МОСКОВСКИЙ АВИАЦИОННЫЙ ИНСТИТУТ
(НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ)

Институт №8 «Компьютерные науки и прикладная математика»
Кафедра 806 «Вычислительная математика и программирование»

Лабораторная работа №4
по курсу «Программирование графических процессоров»

Работа с матрицам. Метод Гаусса.

Выполнил: А. Е. Максимов
Группа: М8О-407Б-19
Преподаватель: А. Ю. Морозов

Москва, 2023

1 Условие

Цель работы:Научиться использовать GPU для классификации и кластеризации изображений. Использование константной памяти и одномерной сетки потоков.

Вариант на "два". Быстрое транспонирование матрицы.

Необходимо реализовать транспонирование матрицы с бесконфликым использованием разделяемой памяти и объединением запросов к глобальной памяти.

Ограничения:

- $w * h \leq 10^8$
- все данные выводить с точностью 10^{-10}

2 Программное и аппаратное обеспечение

Характеристики графического процессора:

- Compute capability: 7.5
- Наименование: NVIDIA GeForce GTX 1660 Ti
- Графическая память: 6143mb
- Константная память: 64kb
- Разделяемая память: 48kb
- Количество регистров на блок: 65536
- Максимальное количество потоков на блок: (1024, 1024, 64)
- Максимальное количество блоков: (2147483647, 65535, 65535)
- Количество мультипроцессоров: 6

Характеристики системы:

- Процессор: «Intel(R) Core(TM) i7-10750H CPU @ 2.60GHz 2.59 GHz»
- Память: 16,0 ГБ встроенной памяти ноутбука
- SSD: «Hynix BC511 NVMe SK hynix 512GB »

Программное обеспечение:

- ОС: «Windows 10 Домашняя x64»
- Текстовый редактор: «Notepad++»
- Компилятор: «nvcc: Cuda compilation tools, release 12.0, V12.0.76»

3 Метод решения

Для решения лабораторной работы я использовал двумерный буфер-массив в разделяемой памяти. Поскольку его размерность создавала бы конфликты памяти, одно из измерений буфера было увеличено на 1 для создания смещения данных в банках памяти.

Основным элементом алгоритма являются своевременные синхронизации нитей блока для оптимизации обращений к глобальной памяти, а также аккуратная работа с индексами внутри цикла сдвигов. Благодаря качественной проработке кода, проблем с лабораторной не возникло.

4 Результаты

Замеры времени работы CPU и ядер с различными конфигурациями.

Данные представлены квадратными таблицами размера $n * n$.

Конфигурация	$n = 10$, мс	$n = 100$, мс	$n = 10^3$, мс	$n = 10^4$, мс
CPU	0.0	0.0	7.197	960.349
(1, 1)(1, 32)	0.0177	0.1030	11.8879	1026.7545
(1, 1)(32, 32)	0.0061	0.0164	1.2451	135.2888
(1, 32)(1, 32)	0.0081	0.0347	0.4669	73.1074
(1, 32)(32, 32)	0.0072	0.0081	0.1124	12.8184
(32, 32)(1, 32)	0.0081	0.0102	0.1649	71.0662
(32, 32)(32, 32)	0.05641	0.0553	0.1085	9.0114
(32, 256)(32, 32)	0.4157	0.4150	0.4643	9.5252
(256, 256)(32, 32)	3.2843	3.2829	3.2767	9.9075

5 Профилировка Nsight Compute

Конфигурация ядра — (32, 32)(32, 32), размер выходных данных — $n * m = 10^6$.

Конфликты банков памяти

```
C:\Users\cerma\Desktop\MAI\4c\ППП\lab4>ncu --metrics
l1tex__data_bank_conflicts_pipe_lsu_mem_shared_op_ld.sum,
l1tex__data_bank_conflicts_pipe_lsu_mem_shared_op_st.sum,
smsp__sass_average_branch_targets_threads_uniform.pct,
dram__sectors_read.sum,dram__sectors_write.sum,
l1tex__t_requests_pipe_lsu_mem_global_op_st.sum
a.exe <test2.txt

==PROF== Connected to process 16500 (C:\Users\cerma\Desktop\MAI\4c\PGP\lab4\a.exe)
==PROF== Profiling "kernel" -0: 0%....50%....100% -4 passes
152.7266235 ms
==PROF== Disconnected from process 16500
[16500] a.exe@127.0.0.1
kernel(double *,double *,int,int) (32,32,1)x(32,32,1),Context 1,Stream 7,Device
0,CC 7.5
Section: Command line profiler metrics
-------------------------------------------------------------------------------
Metric Name                                                 Metric Unit Metric
Value
-------------------------------------------------------------------------------
dram__sectors_read.sum                                        sector      250a181
dram__sectors_write.sum                                       sector      249a902
l1tex__data_bank_conflicts_pipe_lsu_mem_shared_op_ld.sum                       0
l1tex__data_bank_conflicts_pipe_lsu_mem_shared_op_st.sum                       0
l1tex__t_requests_pipe_lsu_mem_global_op_st.sum             request       32a000
smsp__sass_average_branch_targets_threads_uniform.pct             %        99,49
-------------------------------------------------------------------------------
```

Как показывают метрики $data\_bank\_conflicts$ - во время выполнения программы не возникают конфликты банков памяти.

Время выполнения

```
C:\Users\cerma\Desktop\MAI\4c\ППП\lab4>ncu --set full a.exe <tests\test1000x1000.txt
==PROF== Connected to process 22004 (C:\Users\cerma\Desktop\MAI\4c\PGP\lab4\a.exe)
==PROF== Profiling "kernel" -0: 0%....50%....100% -31 passes
629.9649048 ms
```

```
==PROF== Disconnected from process 22004
[22004] a.exe@127.0.0.1
kernel(double *,double *,int,int) (32,32,1)x(32,32,1),Context 1,Stream 7,Device
0,CC 7.5
Section: GPU Speed Of Light Throughput
----------------------------------------------
Metric Name               Metric Unit Metric Value
----------------------------------------------
DRAM Frequency          cycle/nsecond         5,92
SM Frequency            cycle/nsecond         1,43
Elapsed Cycles                  cycle      150a295
Memory Throughput                   %        51,41
DRAM Throughput                     %        51,41
Duration                      usecond       104,99
L1/TEX Cache Throughput             %        57,12
L2 Cache Throughput                 %        29,16
SM Active Cycles                cycle   140a224,08
Compute (SM) Throughput             %        26,42
----------------------------------------------


Section: GPU Speed Of Light Roofline Chart
INF   The ratio of peak float (fp32) to double (fp64) performance on this device
is 32:1. The kernel achieved 0% of
this device's fp32 peak performance and 0% of its fp64 peak performance. See
the Kernel Profiling Guide
(https://docs.nvidia.com/nsight-compute/ProfilingGuide/index.html#roofline)
for more details on roofline
analysis.

Section: Compute Workload Analysis
-------------------------------------------
Metric Name          Metric Unit Metric Value
-------------------------------------------
Executed Ipc Active   inst/cycle         1,13
Executed Ipc Elapsed  inst/cycle         1,06
Issue Slots Busy                %        28,24
Issued Ipc Active     inst/cycle         1,13
SM Busy                         %        28,24
-------------------------------------------
```

```
INF   ALU is the highest-utilized pipeline (23.4%) based on active cycles,taking
into account the rates of its
different instructions. It executes integer and logic operations. It is well-utilized,
should not be a
bottleneck.

Section: Memory Workload Analysis
---------------------------------------
Metric Name        Metric Unit Metric Value
---------------------------------------
Memory Throughput Gbyte/second       146,02
Mem Busy                      %        29,16
Max Bandwidth                 %        51,41
L1/TEX Hit Rate               %        60,33
L2 Hit Rate                   %        75,02
Mem Pipes Busy                %        10,85
---------------------------------------

Section: Memory Workload Analysis Tables

------------------------------------------------------------------------------------------



Section: Scheduler Statistics
-----------------------------------------------------
Metric Name                   Metric Unit Metric Value
-----------------------------------------------------
One or More Eligible                    %        28,65
Issued Warp Per Scheduler                         0,29
No Eligible                             %        71,35
Active Warps Per Scheduler           warp         7,91
Eligible Warps Per Scheduler         warp         1,19
-----------------------------------------------------



Section: Warp State Statistics
---------------------------------------------------------------
Metric Name                                Metric Unit Metric Value
---------------------------------------------------------------
```

```
Warp Cycles Per Issued Instruction             cycle        27,60
Warp Cycles Per Executed Instruction           cycle        27,63
Avg. Active Threads Per Warp                                31,96
Avg. Not Predicated Off Threads Per Warp                    27,77
----------------------------------------------------------------------


----------------------------------------------------------------------------------------------------


Section: Instruction Statistics
---------------------------------------------------------------
Metric Name                                  Metric Unit Metric Value
---------------------------------------------------------------
Avg. Executed Instructions Per Scheduler        inst    39a565,08
Executed Instructions                           inst    3a798a248
Avg. Issued Instructions Per Scheduler          inst    39a605,08
Issued Instructions                             inst    3a802a088
---------------------------------------------------------------

Section: Launch Statistics
---------------------------------------------------------------
Metric Name                             Metric Unit    Metric Value
---------------------------------------------------------------
Block Size                                                   1a024
Function Cache Configuration                        CachePreferNone
Grid Size                                                    1a024
Registers Per Thread             register/thread              46
Shared Memory Configuration Size           Kbyte           32,77
Driver Shared Memory Per Block        byte/block               0
Dynamic Shared Memory Per Block       byte/block               0
Static Shared Memory Per Block       Kbyte/block            8,45
Threads                                    thread       1a048a576
Waves Per SM                                                42,67
---------------------------------------------------------------

Section: Occupancy
-----------------------------------------------------
Metric Name                     Metric Unit Metric Value
-----------------------------------------------------
Block Limit SM                        block          16
```

```
Block Limit Registers                 block            1
Block Limit Shared Mem                block            3
Block Limit Warps                     block            1
Theoretical Active Warps per SM        warp           32
Theoretical Occupancy                     %          100
Achieved Occupancy                        %        98,38
Achieved Active Warps Per SM           warp        31,48
----------------------------------------------------------

INF   This kernel's theoretical occupancy is not impacted by any block limit.

Section: Source Counters
-------------------------------------------------
Metric Name               Metric Unit Metric Value
-------------------------------------------------
Branch Instructions Ratio           %         0,08
Branch Instructions               inst      295a912
Branch Efficiency                   %        99,49
Avg. Divergent Branches                      10,42
-------------------------------------------------
```

6 Выводы

Разделяемая память является очень нужным и сложным в освоении инструментом CUDA. Наибольшие проблемы с ней заключаются в оптимизации обращений и избегании конфликтов при обращении к банкам памяти. Однако, поскольку разделяемая память - вторая по скорости после регистров, невозможно избежать её использования, ведь при правильном использовании она даёт кратную выгоду по времени относительно простого обращения к глобальной памяти.

Сложность алгоритма $O(w * h)$.

Список литературы

[1] *CUDA C++ Programming Guide*
URL: https://docs.nvidia.com/cuda/cuda-c-programming-guide/index.html#

[2] *Manual for the NVIDIA Nsight Compute Command Line Interface*
URL: https://docs.nvidia.com/nsight-compute/NsightComputeCli/index.html