УДК 621.386:004.021

Повышение эффективности вычислений результатов двухволновой рентгеновской рефлектометрии многослойных структур при использовании графических процессоров и технологии CUDA

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

Аннотация

Рассмотрена эффективность компьютерной обработки результатов относительной двухволновой рентгеновской рефлектометрии многослойных структур на графических процессорах с применением платформы NVIDIA CUDA. Впервые предложена эффективная реализация вычислений по моделированию метода относительной двухволновой рентгеновской рефлектометрии на основе технологии CUDA. В работе проведено сопоставление скорости расчета на видеокарте NVidia GeForce 9600 GT и на четырёхядерном центральном процессоре Intel Core 2 Quad 9300. На основании проведённого исследования предложен способ реализации алгоритма решения прямой задачи на платформе CUDA, позволяющий дополнительно ускорить вычислительный процесс.

Ключевые слова

относительная двухволновая рентгеновская рефлектометрия; многослойные структуры; эффективность вычислений; графический процессор; технология CUDA.

Технология NVIDIA 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 [1]. Современные видеоадаптеры содержат сотни математических исполнительных блоков, и эта мощь может использоваться для значительного ускорения множества вычислительно интенсивных приложений. Вместе с тем нынешние поколения GPU обладают 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 [1, 2].

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 и не предназначались. В последние годы исследования в данной области стали значительно интенсивнее [3-7]. Поскольку целью нашего исследования является использование GPU и технологии CUDA для увеличения эффективности вычислений по интерпретации результатов двухволновой рентгеновской

рефлектометрии многослойных структур на кремниевых подложках, более детальную информацию о технологии CUDA можно найти в [1], а также на сайте компании NVIDIA.

Расчет параметров многослойных структур

Процедуру определения параметров МС можно разбить на несколько этапов: экспериментальная съемка угловой зависимости коэффициента отражения; выбор адекватной модели МС для расчета; численное определение параметров МС.

Формирование модели рассчитываемой МС является ключевым моментом в процессе решения обратной задачи рентгеновской рефлектометрии. Для корректного моделирования МС необходим предварительный анализ исследуемых образцов с целью определения степени четкости границ раздела, определения в кристаллическом или аморфном состоянии 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 рассчитываемых параметров структуры. Зачастую на практике используют самую простую модель расчета МС – модель однородных плоских слоев. Это обусловлено тем, что такая структура является целью при изготовлении МС. С ростом технологических возможностей модель однородных плоских слоев становится все более актуальной.

В работе [8] впервые была предложена модель для расчета коэффициента отражения рентгеновского пучка от поверхности МС. Рекуррентные соотношения, описывающие модель могут быть записаны как

$$r(z_j) = \frac{r_j^F + r(z_{j+1})e^{2i\chi_{j+1}l_{j+1}}}{1 + r_j^F r(z_{j+1})e^{2i\chi_{j+1}l_{j+1}}}; j = 0,1,...,n; \qquad (1)$$

где r_j^F и r_{j+1}^F амплитудные коэффициенты отражения для слоев j и $j+1$ соответственно, σ_j – среднеквадратичная шероховатость (ширина границы раздела) слоя.

Рассмотрим модели расчета шероховатости и межслойных границ раздела. В обзорах [9,10] проводится сравнительный анализ нескольких моделей для учета шероховатости в рефлектометрических исследованиях. Учет шероховатости очень важен для корректного

анализа отражения рентгеновского излучения. Обычно он учитывается как фактор шероховатости Q_j при расчете коэффициента зеркального отражения:

$$r(z_j) = Q_j \cdot r(z_j), \qquad (2)$$

где $r(z_j)$ – коэффициент отражения.

Различные модели учитывают фактор шероховатости по-разному: $Q_j = 1$ (идеально гладкая поверхность); $Q_j = e^{-2\sigma_j^2 r_j^F r_j^F}$ (Фактор Дебая-Валера [11]); $Q_j = e^{-2\sigma_j^2 r_j^F r_{j+1}^F}$ (Фактор Нево-Кроса [11])

Другой часто используемой моделью учета шероховатости является модель промежуточных слоев, в рамках которой шероховатость рассматривается как набор однородных слоев (как правило, 20 слоев) с абсолютно гладкой поверхностью, у которых коэффициент преломления рентгеновского излучения изменяется в соответствии с заданной функцией (например, в соответствии с функцией Лапласа). Очевидным недостатком этой модели является увеличение количества рассчитываемых слоев, что приводит к задержкам в расчете угловой зависимости коэффициента зеркального отражения.

В данной работе для расчета шероховатостей использовалась модель промежуточных слоев, изложенную в [10], которая наиболее точно описывает зеркальное отражение от шероховатой поверхности.

Использование любой из моделей подразумевает сравнение данных, полученных экспериментально и в результате компьютерного расчета. Однако, если в случае достаточно тривиальных моделей возможен визуальный анализ, то в случае сложных моделей необходим критерий оценки совпадения расчетных данных с результатами, полученными экспериментально. Таким критерием может служить некоторая функция невязки, рассчитанная, например, как хи-квадрат. На практике численные параметры структуры определяют минимизацией функционала невязки. Поиск глобального минимума функции многих переменных представляет известную проблему. В книге [12] описан целый ряд алгоритмов поиска глобального минимума, среди них мультистарт, туннельные алгоритмы, методы перехода из одного локального минимума в другой, метод отжига, Монте-Карло, метод тяжелого шарика и т.д. Однако в своей работе мы остановились на популярном в последнее время алгоритме поиска – так называемом генетическом алгоритме [13].

Методика эксперимента

Исследования осуществлялись на рентгеновском многоволновом рефлектометре “X-

Ray MiniLab” фирмы ООО “ИРО”. Как показано в [14] 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, А	Толщина слоя, Å	Шероховатость по нижней границе слоя, А
Pt	<10	35	<10

Для проведения математических расчетов использовалась следующая модель видеокарты: NVidia GeForce 9600 GT. Число одновременно обрабатываемых потоков составляет 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 общёго кэша второго уровня.

Основные результаты

В работе проводилось 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 центрального процессора компьютера (CPU).

На рисунке 1 представлены временные графики расчета на CPU Intel Core 2 Quad 9300 и GPU NVidia GeForce 9600 GT по математической модели для МС, состоящей из 10 слоев.

Из данных, представленных на рисунке 1 видно, что максимальное увеличение производительности достигается при количестве экспериментальных точек более 100 000 и

составляет 30 раз. При количестве угловых точек, равным 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 ещё большего увеличения производительности вычислений на GPU.

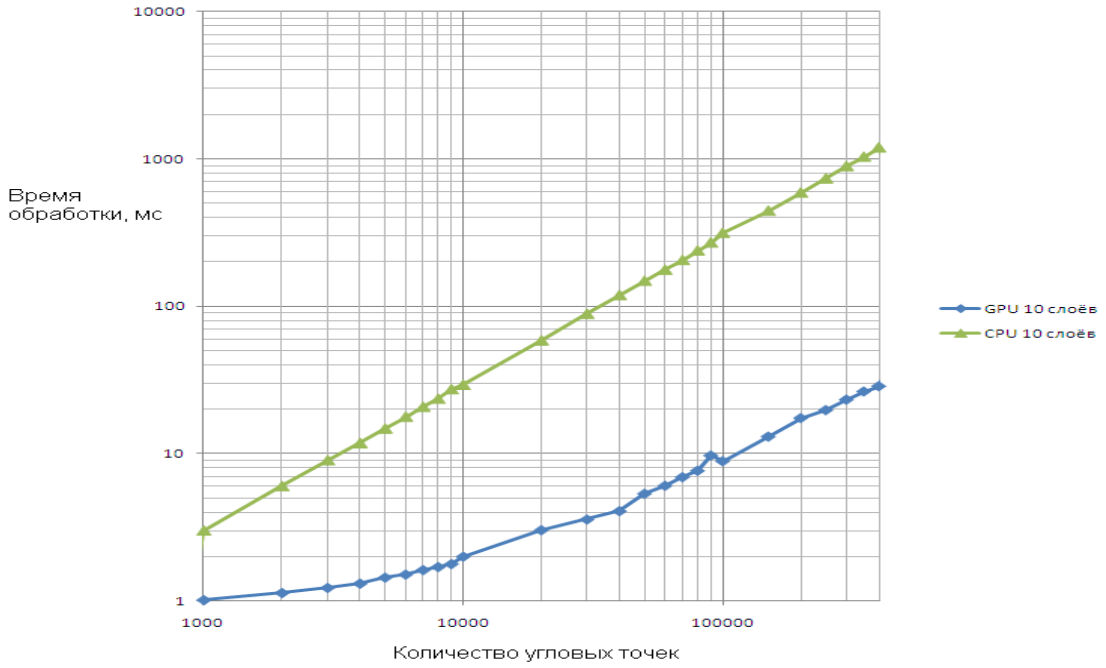

Рисунок 1. Время обработки входного массива данных (прямая задача) от количества угловых точек.

Из результатов, представленных на рисунке 2 видно, что зависимости целевых функций от числа итераций при расчётах на CPU и GPU немного отличаются на начальном участке, что связано с наличием со случайной величины в алгоритме оптимизации. При количестве итераций, равным 1000, значения ошибок, получаемые на CPU и GPU для одного и того же образца не отличаются, что свидетельствует о корректности вычислений на GPU.

Рисунок 2. Зависимость целевой функции от количества итераций.

Результаты, представленные на рисунке 3 также подтверждают тот факт, что рефлектограммы, полученные при вычислении на CPU и GPU хорошо совпадают с экспериментальной рефлектограммой. Различие между двумя вычисленными рефлектограммами невелико.

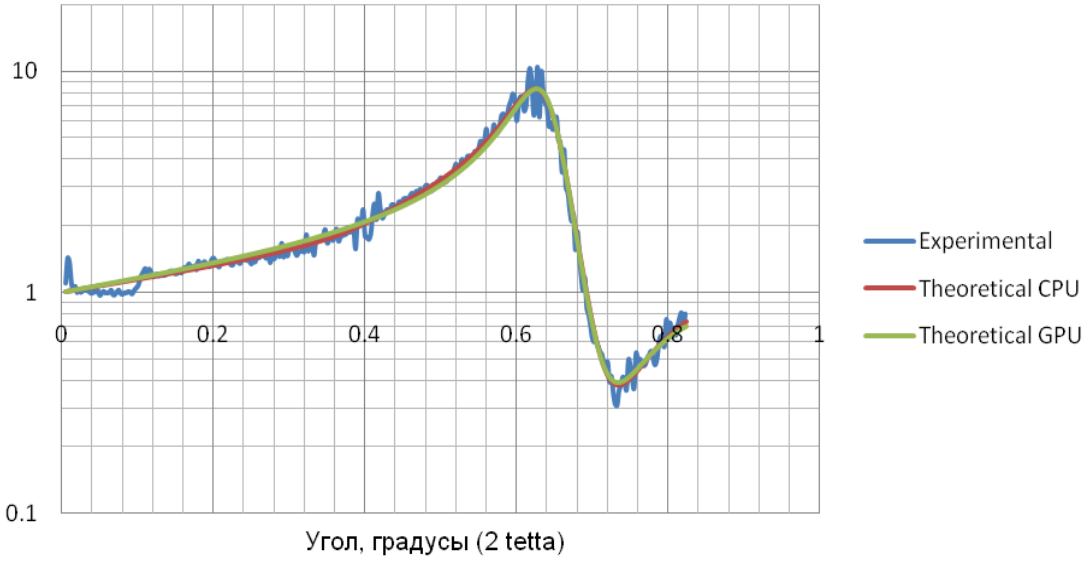

Рисунок 3. Относительные рефлектограммы (Rα/Rβ), полученные при расчёте на CPU Intel Core 2 Quad 9300 и на GPU NVidia GeForce 9600 GT.

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

Таблица 3. Параметры, получаемые при решении обратной задачи.

Толщина, А	Шероховатость по верхней границе слоя, А	Шероховатость по нижней границе слоя, А	Плотность, г/см^3
37.74409	7.023092	8.671289	20.77719

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 35А. Количество особей в популяции составляет 256, количество итераций равно 2560, что соответствует 640 тыс. процедурам решения прямой задачи

Таблица 4. Время решения обратной задачи.

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

Применив свойство нелинейности времени расчёта от количества входных данных на GPU (рисунок 1), оказалось возможным дополнительно ускорить вычислительный процесс на GPU, что отображено в таблице 5. Количество угловых точек составляет 330. Количество особей в популяции 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 ГА	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 ГА составляет 256. Время вычислений на CPU: 2146 секунд, а при вычислении на GPU сокращается до 136 секунд.

Выводы

В данной работе были определены параметры многослойной структуры с использованием описанной технологии CUDA. В результате реализации алгоритма, было получено значительное ускорение вычислений по сравнению с аналогичной реализацией на центральном процессоре, что является критически важным для численного решения обратных задач рентгеновской рефлектометрии.

Современные графические процессоры имеют большие возможности эффективного распараллеливания операций, что позволяет получить решение на 1-2 порядка быстрее, чем на центральном процессоре, что, безусловно, увеличивает эффективность вычислений результатов рентгеновской рефлектометрии, и соответственно, значительно увеличивает перспективность применения данного метода для анализа многослойных структур в режиме on line.

Применяя групповой метод расчёта рефлектограмм на видеокарте, удаётся уменьшить время вычислений ещё в несколько раз по сравнению с одиночным методом расчёта рефлектограмм на GPU. С увеличением объёма данных в вычислительной модели время обсчета на CPU растет линейно, а на GPU нелинейно, что было использовано авторами для увеличения эффективности использования графических процессоров.

Также нужно иметь в виду, что в настоящее время графические процессоры являются оптимальной по соотношению цена-производительность параллельной архитектурой с общей памятью [4]. 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.

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).

Библиографический список

1. NVIDIA CUDA Compute Unified Device Architecture. Programming Guide. http:// developer.download.nvidia.com/compute/cuda/2_0/NVIDIA_CUDA_Programming_Guide_2.0.pdf

2. Аляутдинов М.А., Троепольская Г.В. Использование современных многоядерных

процессоров в нейрокомпьютерах для решения задач математической физики //Нейрокомпьютеры: разработка, применение, 2007. - № 9. - С. 71-80.

3. 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. - V.66 - P.477-492.

4. Боярченков А.С., Поташников С.И. Использование графических процессоров и технологии CUDA для задач молекулярной динамики // Вычислительные методы и программирование, 2009. - Т. 10. - С. 9-23.

5. Боярченков А.С., Поташников С.И. Параллельная молекулярная динамика с суммированием Эвальда и интегрированием на графических процессорах // Вычислительные методы и программирование, 2009. - Т. 10. - С. 158-168.

6. Матвеева Н.О., Горбаченко В.И. Решение систем линейных алгебраических уравнений на графических процессорах с использованием технологии CUDA // Известия ПГПУ, Физико-математические и технические науки, 2008. - № 8(12). - С.115-120.

7. Евстигнеев Н.М. Интегрирование уравнения Пуассона с использованием графического процессора технологии CUDA // Вычислительные методы и программирование, 2009. - Т. 10. - С. 268-274.

8. Parratt L.G. Surface Studies of Solids by Total Reflection of X-Rays. // Phys. Rev., 1954. - V. 95. - P. 359-369.

9. Stoev K. and Sakurai K. // Rigaku J., 1997. - V. 14. - P. 22.

10. Stoev K., Sakurai K. Review on Grazing Incidence X-Ray Spectrometry and Reflectometry // Spectrohimika Acta B., 1999. - V. 54. - P. 41-82.

11. Croce P., Nevot L. // Revue Phys Appl., 1976.- V. 11. - P. 113.

12. Жиглявский А.А., Жилинскас А.Г. Методы поиска глобального экстремума. М.: Наука, 1991.

13. Wormington M., Panaccione C., Matney K.M., Bowen D.K. Characterization of structures from X-ray scattering data using genetic algorithms. //Phil. Trans. R. Soc. Lond. A., 1999. - V. 357. - P. 2827.

14. А.Г. Турьянский, А.В. Виноградов, И.В. Пиршин. "Двухволновой рентгеновский рефлектометр". Приборы и техника эксперимента, № 1, (1999) с.105-111.

Сведения об авторах

Карташов Дмитрий Александрович, ОАО "НИИ Молекулярной электроники и завод Микрон", инженер-технолог отдела исследований перспективных технологий наноэлектроники, аспирант НИЛ «РМТиА», электронная почта: dmitry_kartashov@mail.ru.

Медетов Нурлан Амирович, Московский государственный институт электронной техники, к.ф.м.н., докторант НИЛ «РМТиА», рабочий телефон: 8-499-734-30-11, электронная почта: medetov@rambler.ru

Смирнов Дмитрий Игоревич, Московский государственный институт электронной техники, аспирант, рабочий телефон: 8-499-734-30-11, электронная почта: rmta@miee.ru

Орлов Роман Сергеевич, Московский государственный институт электронной техники, студент, электронная почта: rmta@miee.ru

Иващенко Олег Валерьевич, ОАО "НИИ Молекулярной электроники и завод Микрон", руководитель группы АСУ ТП, электронная почта: oivashchenko@sitronics.com.