**МИНОБРНАУКИ РОССИИ**

**Санкт-Петербургский государственный**

**электротехнический университет**

**«ЛЭТИ» им. В.И. Ульянова (Ленина)**

**Кафедра МО ЭВМ**

**Отчет**

**по лабораторной работе №6**

**по дисциплине «Параллельные алгоритмы»**

**Тема: Оптимизация доступа к памяти в модели OpenCL**

|  |  |  |
| --- | --- | --- |
| Студент гр. 0304 |  | Аристархов И.Е. |
| Преподаватель |  | Сергеева Е.И. |

Санкт-Петербург

2023

**Цель работы.**

Используя OpenCL реализовать параллельное умножение матриц.

**Выполнение работы.**

Общий принцип работы с OpenCL был описан в предыдущей работе, поэтому данная часть будет описана кратко с подробностями в различиях.

Сперва получается устройство (в данном случае GPU), создается на него контекст и очередь команд. Затем загружается исходный код для kernel, он собирается и с ним формируется Kernel. Далее идет формирование переменных. Именно здесь и лежит основное отличие с предыдущей работой, где использовалось 2D изображение, когда как тут требуется создать 3 двумерных массива. Это можно сделать, создав одномерные буферы с ручным контролем перебора индексов у них. Для первых двух буферов (для матриц перемножения) передается не только размер, но и начальное значение, когда последний (для результата) остается без последнего.

Теперь буферы и переменная размера привязываются к Kernel и создается задача на вычисление в очереди. Остается дождаться конца вычислений и очистить память.

Далее перейдем к программе. В ней было решено использовать локальную память, которая общая в пределах рабочей группы. Это позволит переиспользовать память и уменьшить количество запросов в глобальную память.

Каждый рабочий элемент в группе ответственен за загрузку своего участка локальной памяти (которая имеет форму квадрата). Далее производится операция перемножения в пределах этого блока, после чего загружается следующий блок.

Остается добавить, что производимые вычисления в данной программе сравнивались с правильными результатами, рассчитанными с помощью прошлой работы.

**Анализ.**

Вычисления производились на Nvidia GTX 1650 4 Гб. Рабочая группа выставлялась в значение 32x32 (большие значения не поддерживаются).

Были получены следующие результаты:

|  |  |  |
| --- | --- | --- |
| Размер матрицы | Время на CPU, сек | Время на GPU, сек |
| 128 | 0,064 | 0,008 |
| 256 | 0,486 | 0,022 |
| 512 | 1,902 | 0,098 |
| 1024 | 14,267 | 0,463 |
| 2048 | 119,558 | 1,709 |

CPU реализация брала количество потоков по размеру матрицы (в одной из предыдущих работ был сделан вывод, что это лучший подход по скорости выполнения).

Из таблицы прекрасно видно, что GPU реализация выигрывает в десятки раз, относительно CPU реализации.

**Выводы.**

В результате выполнения лабораторной работы была разработана программа производящая умножение матриц, с использованием GPU на OpenCL. Сравнение данной программы с ранней реализацией на CPU показало, что GPU во много раз быстрее производит умножение.