USING OF OPPORTUNITIES OF GRAPHIC PROCESSORS FOR ACCELERATION OF SCIENTIFIC AND TECHNICAL CALCULATIONS

V.A.Dudnik,* V.I.Kudrjavtsev, T.M.Sereda, S.A.Us, M.V.Shestakov

National Science Center "Kharkov Institute of Physics and Technology", 61108, Kharkov, Ukraine

(Received October 10, 2008)

The new opportunities of modern graphic processors (GPU) for acceleration of the scientific and technical calculations with the help of paralleling of a calculating task between the central processor and GPU are described. The description of using the technology NVIDIA CUDA for connection of parallel computing opportunities of GPU within the programme of the some intensive mathematical tasks is resulted. The examples of comparison of parameters of productivity in the process of these tasks' calculation without application of GPU and with use of opportunities NVIDIA CUDA for graphic processor GeForce 8800 are resulted.

PACS: 89.80.+h, 89.70.+c, 01.10.Hx

1. INTRODUCTION

Using of opportunities of modern graphic processors (GPU - Graphical Processor Unit) for acceleration of scientific and technical calculations is the new perspective tendency of development of computing systems. In this connection research of the opportunities of existing graphic processors for programming the some intensive mathematical tasks is represented. The purpose of the given work is research of the efficiency of application of the graphic processor (in particular, processor GeForce 8800 of the firm NVIDIA) for scientific and technical calculations.

2. ORIGIN OF GRAPHIC PROCESSORS

Graphic processors have appeared as a result of evolutionary development of graphic adapters (or videoadapters) which always were an integral part of a personal computer. Originally the videoadapter served only for displaying to the monitor of the raster image generated in a computer. In the process of complication of the tasks of formation of images more parts of the most labour-consuming work was transferred to the videoadapter which has gradually turned to specialized computer's unit GPU (Graphical Processor Unit). At present time it doesn't concede on complexity to the basic processor (CPU - Central Processor Unit). The graphic processor consists of the several computing conveyors consisting from the superscalar ALU (Arithmetic and Logic Unit - ALU, which is the basic unit carrying out the basic operations with the graphic objects), working in parallel.Superscalar processors can give out on performance in each step a variable number of commands. The work of their conveyors can be planned statically by means of the compiler as well as by means of hardware of dynamic optimization.

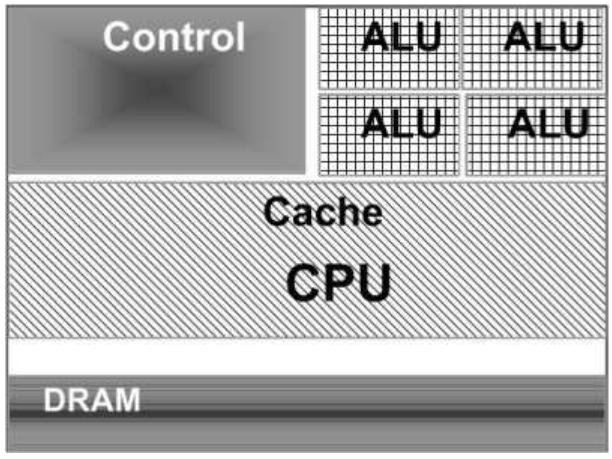

Fig.1. *CPU structure*

These processors use parallelness at a level of commands by sending of the several commands from a usual stream of commands in some functional devices. In addition, such processors use the mechanisms of the extraordinary delivery and extraordinary ending of commands, forecasting of transitions, caches of target addresses of transitions and conditional (under the assumption) performance of commands for taking off the restrictions of the consecutive performance of commands. Generally, the more conveyors contain the graphic processor the more productive this GPU. Used by the authors the graphic processor GPU G80, which is manufactured by the firm NVIDIA, has 128 unified processors which represent

*Corresponding author. E-mail address: dudnik@mail.ru

superscalar processors of a general purpose for data processing with a floating point. They are collected into 8 big blocks which can work independently from each other.

3. SHADERS

Substantially formation of the image in GPU is done by the means of shaders which are the small programmes allowing to program the graphic accelerator. Usually in practice shader is a short sequence of machine codes of GPU which is described by the developer on a special version of the assembler. Shaders allow to build high qualitative 3-D images on the basis of simple models.

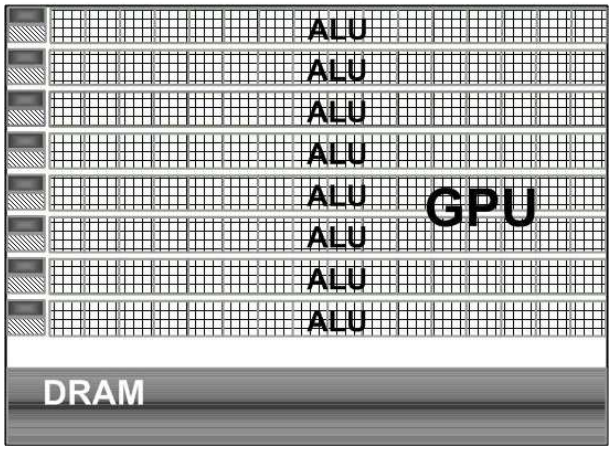

Fig.2. *GPU structure*

Initially the conveyors GPU could carry out only very simple operations in the structure of shaders with low accuracy of calculations. The turning-point has come in the end of 2002 when videomaps GeForce FX from NVIDIA and Radeon 9500 from ATI, created on the basis of GPU G70 and R520 were appeared in sale. The support of the shaders "Shader Model 2.0" was done in their structure. These shaders could contain more difficult programmes both the quantity of instructions and the number of references to data. Besides all intermediate operations could carried out with the numbers of higher accuracy (in comparison with previous GPU). And also there was a support of operations with a floating point. Because of the increased complexity of the shaders it was not convenient to use the assembler. And approximately at the same time C-languages of a high level were appeared. One of them became the C-compiler of shader Cg which was offered by the firm NVIDIA. Microsoft has also included the standard High-Level Shader Language (HLSL) in the structure of the package DirectX 9.0c SDK. With the appearance of "Shader Model 3" with the increase of productivity of videomaps there was an opportunity to carry out on GPU more complex programmes: with cycles, dynamical and conditional transitions, and so on. Accuracy both integer calculations and operations with the floating point has been increased, but finally computing opportunities of GPU were generated only after an output of the fourth version of shaders. After the output of the specifications "Shader Model 4" the graphic processor has received the complete support of standard IEEE 754 (i.e. support of accuracy in operations with the floating point (FP32)). These means have opened the opportunities of using of GPU G80 and R600 for creation of programmes practically of any calculations and any complexity (certainly with some restrictions). Comparison of dynamics of growth of productivity CPU and GPU for operations with a floating point also shows the significant superiority of GPU on this parameter. At present time according to the criterion of cumulative computing capacity the graphic processors advance the most productive processors of a general purpose, and this gap continues to increase. For example, if we compare the peak productivity of one of the last four-nuclear server processors Intel Quad-Core Xeon E5345 (with frequency of 2,66 GHz) - 42 GFlops and productivity GeForce 8800 Ultra - 510 GFlops we can see that GPU is more advanced on this parameter than one of the most powerful new server CPU. The development of GPU is in the continue process. And in the next graphic processor G92 of company NVIDIA the support of double accuracy in operations with a floating point (FP64) will be provided. It will expand the area of effective using of GPU for scientific and technical calculations even more.

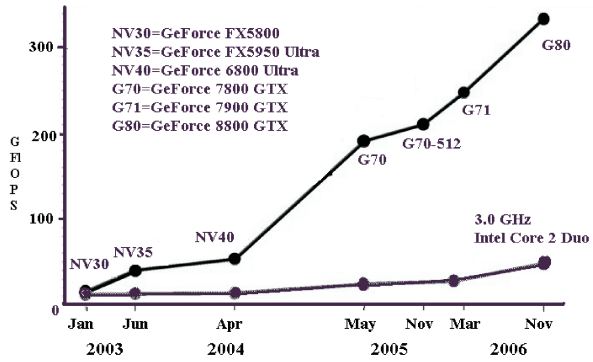

Fig.3. *Compare CPU vs GPU performance*

4.CALCULATIONS OF GENERAL PURPOSE BY MEANS OF GRAPHIC PROCESSORS (GPGPU)

Rapid development and complication of GPU, their growing productivity and simplicity of programming have created the conditions for the active works with using GPU for calculations of a general purpose (GPGPU - General-Purpose Computation on GPUs) since 2003. However, despite of impressing computing capacity, today's GPU can surpass CPU only in a narrow class of tasks because they have not powerful means for a prediction of branching of performance of commands, first of all massive caches for storage giv-

ing in and not giving in to a prediction of the order of processing of instructions. The best results of GPU show for applications in which probably to organize parallel data processing (with the same sequence of the actions applied to the big volume of data), or with the high density of arithmetics (enough big attitude of a number of an arithmetic instructions to the number of the references to the memory). Therefore using of GPU provides the greatest efficiency for enough narrow circle of applications:
"Simulation of physics "Processing signals "The computing mathematics/geometry "Operations with databases "Computing biology "Computing economy "Computer sight
However, for these classes of tasks the graphic processor can provide in more times (up to 10 and more) bigger productivity, than the most powerful and expensive CPU. Though the number of the tasks for which using if GPU can be effective is limited because of a specificity of the architecture of graphic processors, but their number constantly grows. The firms-manufacturers of graphic processors (first of all NVIDIA and ATI) were interested in the new opportunities of use of the graphic accelerators which allowed to expand the market of graphic maps due to attraction of attention to the production of new users - researchers and developers. For example, the firm NVIDIA in 2007, except for videoaccelerators GeForce 8800, in the specialized line of the Tesla products: processor S870 and computers on its base, servers D870 and S870. GPU Tesla C870 is an internal board of expansion on the videoprocessor of series GeForce 8 (it is externally practically identical "usual" 8800, only without sockets for connection of displays). It is installed in socket PCI-E x16 and contains 128 blocks with a floating point, however only single accuracy (64-digit calculations will be supported in following generation of products Tesla). Its peak productivity exceeds 500 GFLOPS. Own operative memory for GPU has capacity of 1,5 Gbytes at throughput of 76,8 Gbytes/sec. and channel PCI-E is applied for communications with GPU x16. Compact 19 "1U" the thin server " Tesla D870 includes two GPU (1 GFLOPS), desktop server Tesla S870 - four GPU (2 GFLOPS).

5. CUDA - HARDWARE-SOFTWARE ARCHITECTURE FOR CALCULATIONS ON GPU

Company NVIDIA (with the purpose of simplification of use of the GPU for calculations of a general purpose) has given the technology CUDA (Compute Unified Device Architecture). It is the environment of development for GPU G80 based on C-language allowing programmers to realize algorithms doing on the graphic processors of accelerators GForce of the eighth generation (G8x) companies NVIDIA. CUDA gives to the developer of applications following opportunities:
" The Standard language of programming C for GPU " The Unified hardware-software decision for parallel calculations on GPU from NVIDIA, supporting CUDA " Wide spectrum CUDA compatible GPU (from economic GPU for notebooks up to high-efficiency systems on the basis of several GPU) " GPU with CUDA support Parallel Data Cache and Thread Execution Manager " Standard libraries of numerical analysis FFT (fast transformation by Furie) and BLAS (base subroutines of linear algebra) " The Special driver for calculations " The Optimized data exchange between CPU and GPU with support CUDA " Joint jobs with graphic drivers OpenGL and DirectX " Support of operational systems Linux bit-bit and Windows XP 32/64 bit " Direct access to the driver and an opportunity of development at a level of the assembler for creation of modern languages and environments of development.
CUDA enables to the developer the opportunity to organize at own discretion access to a set of instructions of the graphic accelerator and to operate with its memory, to organize on it complex parallel calculations. The graphic accelerator with support CUDA becomes the powerful programmed open architecture similarly to today's central processors. All these put at disposal of the developer the high-leveled, operated and high-speedy access to the equipment which make CUDA an effective basis for construction of serious applications.

6. RESEARCH OF EFFICIENCY OF USE GPU FOR SOME TYPES OF SCIENTIFIC AND TECHNICAL CALCULATIONS

The object of the research was the computer system generated in the server center on the basis of platform ASUS P5WD2 with processor Intel 3.0 GHz and GPU GForce 8800, working under the control of operational systems (OS) Windows Server 2003 EE in a mode of a terminal server. Comparisons of execution time of the test examples were realized by means of CUDA for the graphic processor and for usual CPU which results were shown in the table.

Tests samples execution times

Test name	CPU	GPU	Compare
Monte Carlo Black Scholes	t=450 ms	t=5 ms	90
Monte Carlo (106 samples)	658 sec	33 sec	19.9
Fast Walsh	t=3213,66 ms	t=70,06 ms	45
Transform1	v=0, 049 Gfl	v=8,62 Gflps	176
Eigenvalues	t1=446,29 ms	t1=17,76 ms	25
	t2=317,87 ms	t2=5,5 ms	57.8

It is necessary to understand that with the help of use GPU as calculators their using for display of images is not usefull. Differently the maximal operating time of computing programs in GPU will be limited approximately by 5 seconds by virtue of feature of their use by their operational system Windows. For a conclusion of the generated images it is necessary to use the separate videoadapter installed on the same computer to which used monitors should to be connected. Videoadapters used for calculations should be supported in operational system Windows (i.e. for them should be installed all necessary drivers, but they should not be specified as the device of display of "desktop" Windows. Besides for some motherboards the slots in which these adapters are installed, should not be specified as primary at loading system. Otherwise it led to fatal failures at loading Windows.

7.SUMMARY

As a result of performed work the measurements of execution time of examples of typical calculations on CPU and GPU are done. The comparison of these parameters is realized. The opportunities of using of graphic processors for programming tasks of scientific and technical calculations are described. The results of measurement of execution time of typical calculations on CPU and GPU are presented and the comparison of these parameters is done, which allows recommending of the use GPU for acceleration of performance of the tasks demanding the big volume of mini-structured intensive calculations. The analysis of parameters of productivity shows the essential acceleration for the calculations by means of GPU that allows to recommend their using for programming the tasks demanding the big volume of the mini-structured intensive calculations.

References

1. A.Zubinsky. NVIDIA Cuda: graphics and calculations unification. 2007 (http://itc.ua/node/27969).

2. David Luebke. Graphics CPU-not only for graphics (http://www.osp.ru/os/2007/02/4106864/).

3. David Luebke, Greg Humphreys. How GPUs Work// *IEEE Computer. February 2007.* IEEE Computer Society, 2007.

ИСПОЛЬЗОВАНИЕ ВОЗМОЖНОСТЕЙ ГРАФИЧЕСКИХ ПРОЦЕССОРОВ ДЛЯ УСКОРЕНИЯ НАУЧНО-ТЕХНИЧЕСКИХ РАСЧЕТОВ

В.А. Дудник, В.И. Кудрявцев, Т.М. Середа, С.А. Ус, М.В. Шестаков

Описаны новые возможности современных графических процессоров (GPU) по ускорению научно-технических расчётов за счёт распараллеливания вычислительной задачи между центральным процессором и GPU. Приведено описание использования технологии NVIDIA CUDA для подключения параллельных вычислительных возможностей GPU при программировании некоторых математически интенсивных задач. Приведены примеры сравнения показателей производительности при решении этих задач без применения GPU и с использованием возможностей NVIDIA CUDA для графического процессора GeForce® 8800.

ВИКОРИСТАННЯ МОЖЛИВОСТЕЙ ГРАФІЧНИХ ПРОЦЕСОРІВ ДЛЯ ПРИСКОРЕННЯ НАУКОВО-ТЕХНІЧНИХ РОЗРАХУНКІВ

В.О. Дудник, В.І. Кудрявцев, Т.М. Середа, С.О. Ус, М.В. Шестаков

Описано нові можливості сучасних графічних процесорів (GPU) по прискоренню науково-технічних розрахунків за рахунок розпаралелювання обчислювального завдання між центральним процесором і GPU. Приведено опис використання технології NVIDIA CUDA для підключення паралельних обчислювальних можливостей GPU при програмуванні деяких математично інтенсивних завдань. Приведені приклади порівняння показників продуктивності при вирішенні цих завдань без застосування GPU і з використанням можливостей NVIDIA CUDA для графічного процесора GeForce® 8800.