ГОДИШНИК НА МИННО-ГЕОЛОЖКИЯ УНИВЕРСИТЕТ “СВ. ИВАН РИЛСКИ”, Том 54, Св. I, Геология и геофизика, 2011
ANNUAL OF THE UNIVERSITY OF MINING AND GEOLOGY “ST. IVAN RILSKI”, Vol. 54, Part I, Geology and Geophysics, 2011

СРАВНЯВАНЕ СКОРОСТТА НА ОБРАБОТКА НА ГОЛЕМИ МАСИВИ ОТ ДАННИ ПРЕЗ ЦЕНТРАЛНИЯ ПРОЦЕСОР (CPU) И ГРАФИЧНИЯ ПРОЦЕСОР (GPU) ПРИ РЕШАВАНЕТО НА ПРИЛОЖНИ ГЕОФИЗИЧНИ ЗАДАЧИ

Християн Цанков

Минно-геоложки университет "Св. Иван Рилски", 1700 София, ch.tzankov@gmail.com

РЕЗЮМЕ. Тема на настоящия доклад е едно сравнение на изчислителните възможности на GPU спрямо CPU. За тази цел, с нарастваща гъстота на изчислителните точки, е решена правата гравиметрична задача за точков източник в тримерно пространство. Изчисленията са извършени в MATLAB среда като програмно са насочени веднъж през централния процесор и веднъж през графичните процесори като е отчетено компютърното време необходимо за решаване на задачата.

COMPARING THE CPU AND THE GPU PERFORMANCE DURING COMPUTATION OF APPLIED GEOPHYSICAL PROBLEMS INVOLVING LARGE DATABASES

CHRISTIAN TZANKOV
University of Mining and Geology “St. Ivan Rilski”, 1700 Sofia, ch.tzankov@gmail.com

ABSTRACT. The current report presents a comparison test of the CPU and the GPU performance during computation of applied geophysical problems involving large databases. In order to solve the task a direct 3D gravity problem for one point mass model is solved. This procedure is carried out after increasing several times the calculation grid density which forced the necessary computational operations to increase too. The test is made in MATLAB environment where the computing process is directed once through the CPU and once through the GPU. The computer time for both operations is compared.

Въведение

През последните няколко години графичните процесори (GPU) намират все по-широко приложение при провеждане на компютърни изчисления в различни сфери на науката. С развитието на съвременните технологии и най-вече внедряването през 2007 г. на програмно машинната архитектура за паралелно смятане CUDA (Compute Unified Device Architecture) и последвалата я през 2009 г. архитектура от ново поколение с наименованието Fermi, изчисленията с помощта на графични процесори достигнаха ново ниво на развитие (NVIDIA, 2009, 2010, 2011).

Днес редица изследователски колективи и софтуерни компании (Adobe, Autodesk, MathWorks, Microsoft и др.) разработват успешно различни програмни продукти, в които сполучливо се използва огромния изчислителен потенциал на GPU (Yuen et al., 2010).

Достъпът до ресурсите на GPU може да бъде осъществен чрез подходящо написан софтуерен код на CUDA C/C++, OpenCL, DirectCompute, CUDA Fortran и др., чрез използване на поддържащи CUDA програмни системи (MATLAB, Mathematica) или чрез интегрирането на готови програмни библиотеки и инструменти в подходяща работна среда (CUDA Toolkit, Parallel Nsight, Jacket, GPUmat, и др.).

Внедряването на графичните процесори значително съкращава компютърното време при работа с големи масиви от данни като по този начин GPU се налагат като един бърз, евтин и надежден метод за решаване на задачи свързани със структурна механика, биоинформатика, електродинамика, геофизика, магнетизъм, динамика на флуиди, молекулярни модели, финансов анализ, сеизмика, метеорологични модели, цифров анализ, медицинска визуализация, астрофизика и др.

Основната идея на настоящия доклад е провеждането на сравнителен тест на производителността на централния процесор и тази на графичен ускорител, съвместим с CUDA архитектурата с кодово име Fermi. Подобни изследвания са важни от гледна точка на подходящото използване на съвременен хардуер и софтуер и постигане на максимално бързодействие, тъй като решаването на оптимизационните задачи в геофизиката изисква обработката на огромни масиви от данни свързани с

решаването на системи нелинейни уравнения, определяне на градиенти $\nabla f(x)$ Хесиана $H(x)$ и т.н.

За решаването на тази задача е използван модела на точкова маса разположена в декартова координатна система.

1. **Описание на работната станция**

Работната станция, на която е проведен сравнителният тест е закупена от катедра „Приложна геофизика“ по договор с фонд „Научни изследвания“ към МОМН с ръководител доц. дгн Ст. Димовски. Темата на проекта е „Построяване на национален гравиметричен квазигеоид чрез оптимизирани модели от краен брой точкови маси“.

Станцията има следните основни технически характеристики:

Дънна платка: Supermicro SYS-7046GT-TRF;
Процесор: 2 x Intel Xeon E5620, 4 Cores x 2.40GHz;
Графичен процесор: 2 x Tesla C2050, 448 CUDA Cores;
Видео карта: NVIDIA Quadro 600;
Памет: 6 x Kingston 4GB, DDR3-1333;
Твърд диск: 2 x 1000GB WD Raid Edition4, SATA2;
ОС: 64-bit Windows Server 2008 R2 Enterprise SP1

2. **Избор на геофизична задача за провеждане на сравнителния тест**

Изпълнението на задачата за съпоставяне производителността на централния процесор спрямо тази на графичния ускорител е свързано с избора на подходящо уравнение, което освен с простота на аналитичния израз да включва няколко различни прости математически действия. За тази цел е подбран възможно най-елементарният геофизичен модел – точковата маса.

2.1. **Постановка на задачата**

Както е известно от теорията на потенциала, гравитационното поле на сферична маса с хомогенна плътност в която и да е точка вън от масата може да се разглежда като поле на същата маса, съсредоточена в центъра на сферата (Димитров, 1976).

Ако точката $C(\xi, \eta, \zeta)$ е център на сфера с маса M, гравитационния потенциал в точка $P(x, y, z)$ е

$$V = G\frac{M}{r}, \tag{1}$$

а неговата първа производна е

$$V_z = \Delta g = GM\frac{\zeta - z}{r^3}, \tag{2}$$

където $G = 6{,}674.10^{-11}$ m³/kg.s²

$$r = \sqrt{(\xi - x)^2 + (\eta - y)^2 + (\zeta - z)^2}\ .$$

В конкретния пример задачата за полето на точкова маса (2) е решена в декартова координатна система. Точка с маса $M = 1$ kg и координати C (0, 0, 1) е разположена на дълбочина h = 0.001 km под началото на координатната система. Гравитационният потенциал на точката е изчислен в последователно сгъстяваща се квадратна мрежа с размер 1 x 1 km, съвпадаща с равнината $X0Y$. Изчислителните точки имат координати P_i (x_i, y_i, 0), където индексите i и j приемат стойности от 1 до 64, 128, 256, 512, 1024, 2048, 4096, 8192, образувайки поредица от изчислителни мрежи представени в Таблица 1.

Таблица 1.
Параметри на изчислителните мрежи

Стъпка на изчислителната мрежа в km	Размер на изчислителната мрежа n	Брой изчислителни точки n^2
1 / 64	64 x 64	4 096
1 / 128	128 x 128	16 384
1 / 256	256 x 256	65 536
1 / 512	512 x 512	262 144
1 / 1024	1024 x 1024	1 048 576
1 / 2048	2048 x 2048	4 194 304
1 / 4096	4096 x 4096	16 777 216
1 / 8192	8192 x 8192	67 108 864

2.2. **Определяне броя на изчислителните операции**

С цел по-точното определяне времето и скоростта, необходими на централния процесор и графичния ускорител, при изчисляване на задачата за точкова маса, компютърните операции са сведени до минимум. Това е постигнато като разстоянието r от точките на наблюдение до източника, произведението $k = GM$ и разликата $h = \zeta - z$, са определени предварително. Така изразът (2) добива следния вид:

$$\Delta g = k\frac{h}{r^3}, \tag{3}$$

В израза (3) константата k има измерение 10^{-11} m³/s², а разстоянията h и r са в километри, така че полето Δg се получава в милигали.

Изчислителните операции в горния израз (3) са три. Това са едно умножение, едно деление и едно повдигане на трета степен. Броя на тези операции зависи от количеството на изчислителните точки, които са в пряка зависимост от размера на изчислителната мрежа n. В MATLAB, броят операции с плаваща запетая (FLOP) извършени от компютъра се определят по следния начин (Gloub, 2005):

- За поелементно умножение или деление:

$$flop = n^2, \tag{4}$$

- За поелементно степенуване:

$$flop = (m-1)n^2, \tag{5}$$

където m е степенният показател.

Вземайки предвид формули (4) и (5) за определяне броя на операциите в израза за точкова маса (3) и след отчитане на компютърното време t е подходящо броят операции да бъдат превърнати в гигафлопс, т.е.:

$$gigaflops = \frac{4n^2}{t}10^{-9}. \quad (6)$$

Представените изрази са записани в автоматизиран програмен код, работещ в MATLAB среда, с помощта на който се засича процесорното време за пресмятане на всяка една от изчислителните мрежи представени в Таблица 1. Освен това се определят и броя операции с плаваща запетая извършвани от компютъра за една секунда.

3. **Програма и алгоритми за оценка производителността на** CPU **и** GPU

Оценката на производителността на централния процесор и графичния ускорител е извършена в MATLAB среда. За целта е преработен програмен код (Moler, 2004, 2009), и са написани алгоритми, които съвместно решават поставената задача.

При стартиране на програмата е необходимо задаването на максималната по размер изчислителна мрежа (в случая 8192 x 8192), съобразена с процесорните и графични възможности на компютъра. Останалите изчислителни мрежи се изчисляват автоматично като всяка следваща е два пъти по-малка от предишната. За да бъде избегнато натрупването на незначителни и непредставителни резултати, е наложено ограничение за минималната възможна мрежа – 50 x 50.

След генериране на поредица от изчислителни мрежи програмата преминава към създаване и записване на променливите, описани в представения в глава 2 модел. Като се започне от най-малката мрежа, този процес се извършва в два независими, идентични цикъла за всяка една от тях. В първия цикъл променливите се задават с единична точност (single precision), а при втория – с двойна (double precision).

За всеки един от двата независими програмни цикъла, опростената задача за точкова маса (3) се решава веднъж през централния процесор и втори път, след трансформирането на променливите в графични променливи, през графичната карта. Началото и края на изчислителните операции провеждани при изпълнението на уравнение (3) се засичат от таймер, след което компютърното време и броя операции с плаваща запетая извършвани от компютъра за една секунда (6) се записват в нова променлива.

За да бъдат избегнати смущения в производителността при протичане на теста, предизвикани от странични процеси извършвани от работната станция, за всеки текущ изчислителен модел, операцията по решаване на задачата се повтаря многократно като се записва най-доброто време. След изчерпване на сравнителните тестове с единична и двойна точност за всяка наблюдателна мрежа през централния и графичния процесор, резултатите се записват на твърдия диск (Табл. 2) и се представят графично (виж фигурите).

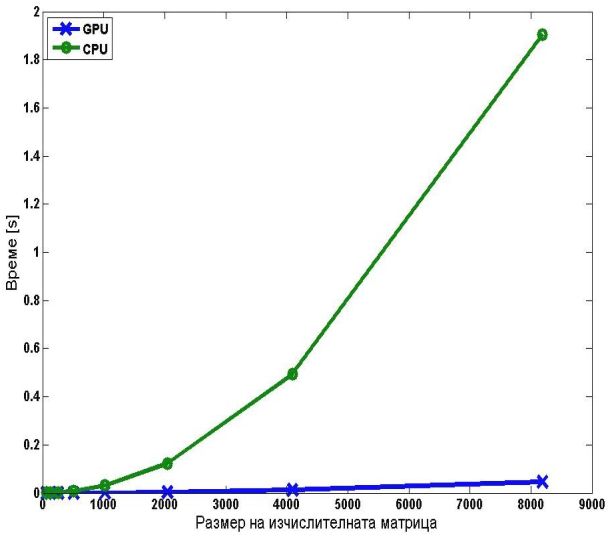
Всички действия и операции по изпълнение на програмния код, освен тестовата задача (3), са съпътстващи и не участват в оценката на производителността на процесорите.

4. **Резултати от сравнителния тест**

За да бъдат избегнати неточности породени от скрити системни процеси по време на тестването, данните представени в таблиците по-долу са получени като оптимален резултат при стократно повторение на операцията за оценка на производителността (3).

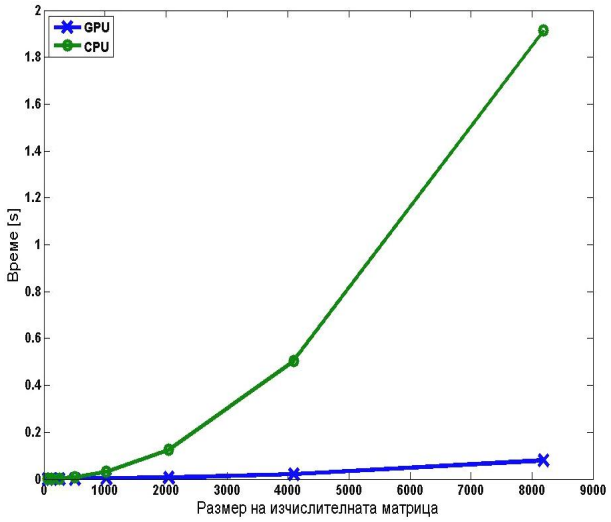
Таблица 2.
Компютърно време за провеждане на сравнителния тест в MATLAB среда

Мрежа	Единична точност		Двойна точност	
n	CPU	GPU	CPU	GPU
64	3,6E-04	6,4E-04	3,6E-04	6,5E-04
128	7,7E-04	6,1E-04	7,2E-04	6,2E-04
256	2,4E-03	6,2E-04	2,4E-03	7,5E-04
512	7,8E-03	9,9E-04	7,9E-03	1,1E-03
1024	3,1E-02	1,5E-03	3,1E-02	2,1E-03
2048	1,2E-01	3,7E-03	1,3E-01	5,9E-03
4096	4,9E-01	1,2E-02	5,0E-01	2,1E-02
8192	1,9E+00	4,5E-02	1,9E+00	8,0E-02



Фиг. 1. **Време на изчислителния процес през** CPU **и** GPU **за мрежи с нарастваща размерност** (**променливи с единична точност**)

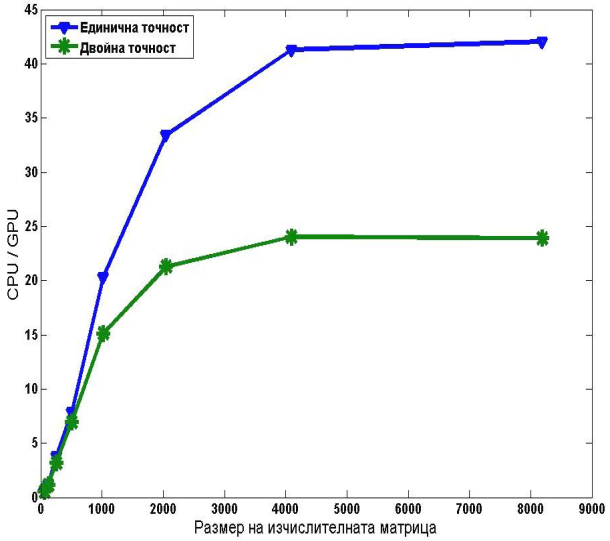
От получените сравнителни резултати, представящи времето необходимо за решаване на задачата за точкова маса през централния и графичния процесор (Табл. 1) и графичните зависимости (фиг. 1-2), е видно, че по своя характер операциите върху променливи с единична и двойна точност не се различават значително (фиг. 1-2). При изчислителните мрежи с малки размери (до 128 x 128) скоростта на двата процеса (през CPU и през GPU) вървят успоредно като преимущество има графичният процесор. С увеличаване размера на наблюдателните мрежи, предимството на графичния ускорител се запазва, при което централният процесор забавя значително своята производителност, достигайки 1,9 s за мрежа с размер 8192 x 8192, докато компютърното време на графичния процесор нараства едва до 0,045 s за променливи с единична точност и до 0,08 s за такива с двойна.



Фиг. 2. **Време на изчислителния процес през** CPU **и** GPU **за мрежи с нарастваща размерност** (**променливи с двойна точност**)

На фиг. 3 е показана зависимостта на забавяне на CPU спрямо GPU като функция на размера на изчислителната матрица. Видно е, че достигайки размери на изчислителната мрежа 1024 x 1024, и при двата вида променливи забавянето е линейно, след което при размери 4096 x 4096 отношението CPU/GPU достига плато където забавянето, съответно ускоряването на GPU спрямо CPU, за променливи с единична точност е около 42,22 пъти, а при двойни – около 23,75 пъти.

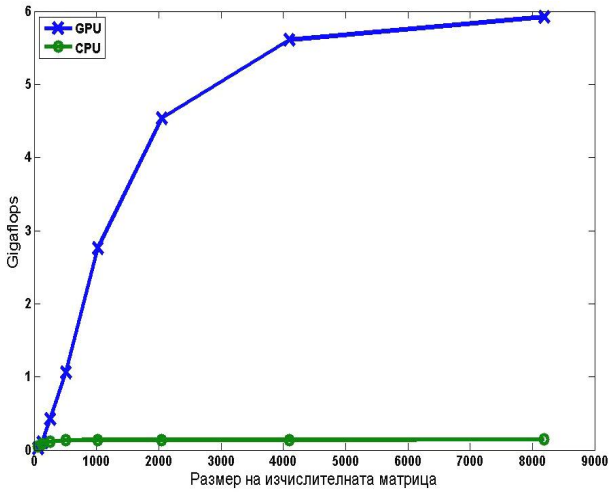
Получените по формула (6) резултати за броя операции с плаваща запетая извършени от компютъра са представени в Табл. 3. Както от таблицата, така и от фигурите (фиг. 4-5) е видно, че при достигане на мрежа с размер 1024 x 1024 броят операции за секунда през процесора клонят към своя лимит, който за променливите от единия и другия тип е около 0,14 GFLOPS. Трябва да се има предвид, че достигната граница от 0,14 GFLOPS е показателна единствено за провеждания тест.



Фиг. 3. **Забавяне на необходимото компютърно време на** CPU **спрямо** GPU **при изчисляване на мрежи с нарастваща размерност**

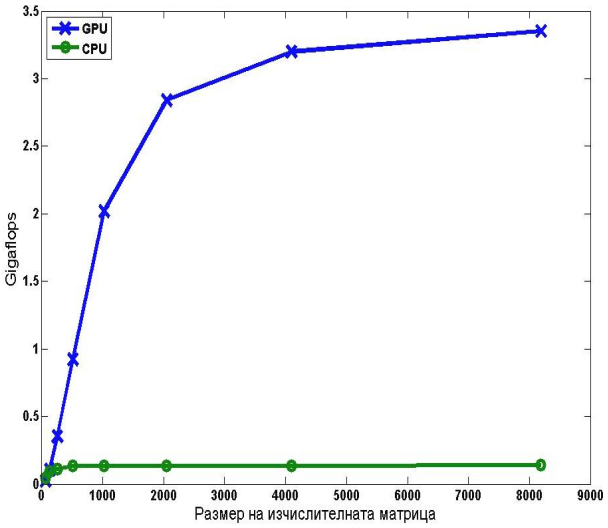
Таблица 3.
Брой операции с плаваща запетая в секунда за провеждане на сравнителния тест в гигафлопс

Мрежа	Единична точност		Двойна точност	
n	CPU	GPU	CPU	GPU
64	4,6E-02	2,6E-02	4,6E-02	2,5E-02
128	8,5E-02	1,1E-01	9,1E-02	1,1E-01
256	1,1E-01	4,2E-01	1,1E-01	3,5E-01
512	1,3E-01	1,1E+00	1,3E-01	9,2E-01
1024	1,4E-01	2,8E+00	1,3E-01	2,0E+00
2048	1,4E-01	4,5E+00	1,3E-01	2,8E+00
4096	1,4E-01	5,6E+00	1,3E-01	3,2E+00
8192	1,4E-01	5,9E+00	1,4E-01	3,4E+00



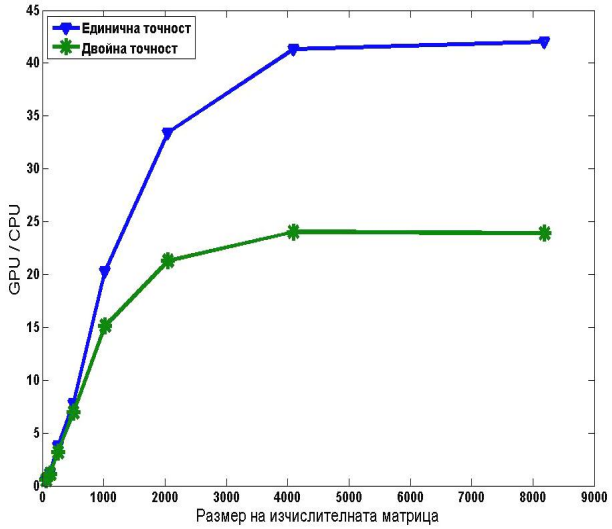
Фиг. 4. **Производителност на** CPU **и** GPU **за мрежи с нарастваща размерност** (**променливи с единична точност**)

За разлика от централния процесор, производителността на графичния значително нараства с увеличаване размера на изчислителната мрежа като за променливи с единична точност тя достига 5,9 GFLOPS, а при променливи с двойна точност – 3,4 GFLOPS.



Фиг. 5. **Производителност на** CPU **и** GPU **за мрежи с нарастваща размерност** (**променливи с двойна точност**)

Същевременно тенденцията към плавно нарастване на производителността се запазва при различните типове променливи. Тук също е необходимо да се отбележи, че получените резултатите имат качеството на оценка главно за конкретния тестов пример.



Фиг. 6. **Ускорение на броя операции с плаваща запетая за една секунда извършени от** GPU **спрямо** CPU

Поради наличието на обратнопропорционална връзка между гигафлопс и компютърното време (6), графиката на ускорението на броя операции с плаваща запетая за една секунда извършени от GPU спрямо CPU (фиг. 6) изключително наподобява тази на фигура 3. След изчисляване на матрицата с размери 4096 x 4096 се вижда, че отношението на гигафлопс за GPU и CPU достига палто. За променливи с единична точност, броят операции през графичния процесор за единица време е с 42,14 пъти по-голям от този на централния процесор, докато при променливи с двойна точност отношението GPU/CPU е почти наполовина – 24,26.

Заключение

В резултат на проведения тест на компютърната производителност става ясно, че графичните процесори базирани на CUDA архитектурата значително оптимизират изчислителния процес и намаляват компютърното време.

Постигнатите резултати при решаване на поставената геофизична задача дават основание да се направи извода, че използването на графични ускорители при решаване на задачи, изискващи обработката на огромни масиви от данни, каквито често се срещат в геофизичната практика, би довело до значително ускоряване на тяхното решение. Такива са задачите свързани с анализа и обработката на сеизмични данни, създаване на най-различни геофизични модели (сеизмологични, геомагнитни, гравитационни, океноложки, атмосферни и др.), интерпретация и оптимизация на обратни задачи и др.

Трябва да се има предвид, че получените резултати имат ориентировъчен характер, тъй като различните аналитични изрази (събиране, изваждане, умножение, деление, степенуване, логаритмуване и т.н.) и процедури (логически операции, задаване на цикли, четене и запаметяване на променливи и др.) изискват различно компютърно време, а на този етап някои от тях са все още практически неизпълними от графичните ускорители.

При създаването на софтуерен код, описаните по-горе особености трябва да бъдат взети под внимание, като по този начин бъдат разработени най-подходящи алгоритми, позволяващи оптимално и пълноценно използване на огромния потенциал на графичните ускорители.

Литература

Гравиразведка. 1981. *Справочник геофизика*. М., Недра, 398 с.

Димитров, Л. В. 1976. *Гравипроучване.* С., Техника, 294 с.

Тончев, Й. 2007. *MATLAB: Преобразувания, изчисления, визуализация. Част I.* С., Техника, 220 с.

Тончев, Й. 2008. MATLAB: *Преобразувания, изчисления, визуализация. Част II.* С., Техника, 220 с.

Тончев, Й. 2009. *MATLAB: Преобразувания, изчисления, визуализация. Част III.* С., Техника, 333 с.

Gloub, G., L. H. Lim. 2005. Counting flops in MATLAB. Numerical Linear Algebra (Lecture 7). ICME and DCS, Stanford University. CME 302/CS 237A. http://www.stat.uchicago.edu/~lekheng/courses/302/.

Moler, Cl. 2009. Experiments with MATLAB. Electronic edition: *The MathWorks, Inc., Natick, MA, 2004*.

Moler, Cl. 2004. Numerical Computing with MATLAB (Revised Reprint, 2008). Electronic edition: *The MathWorks, Inc., Natick, MA, 2004;* SIAM, Philadelphia, xii+336 p.

NVIDIA. 2011. *Release 270 Tesla Driver for Windows – Version 270.90. RN-05247-270-90v01, June 2011. www.nvidia.com/tesla*

NVIDIA. 2010. *Tesla™ C2050 and C2070 GPU Computing Processors. Supercomputing at 1/10th the Cost. www.nvidia.com/tesla.*

NVIDIA. 2009. *NVIDIA's Next Generation CUDA™ Compute Architecture: Fermi™. www.nvidia.com/tesla.*

Yuen, D. A., G. B. Wright, D. A. Sanchez, G. A. Barnett, Jr. 2010. The coming role of GPU in computational geosciences. – *28th IUGG Conference on Mathematical Geophysics, Pisa, Italy 7-11 June 2010, Book of Abstracts*, 187.

Препоръчана за публикуване от
Катедра "Приложна геофизика", ГПФ