GPUBenchmark results for tesla2

May 14, 2012

Abstract

This report shows the GPUBenchmark results obtained on tesla2 on May 14, 2012.

Contents

1 Introduction

GPUBenchmark has been used to evaluate different aspects of the tesla2 computer. Depending on its hardware architecture and the libraries available when GPUBenchmark was run, some or all of the following aspects will be reported in this document:

- Transfer speed between hard disk and main memory.
- Transfer speed between GPU and main memory.
- Transfer speed between two GPUs.
- Matrix-matrix multiplication performance.
- Matrix-vector multiplication performance.

The next section describes the hardware characteristics of tesla2. Each one of the remainder sections will focus in one of the previously enumerated performance aspects.

2 Hardware description

This section shows the characteristics of the CPUs and GPU of tesla2. The CPUs available at tesla2 have the next characteristics:

```
Intel(R) Xeon(R) CPU E5440 @ 2.83GHz
 cpu MHz    : 2833.538
 cache size : 6144 KB
 cpu cores  : 4
 bogomips   : 5667.07

Intel(R) Xeon(R) CPU E5440 @ 2.83GHz
 cpu MHz    : 2833.538
 cache size : 6144 KB
 cpu cores  : 4
 bogomips   : 5666.10

Intel(R) Xeon(R) CPU E5440 @ 2.83GHz
 cpu MHz    : 2833.538
 cache size : 6144 KB
 cpu cores  : 4
 bogomips   : 5666.08

Intel(R) Xeon(R) CPU E5440 @ 2.83GHz
 cpu MHz    : 2833.538
 cache size : 6144 KB
 cpu cores  : 4
 bogomips   : 5666.08

Intel(R) Xeon(R) CPU E5440 @ 2.83GHz
 cpu MHz    : 2833.538
 cache size : 6144 KB
 cpu cores  : 4
 bogomips   : 5666.07

Intel(R) Xeon(R) CPU E5440 @ 2.83GHz
 cpu MHz    : 2833.538
 cache size : 6144 KB
 cpu cores  : 4
 bogomips   : 5666.09

Intel(R) Xeon(R) CPU E5440 @ 2.83GHz
 cpu MHz    : 2833.538
 cache size : 6144 KB
 cpu cores  : 4
 bogomips   : 5666.08

Intel(R) Xeon(R) CPU E5440 @ 2.83GHz
 cpu MHz    : 2833.538
 cache size : 6144 KB
```

```
 cpu cores  : 4
 bogomips   : 5666.10
```

The GPU Device installed on tesla2 that has been used to perform some of the tests has the next characteristics:

```
Tesla T20 Processor
 CUDA driver version             : 4000
 CUDA Runtime version            : 4000
 Multiprocessors                 : 14
 Global memory (total)           : 2817982464 bytes
 Constant memory (total)         : 65536 bytes
 Shared memory per block (total) : 49152 bytes
 Available registers per block   : 32768
 Threads per block               : 1024
 Max. dimension of a block       : 1024 x 1024 x 64
 Max. dimension of a grid        : 65535 x 65535 x 65535
 Clock rate                      : 1.15 GHz
```

3 Transfer speed between hard disk and main memory

To obtain the transfer speed from hard disk to main memory, and from main memory to hard disk, several matrices of floats with different number of rows and columns have been created, and the time required to transfer these matrices from hard disk to main memory, and viceversa, has been measured.

In order to obtain a more accurate measure, for each number of rows and columns, ten transfers were carried out in both directions. The median of the results for each case is reported.

Table 1 shows the transfer speed obtained for several numbers of rows and columns, from hard disk to main memory, and from main memory to hard disk. Note that the transfer speed is given in Mebibytes per second (MiB/s)[1]. These results are reported graphically in Figure 1.

Rows	Columns	Size (MiB)	Hard disk → Main memory (MiB/s)	Main memory → Hard disk (MiB/s)
128	128	0.06	244.14	113.22
256	256	0.25	411.18	200.16
512	512	1.00	451.67	274.12
1024	1024	4.00	410.21	271.65
2048	2048	16.00	588.04	275.00
4096	4096	64.00	529.61	235.49
8192	8192	256.00	167.95	272.18
10240	10240	400.00	170.60	268.33

Table 1: Transfer speed from hard disk to main memory, and from main memory to hard disk, for different matrix sizes

[1] A Mebibyte is defined as 2^{20} bytes

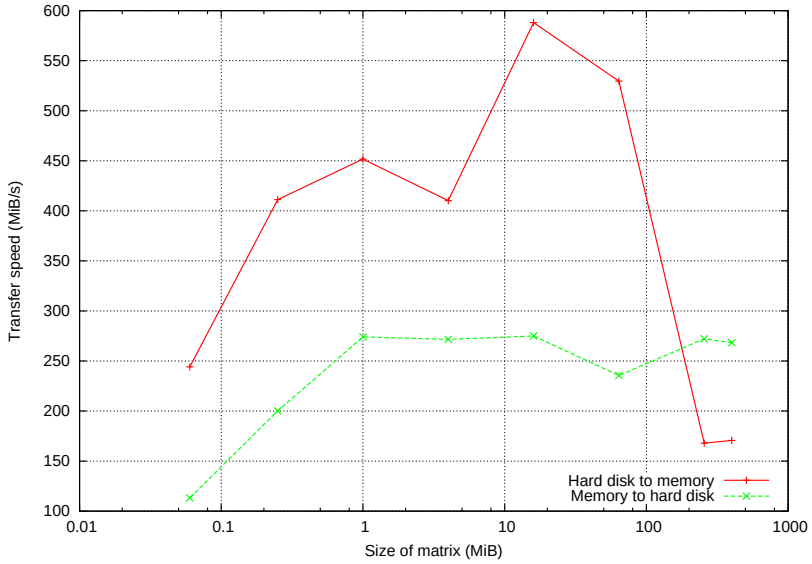


Figure 1: Transfer speed from hard disk to main memory, and from main memory to hard disk, for different matrix sizes

4 Transfer speed between GPU and main memory

To obtain the transfer speed from GPU internal memory to main memory, and from main memory to GPU internal memory, several matrices of floats with different number of rows and columns have been created, and the time required to transfer these matrices from GPU internal memory to main memory, and viceversa, has been measured.

In order to obtain a more accurate measure, for each number of rows and columns, ten transfers were carried out in both directions. The median of the results for each case is reported.

Table 2 shows the transfer speed obtained for several numbers of rows and columns, from GPU to main memory, and from main memory to GPU. Note that the transfer speed is given in Mebibytes per second (MiB/s)[2].

Rows	Columns	Size (MiB)	GPU → Main memory (MiB/s)	Main memory → GPU (MiB/s)
128	128	0.06	735.36	1009.89
256	256	0.25	1151.61	1807.19
512	512	1.00	1329.22	2553.52
1024	1024	4.00	1489.48	3198.49
2048	2048	16.00	1903.83	1894.31
4096	4096	64.00	828.10	2030.35
8192	8192	256.00	837.69	1789.22
10240	10240	400.00	808.38	2070.54

Table 2: Transfer speed from GPU to main memory, and from main memory to GPU, for different matrix sizes

[2] A Mebibyte is defined as 2^{20} bytes

The same tests have been done using padding to allocate the matrices in the GPU internal memory. When padding is applied, it may be necessary to reserve additional storage to ensure that corresponding pointers in any given row will continue to meet the alignment requirements for coalescing. Padding is the recommended allocation method for 2D arrays.

Table 3 shows the transfer speed obtained for several numbers of rows and columns, from GPU to main memory, and from main memory to GPU, when padding is in use. Note that the transfer speed is given in Mebibytes per second.

Rows	Columns	Size (MiB)	GPU → Main memory (MiB/s)	Main memory → GPU (MiB/s)
128	128	0.06	1265.80	1269.91
256	256	0.25	2883.91	2896.74
512	512	1.00	4219.55	4269.71
1024	1024	4.00	4864.38	4851.73
2048	2048	16.00	5966.30	5604.82
4096	4096	64.00	6204.19	5788.54
8192	8192	256.00	6228.63	5837.62
10240	10240	400.00	6230.73	5869.10

Table 3: Transfer speed from GPU to main memory, and from main memory to GPU, when using padding, for different matrix sizes

Figure 2 shows the transfer speed from GPU to main memory, and from main memory to GPU, with and without padding.

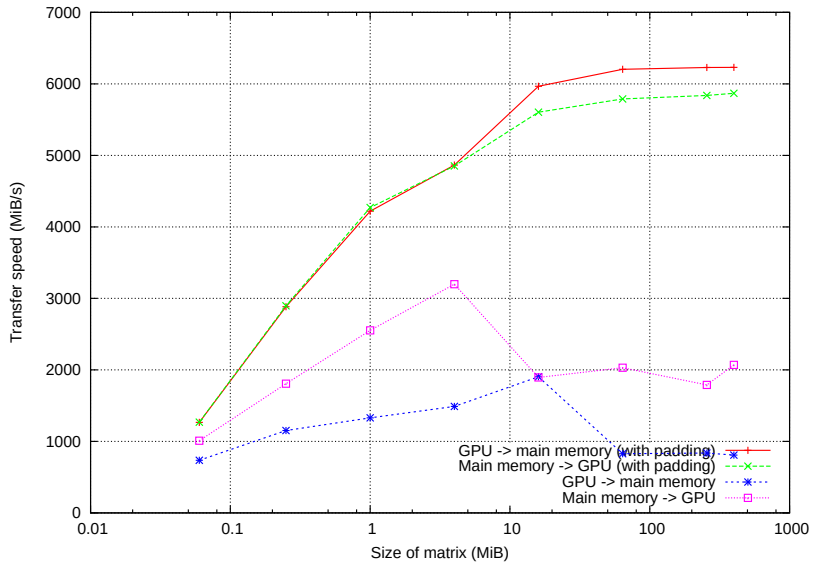


Figure 2: Transfer speed from GPU to main memory, and from main memory to GPU, with and without padding, for different matrix sizes

5 Transfer speed between two GPUs

To obtain the transfer speed between two GPUs using GPUDirect or via main memory, several matrices of floats with different number of rows and columns have been created, and the time required to transfer these matrices from one GPU to another, both using GPUDirect and via main memory, has been measured.

In order to obtain a more accurate measure, for each number of rows and columns, ten transfers were carried out in both directions. The median of the results for each case is reported.

Table 4 shows the transfer speed obtained for several numbers of rows and columns, from one GPU to another both using GPUDirect and via main memory. Note that the transfer speed is given in Mebibytes per second (MiB/s)[3]. These results are reported graphically in Figure 3.

Rows	Columns	Size (MiB)	GPU1 GPU2 (MiB/s)	GPU1 MM GPU2 (MiB/s)
128	128	0.06	690.39	442.48
256	256	0.25	1628.96	680.53
512	512	1.00	2469.77	832.87
1024	1024	4.00	3877.65	875.07
2048	2048	16.00	4559.55	998.41
4096	4096	64.00	4808.27	1026.25
8192	8192	256.00	4872.61	955.65
10240	10240	400.00	4881.14	981.65

Table 4: Transfer speed from one GPU to another using GPUDirect and via main memory

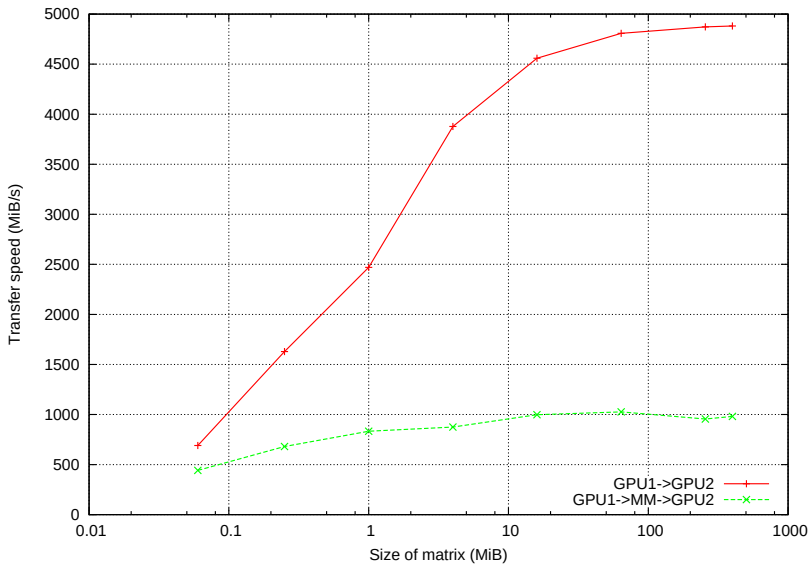


Figure 3: Transfer speed from one GPU to another using GPUDirect and via main memory

[3] A Mebibyte is defined as 2^{20} bytes

6 Matrix-matrix multiplication performance

This section shows both the time required and the GFLOPS obtained by the CPU and the GPU on tesla2 when computing the operation:

$$C = \alpha AB + \beta C,$$

where $A \in \mathcal{R}^{m \times k}$, $B \in \mathcal{R}^{k \times n}$, $C \in \mathcal{R}^{m \times n}$, and α and β are scalars.

This operation has been performed on the CPU by calling the CBLAS `cblas_sgemm()` function, and on the GPU by calling the CUBLAS `cublasSgemm()` function. Different m, n, and k values have been used. For each value of m, the n and k values have been obtained as 25, 50, 75, and 100% of m.

In order to obtain a more accurate measure, for each combination of m, n, and k, ten operations have been carried on in both CPU and GPU. Then, the median of the outcomes for each case has been computed.

The GFLOPS for each combination of m, n, and k has been computed as:

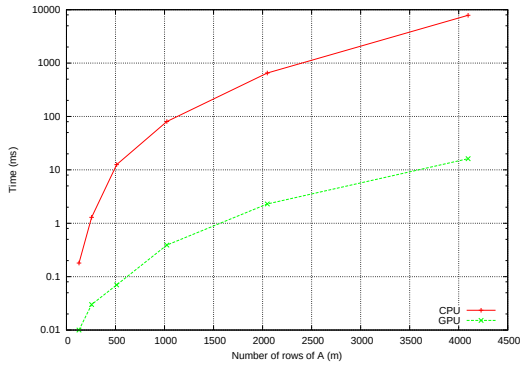
$$GFLOPS = \frac{2mnk \cdot 10^{-9}}{time}$$

Table 5 shows the time and GFLOPS results obtained for the given m, n, and k values.

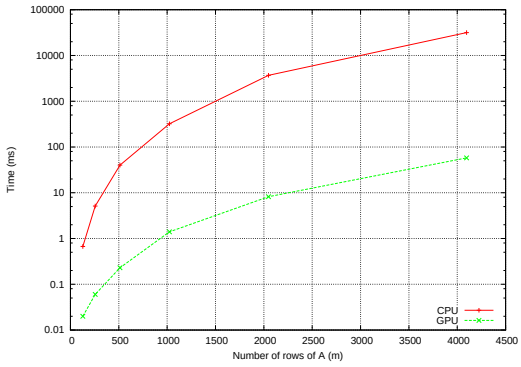
m	n	k	CPU (ms)	GPU (ms)	CPU GFLOPS	GPU GFLOPS
128	32	32	0.18	0.01	1.48	18.12
128	64	64	0.67	0.02	1.57	50.41
128	96	96	1.50	0.03	1.57	94.28
128	128	128	2.67	0.04	1.57	98.48
256	64	64	1.29	0.03	1.63	80.31
256	128	128	5.12	0.06	1.64	147.85
256	192	192	11.56	0.07	1.63	252.28
256	256	256	20.53	0.12	1.63	282.18
512	128	128	12.56	0.07	1.34	253.65
512	256	256	40.16	0.23	1.67	289.10
512	384	384	91.44	0.43	1.65	354.73
512	512	512	162.83	0.74	1.65	361.55
1024	256	256	80.53	0.39	1.67	341.56
1024	512	512	322.21	1.40	1.67	384.51
1024	768	768	726.89	2.44	1.66	494.90
1024	1024	1024	1432.08	4.97	1.50	432.12
2048	512	512	652.88	2.30	1.64	467.01
2048	1024	1024	3678.74	8.21	1.17	523.38
2048	1536	1536	8934.51	15.97	1.08	605.23
2048	2048	2048	15943.65	29.57	1.08	580.92
4096	1024	1024	7851.95	16.20	1.09	530.21
4096	2048	2048	31604.70	57.83	1.09	594.18
4096	3072	3072	71247.08	125.18	1.09	617.56
4096	4096	4096	126675.19	229.81	1.08	598.05

Table 5: Time and GFLOPS for the operation $C = \alpha AB + \beta C$ on the CPU and on the GPU for different m, n and k values

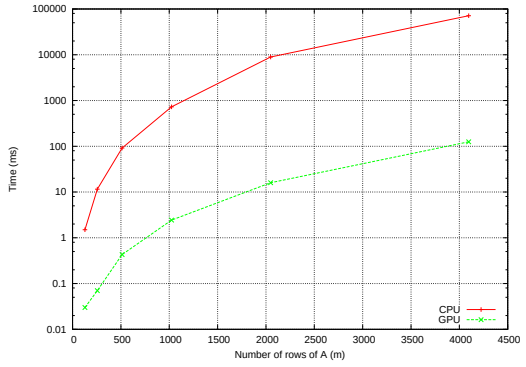
Figure 4 shows the time to compute the operation $C = \alpha AB + \beta C$ on the CPU and on the GPU (notice that n and k are 25, 50, 75, and 100% of m). Figure 5 shows the CPU and GPU GFLOPS for these values of m, n and k.



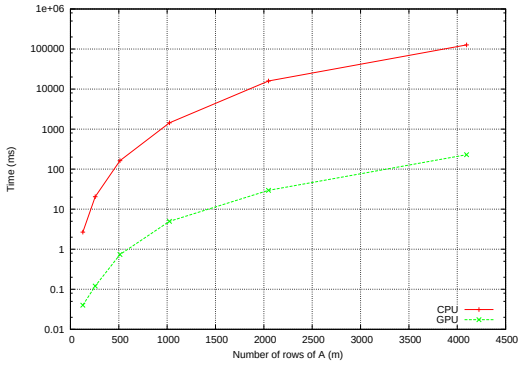
(a) n and k are 25% of m



(b) n and k are 50% of m

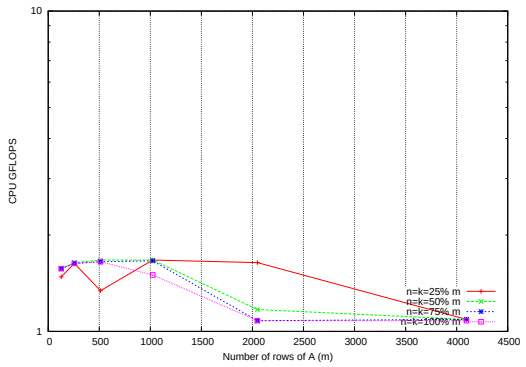


(c) n and k are 75% of m



(d) n and k are 100% of m

Figure 4: Time to compute the operation $C = \alpha AB + \beta C$ on the CPU and on the GPU



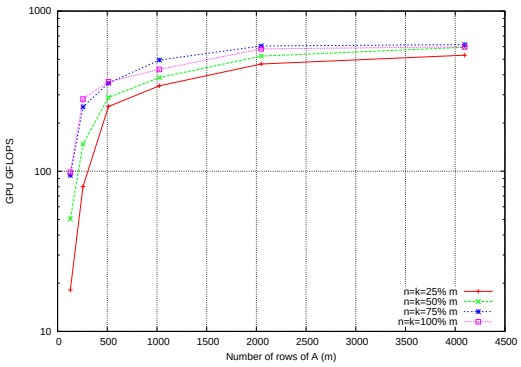


Figure 5: CPU and GPU GFLOPS for $C = \alpha AB + \beta C$ with several m, n and k values (n and k are 25, 50, 75, and 100% of m)

7 Matrix-vector multiplication performance

This section shows both the time required and the GFLOPS obtained by the CPU and the GPU on tesla2 when computing the operation:

$$y = \alpha Ax + \beta y,$$

where $A \in \mathcal{R}^{m \times n}$, $x \in \mathcal{R}^n$, $y \in \mathcal{R}^m$, and α and β are scalars.

This operation has been performed on the CPU by calling the CBLAS `cblas_sgemv()` function, and on the GPU by calling the CUBLAS `cublasSgemv` function. Different m and n values have been used. For each value of m, the n values have been obtained as 25, 50, 75, and 100% of m.

In order to obtain a more accurate measure, for each combination of m and n, ten operations have been carried on in both CPU and GPU. Then, the median of the outcomes for each case has been computed.

The GFLOPS for each combination of m and n has been computed as:

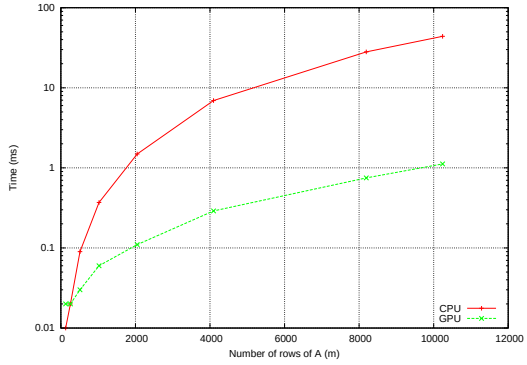
$$GFLOPS = \frac{2mn \cdot 10^{-9}}{time}$$

Table 6 shows the time and GFLOPS results obtained for the given m and n values.

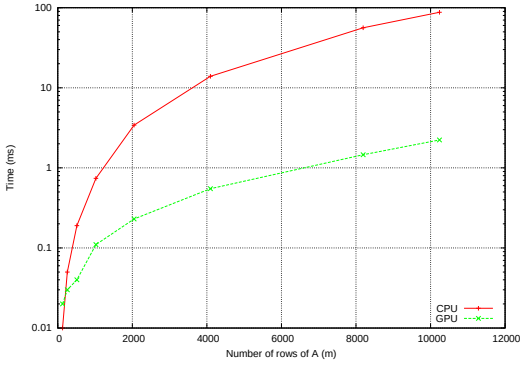
Figure 6 shows the time to compute the operation $y = \alpha Ax + \beta y$ on the CPU and on the GPU (notice that n is 25, 50, 75, and 100% of m). Figure 7 shows the CPU and GPU GFLOPS for these values of m and n.

m	n	CPU (ms)	GPU (ms)	CPU GFLOPS	GPU GFLOPS
128	32	0.01	0.02	1.17	0.53
128	64	0.01	0.02	1.26	0.86
128	96	0.02	0.02	1.29	1.09
128	128	0.02	0.03	1.31	1.27
256	64	0.02	0.02	1.31	1.71
256	128	0.05	0.03	1.37	2.52
256	192	0.07	0.03	1.38	2.99
256	256	0.09	0.04	1.39	3.30
512	128	0.09	0.03	1.41	4.99
512	256	0.19	0.04	1.39	6.52
512	384	0.28	0.07	1.40	5.66
512	512	0.38	0.11	1.40	4.88
1024	256	0.37	0.06	1.41	8.50
1024	512	0.74	0.11	1.42	9.69
1024	768	1.11	0.16	1.41	10.03
1024	1024	1.50	0.21	1.40	10.23
2048	512	1.49	0.11	1.41	18.27
2048	1024	3.42	0.23	1.23	18.09
2048	1536	5.21	0.33	1.21	18.97
2048	2048	6.95	0.43	1.21	19.52
4096	1024	6.93	0.29	1.21	28.55
4096	2048	13.93	0.55	1.20	30.45
4096	3072	20.85	0.81	1.21	31.24
4096	4096	27.87	1.06	1.20	31.58
8192	2048	28.08	0.75	1.19	44.93
8192	4096	56.16	1.46	1.19	45.87
8192	6144	84.37	2.18	1.19	46.14
8192	8192	112.49	2.90	1.19	46.33
10240	2560	43.91	1.12	1.19	46.61
10240	5120	87.73	2.23	1.20	47.11
10240	7680	131.46	3.33	1.20	47.20
10240	10240	175.52	4.44	1.19	47.23

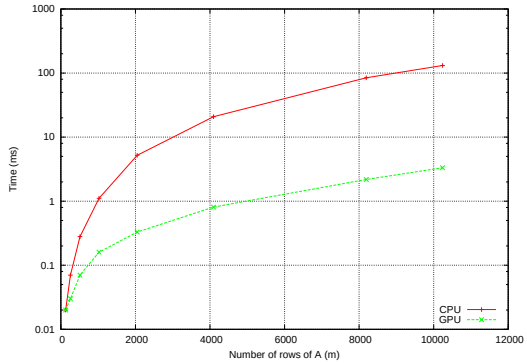
Table 6: Time and GFLOPS for the operation $y = \alpha Ax + \beta y$ on the CPU and on the GPU for different m and n values



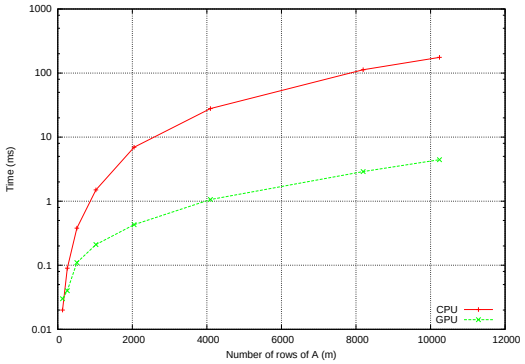
(a) n is 25% of m



(b) n is 50% of m

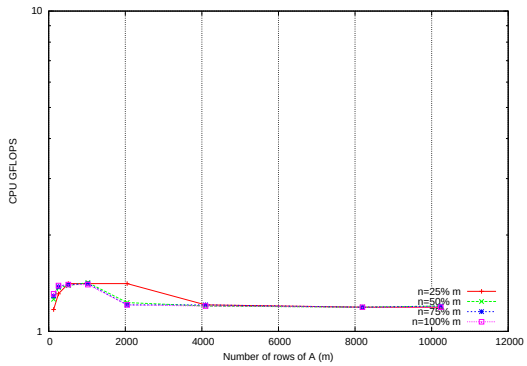


(c) n is 75% of m



(d) n is 100% of m

Figure 6: Time to compute the operation $y = \alpha Ax + \beta y$ on the CPU and on the GPU



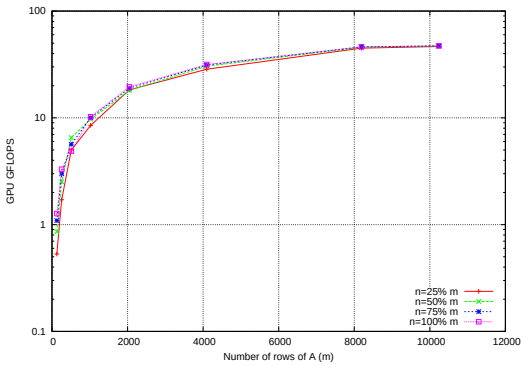


Figure 7: CPU and GPU GFLOPS for $y = \alpha Ax + \beta y$ with several m and n values (n is 25, 50, 75, and 100% of m)