ФЕДЕРАЛЬНОЕ АГЕНТСТВО СВЯЗИ
ГОСУДАРСТВЕННОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ
“СИБИРСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ТЕЛЕКОММУНИКАЦИИ И ИНФОРМАТИКИ”

Кафедра ВС

Лабораторная работа № 3, 4
«Программирование графических ускорителей»

Выполнил:
студент группы МГ-165
Марков В.А.

Проверил:
Малков Е.А.

Новосибирск 2017

Лабораторная 3

Цель лабораторной работы: изучить модель выполнения CUDA, warps, совместный доступ к глобальной памяти.

Задание 1:

- определить для своего устройства зависимость теоретической заполняемости мультипроцессоров от числа нитей в блоке;
- для программы инициализации вектора определить достигнутую заполняемость в зависимости от длины вектора.

Name	Start Time	Duration	Grid Size	Block Size	Regs	Static SMem	Dynamic SMem	Achieved Occupar
vector_init(int*)	524,177 ms	65,002 µs	[512,1,1]	[256,1,1]	7	0	0	1,258
vector_init(int*)	516,717 ms	127,722 µs	[512,1,1]	[512,1,1]	7	0	0	1,253
vector_init(int*)	511,867 ms	64,167 µs	[512,1,1]	[256,1,1]	7	0	0	1,243
vector_init(int*)	516,907 ms	128,192 µs	[512,1,1]	[512,1,1]	7	0	0	1,232
vector_init(int*)	508,545 ms	128,279 µs	[512,1,1]	[512,1,1]	7	0	0	1,195
vector_init(int*)	512,915 ms	48,455 µs	[512,1,1]	[192,1,1]	7	0	0	1,182
vector_init(int*)	510,969 ms	65,189 µs	[256,1,1]	[512,1,1]	7	0	0	1,179
vector_init(int*)	525,077 ms	48,872 µs	[512,1,1]	[192,1,1]	7	0	0	1,161
vector_init(int*)	508,946 ms	128,101 µs	[512,1,1]	[512,1,1]	7	0	0	1,158
vector_init(int*)	524,383 ms	64,2 µs	[256,1,1]	[512,1,1]	7	0	0	1,147
vector_init(int*)	525,19 ms	48,71 µs	[192,1,1]	[512,1,1]	7	0	0	1,049
vector_init(int*)	512,796 ms	49,069 µs	[192,1,1]	[512,1,1]	7	0	0	1,024
vector_init(int*)	513,987 ms	32,855 µs	[128,1,1]	[512,1,1]	7	0	0	0,893
vector_init(int*)	525,856 ms	32,397 µs	[128,1,1]	[512,1,1]	7	0	0	0,879
vector_init(int*)	516,058 ms	8,264 µs	[32,1,1]	[512,1,1]	7	0	0	0,793
vector_init(int*)	515,413 ms	12,306 µs	[48,1,1]	[512,1,1]	7	0	0	0,732
vector_init(int*)	527,491 ms	12,156 µs	[48,1,1]	[512,1,1]	7	0	0	0,713

Рисунок 1 — Результат выполнения задания 1

Оптимальное количество нитей в блоке равно **192**.

Примечание: *использовать nvprof (пример: nvprof --metrics achieved_occupancy ./lab3) или nvvp, добавив метрику achieved_occupancy.*

Задание 2:

- применяя двумерную индексацию нитей в блоке и блоков в гриде написать программу инициализации матрицы, сравнить эффективность кода ядра при двух различных линейных индексациях массива;
- написать программу транспонирования матрицы.

```
$ ./matrix_init 1024 192 512
matrix_init_by_row took 0.054112
matrix_init_by_col took 8.14934
$ ./matrix_init 4096 192 512
matrix_init_by_row took 0.054304
matrix_init_by_col took 33.8086
```

Рисунок 2 — Результат выполнения задания 2

Примечание: *для профилирования программы использовать nvprof и nvpp.*

Лабораторная 4

Цель лабораторной работы: научиться использовать разделяемую память.

Задание:

- написать программу транспонирования матриц, реализующую алгоритм без использования разделяемой памяти, наивный алгоритм с использованием разделяемой памяти и алгоритм с разрешением конфликта банков разделяемой памяти;
- провести профилирование программы с использованием nvprof и nvpp — сравнить время выполнения ядер, реализующих разные алгоритмы, и оценить эффективность использования разделяемой памяти.

```
./matrix_transpose
Device : GeForce 820M
Matrix size: 4096 4096, Block size: 32 8, Tile size: 32 32
dimGrid: 128 128 1. dimBlock: 32 8 1
                  Routine         Bandwidth (GB/s)
                     copy              13.20
       shared memory copy              13.21
          naive transpose               8.03
      coalesced transpose              12.25
  conflict-free transpose              13.09
```

Рисунок 3 — Результат выполнения задания транспонирования матрицы

Device 0: "GeForce 820M"

CUDA Driver Version / Runtime Version	9.0 / 8.0
CUDA Capability Major/Minor version number:	2.1
Total amount of global memory:	964 MBytes (1011286016 bytes)
(2) Multiprocessors, (48) CUDA Cores/MP:	96 CUDA Cores
GPU Max Clock rate:	1250 MHz (1.25 GHz)
Memory Clock rate:	900 Mhz
Memory Bus Width:	64-bit
L2 Cache Size:	**131072 bytes**
Maximum Texture Dimension Size (x,y,z)	**1D=(65536),** **2D=(65536, 65535),** **3D=(2048, 2048, 2048)**
Maximum Layered 1D Texture Size, (num) layers	**1D=(16384), 2048 layers**
Maximum Layered 2D Texture Size, (num) layers	**2D=(16384, 16384), 2048 layers**
Total amount of constant memory:	**65536 bytes**
Total amount of shared memory per block:	**49152 bytes**
Total number of registers available per block:	**32768**
Warp size:	32
Maximum number of threads per multiprocessor:	1536
Maximum number of threads per block:	1024
Max dimension size of a thread block (x,y,z):	(1024, 1024, 64)
Max dimension size of a grid size (x,y,z):	(65535, 65535, 65535)

Рисунок 4 — Конфигурация графической карты