UPPSALA
UNIVERSITET

Assignment 2:
Programming in CUDA

Accelerator-based Programming

Oskar Tegby

October 2022

1 Introduction

This assignment studies the computational properties of `CUDA` implementations of matrix-vector and matrix-matrix multiplication. The former is parallelized only over the rows, and the latter over both rows and columns, which is assumed to be faster than only parallelizing over the rows in the latter case. Notably, these operations are completely parallelizable since they are fully data parallel. Thus, we will dedicate this report almost exclusively to reporting and discussing the achieved bandwidth of the different implementations. Lastly, all tests are run on a `Nvidia Tesla T4` graphics card running on the `Snowy` server on the `UPPMAX` cluster.

The veracity of the code is tested by comparing the results for a matrix, A, and a vector, x, with strictly increasing values, since that is fully asymmetric. That is, we tested the code with e.g.

$$A = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \\ 9 & 10 & 11 & 12 \\ 13 & 14 & 15 & 16 \end{bmatrix} \quad \text{and} \quad x = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{bmatrix},$$

as input values, which gives us the output

$$b = \begin{bmatrix} 30 & 70 & 110 & 150 \end{bmatrix}.$$

This indicates that the implementation is indeed correct.

2 Tasks

2.1 Task 1

2.1.1 Part A

Figure 1 shows the throughput of the single-precision matrix-vector multiplication run with N = M ranging from 104 to 9576 obtained by running the code with `block_size` = 128. These results are the averages of 20 tests each run with 20 repeats to decrease the influence of computational noise, which allows us to rely on smaller details in the obtained graphs. The bandwidth is computed as

$$(MN + N + M) \cdot \texttt{sizeof(float)} \cdot 10^{-9}/t_{\text{best}}.$$

2.1.2 Part B

Figure 1 shows a linear increase in bandwidth with the tensor sizes. The increase in bandwidth is almost exactly the double from $M = N \approx 5000$ to $M = N \approx 10000$. Namely, we go from about 10 GB/s to about 20 GB/s. This indicates that we are not compute-bound, as expected for such small tensor sizes.

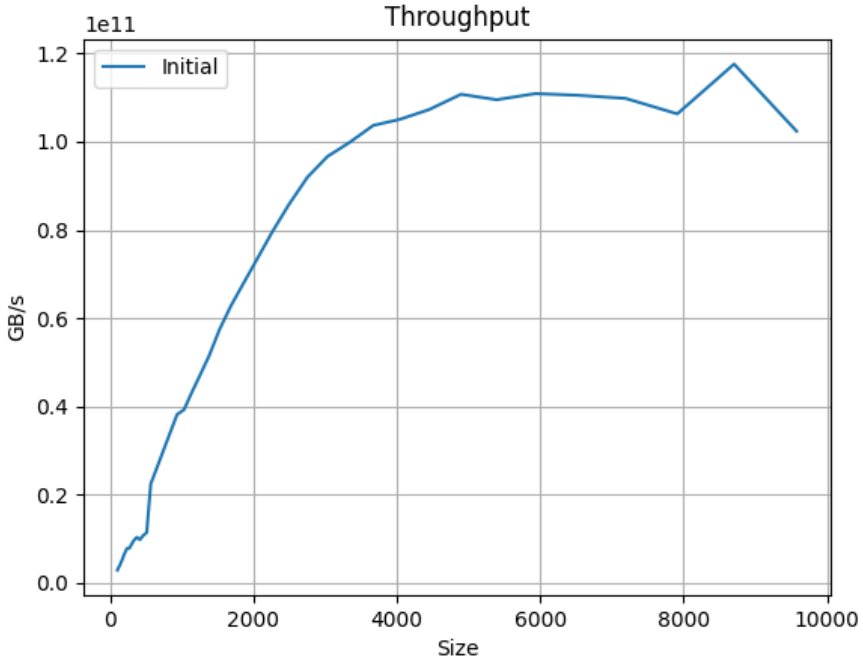

Figure 1: Matrix-vector multiplication with $M = N$ ranging from 100 to 10000.

2.2 Task 2

2.3 Task 3

2.4 Task 4