Разделение NDRange на рабочие группы разных размеров.

2.2 Настройка среды разработки

В данном разделе будет рассмотрен процесс настройки среды разработки и создания первого OpenCL-проекта.

В качестве среды разработки для программирования с использованием OpenCL выбрана Microsoft Visual Studio, язык программирования — C++.

На компьютер была установлена реализация OpenCL от Nvidia — Nvidia GPU Computing SDK. А также программа CMake, являющаяся независимым от платформы инструментом для сборки проектов 5.

С помощью графического интерфейса выберем расположение файлов исходного кода и места сборки проекта. Директория с исходными файлами должны содержать текстовые файлы CMakeLists из приложения Б Если OpenCL установлен корректно, то нажатие кнопки «Configure» выведет найденные на компьютеры файлы, связанные с OpenCL. Нажмем Generate, и перейдем в папку с проектом, в котором можно увидеть созданный файл .sln проекта Microsoft Visual Studio, сконфигурированного под OpenCL.

2.3 Инициализация OpenCL программы

В данном разделе будут рассмотрены базовые функции, необходимые для инициализации параллельной программы с использованием OpenCL. Данные функции будут предварять задачи из следующих разделов.

2.4 Задачи на одномерных массивах

(сумма ряда, какая-нибудь префикс-функция)

2.5 Задачи на двумерных массивах

(транспонирование, переменожение матриц)

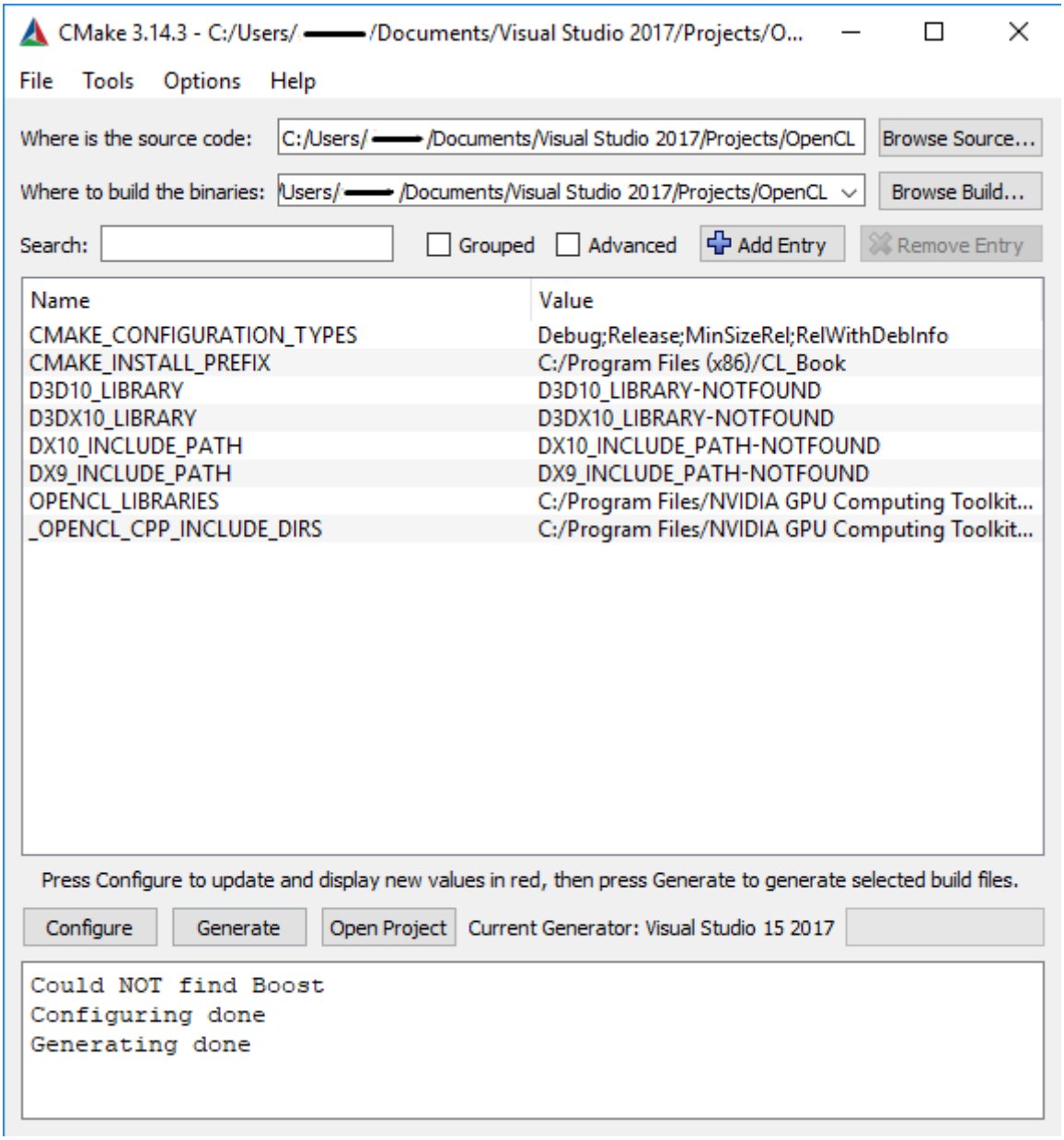

Рисунок 5 – Окно программы CMake-gui.

ЗАКЛЮЧЕНИЕ

В настоящей работе

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1 [NVIDIA's Next Generation CUDA Compute Architecture: Fermi]. — URL: https://www.nvidia.com/content/PDF/fermi_white_papers/NVIDIA_Fermi_Compute_Architecture_Whitepaper.pdf (Дата обращения 12.05.2019). Загл. с экр. Яз. англ.

2 [Введение | OpenCL]. — URL: http://opencl.ru/node/8 (Дата обращения 12.05.2019). Загл. с экр. Яз. рус.

3 [The OpenCL Specification]. — URL: https://www.khronos.org/registry/OpenCL/specs/ (Дата обращения 12.05.2019). Загл. с экр. Яз. англ.

4 [Введение в OpenCL. Архитектура видеокарты]. — URL: https://compscicenter.ru/courses/video_cards_computation/2018-autumn/classes/3980/ (Дата обращения 12.05.2019). Загл. с экр. Яз. рус.

5 [Achieved Occupancy]. — URL: https://docs.nvidia.com/gameworks/content/developertools/desktop/analysis/report/cudaexperiments/kernellevel/achievedoccupancy.htm (Дата обращения 12.05.2019). Загл. с экр. Яз. англ.

6 [Неоднородные рабочие группы OpenCL 2.0]. — URL: https://software.intel.com/ru-ru/articles/opencl-20-non-uniform-work-groups (Дата обращения 12.05.2019). Загл. с экр. Яз. рус.

ПРИЛОЖЕНИЕ А
Листинг программы

Код

ПРИЛОЖЕНИЕ Б

Листинг сборочных файлов CMake

Код