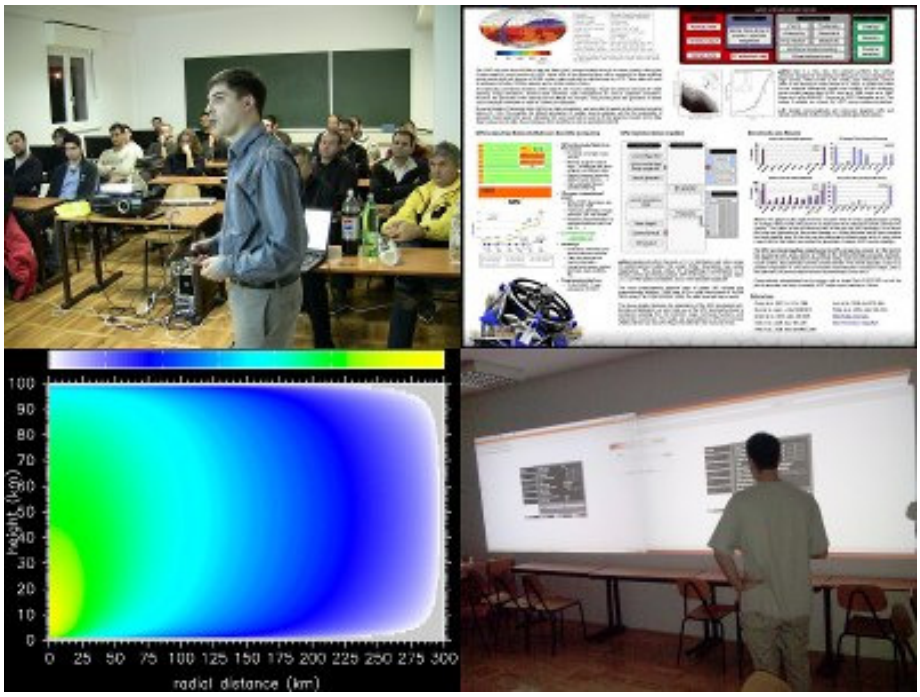
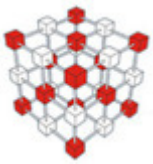
Project:

Constructing, testing, and utilizing a next-generation multi-TFLOP hybrid GPUCPU cluster

Dejan Vinković
Physics Department, Faculty of Natural Sciences and Mathematics,
University of Split, Croatia





2nd report to the National Foundation for Science, Higher Education and Technological Development of the Republic of Croatia (NZZ)

November 30, 2009
Split, Croatia

1. Project activities in the second half-year period

All the key goals of the project have been fulfilled: a GPU-CPU cluster has been assembled and utilized, a computer visualization lecture room has been equipped and set for usage, and the first scientific results are already achieved. However, the road toward these goals has not been without difficulties.

Slight delays experienced during the first half of the project were described in detail in our half-year report and include difficulties in purchasing computer components in Croatia and problems with securing a dedicated room for our computer cluster. During the second half-period we had a delay caused by receiving the second installment of grant funds just before the beginning of the summer vacations at the University. Instead of using summer time for student projects and various activities that can benefit from the vacant university premises, the University closes its doors. This brought our project to a standstill. We had only a month to wrap up the project in September when the administration staff returned. Although we managed to complete the majority of tasks on time, some activities dragged into October. It has to be noted that we again had difficulties in purchasing equipment due to unreliable sellers of computer equipment.

Setting up the visualization lecture room required additional time and was completed in the end of October. Also, this part of the project had a major reallocation of funds approved by the NZZ's steering committee on its 51st meeting on 28. May 2009. This change consisted of replacing one low-resolution computer projector and video wall with two high-definition projectors. The achieved visual characteristics of the image and simplification of the visualization process show that we made the right decision to abandon the video wall option. Delays in setting up the visualization lecture room shifted by one month our plans for organizing open training courses. The first introductory training course on using CUDA on GPUs took place on 3rd of December and it was a success. We already have enough interest from undergraduate and graduate students, research assistants and professors that we have to start planning multiple courses. Some faculty members from the Department of Informatics also expressed their interest in incorporating these lectures into their regular course materials. Details will be arranged once we carry out our initial training courses.

Although the GPU-CPU cluster has entered the full production mode by the end of the project, we experienced some problems that slowed down the work on finalizing the cluster configuration. One of the nodes was producing errors and would freeze on a random base. It took a couple of months until we pinpointed the source of this malfunction, with the help of the node supplier's customer support. Another slow down was caused by the cluster's job management system. During the initial testing phase we

used PBS (Portable Batch Scheduler), but for the final configuration we switched to SGE (Sun Grid Engine). The reason is that SGE provides a better job management. On the other hand, this caused a delay because we had problems with configuring SGE to work with MPICH2 (a portable version of Message Passing Interface – a standard programming interface for utilizing multiple computers directly from a computer code).

Finally, our project partners have used the cluster with various levels of commitment. Some have used it extensively and already adapted parts of their codes to GPU, while others have been working on the first steps toward learning how to use CUDA and run codes on GPUs. Thus, we do not have results form all the listed project partners. However, we had interest from other people more than initially anticipated. We grant access to the cluster to users who provide information on their particular programming goals on GPUs, no matter if they are students, professors or researchers. This provides sustainability for our project and helps us to secure its future growth. An example is a research support sub-award of $14.000 that we received from the Large Synoptic Survey Telescope project (LSST; http://www.lsst.org), under which they co-finance one programmer for a year. LSST will create the largest scientific database ever assembled and it requires many technological advances in data-intensive science and computing. This is the most ambitious ongoing project in the US astronomy.

We also received support from the Ministry of Science, Education and Sport in form of financing one research assistant (PhD student - "znanstveni novak"). The position was awarded to dipl.inž. Jurica Teklić who received its undergraduate degree in computer engineering. The project team now includes doc.dr.sc. Dejan Vinković (the PI), dr.sc. Mario Jurić (the co-PI, Hardvard University), dipl.inž. Jurica Teklić (research assistant), Dubrako Balić (system administrator) and dipl.inž. Krešimir Ćosić (project assistant).

2. Project timetable

Activity	Time (start-end months)	Results
First half-year report		
Hiring of a student novice (PhD student).	Planned: 1. – 2. Realized: 1. – 7.	A novice (PhD student) hired
Collecting offers for the purchase of cluster components. Purchasing the cluster components.	Planned: 1. – 2. Realized: 1. – 6.	Cluster components purchased for the best price available.
Assembling the cluster	Planned: 2. – 5. Realized: 3. – 6.	The cluster in basic operational mode.

Development of CPU-GPU applications for the cluster	5.-12.	The first codes are under development
Second half-year report		
Purchase and installation of a workstation for the project leader, the novice and visitors (project partners)	Planed: 1., 6. Realized: 9. - 10.	Workstations operational
Purchase of the computers the visualization lecture room	Planed: 7. Realized: 10 - 12.	The visualization lecture room equipped with computers
Purchase of equipment for lectures and visualization	Planed: 7. Realized: 10 – 12.	The visualization lecture room operational
Hiring a research assistant (znanstveni novak)	Planed: 1.-2. Realized: 1.- 8	A research assistant hired
Monitoring of the cluster performance	Planed: 6. - 12. Realized: 7.- 12.	"burn-out" test completed, one node repaired, full-production phase of the cluster achieved
Developing instructions and user manuals for cluster usage	Planed: 5. – 12. Realized: 7.- now	Online tools for monitoring cluster status installed and operational, blog and twitter pages for communication with users established, detailed instructions for users under development
Developing applications for the cluster	Planed: 5. – 12. Realized: 7.- now	The first applications written, the first scientific results achieved

By the end of the project we identify the following goals to be partially unfulfilled:

- Visits by project partners:
  - it was planned to start at the very end of the first six months, but inability to prepare working conditions for visitors left us with no choice

but to provide only remote access to our facilities and user support. The first visitors are arranged for the beginning of 2010.

- Usage of the visualization lecture room
  - problems described above created delays that resulted in the computer room being ready more than a month after the end of the project

- Developing manuals for cluster and GPU usage:
  - we had to focus more than planned on purchasing and setting up the equipment, which resulted in delays in writing extended manuals and instructions. However, the growing number of users and collaborating groups is putting pressure on us to complete this task as soon as possible.

3. Examples of research activities on Hybrid

Reallocation of funds from a video wall to high-definition projectors proved to be a good decision. Running both projectors concurrently on one computer enables an easy spread of a visualization application over both projectors, which results in the overall resolution of 3840x1080 pixels. In our lecture room this produced an interactive 3D visualization of more than 5 meters in length (see Figure 1). The room is also equipped with desktop computers that will be used during training courses by course participants and during other periods by students interested in programming on GPUs.

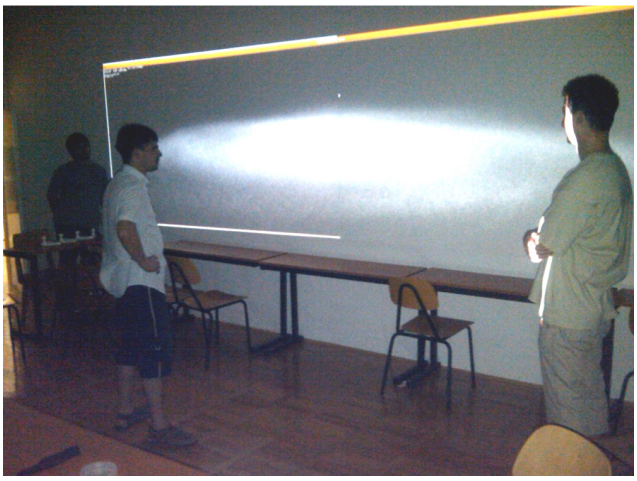


Figure 1: An interactive 3D visualization of a galaxy with 7 million stars (from numerical simulations) projected with our high-definition projectors to produce a composite image of more than 5 meters in length.

Figure 2: Our computer room will used by researchers for visulazation experiments and students who want to learn how to use GPUs

In order to demonstrate the full-production mode of our cluster, I am showing some testimonies from research groups who used the cluster and examples of scientific results. We start with results from two of my research projects.

My first research attempt of using GPUs was to make faster calculations of artificial astronomical images of dusty outflow from protoplanetary disks around young stars. This is a follow up research on the work that I published in Nature, v.459, p.227 (2009). An example of such an image is shown in Figure 3. Interestingly enough, the speed-up was only about 20%. The reason for such a small improvement lies in small data arrays sent to GPUs where the data was analyzed with only a few thousands of parallel threads. This is not enough for a significant speed-up.

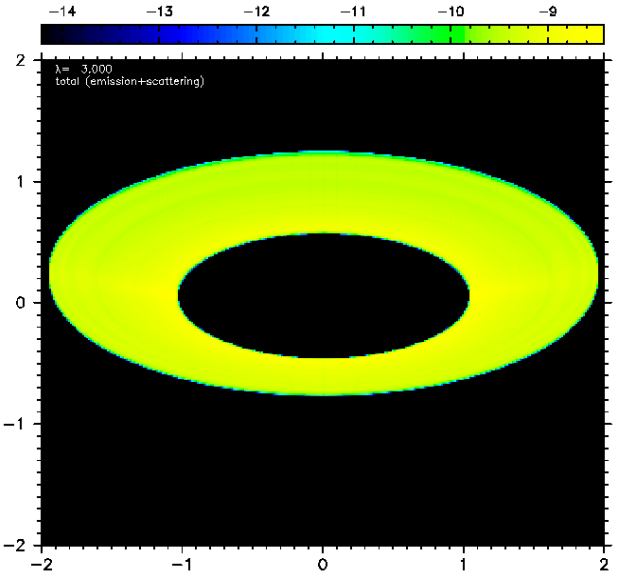


Figure 3: An artificial astronomical image of a dusty disk around a star where GPUs produced only 20% speed-up.

In comparison, my next GPU project is a complete success. The goal was to calculate the electric potential (and electric field) between the ground and ionosphere at 100km altitude for a given distribution of electric charge in the atmosphere. The computational domain was specified in cylindrical coordinates, with 300 radial grid points and 100 vertical (total of 30.000 parallel threads). An iterative numerical method was used, where a value at one grid point is calculated form neighboring points. The solution converges after ~20.000 iterations, which takes less than a second on GPUs, while it is 50 times slower on a CPU. An example of the solution is shown in Figure 4.

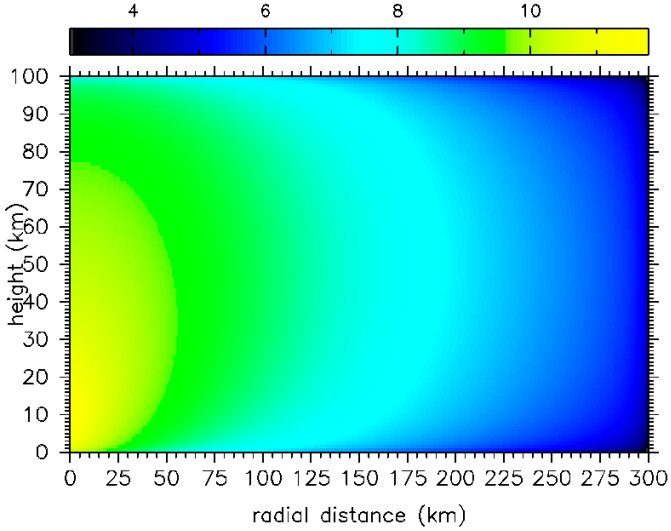


Figure 4: Electrostatic potential in the Earth atmosphere between the ground (0km) and ionosphere (100km) from a charge of 200C at

The influence of the grid size on the GPU performance is nicely demonstrated in an example used by Matija Piškorec from the Faculty of Electrical Engineering at the University of Zagreb. He wrote a user manual in Croatian for programming in CUDA on GPUs and used our cluster for test codes. Figure 5 is taken from his manual and it shows a performance comparison of multiplying a pair of matrices on CPU and GPU. Although the multiplying algorithms used here are non-optimized, the advantage of running on GPU is apparent. While a CPU stalls at matrices of about

300x300 in size, GPUs can handle size of about 3.000x3.000. Notice, however, that GPUs are slower for matrices smaller than about 200x200. CUDA version of the algorithm is taken from "NVIDIA CUDA Programming Guide 2.1" manual and it is described in Croatian in Matija's manual (available at: http://gpuhybrid.org/documents/Seminar_2009_Piskorec.pdf).

The work on Hybrid that led to the research sub-award from LSST started already during the testing phase of our cluster. This initial work focused on speeding-up the code that calculates realistic distribution of starts in our Milky Way galaxy and creates a catalogue of stars that a telescope would see. The work was initiated by dr.Mario Jurić (now at Harvard University), the co-PI on our project. He worked with Krešmir Ćosić, our project assistant. The result is shown in Figure 6. The total runtime was shortened by a factor of 200 thanks to GPUs. This success convinced the LSST management that GPUs could be of a great help in their efforts to produce simulated images for the future LSST telescope. Currently it takes a week to run one LSST simulation on CPUs. In October we signed an official collaboration agreement with the LSST Corporation to work on adapting parts of their code to work on GPUs. These first results on improving the Milky Way simulation code are going to be presented at the 215th American Astronomical Society Meeting in Washington, DC, in January 2010.

The Hybrid cluster has been also used by a group led by Simon F. Portegies Zwart from the University of Amsterdam. They are one of the forefront

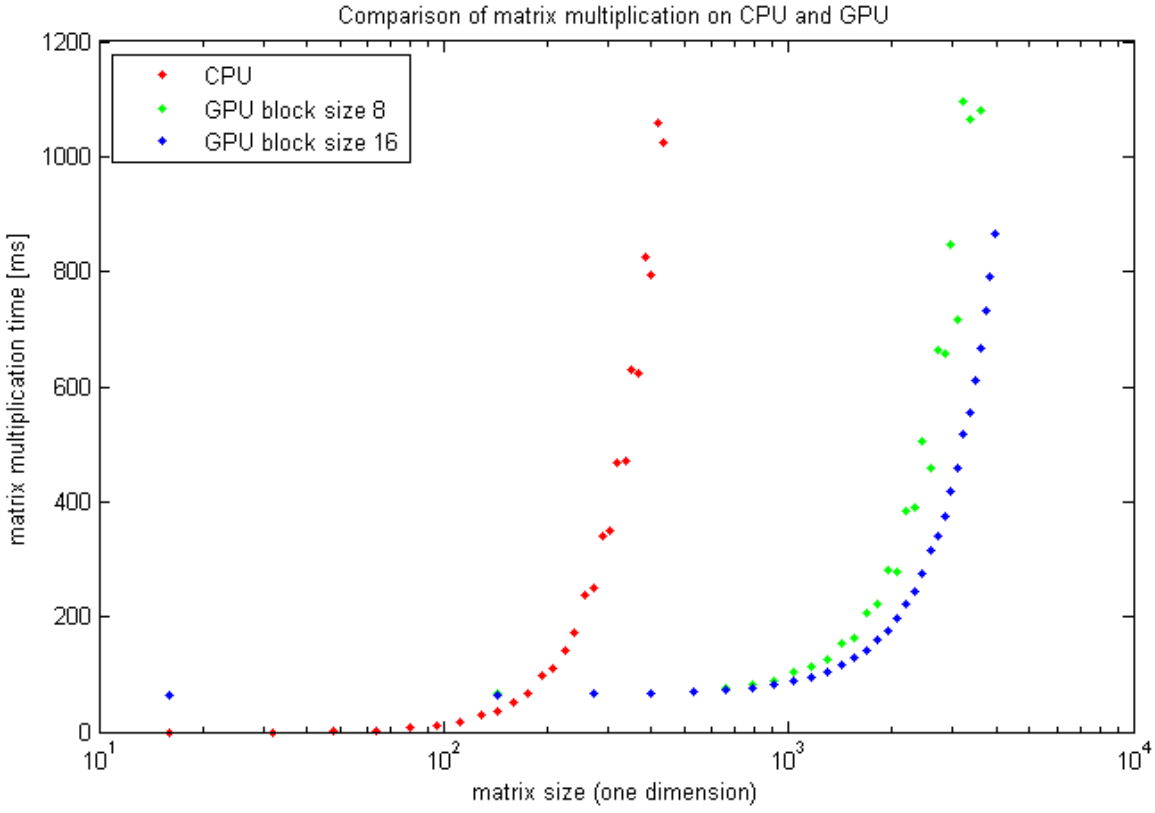


Figure 5: Performance comparison of multiplying a pair of matrices on CPU and GPU.

researchers of general purpose GPU computing. They have developed some of the first large-scale scientific applications using GPUs. Together with his graduate student Jeroen Bédorf, Simon works on direct gravitational N-body simulation of stellar systems on GPUs. The have used Hybrid for scaling tests of their parallel codes, thereby making use of all available nodes and GPUs on the cluster. This gave us the first comparison tests with other GPU-CPU clusters. They reported that the performance of Hybrid seems to be good and comparable to their local machines and other GPU clusters.

Another project that has given us results on an increased code performance on Hybrid, but this time with double precision, is a study of the long term evolution of planetary orbits around a star. Mher Kazandjian from the American University of Beirut, Lebanon, is working on this project in collaboration with Mario Jurić (Harvard) and Jihad Touma (American University of Beirut). In planetary systems, a star is the central object and planets move on elliptic trajectories around it. This applies also to galactic centers, where the dominant object is the super-massive black hole and the stars rotate around it. Thanks to advanced analytic expressions that Kazandjian developed together with S.Tremaine (Princeton University), their code is using 100 to 10.000 times larger time-steps than those in N-body simulations. Their GPU version of the code in double precision has a speedup of about 25 over the serial version. Scaling with the number of GPUs is shown in Figure 7.

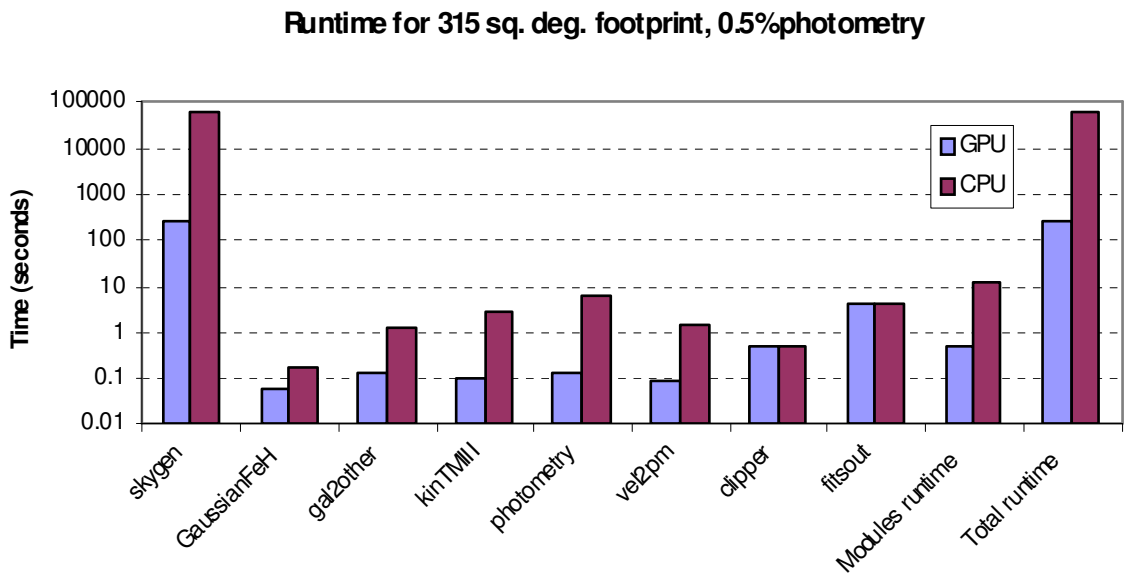


Figure 6: The execution time of various stages of the Milky Way model running on a single Tesla S1070 GPU (blue) vs a single core of an Intel Xeon E5405 2.0GHz CPU (purple). The speed-ups go up to a factor of~200. This work is a part of our

Speed-up is not always achieved when working with GPUs. A large effort was devoted into adapting relativistic magnetohydrodynamic code MRGENESIS. Petar Mimica and Carmen Aloy from the University of Valencia, Spain, have been looking for vectorial operations within the code that might be particularly suited for execution on GPUs. One such module is the Riemann solver used to compute fluxes of conserved hydrodynamic quantities across numerical cell interfaces. They have used the new version of the PGI Fortran compiler and its Fortran extension which enables the programmer to specify the region which will be executed on a GPU. The compiler is quite verbose and aids the programmer in making optimization decisions. Unfortunatelly, this approach has only slowed down the code instead of making it faster. Their continue their work on optimizing this and other modules within the code to run on GPUs.

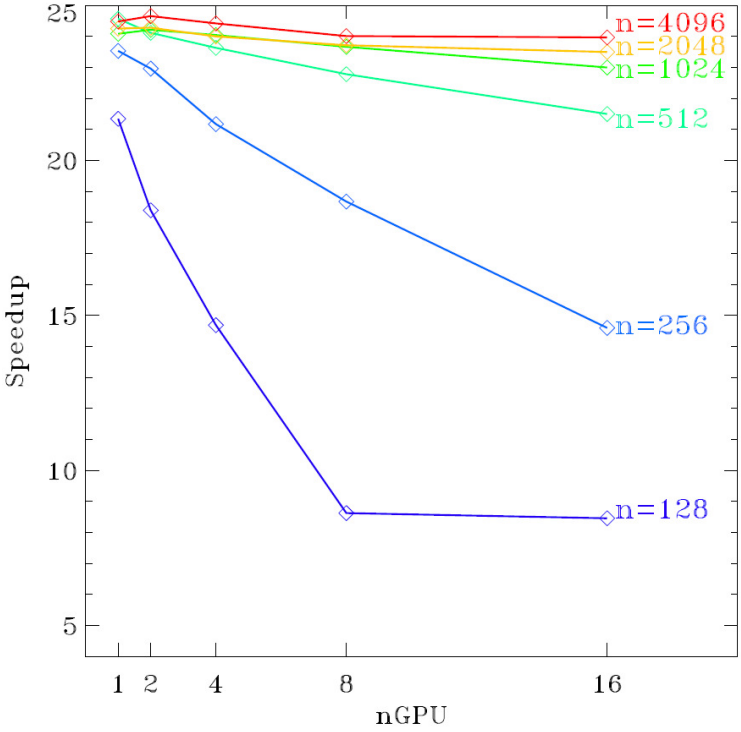


Figure 7: Speed-up in double precision for different number of GPUs of a code that simulates long term evolution of planetary systems.

4. Publications and presentations

In the first half-year report I gave an overview of web-pages describing the project and our GPU-CPU cluster. These pages are now linked together under the domain name http://gpuhybrid.org. Here we show two publications that are

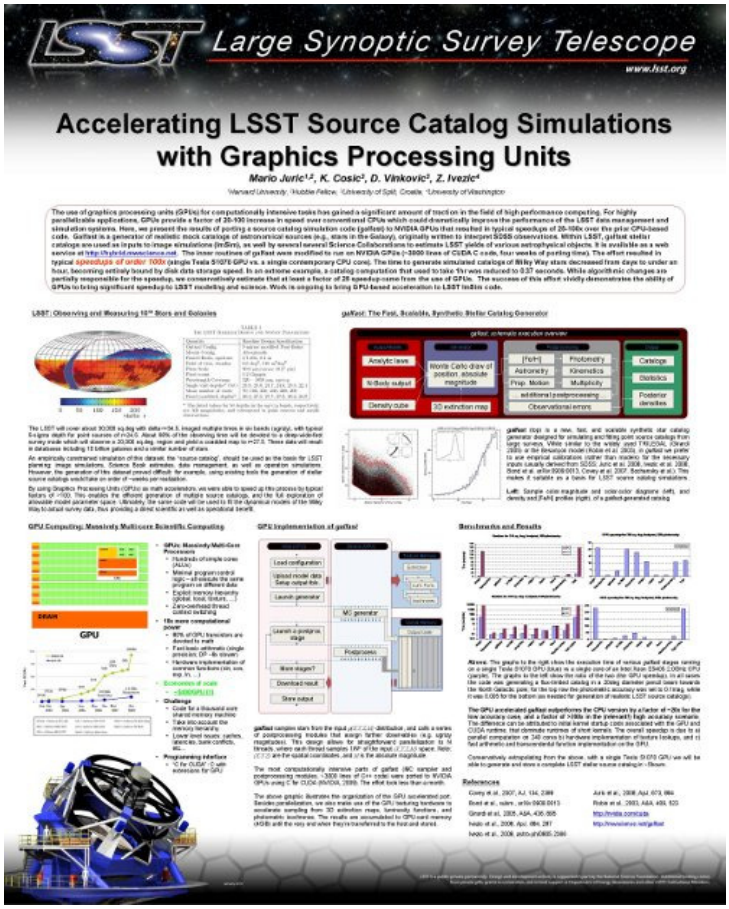


Figure 8: A poster for the American Astronomical Society conference

based on work performed in part on our cluster. The first is a manual on programming in CUDA for GPUs written by Matija Piškorec who used our cluster for testing his codes (Figure 9). The second is a poster (Figure 8) for the American Astronomical Society conference in January (http://aas.org/meetings/aas215has). The authors are M.Jurić (Harvard University), K.Ćosić and D.Vinković (Split) and Ž.Ivezić (University of Washington). It shows a work done in collaboration with the LSST project.

Figure 9: A user manual on programming in CUDA by Matija

5. List of costs

description	0 – 6 months	7 -12 months	total
Computer for the project applicant	9.999,00		9.999,00
Disk storage and backup	8.919,66		8.919,66
The main server	4.559,51		4.559,51
Tesla S1070 system components	57.981,72		57.981,72
Air-conditioning system and installation	17.496,02		17.496,02
36U frame rack	6.771,00		6.771,00
Adaptation and installation of UPS	7.400,00		7.400,00
UPS	33.428,00		33.428,00
Installation of acoustic isolation	8.595,00		8.595,00
Connectors for electric cables	154,49		154,49
Tesla S1070 system components	59.353,00		59.353,00
Tesla S1070 system components	56.012,64		56.012,64
Tesla S1070 system components	62.693,36		62.693,36
Switch, cables	3.384,39		3.384,39
Components for the frontend	24.387,80		24.387,80
Components for the nodes	26.095,80		26.095,80
Monitor for the frontend	850,00		850,00

Small items for the cluster power setup	1.555,57		1.555,57
Components for the nodes	38.548,25		38.548,25
Components for the nodes	29.118,18		29.118,18
Memory for nodes	12.078,00		12.078,00
Monitor, keyboard	4.189,18		4.189,18
Extension cords	54,77		54,77
Disk, graphic card	3.552,03		3.552,03
Graphic card	915,00		915,00
Electric multi-socket with a fuse	466,65		466,65
Monitors for workstations	9.128,53		9.128,53
Color printer	6.441,60		6.441,60
Monitor for a workstation	2.282,13		2.282,13
Workstation (for the PI)		25.437,00	25.437,00
Monitor		2.330,69	2.330,69
Signal splitter		55,28	55,28
Color printer cartridges		5.000,29	5.000,29
2x workstations		21.528,24	21.528,24
2x 1Tb hard disk for workstations		1.512,80	1.512,80
CPU cooler		317,20	317,20
network switch, keyboard, mouse		939,40	939,40
Fans for CPUs		244,68	244,68
2x HD projectors		27.328,00	27.328,00
small video projector		5.856,00	5.856,00
Internet domain		101,68	101,68
plotter, ink, paper rolls		90.408,10	90.408,10
HDMI cables		259,98	259,98
test computer for the room		4.324,53	4.324,53
laser printer		3.803,90	3.803,90
multimedia TV tuner		799,50	799,50
4x computers for the room		16.796,73	16.796,73
4x keyboard		840,58	840,58
multimedia workstation + screen		14.463,12	14.463,12
web cameras for workstations		1.153,15	1.153,15
black ink cartridges for the printer		3.607,34	3.607,34

multimedia camera		9.195,64	9.195,64
network switch		319,80	319,80
tripod for the camera		1.130,14	1.130,14
3x UPS for workstations		1.826,55	1.826,55
PGI compiler for the cluster		3.519,55	3.519,55
potentiometer, printer memory, cooler		1.137,75	1.137,75
headset		109,89	109,89
extension cords		332,29	332,29
office supplies		5.304,72	5.304,72
extension cord		32,00	32,00
TOTAL =	496.411,28	250.246,22	746.657,50