АСПАПТАР ЖӘНЕ ЭКСПЕРИМЕНТ ТЕХНИКАЛАРЫ
ПРИБОРЫ И ТЕХНИКА ЭКСПЕРИМЕНТА

УДК 621.386:004.021

Увеличение производительности компьютерных вычислений рентгеновских рефлектограмм

Increase in performance of computer calculations of x-ray reflectometry model

Карташов Д.А., Медетов Н.А., Смирнов Д.И., Орлов Р.С.

Московский государственный институт электронной техники (технический университет), Зеленоград, Москва, Россия (e-mail: rmta@miee.ru)

Мақалада екі толқынды рентгендік рефлектометрдің тәжірибелік нәтижелерін көп қабыршақты құрылым параметрлерін анықтау дәлдігіне алдын ала түрлендіру әсері қарастырылды. Нәтижелерді компьютерлік өңдеу үшін CUDA технологиясын қолданудың эффективтілігі зерттелді. Сондай-ақ екі толқынды рентгендік рефлектометр әдісін үлгілеу бойынша есептеу жүргізудің эффективтілігін арттыру үшін CUDA технологиясын қолдану ұсынылды.

A novel algorithm of experimental data set transformation is presented in respect of double-wave X-ray reflectometry. This algorithm is realized with computer unified device architecture (CUDA). An impact of concerned algorithm application on accuracy of multi-layer structures parameters and efficiency of respective parameters computation are studied. Effective way of multi-layer structure parameters calculations for double-wave X-ray reflectometry models with the help of CUDA based algorithms is declared

Одним из точных неразрушающих методов контроля параметров слоев (толщина, плотность, шероховатость границы раздела) в многослойных тонкопленочных структурах (МС) с периодическими и непериодическими чередующимися слоями на кремниевых подложках является относительная двухволновая рентгеновская рефлектометрия [1].

Трудность применения метода относительной рентгеновской рефлектометрии связана с тем, что данный метод является косвенным, т.е. по экспериментальным данным нельзя непосредственно оценить параметры МС. Для эффективного использования методики требуется сложная интерпретация экспериментальных данных, когда для точного расчета сравнительно небольшого количества параметров достаточно «простых» исследуемых структур требуются большие вычислительные мощности и время.

Процедура определения параметров МС разделяется на два этапа: экспериментальная съемка угловой зависимости коэффициента отражения, а также численное определение параметров МС по выбранной для расчета математической модели МС.

Для корректного моделирования МС необходим предварительный анализ исследуемых образцов с целью определения степени четкости границ раздела, определения того, в каком состоянии, кристаллическом или аморфном, находятся слои, оценки величины шероховатости и т.д. Математическая модель должна учитывать как можно больше особенностей структуры (взаимодиффузию слоев, шероховатость, наличие оксида на поверхности образца или подложки, наличие слоев с переходными фазами). Чем точнее формируется модель, тем лучше результат расчета должен совпадать с экспериментальными данными. Однако это имеет и побочные эффекты: чем сложнее создаваемая модель, тем больше времени требуется для расчета структуры. Это обусловлено увеличением числа рассчитываемых параметров структуры.

В данной работе экспериментальные исследования осуществлялись на рентгеновском многоволновом рефлектометре «X-Ray MiniLab», разработанном в ООО «Институт рентгеновской оптики». Как показано в [2], преимуществом данного прибора является возможность измерения интенсивностей исследуемого рентгеновского излучения на нескольких длинах волн одновременно за одно сканирование. Рефлектометрические измерения проводились по схеме Θ — 2Θ. В качестве источника используется трубка БСВ-21 с медным анодом и видимой проекцией фокусного пятна на аноде 0,02×8 мм. Мощность рентгеновской трубки 280 Вт. Охлаждение источника излучения осуществляется системой замкнутого водяного охлаждения. В качестве детекторов использовались сцинтилляционные детекторы с люминофором NaI: Tl. При рефлектометрических исследованиях для определения параметров пленок использовался генетический алгоритм. Исследованные образцы были получены путем магнетронного распыления на кремниевые подложки. В таблице 1 приведены технологические параметры исследуемой структуры.

Таблица 1

Параметры исследуемой структуры

Слой	Шероховатость по верхней границе слоя, Å	Толщина слоя, Å	Шероховатость по нижней границе слоя, Å
Pt	<10	35	<10

Предварительная обработка экспериментальных результатов относительной двухволновой рентгеновской рефлектометрии. Для численного определения параметров слоёв в МС необходимо выполнить минимизацию функции невязки теоретической и экспериментальной рефлектограммы по критерию среднеквадратического отклонения (СКО). При подгонке теоретической и экспериментальной кривой возникает трудность, связанная с тем, что если экспериментальная кривая периодическая, то в соответствии с выбранным критерием теоретическая кривая может проходить не через точки локальных максимумов и минимумов, а между ними. При этом периоды экспериментальных и теоретических кривых не будут совпадать, и, соответственно, параметры экспериментальной и расчётной МС будут сильно отличаться. Полученная таким образом расчётная структура будет соответствовать локальному минимуму целевой функции. Ситуация осложняется тем, что выйти из найденного локального экстремума не удаётся без применения специальных методов. Обычно это явление имеет место, если на экспериментальной кривой имеется несколько экстремумов. Применение специального предварительного преобразования экспериментальных данных, описанного в данной работе, позволяет избавиться от описанного выше эффекта. Преобразование состоит из трех этапов.

Первый этап преобразования заключается в разделении исходной экспериментальной рефлектограммы на две: прямой и обратной функции. Полученная при этом преобразованная рефлектограмма представляется в памяти компьютера как массив комплексных чисел, действительная часть каждого элемента которого будет соответствовать прямой функции (f (angle)= Ka (angle)/Kb (angle)), а мнимая — обратной (1/f (angle)=Kb (angle)/Ka (angle)).

Второй этап преобразования заключается в выборе определённых значений из массива, полученном в первом этапе. Таким образом, получается ещё один массив комплексных чисел. Приведем пошаговый алгоритм второго этапа преобразования:

1) вычислить интегрированный вид рефлектограммы (для прямой и обратной функции);
2) вычислить значение интеграла для прямой и обратной функции;
3) разбить вертикальные отрезки, ограничиваемые нулём и значением интегралов прямой и обратной функции на равные части. Количество частей определяется числом неизвестных параметров;
4) найти соответствующее Y1i значение на оси X1i для выбранной кривой;
5) определить значение X2i для другой кривой по ранее заданному значению Y1i;
6) определить значение Y2i для другой кривой по ранее заданному значению X1i;
7) вычислить значение квадратного корня из площади прямоугольника, построенного между точками (X1i, Y1i) и (X2i, Y2i).

Третий этап преобразования заключается в получении целевой функции. По полученному массиву определяется целевая функция, равная сумме квадратов его элементов, как действительных, так и мнимых. Целевая функция будет уже действительным числом. Основное отличие полученной таким образом функции ошибки является то, что она не нормирована на единицу. Поэтому диапазон

поиска равен квадратному корню отношения минимальной и максимальной ошибки во всём пространстве поиска. Такой подход позволяет учесть расстояние между экстремумами, так как чем больше расстояние по горизонтальной оси между экстремумами, тем значение полученной суммы и, соответственно, целевая функция больше. Следовательно, целевая функция устроена таким образом, чтобы совмещать X и Y координаты экстремумов. Это обеспечивает более точное совпадение периодов расчётной и экспериментальной рефлектограмм.

На рисунке 1 представлены отношения коэффициентов отражения для экспериментальной и теоретических рефлектограмм (для длин волн CuK_α и CuK_β). Результатом, подтверждающим факт эффективности предварительной обработки экспериментальных результатов относительной двухволновой рентгеновской рефлектометрии на точность определения параметров МС, является более строгое совпадение экспериментальной и теоретической рефлектограммы, хотя необходимо признать, что различие между двумя вычисленными рефлектограммами не столь велико.

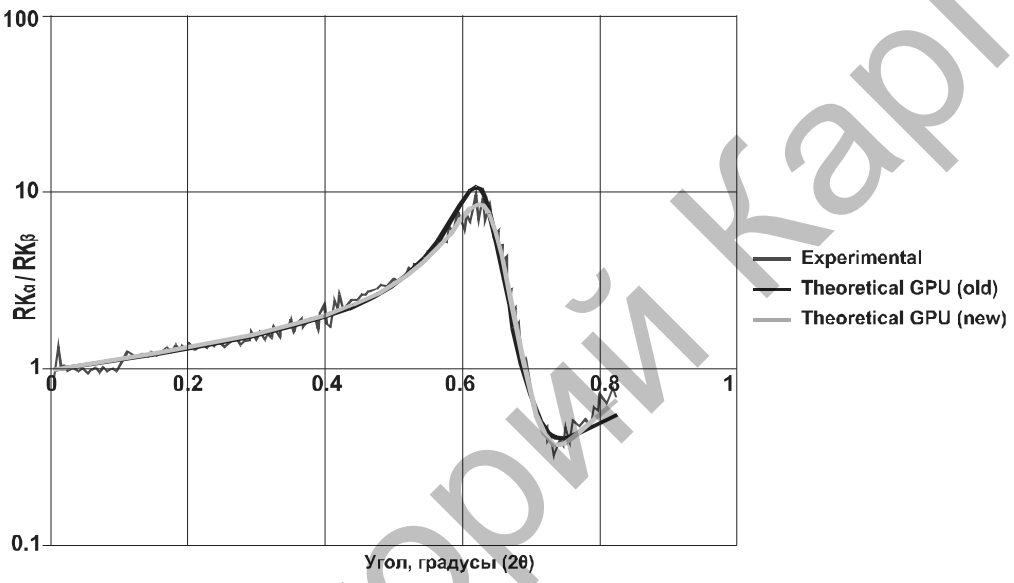

Рис. 1. Угловая зависимость отношения коэффициентов отражения экспериментальной и теоретических рефлектограмм, полученная при расчёте целевой функции без предварительного преобразования экспериментальных данных (old) и с таковым (new)

Таким образом, можно утверждать, что применение описанного нами предварительного преобразования экспериментальных результатов относительной двухволновой рентгеновской рефлектометрии помогает устранить ряд недостатков в численном определении параметров МС, в частности, позволяет более эффективно решить проблему уменьшения шумов в области больших углов, а также добиться более точного совмещения угловых координат максимумов и минимумов расчётных и экспериментальных рефлектограмм.

Использование графических процессоров и технологии CUDA. В настоящее время развитие параллельных вычислительных технологий достигло значительного прогресса, так или иначе связанного с трёхмерными играми. Уже в течение нескольких лет графические процессоры (GPU) используются для неграфических вычислений, выполняя на них сложные математические расчеты. Универсальные устройства с многоядерными процессорами для параллельных векторных вычислений, используемых в 3D-графике, достигают высокой пиковой производительности, которая центральным процессорам (CPU) не под силу. Это связано с тем, что видеокарты состоят из множества мультипроцессоров, которые управляют высокоскоростной памятью, что делает их использование эффективным как для графических, так и для неграфических вычислений.

Применение GPU позволяет значительно ускорить расчеты на обычных персональных компьютерах малой стоимости за счет использования общей памяти и значительного параллелизма [3]. Эффект распараллеливания вычислений и увеличения эффективности достигается за счет того, что на CPU обрабатываются данные, во-первых, последовательно, а во-вторых, не более чем в 8 потоках одновременно. Процессоров же на GPU значительно больше, за счет чего одновременно обрабатывается гораздо большее количество данных. Также производительность GPU и CPU отличает-

ся за счет количества набора операций (в CPU их больше, так как GPU предназначены для решения более узких задач). Современные видеоадаптеры содержат сотни математических исполнительных блоков, и эта мощь может использоваться для значительного ускорения множества вычислительно интенсивных приложений.

Вместе с тем нынешнее поколение GPU обладает достаточно гибкой архитектурой, что вместе с высокоуровневыми языками программирования и программно-аппаратными архитектурами раскрывает эти возможности и делает их значительно более доступными. До недавнего времени эффективное использование вычислительных возможностей видеокарт для неграфических вычислений оставалось сложным из-за возможности управления GPU только через интерфейс прикладного программирования. Именно поэтому компания NVIDIA выпустила технологию программирования Compute Unified Device Architecture (CUDA). Это программно-аппаратная вычислительная архитектура NVIDIA, основанная на расширении языка СИ со своим компилятором и библиотеками для вычислений на GPU.

Технология CUDA обеспечивает быструю разработку и адаптацию программ для исполнения на GPU, а также даёт возможность организации доступа к набору инструкций GPU и управления его памятью при организации параллельных вычислений. Важно, что поддержка NVIDIA CUDA есть у всех чипов G8x, G9x, GT2xx и GF1xx, применяемых в видеокартах GeForce серий 8, 9, 200 и 400, которые очень широко распространены [3, 4].

Конечно, максимальная скорость вычислений на GPU достигается лишь в ряде удобных задач и имеет некоторые ограничения, но такие устройства уже начали довольно широко применять в сферах, для которых они изначально не предназначались. В последние годы исследования в данной области стали значительно интенсивнее [5–9].

Поскольку целью нашего исследования является использование GPU и технологии CUDA для увеличения эффективности вычислений по интерпретации результатов относительной двухволновой рентгеновской рефлектометрии МС на кремниевых подложках, более детальную информацию о технологии CUDA можно найти в [3], а также на сайте компании NVIDIA.

В работе проводились вычисления по интерпретации экспериментальных данных, полученных методом относительной двухволновой рентгеновской рефлектометрии, на базе GPU по технологии CUDA. В качестве сравнения был проведен расчет тех же экспериментальных данных с использованием CPU. Для проведения математических расчетов использовалась следующая модель видеокарты: NVidia GeForce 9600 GT. Число одновременно обрабатываемых потоков 128, максимальное количество потоков может составлять 512. Эта видеокарта обладает 64 процессорами с частотой 1625 МГц и 1024 Mb памяти частотой 1800 МГц. В качестве центрального процессора использовался Intel Core 2 Quad 9300 с четырьмя ядрами, частотой 2,5 GHz каждое, с кэшем первого уровня 64Кб на каждое ядро процессора и 6 Мб общего кэша второго уровня.

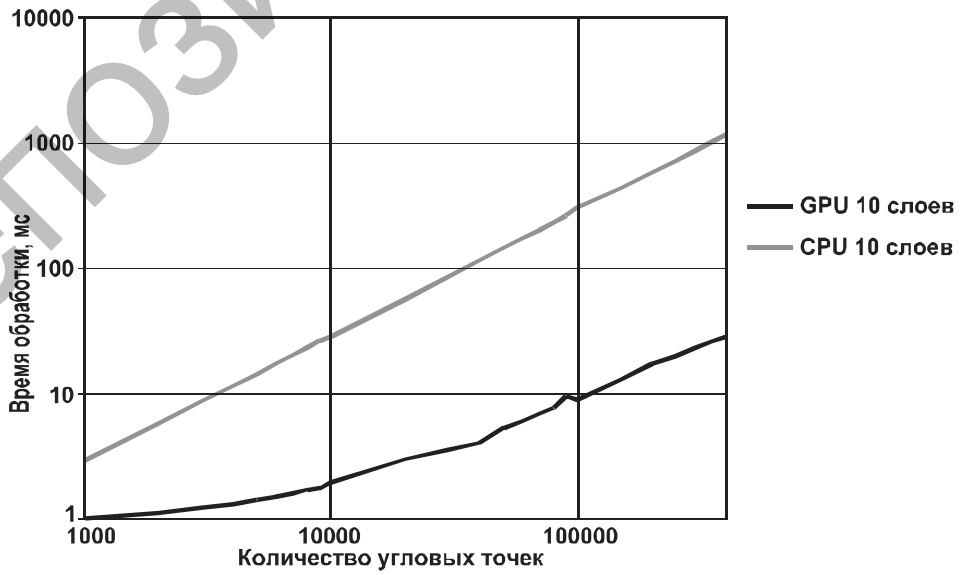

Рис. 2. Время обработки входного массива данных (прямая задача) от количества угловых точек. GPU — NVidia GeForce 9600 GT; CPU — Intel Core 2 Quad 9300

На рисунке 2 представлены временные графики расчета по математической модели для МС, состоящей из 10 слоев. Из них видно, что максимальное увеличение производительности достигается при количестве экспериментальных точек более 100 000 и равно 30. При количестве угловых точек, равном 1000, время обработки результатов на GPU составляет 1 мс, а на CPU — 3 мс, т.е. время вычислений сокращается в 3 раза.

С увеличением количества параметров в вычислительной модели время обсчета на CPU растет линейно, а на GPU нелинейно, что свидетельствует об увеличении эффективности использования графических процессоров для расчёта моделей с большим количеством параметров. Это свойство будет в дальнейшем использовано для увеличения производительности вычислений на GPU.

Из результатов, представленных на рисунке 3, видно, что зависимости целевых функций от числа итераций при расчётах на CPU и GPU немного отличаются на начальном участке, что связано с наличием случайной величины в алгоритме оптимизации. При количестве итераций, равном 1000, значения ошибок, получаемые на CPU и GPU для одного и того же образца, не отличаются, что свидетельствует о корректности вычислений на GPU.

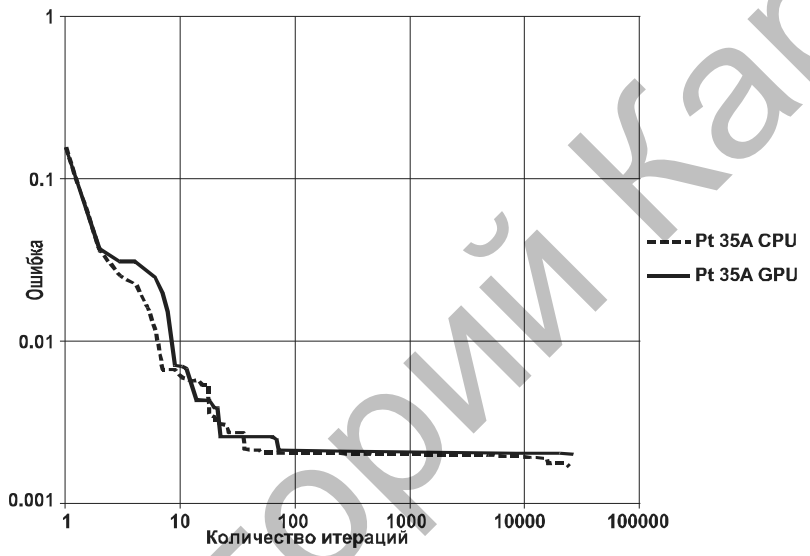

Рис. 3. Зависимость целевой функции от количества итераций

Результаты, представленные на рисунке 4, также подтверждают тот факт, что теоретические рефлектограммы, полученные при вычислении на CPU и GPU, хорошо совпадают с экспериментальной рефлектограммой. Различие между двумя вычисленными рефлектограммами невелико.

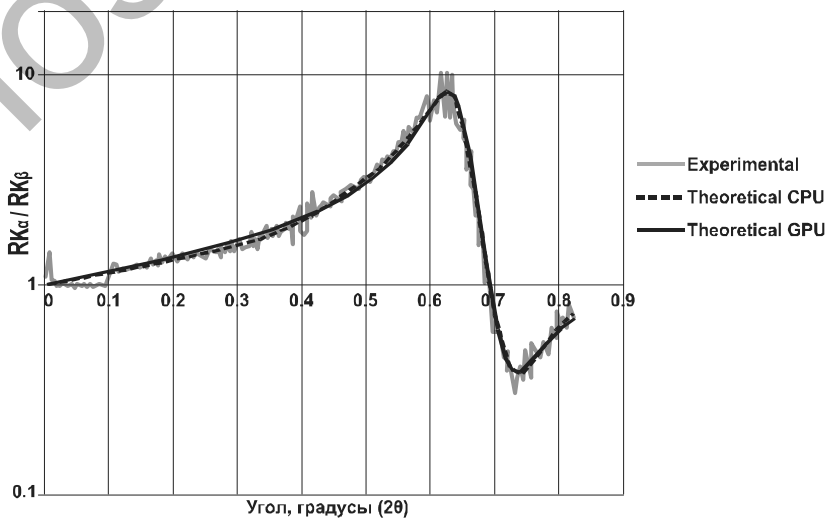

Рис. 4. Рефлектограммы, полученные при расчёте на CPU — Intel Core 2 Quad 9300 и на GPU — NVidia GeForce 9600 GT

Полученные в результате расчётов сведения о параметрах слоёв приведены в таблице 3 и соответствуют действительности. В частности, толщина и плотность слоя платины близка к ожидаемой. Шероховатости границ раздела не превышают 10 Å.

Таблица 3

Параметры, получаемые при решении обратной задачи

Толщина, Å	Шероховатость по верхней границе слоя, Å	Шероховатость по нижней границе слоя, Å	Плотность, г/см^3
37,74	7,02	8,67	20,78

Данные, представленные в таблице 4, показывают, что решение обратной задачи с применением GPU 9600 GT (64 процессора, 1,6 ГГц) проходит в 4,7 раза быстрее, чем на CPU Q9300 (1 процессор, 2,5 ГГц), для слоя платины толщиной 35 Å. Количество особей в популяции генетического алгоритма составляет 256, количество итераций — 2560, что соответствует 640 тыс. процедурам решения прямой задачи.

Таблица 4

Время решения обратной задачи

Толщина, Å	CPU Q9300 (2,5 ГГц), с	GPU 9600 GT, с	Прирост скорости вычислений, разы
35A	2102	448	4,7

Применив свойство нелинейности времени расчёта от количества входных данных на GPU (рис. 2), оказалось возможным дополнительно ускорить вычислительный процесс на GPU, что отображено в таблице 5. Количество угловых точек составляет 330. Количество особей в популяции генетического алгоритма 1–1000. Количество итераций равно 2560, что соответствует 2,56 тыс. — 2,56 млн. процедурам решения прямой задачи.

Таблица 5

Зависимость времени решения обратной задачи от числа особей в генетическом алгоритме, рассчитываемых на видеокарте за один вызов функции

Число особей в генетическом алгоритме	CPU Q9300 (2,5 ГГц), с	GPU 9600 GT, с	Прирост скорости вычислений, разы
1	9	27	0.333
2	17	28	0.6
3	26	29	0.89
5	43	30	1.43
10	85	31	2.74
25	210	38	5.52
100	847	77	11
256	2102	136	15.45
1000	8468	511	16.57

В программе обработки данных количество особей для генетического алгоритма составляет 256. Время вычислений на CPU — 2146 с, а при вычислении на GPU сокращается до 136 с.

Анализируя полученные результаты, можно сделать следующие выводы.

Предложенное предварительное преобразование экспериментальных результатов позволяет улучшить совпадение теоретической и экспериментальной рефлектограмм в процессе вычисления и, соответственно, повысить точность обработки экспериментальных данных.

В результате реализации алгоритма по технологии CUDA на графических процессорах был получен значительный прирост производительности вычислений данных по сравнению с аналогичной реализацией на центральном процессоре компьютера.

С увеличением количества параметров в вычислительной модели время обсчета на базе центрального процессора растет линейно, а на базе графических процессоров — нелинейно, что говорит о большей эффективности использования графических процессоров для обсчета моделей с большим количеством параметров.

Групповой метод расчёта рефлектограмм на графическом процессоре уменьшает время вычислений ещё в несколько раз по сравнению с одиночным методом расчёта рефлектограмм на графическом процессоре. В то же время при малом количестве особей в популяции генетического алгоритма и небольшом числе слоёв эффективность использования центрального процессора для этой задачи выше, но при реальных расчётах такие параметры используются редко.

Также нужно иметь в виду, что в настоящее время графические процессоры являются оптимальной по соотношению «цена — производительность» параллельной архитектурой с общей памятью [6]. При своей относительно невысокой стоимости по вычислительным мощностям они сравнимы с более дорогими небольшими кластерами, реализованными на центральных процессорах. Данный факт увеличивает перспективность использования технологии CUDA в решении задач по интерпретации результатов относительной двухволновой рентгеновской рефлектометрии наноструктур.

Таким образом, описанные в данной работе результаты демонстрируют увеличение точности обработки экспериментальных результатов и сокращение временных затрат на компьютерные вычисления при широкодоступных вычислительных мощностях, что, безусловно, увеличивает эффективность вычислений результатов относительной двухволновой рентгеновской рефлектометрии, а соответственно, значительно увеличивает перспективность применения данного метода для анализа многослойных структур.

Коллектив авторов выражает благодарность профессору Н.Н.Герасименко за идеи и плодотворное обсуждение материалов. Исследования выполнены при финансовой поддержке Федеральной целевой программы «Научные и научно-педагогические кадры инновационной России на 2009–2013 годы» (шифр заявки НК-419П/12, контракт № П2426).

References

1. *Wainfan N., Scott N.J., Parratt L.G.* Density measurements of some thin copper films // J. Appl. Phys. — 1959. — Vol. 30. — No. 10. — P. 1504–1609.
2. *Tur'yanskii A.G., Vinogradov A.V., and Pirshin I.V.* A Two-Wave X-ray Reflectometer // Instruments and experimental techniques. — Vol. 42. — No. 1. — 1999. — P. 105–111.
3. NVIDIA CUDA Compute Unified Device Architecture. Programming Guide. http://developer. download.nvidia.com/compute/cuda/2_0/NVIDIA_CUDA_Programming_Guide_2.0.pdf
4. *Aliautdinov M.A. Troepolskaia G.V.* Using of modern multicore microprocessors for mathematical physics neural network algorithms speed-up // Neural computers: development and using. — 2007. — No. 9. — P. 71–80.
5. *Hagiwara K., Kanzaki J., Okamura N., Rainwater D., Stelzer T.* Fast calculation of HELAS amplitudes using graphics processing unit (GPU) // Eur. Phys. J.C. — 2010. — Vol. 66. — P. 477–492.
6. *Boyarchenkov A.S., Potashnikov S.I.* Molecular dynamics using graphics processors and CUDA technology // Numerical methods and programming. — 2009. — Vol. 10. — P. 9–23.
7. *Boyarchenkov A.S., Potashnikov S.I.* Parallel molecular dynamics with Ewald summation and integration on GPU // Numerical methods and programming. — 2009. — Vol. 10. — P. 158–168.
8. *Matveeva N.O., Gorbachenko V.I.* The decision of systems of the linear algebraic equations on graphic processors with use of technology CUDA // Izvestia Penzenskogo gosudarstvennogo pedagogicheskogo universiteta imeni V.G. Belinskogo. — 2008. — № 8 (12). — P. 115–120.
9. *Evstigneev N.M.* Numerical integration of Poisson's equation using a graphics processing unit with CUDA-technology // Numerical methods and programming. — 2009. — Vol. 10. — P. 268–274.