Preliminary report on GPU performance comparison

Massimiliano Zanin, Alex Gimenez, Pablo Moreno, Pere Colet

Instituto de Física Interdisciplinar y Sistemas Complejos IFISC (CSIC-UIB),
Campus UIB, 07122 Palma de Mallorca, Spain

February 2, 2021

1 Introduction

We report the preliminary results of performance tests carried out on NVIDIA's latest generation GPUs installed at a server at SIE headquarters. The base server is a Gigabyte G492-Z50 with two AMD EPYC 7282 processors (16 cores, 2.8GHz) and 256 GB of DDR4-3200 memory. The GPU models tested are:

- RTX 3090. Uses a GA102 graphics processor with 10496 cuda cores, base clock 1.4GHz, boost clock 1.7GHz, 24GB of memory GDDR6X and a memory bandwidth of 936 GB/s. Theoretical peak performance: 35.58 TFLOPS in single precision, 0.56 TFLOPS in double precision (1:64 with respect to single precision).
- RTX A6000 (which for simplicity will call A6000 in what follows). Uses a GA102 graphics processor with 10752 cuda cores, base clock 1.45GHz, boost clock 1.86GHz, 48 GB of memory GDDR6 and a memory bandwidth of 672 GB/s. Theoretical peak performance: 40.0 TFLOPS in single precision, 1.25 TFLOPS in double precision (1:32 with respect to single precision).
- A100. Uses a GA100 graphics processor with 6912 cores, base clock 0.765 GHz, boost clock 1.41GHz, 40GB of memory HBM2e and a memory bandwidth of 1555 GB/s. Theoretical peak performance: 19.49 TFLOPS in single precision, 9.746 TFLOPS in double precision (1:2 with respect to single precision).
- For comparison tests were also run on a TITAN X Pascal available at alcaufar, a personal computer at IFISC wuth an Intel i7-4790K processor and 32GB of memory. The TITAN X uses a GP102 graphics processor with 3584 cuda cores, base clock 1.4GHz, boost clock 1,53 GHz, 12GB of memory GDDR5X and a memory bandwidth of 480 GB/s. Theoretical peak performance: 10.97 TFLOPS in single precision, 0.34 TFLOPS in double precision (1:32 with respect to single precision)

A few remarks are in order:

- The RTX3090, A6000, and A100 belong to the latest generation Ampere architecture released in September-October 2020. The TITAN X is based on the Pascal architecture released in August 2016, which was followed by the Turing architecture (August 2018) and later by Ampere. Therefore the TITAN X is two generations older than the RTX3090, A6000 and A100.
- Regarding market segments, the RTX3090 and the TITAN X are intended for high-end desktop gaming. The A6000 is intended for professional workstations and the A100 is intended for High Performance Computing and does not have graphics ports.
- Regarding memory: The RTX 3090 uses a faster memory than the A6000, the reason being that the memory modules of the A6000 are larger (it has the double of memory) and are not

available at larger frequencies. The A100 uses a different kind of memory, HBM2e and also a very different memory bus, which is much wider (5120 bits while the others use a 384 bit memory bus) so that despite it operates at a lower frequency it provides a larger memory bandwidth.

2 Test 1: classification of time series using TensorFlow / Keras

This first test is aimed at comparing the performance of the different GPUs in a large scale classification problem, specifically in the classification of time series representing electroencephalography (EEG) time series of brain activity. The data set comprises 5.760 time series of 1.000 points each, organised in two groups - the same data have been used in all analyses, to ensure homogeneity of results. The classification has been performed using a standard artificial neural network *resnet* model, using 11 convolutional layers, 128 filters in each one of them, and a *relu* activation function. The optimisation has been performed using the Adam algorithm, monitoring the categorical cross-entropy of the results. Finally, the training has been performed with batch of 64 time series and 200 epocs (results shown below are normalised per epoc). Five configurations are here compared:

- An AMD EPYC 7402 CPU at 2.8 GHz available at IFISC (limited to 1 core).
- A TITAN X.
- All GPUs that were available at the SIE server at the time of the test, which were two A100 and two A6000. The configuration (e.g. how they are interconnected) has been left as per default.
- One single A6000 GPU.
- One single A100 GPU.

The RTX 3090 was not installed at the server at the time in which this test was performed.
The results, in terms of number of training epochs per second, are shown in Fig. 1. Several important points can be appreciated:

- GPUs are clearly faster than the use of a single core of the CPU, providing a performance ≈ 100 time higher. This is, of course, to be expected, as GPUs are designed to handle the kind of problem here considered; and additionally, as the CUDA libraries are specifically optimised for this types of GPUs.
- The GPU available at IFISC is $2-3$ time slower than the proposed ones; again, this is to be expected, given that the TITAN X belongs to a previous generation.
- The A100 is $\approx 32\%$ faster than the A6000.
- Finally, using multiple GPUs at the same time is not beneficial, and the speed corresponds to the one of the fastest GPU. This is probably due to the size of the problem that, while being large, is not enough for benefiting from a larger computational power.

3 Test 2: Classification of *Posidonia oceanica* from satellite imagery using TensorFlow/Keras

In this second test a large data set is built from World-View 2 (WV2) imagery in order to train a neural network to identify *P. oceanica.* The data set contains around 15 M points with information about 8 spectral band each, giving rise to a $(1.5\cdot 10^7, 8)$ feature matrix used for training the model. The model itself is formed by an Artificial Neural Network (ANN) with 4 layers of (N, 500, 100, 1) nodes each, where N varies from 10^3 to $5\cdot 10^5$ in order to test the GPU performance over a given

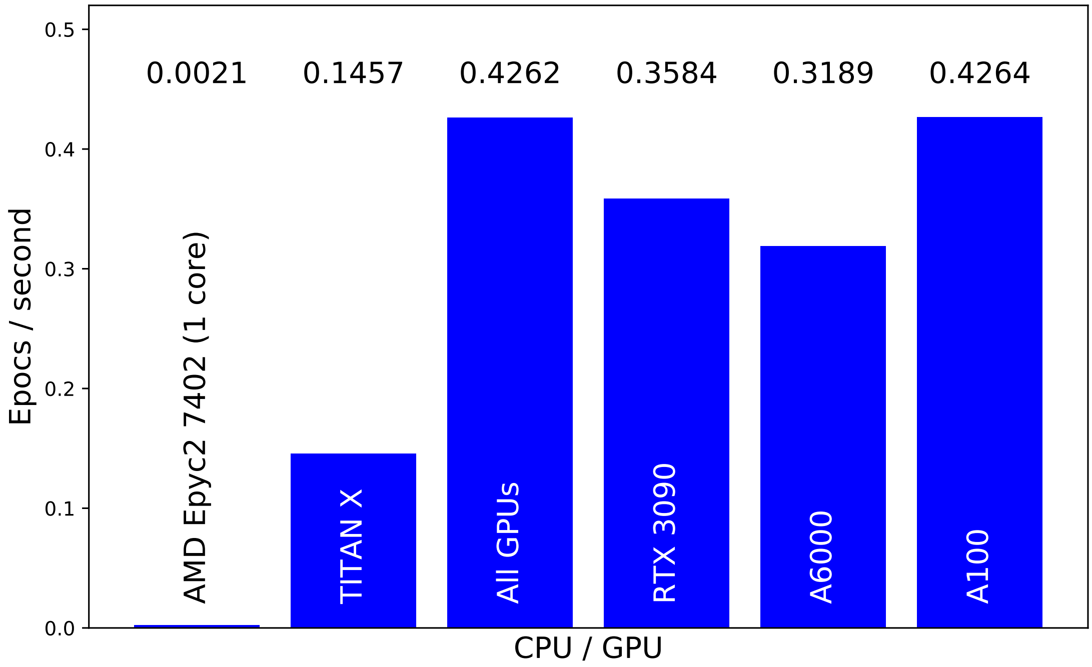

Figure 1: Comparison of the number of epochs in the training per second, across the different configurations.

CPU. We should note that for this specific problem the largest values for N do not provide a better accuracy in the trained model. The size of N was artificially increased in order to appreciate the speed-up provided by the GPUs. The tested configurations are the following:

- The CPU at the SIE. No limitations on core number are imposed. TensorFlow makes use of an increasing number of cores depending on the size of N. For each run we check the number of cores used and the CPU time is reported as the equivalent time on 10 cores, so for instance for a test that takes 45 seconds using 20 cores report the time as 80 seconds.
- TITAN X (only F64 was tested).
- A6000 GPU
- A100 GPU
- RTX 3090

In Fig. 2 we can clearly see the speed up provided by the GPUs for all system sizes. The computation times on the A100 and A6000 are very similar for all N. In Fig. 3 a comparison is provided using the epochs per second metric for the case $N = 10^5$, where we observe that both A100 and A6000 give rise to a 1.6x speed up in comparison to TITAN X and 60x with respect to the CPU. The RTX 3090 was not installed at the time this test was performed.

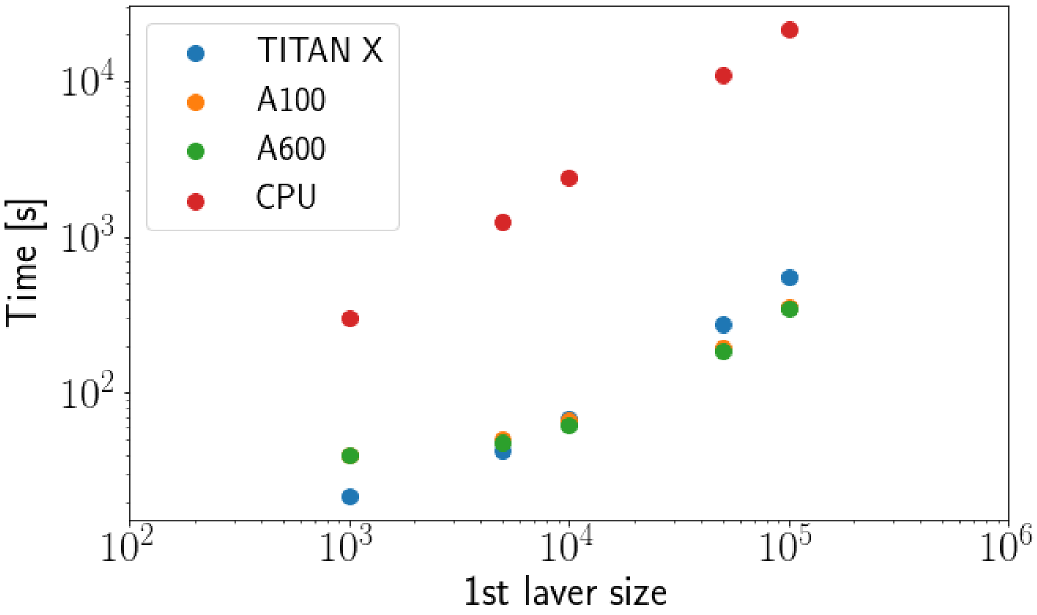

Figure 2: Training time (1 epoch) for different devices and system size.

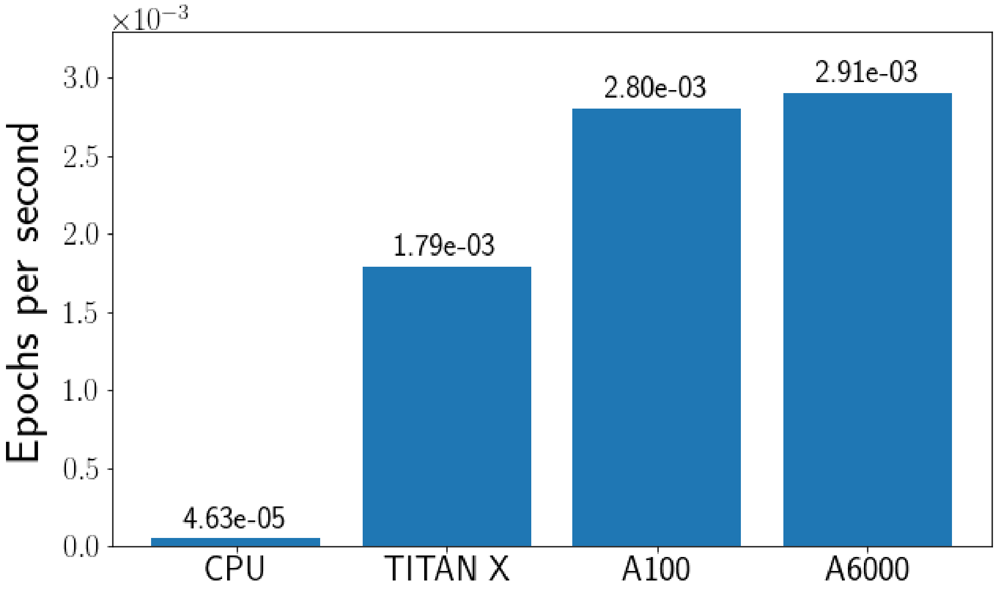

Figure 3: Comparison of the number of epochs in the training per second, when considering $N = 10^5$.

4 Test 3: Integration of a partial differential equation using pseudo-spectral methods

In the third test we have compared the performance of the different GPUs in the solution of 1D partial differential equation. In particular we have compute the evolution of the Complex Ginzburg-Landau equation with complex parameters using a Pseudo-spectral Method. The simulation have been performed for 2000 time-steps with system size from 2^{11} to 2^{20} points using a exponential time differencing fourth-order Runge-Kutta method (ETD RK4). This method do four computations of the non-linear part of the equation, four forward-backward Fourier transformations per time-step. The simulations have been performed using the FourierFlows package, which uses FFTW for the Fourier transformation.
Single precision (F32) and double precision (F64) calculations were tested:

- The CPU at the SIE server. No limitations on core number are imposed. Nevertheless we have check that all runs make use of two 2 cores.
- TITAN X (only F64 was tested).
- A6000 GPU
- A100 GPU
- RTX 3090

The wall-clock time to complete the test for different configurations and different system sizes is given in table 1. The column for "CPU" refers to the wall-clock time for the program running on the server CPU and using 2 cores.

N (log_2)	F32				F64				
	CPU	A100	A6000	RTX3090	CPU	A100	A6000	TitanX	RTX3090
11	1	0.6	0.6	0.5	2.74	0.6	0.6	0.56	0.58
12	1.8	0.6	0.6	0.5	4.6	0.6	0.6	0.56	1
13	3	0.6	0.6	0.5	6.44	0.6	0.6	0.56	2.28
14	6	0.6	0.6	0.54	11	0.6	0.94	0.56	0.98
15	12.5	0.6	0.6	0.64	20	0.6	1	0.67	0.98
16	24	0.6	0.6	0.52	35	0.6	1	1.17	1.047
17	50	0.6	0.6	0.66		0.66	1.44	2.3	1.38
18		0.6	0.6	0.65		0.92	2.93		2.75
19		0.95	1.42	1.28		1.56	6.2		5.8
20		1.5	3.32	2.8		3.4	11.7		10.9

Table 1: Wall-clock time required to complete the test.

First we analyse the case of using single precision, F32. Results for the time to complete the test and for the speed up provided by GPUs as compared to using only the CPU are shown in Figs. 4

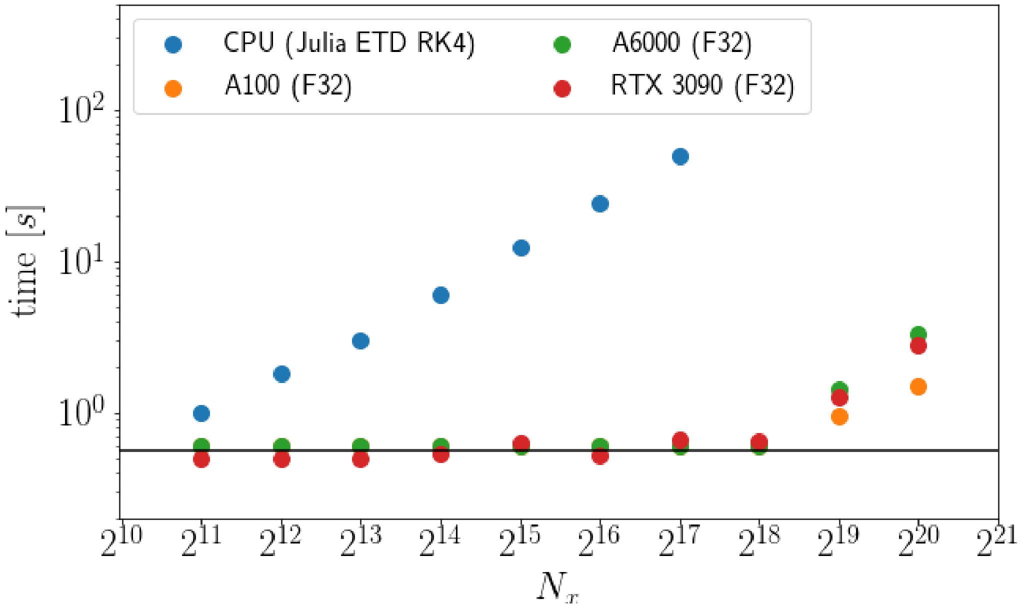

Figure 4: Time required complete the task for F32.

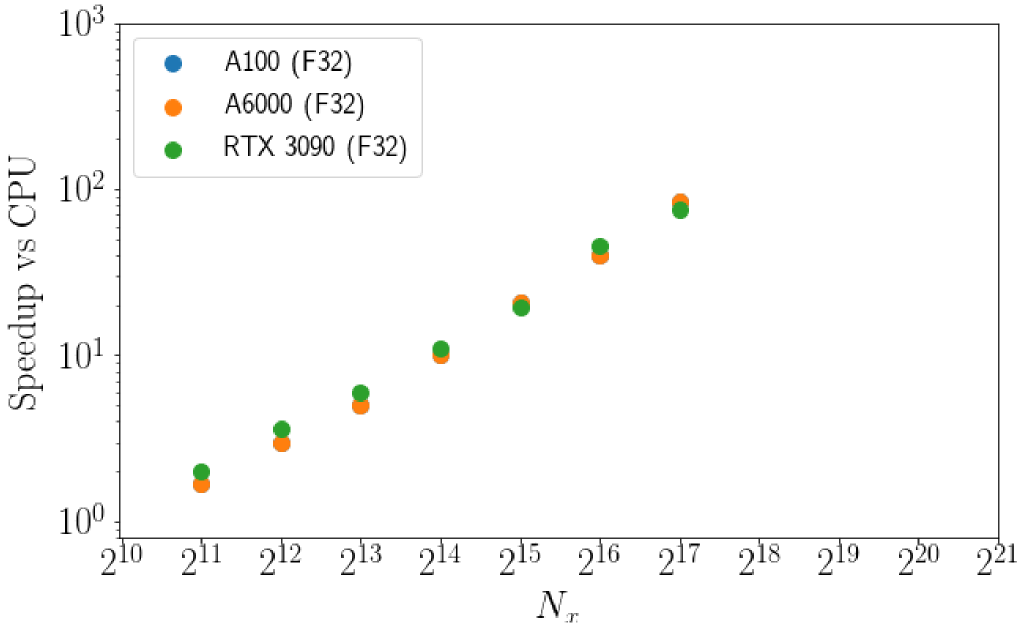

Figure 5: Speedup of the GPU vs CPU using F32

GPUs provides a substantial speed-up over CPU only calculations and the speed-up increases with the system size. For $N = 2^{17}$, the largest system for which test on CPU were performed, the speed up is a factor 100. We should note that while the execution on CPU was not thread limited we have not check how many CPU cores were effectively used by the FourierFlows package during the test.

For all GPUs, the computation time remains practically constant as the system size increases up to $N = 2^{18}$. This may be an indication that systems smaller than that are too small to take full advantage of the parallelism provided by GPUs. For system size $N = 10^{19}$ or larger the time grows with the system size, as expected. For these large systems the RTX 3090 is slightly faster than the A6000 (18%) while the A100 is about twice as fast, providing the minimal computation time. We now consider the case of using double precision, F64. The time to complete the test is shown in Fig. 6 and the GPU speedup in Fig. 7.

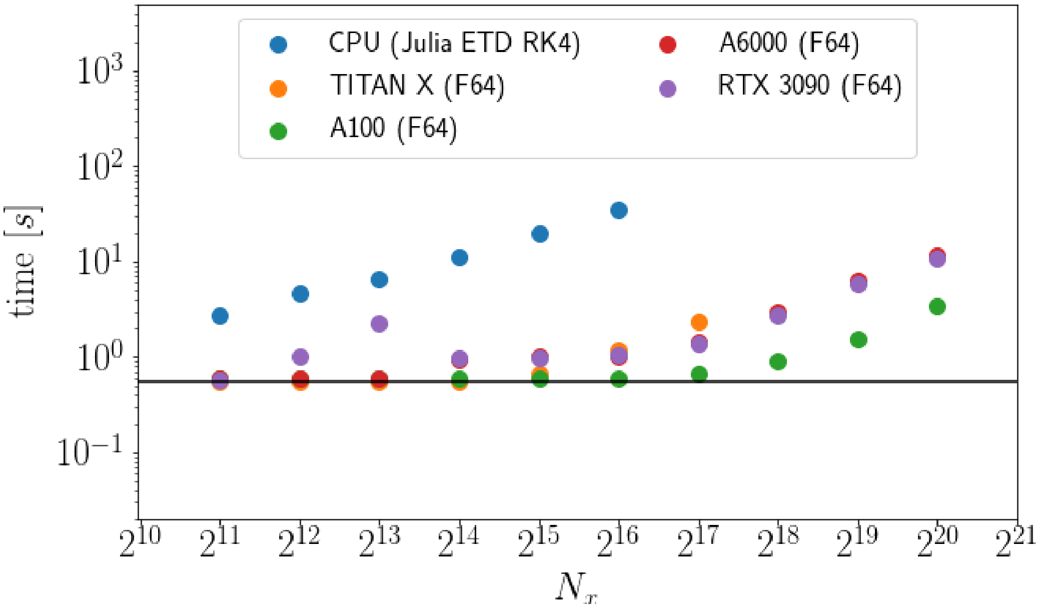

Figure 6: Time required complete the task for F64.

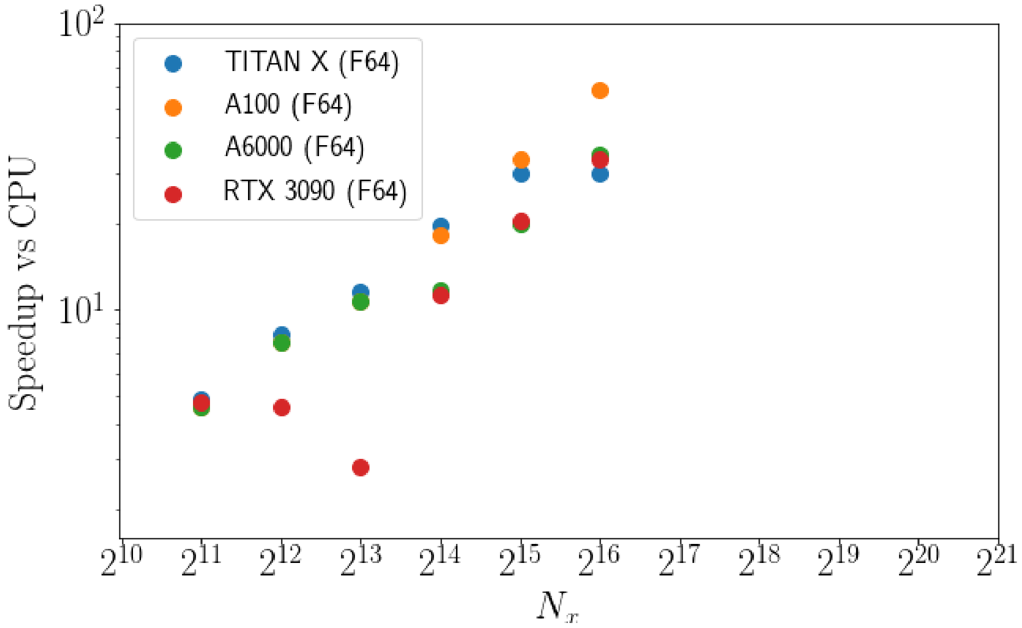

Figure 7: Speedup of the GPU vs CPU using F64

GPUs still provides a speed-up over CPU only calculations but it is not as substantial as for F32. For $N = 2^{16}$, the largest system for which test on CPU were performed, the speed up is a factor 58 for the A100 and 35 for the A6000. The results for the A6000 and RTX 3090 are very similar for system sizes $N = 2^{14}$ or larger, however, for systems of size $N = 2^{12}$ and $N = 2^{13}$ the RTX is significantly slower, in fact these tests take more time than the case $N = 2^{14}$. It is not clear the reason for this.

For large system sizes $N = 2^{19}$, $N = 2^{20}$ the A100 is faster than the A6000 and RTX 3090 by a factor in between 3 and 4. It should be noted that the theoretical peak performance on F64 of the A100 is 16 time bigger than that of A6000. Despite the large difference on theoretical peak performance is much reduced in practice, at least for this test.

Finally Fig.8 shows the amount of memory used by the program as function of the system size N as reported by the command nvidia-smi. Systems larger than $N = 2^{18}$ make use of all memory available on the GPU. It would be noted that the same test for $N * 2^{18}$ running on a standard CPU requires 1.8 GB of RAM memory which is about 20 times less than the memory reserved on the GPU. Thus it may be that more memory is reserved in the GPU in order to expedite calculations. In fact a similar behavior is observed for the programs of test 1. These programs take practically all memory available in the TITAN X (12 GB) while when running on CPU they require much smaller amounts of memory.

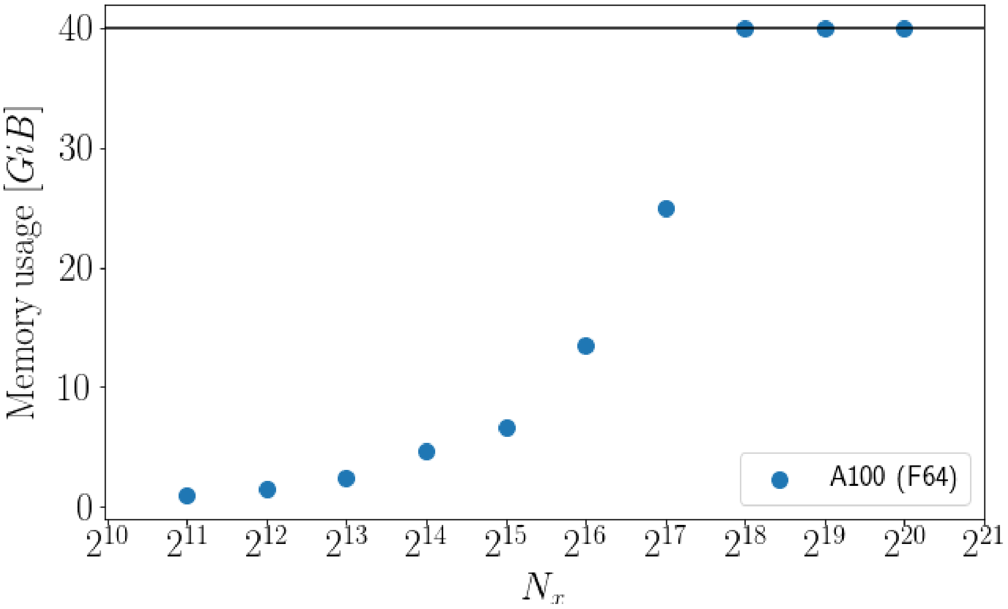

Figure 8: Memory usage of the GPU as function of the system size