ECONOMIC CONSIDERATIONS REGARDING THE OPPORTUNITY OF OPTIMIZING DATA PROCESSING USING GRAPHICS PROCESSING UNITS

Dana-Mihaela Petroşanu[1]
Alexandru Pîrjan[2]

Abstract

In this paper, we have researched economic considerations regarding the opportunity of optimizing data processing using graphics processing units (GPU) that implement the Compute Unified Device Architecture (CUDA). After analyzing the hardware's costs and a series of experimental tests, the study concludes that using GPU-based configurations for running basic algorithmic functions that optimize data processing, offers an improved performance and a better price/performance ratio than in the case when a CPU-based system is used. This justifies the opportunity of using the CUDA graphics processing units for optimizing data processing.

Keywords: energy efficiency, parallel processing, CUDA, GPGPU, Kepler, algorithmic function.

1. Introduction

After analyzing the parallel processing CUDA (Compute Unified Device Architecture) [1] and solutions for improving data processing using a series of algorithmic functions [2], [3] running on graphic processing units, in this paper we present a study on the economic advantages of using these algorithmic functions in CUDA. Our study takes into account as main factors:

- the costs and benefits resulting from implementing a system that runs on graphics processing units (GPU) from the CUDA architecture;
- the energy consumption;
- the performance per watt consumed.

The GPU's implementation leads to a considerable improvement regarding the energy efficiency in the computers industry, starting with smartphones and up to supercomputers. Today, major information technology (IT) companies are concerned about lowering energy costs and increasing performance. High performance computing systems are special computers or computer networks, being the fastest and most powerful in the world, designed for solving large problems, with high computational complexity. For researchers, scientists and engineers, energy consumption in high performance computing systems requires special attention and it represents a field of great interest. On a small scale, the limited battery life embedded in laptops, tablets or smartphones influences the possibility to access information and the work efficiency for billions of people. More widely, the energy efficiency of computing systems globally affects the entire IT industry.

[1] PhD, Department of Mathematics-Informatics I, University Politehnica of Bucharest, 313, Splaiul Independentei, district 6, code 060042, Bucharest, Romania, e-mail: danap@mathem.pub.ro
[2] Ph D Candidate, Faculty of Computer Science for Business Management, Romanian-American University, 1B, Expozitiei Blvd., district 1, code 012101, Bucharest, Romania, e-mail: alex@pirjan.com

In recent years, graphics processing units have evolved considerably from their initial exclusive graphics applications to various fields like medical imaging, oil exploitation, data prediction, telecommunication control, neuroscience, medical data analysis, image and sound processing or option pricing. In this context, the IT researchers have redesigned and developed the GPUs architectures as to become the most efficient processors on the market. When compared to central processing units, GPUs provide a considerably improvement in terms of energy consumption during the software's execution.
On 24 October 2011, dr. W. J. Bell, professor at Stanford University, said that GPU units are inherently more energy efficient than CPUs as they are optimized specifically for performance per watt and not for absolute performance. The researchers focused mainly on ensuring, through efficient energy usage, an optimum price-performance balance and a remarkable performance per watt. W. J. Bell also noted that most of the limitations in the IT field are related to the energy consumption and not to the memory space. Therefore, researchers have focused, at each design step, on accounting for every dissipated joule of energy and on optimizing the designs as to be more energy efficient. The researchers have improved energy efficiency starting from the level of logic gates (the basic building blocks of chips that convert the "0s" and "1s" of computer language into decisions) and all the way up to the power supply [4].

These researches resulted in the modern graphics processing units. The "Fermi" generation of GPUs uses an average energy of 200 picoJoules (10^{-12} J) to execute one instruction (a single computing task, like adding two numbers). By comparison, the most efficient x86 CPUs require 10x more energy for the same computations. Researchers are concerned to achieve a continuous improvement of the GPU's energy consumption in the next years.

For smartphone, tablet and laptop users, the research on performance per watt will result in increasing the autonomy of communication, in longer-lasting battery life and in a better overall experience. In the high-performance computing (HPC) world, improvements in the performance per watt will save energy, costs and space, facilitating the development of new applications by surpassing the current limits. The supercomputers handle complex computing tasks as seismic imaging, weather forecasting or options pricing in finance. In order to manage such extremely compute-intensive tasks, GPUs use parallel processing to break down these complex problems into many smaller tasks that can be processed simultaneously. Implementing the high computational power of parallel processing, an increasing number of corporations and research institutions have concluded that, in appropriate parallel processing problems, the CPUs can be successfully replaced by GPUs consuming less energy for power and cooling.

In the top of the world's supercomputers[3], a biannual ranking of supercomputer systems around the world, the number of the GPU-powered systems is rapidly growing. Today, three of the five fastest supercomputers in the world are GPU-powered and are systems with optimized energy efficiency. For example, one of the world's fastest supercomputers, China's Tianhe-1A, which uses more than 7,000 NVIDIA Tesla GPUs, consumes about

[3] http://www.top500.org/

half as much power as the CPU-powered Jaguar, number three on the list. Another GPU-powered system, Tsubame 2.0, is the fourth world's fastest supercomputer and the second most energy-efficient supercomputer in the world, according to the latest top of the world's supercomputers. It is located at the Tokyo Institute of Technology and it achieves a performance nearly to that of Jaguar, having the advantages that it uses 92% less servers and consumes only 1/7 of the Jaguar's power. Tsubame is used by scientists to solve varied and complex subjects such as pulmonary airflow or typhoon simulation. GPU-powered supercomputers represent a real standard regarding the energy efficiency. More and more research institutions use GPU-based computer systems and among them the TeraDRE at Purdue University, the Lincoln cluster at the University of Illinois, the Nautilus at the National Institute for Computational Sciences in Oak Ridge, Tennessee and the Keeneland Project at Georgia Tech [5].

2. Achieving Energy Efficiency In Industry Using Gpu-Based Systems

Many companies are striving to reduce the amount of consumed energy by running their basic applications through GPU-based systems that provide an increased performance for each watt consumed. Across a variety of industries, companies are implementing GPU-based systems in order to obtain spectacular gains in energy efficiency for their high-performance computing needs and for processing huge computational data volumes.

In finance, high-performance computing (HPC) systems handle the complex transactions underlying the global markets. In order to reduce costs and save energy, the Bloomberg company replaced, for running an application related to bond pricing, a 2,000 CPU-based server with a 48 GPU rack of Tesla GPUs. The CPU system costs 4 million $ and involves annual energy bills of 1.2 million $. By implementing the GPU-based system, the costs have been significantly reduced, as the graphics processing units have cost under 150,000 $ and the annual energy consumption cost was 30,000 $ [4] (**Table 1**).

Table 1. Economic advantages of the Bloomberg's GPU-based system

Specifications	CPU based system	GPU – CUDA based system
The number of processing units	2.000	48 x Nvidia Tesla
The system's cost	4,000,000 $	< 150,000 $
The annual energy consumption's cost	1,200,000 $	30,000 $

Similarly, the French bank BNP Paribas has replaced a 500 CPU cores system, consuming 25 kW, with one based on two Nvidia Tesla S1070 systems consuming only 2 kW. By using the Tesla GPU processors, BNP Paris has obtained, besides a considerable improvement in performance, an energy consumption of 190 times lower than that of the previous system. By implementing graphics processors based on the CUDA architecture, the volume of computations has increased 100 times for each Watt consumed [4] (**Table 2**).

Table 2. Economic advantages of the BNP Paribas' GPU-based system

Specifications	CPU based system	GPU – CUDA based system
The number of processing units	500	2 x Nvidia Tesla S1070
The energy consumption	25 kW	2 kW
The annual energy consumption	c	$c/190$
The volume of computations per each Watt consumed	v	$100\ v$

In the oil and gas industry, the exploration for new energy resources requires to process sonic images of very large areas. This large volume of collected data requires a huge computational power. One of the major oil and gas firms in the United States, the HESS company, has replaced a 2,000 CPU cluster system with 32 Tesla S1070 servers. The GPU-based system consumes only 47 kilowatts, while the old system consumed 1.34 megawatts. The annual energy bill dropped from 2.3 million $ to only 82,000 $. Today, more than 20 energy firms are in the process of transition to GPU-based processing systems, among them being the Chevron, Schlumberger and BR Petrobras companies [4] (**Table 3**).

Table 3. Economic advantages of the HESS' GPU-based system

Specifications	CPU based system	GPU – CUDA based system
The number of processing units	2.000	32 x Nvidia Tesla S1070
The energy consumption	1.34 MW	47 kW
The annual energy consumption's cost	2,300,000 $	82,000 $

Using GPUs to research energy efficiency has also led to remarkable results in the packaged good industry. The Procter & Gamble company and the researchers at Temple University, Philadelphia have made molecular dynamics simulations to improve their products' quality. In order to increase the energy efficiency, they replaced 32 CPU servers with a single server implementing NVIDIA Tesla C2050 GPUs. The power consumption decreased from 21 kW to 1 kW and the energy costs were cut down from 37,000 $ to just 2,000 $ per year [4] (**Table 4**).

Table 4. Economic advantages of the Procter & Gamble's GPU-based system

Specifications	CPU based system	GPU – CUDA based system
The number of processing units	32	1 x Nvidia Tesla C2050
The energy consumption	21 kW	1 kW
The annual energy consumption's cost	37,000 $	2,000 $

The energy efficiency of GPUs is of major importance in high-performance computing, both for researchers and for the industry as a whole. Whether one analyzes the improvement in supercomputers' performance worldwide, the reduction of energy consumption for high performance computing (HPC) systems or overcoming the limits for mobile computing systems, the CUDA-based graphics processing units lead to the improvement of energy efficiency across the entire computing industry. The increasingly importance of GPUs in the IT field and the improvement of energy efficiency through their use proves that GPUs represent economically efficient computing solutions.

3. Economic Study Regarding The Optimization Of Data Processing In Cuda

In the following, we analyze a number of issues that must be taken into account when analyzing economic issues related to the use of CUDA graphics processors. In analyzing the involved costs, one should consider expenses related to:

- the cost of the system's acquisition;
- the cost of migrating the software component (if application is already developed on a standard CPU architecture);
- the cost of acquisitioning specific software;
- the cost of training the staff;
- the cost of energy for power and cooling;
- the costs of system's maintenance and support.

Adding in the system a powerful GPU that implements CUDA results in increased energy consumption and system's cost of acquisition. However, the costs involved in purchasing such a unit are amortized during its exploitation due to the following main advantages:

- decreasing of the execution time and obtaining therefore the energy efficiency for a given task
- high computational power that reduces the need for future hardware acquisitions.

If the application is already developed on a standard CPU architecture, migration costs for the software component are also involved and these include costs for training, for specific software development and for porting. The main benefit of migration lies in the possibility of using massive parallelism and depends on the main features of the developed application.

The developers who decide to develop their applications using the algorithmic functions [2], [3], running on graphic processing units, benefit from a reduced cost of the software's components migration, as the basic algorithmic functions are designed and optimized in CUDA. Powerful algorithmic functions previously developed and optimized in CUDA represent a viable solution for the applications development in many GPU hardware generations, offering to the developers a powerful tool to implement applications, requiring only minor adjustments to the software component. By using these functions, both the development time and execution time for applications that rely on this solution are considerably reduced. Therefore, the system is able to process a larger volume of information during its life cycle. When evaluating the migration from a CPU-based system to a GPU-based one, the developer should consider the main characteristics of the

developed application; choosing the most favorable configuration in terms of performance and costs; reducing the costs.

In this study we have considered only the costs of the equipment and of the power consumption, as the remaining costs depend on the application in which they are implemented and on the user's configuration. A brief description of the latest three CUDA-enabled graphic cards is presented in **Table 5**.

Table 5. The main characteristics of the graphics cards used

Graphics Card	**GTX 280**	**GTX 480**	**GTX 680**
GPU	GT200 Tesla	GF100 Fermi	GK104 Kepler
Release Date	16.06.2008	26.03.2010	22.03.2012
Fabrication Node (nm)	65	40	28
Number of Transistors	1.4 Billion	3.2 Billion	3.54 Billion
Shader Processors (Cuda Cores)	240	480	1536
Streaming Multiprocessors (SM)	30	15	8
Graphics Clock (MHz)	602	700	1006
Processor Clock (MHz)	1296	1401	-
Boost Clock (MHz)	-	-	1058
Texture Fill Rate (billion/sec)	48.2	42	128.8
Texture Units	80	60	128
ROP Units	32	48	32
Memory Clock (effective MHz)	1107	3700	6000
Standard Memory Config (MB)	1024	1536	2048
Memory Interface Width	512-bit gDDR3	384-bit gDDR5	256-bit gDDR5
Memory Bandwidth (GB/sec)	141.7	177.4	192.2
Max Board Power (TDP	236 Watts	250 Watts	170 Watts

When writing this paper, the latest Nvidia GTX 680 graphics card from the Kepler architecture, was not yet commercially available and, therefore, our study uses the GTX 280 from the Tesla architecture and the GTX 480 from the Fermi architecture.

In the following we present a comparison between the costs of the processors used in running a set of experimental tests for the parallel prefix sum algorithmic function [2], namely the central processing unit i7-2600K and the graphics processing units GTX 280 and GTX 480. For each of the three processing units mentioned above, we have considered the selling price of the online shop Emag[4], available at 01/29/2012. Analyzing the components' prices we have found that the price of the central processing unit i7-2600K is 2.38 times greater than the price of the GTX 280 (653.4 lei more) respectively, 1.32 times greater than the price of the GTX 480 (273.9 lei more) (**Table 6, Figure 1)**.

[4] www.emag.ro

Table 6. The hardware's prices comparison

Specifications	CPU i7-2600K	GPU GTX 280	GPU GTX 480
The hardware's price (lei) at 01/29/2012	1125.3	471.9	851.4
The hardware's prices comparison GPU vs CPU		2.38 x smaller	1.32 x smaller

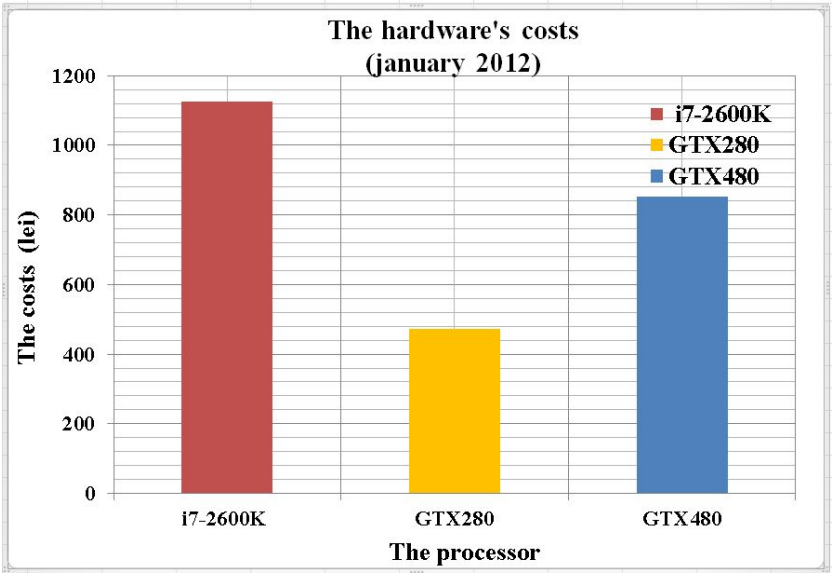

Figure 1. The hardware's prices comparison - CPU vs GPU

In this analysis, it should be mentioned that the central processing unit can run only a specific CPU test suite, without having to add a graphics card, as the CPU has the possibility to use the video core Intel HD Graphics 3000 incorporated in the i7-2600K processor, but this video core cannot run CUDA code.

When the optimized algorithmic function developed in CUDA is running on the GTX 280 and GTX 480 graphics processors, the system requires a central processing unit. This may be less performant (and less expensive) than the i7-2600K CPU. The processing of the algorithmic function is performed mostly by the GPU, the central processing unit having a minimal role in data processing, because the computational load is processed by the CUDA architecture (**Table 7**). For each component, we have considered the selling price of the online shop Emag, available at 01/29/2012.

Table 7. The comparison of the prices and of the total execution times

CPU	CPU's price (lei)	The CPU + GPU price (lei)		The total execution time (h)	
		GPU GTX 280	GPU GTX 480	GPU GTX 280	GPU GTX 480

		(471.9 lei)	(851.4 lei)		
CPU i3-2100	519.99	991.89	1371.39	0.027	0.016
CPU i7-2600k	1,125.3	1,597.2	1976.7	0.025	0.013
CPU i7-3960x	4,709.99	5,181.89	5,561.39	0.024	0.011

When analyzing the total execution time obtained by running a benchmark suite for the parallel prefix sum algorithmic function on different CPU + GPU configurations, we noted very low variations in performance, regardless of the chosen central processing units (we have tested CPU covering the three segments of performance and price: low, medium, high).

The price of the three configurations (consisting of central processing unit and graphics processing unit) covers a wide range of values between 991.89-5,181.89 lei for the GTX 280 and 1,371.39-5561.39 lei for the GTX 480. Choosing a 89% more expensive central processing unit (i7-3960X vs i3-2100) resulted in a performance improvement of only 11.1% for the GTX 280 and 31.25% for the GTX 480. Choosing a 45% more expensive graphics processor (GTX 480 vs GTX 280) caused a 54% performance improvement for the i7-3960X and 41% for the i3-2100. Therefore, investing in a more efficient GPU is completely justified because considerable performance improvements are achieved with much lower costs than those of a more efficient central processing unit.

In the following we present some experimental results highlighting the power consumption, the total execution time, the energy consumption and the running costs for a benchmark suite running the parallel prefix sum algorithmic function on the GPUs GeForce GTX 280 and GeForce GTX 480 (from the Fermi architecture) and on the central processing unit. For each of the three above mentioned cases (when the tests were run on the GTX 280, GTX 480 and on the CPU), we have calculated the consumed energy and then the running costs, taking into account a price of 0.3247 lei/kWh (available at 01/29/2012). When running on the CPU we have used the Intel HD Graphics 3000 video core, incorporated in the i7-2600K processor and there wasn't any graphics card installed in the system.

The CPU used was Intel i7-2600K operating at 4.6 GHz with 8 GB (2x4GB) of 1333 MHz, DDR3 dual channel. We have used the Windows 7 64-bit operating system. In the benchmark suite of 22 tests, the number of input vector's elements ranged between 35 and 8,388,600 and we have used 10,000 iterations for each test. In order to evaluate the total execution time for all of the tests and iterations we have used the same methodology as in the experimental tests of the parallel prefix sum algorithmic function [2].

We have used the CUDA toolkit 4.0, with the NVIDIA driver version 270.81 for programming and access to the GPUs. In addition, all the processes related to the graphical user interface have been disabled to reduce the external traffic to the GPU.

For the GeForce GTX 280 graphics card, the whole system's power consumption in idle was 0.175 kW. When the system ran the benchmark suite of the parallel prefix sum function, the registered consumption was 0.306 kW and the total execution time was 0.025 h. We have recorded the whole system's power consumption in idle, when using the

GeForce GTX 480 graphics card and got a consumption of 0.183 kW. When the system ran the benchmark suite for the parallel prefix sum function, we have recorded a 0.358 kW consumption, while the total execution time was 0.013 h. When the system ran the benchmark suite using the central processing unit (CPU), the recorded consumption was 0.283 kW and the total execution time was 0.196 h (**Table 8**).

Table 8. The energy efficiency of the GPU-based systems vs the CPU-based system

Specifications	CPU i7-2600K	GPU GTX 280	GPU GTX 480
The energy consumption (kW)	0.283	0.306	0.358
The total execution time (h)	0.196	0.025	0.013
The energy consumption (kWh)	0.055	0.008	0.004
The running cost (lei)	0.018	0.002	0.001
The cost of running on the GPU vs on the CPU		9 x lower	18 x lower

Analyzing the obtained experimental results we have observed that, in terms of energy consumption, running algorithms on graphics processors is more effective than running on the central processing unit, for the both GPUs. The best results were obtained for the GeForce GTX 480 graphics card.

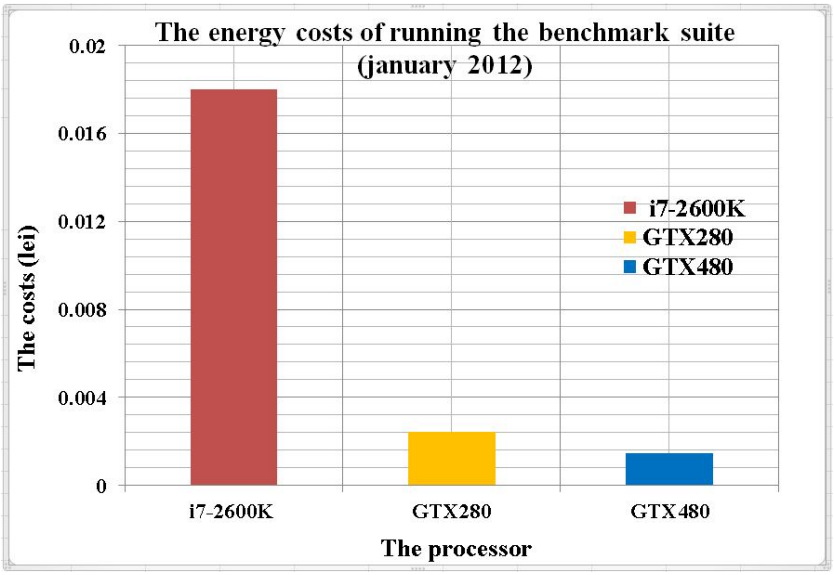

Figure 2. The comparison of costs for running the benchmark suite on CPU and GPUs

Although the total system's consumption is higher when running the benchmark suite of the parallel prefix sum algorithmic function on graphics processors (0.306 kW for GTX 280 and 0.358 kW for GTX 480) than for running on the central processing unit (0.283 kW), the energy consumption, and thus the costs, are considerably lower for the GPU than for the CPU as the total execution time is significantly reduced for the GPUs (0.025 h for the GTX 280 and 0.013 h for the GTX 480) than for the CPU (0.196 h). Consequently, the cost of running the benchmark suite on GPUs is 9 times lower when running on GTX 280

than when running on the CPU and, respectively, 18 times lower when running on GTX 480 than when running on the CPU (**Figure 2**).

We have observed that in a 1-hour time frame, the benchmark suite of the parallel prefix sum algorithmic function runs 5 times on the central processing unit with an energy cost of 0.09 lei, 40 times on the GTX 280 with an energy cost of 0.08 lei and 76 times on the GTX 480 with an energy cost of 0.076 lei.

Considering a period of one year, we have noticed that in this time the benchmark suite of the parallel prefix sum algorithmic function runs 43,800 times on the central processing unit, with an energy cost of 788.4 lei, 350,400 times on the graphics processor GTX 280, with an energy cost of 700.8 lei and 665,760 times on the GTX 480 GPU, with an energy cost of 665.76 lei.

Analyzing the obtained results, we have noted a significant energy saving factor regarding the costs when running the algorithmic function on the GPU than on the CPU. Thus, the cost is 9 times lower when running on GTX 280 than when running on the CPU and, respectively, 18 times lower when running on the GTX 480. By using this function in various applications that require high complexity runs for long periods of time, the execution times and the involved costs are considerably reduced when running on the GPU. Using GPU-based configurations for running the parallel prefix sum (and other algorithmic functions) offers an improved performance and a better price/performance ratio than in the case of CPU-based systems.

Conclusions

In this paper, we have first highlighted a series of economic advantages of using GPU-based systems for improving energy efficiency in the computer industry. We have presented the main fields in which the GPU-based systems have brought a considerable improvement regarding both execution time and energy expenses. Taking into account this analysis and the increasing importance of GPUs in the industry, we have concluded that GPUs represent economically efficient computing solutions.

Then, we have presented an economic study regarding the use of CUDA technology in developing basic algorithmic functions, highlighting a number of issues to be considered when evaluating economic issues related to graphics processors that implement CUDA. We have studied the economic benefits of choosing CUDA when implementing algorithmic functions, taking into account the energy consumption cost and the cost of the hardware equipment, as the remaining costs depend on the implemented applications and on the user's configuration.

We have first compared the hardware's costs (CPU and GPUs), then we have analyzed the performance variations depending on the central processing units that were chosen to cover all the performance and price segments: low, medium, high. Analyzing the performance and costs of various configurations, we have found that the investment in a more efficient GPU is fully justified, since it creates a considerable performance increase, at a much lower cost than investing in a more efficient central processing unit. Thus, by choosing a 89% more expensive CPU (i7-3960X compared to i3-2100), one obtains a

11.1% performance improvement for GTX 280 and 31.25% for GTX 480, while choosing a 45% more expensive GPU (GTX 480 compared to GTX 280), causes a 54% performance improvement for i7-3960X and 41% for i3-2100.

Then, we have run a series of experimental tests in order to highlight the power consumption, the total execution time, the energy consumption and the running cost of a benchmark suite of the parallel prefix sum algorithmic function running on the graphics processing units GeForce GTX 280 and GeForce GTX 480 and on the central processing unit i7-2600K. Analyzing the obtained experimental results we have observed that in terms of energy consumption, running on the GeForce GTX 280 and GeForce GTX 480 graphics processing units is more efficient than running on the central processing unit and the best results were obtained for the GeForce GTX 480 GPU.

Although the tasks are of small complexity compared to the system's capacity, the energy saving costs when running the algorithmic function on the GPU are significantly reduced compared to those on the CPU. Thus, the cost is 9 times lower when running the basic algorithmic functions on the GTX 280 than on the CPU and, respectively, 18 times lower when running on the GTX 480. The execution times and the involved costs are considerably reduced when running the algorithmic function on the GPU, in various applications requiring high complexity runs for long periods of time. Therefore, running the algorithmic functions on CUDA graphics processing units offers both a better price/performance ratio and an improved performance than using only CPU-based systems for running the functions, thus optimizing the efficiency of data processing.

The context of this research is extremely favorable, given the official launch in March 2010 of the Fermi architecture that is aimed specifically on general purpose graphics processing computations. This facilitates tremendous opportunities for developing solutions that optimize data processing. The importance and actuality of our research are highlighted by the launch of the GPU Kepler architecture in March 2012, that offers potential improvements in performance, energy efficiency and processing capabilities for general purpose computations using graphics processing units.

References

1. Sanders J., Kandrot E., CUDA by Example: An Introduction to General-Purpose GPU Programming, Addison-Wesley Professional, New Jersey, 2010.
2. Lungu I., Petrosanu D., Pîrjan A., Solutions for optimizing the data parallel prefix sum algorithm using the Compute Unified Device Architecture, Journal of Information Systems & Operations Management, pg. 465-477, Vol. 5 No. 2.1 /December 2011.
3. Pîrjan A, Optimization Techniques for Data Sorting Algorithms, Annals of DAAAM for 2011 & Proceedings of the 22nd International DAAAM Symposium, Editor B. Katalinic, DAAAM International, Vienna, 2011.
4. http://www.nvidia.com/object/gcr-energy-efficiency.html
5. Hwu W. W., GPU Computing Gems Jade Edition, Morgan Kaufmann, 2011.