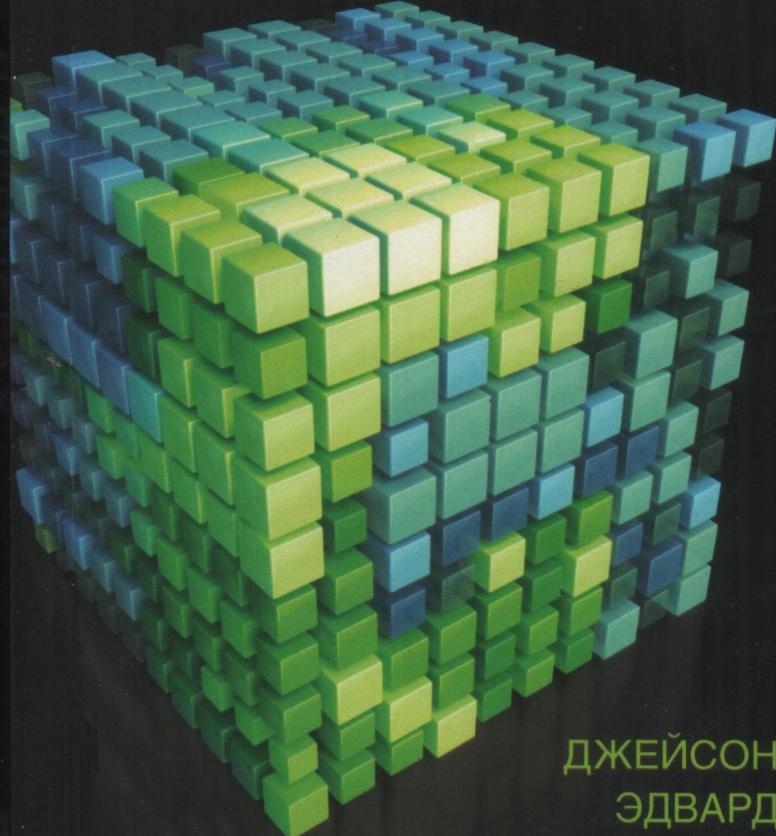




ТЕХНОЛОГИЯ **CUDA** В ПРИМЕРАХ

Введение в программирование
графических процессоров



ДЖЕЙСОН САНДЕРС
ЭДВАРД КЭНДРОТ

Джейсон Сандерс, Эдвард Кэндрот

Технология CUDA в примерах

**Введение в программирование
графических процессоров**

Jason Sanders, Edward Kandrot

CUDA by Example

An introduction to general-purpose GPU programmong

 Addison-Wesley

Upper Saddle River, NJ • Boston • Indianapolis • San Francisco
New York • Toronto • Montreal • London • Munich • Paris • Madrid
Capetown • Sydney • Tokyo • Singapore • Mexico City

Джейсон Сандерс, Эдвард Кэндрот

Технология CUDA в примерах

Введение в программирование графических процессоров

Предисловие Джека Донгарра



Москва, 2013

УДК 004.3'144: 004.383.5CUDA
ББК 32.973.26-04
C18

- Сандерс Дж., Кэндрот Э.
- C18 Технология CUDA в примерах: введение в программирование графических процессоров: Пер. с англ. Слинкина А. А., научный редактор Боресков А. В. – М.: ДМК Пресс, 2013. – 232 с.: ил.
- ISBN 978-5-94074-889-2

CUDA – вычислительная архитектура, разработанная компанией NVIDIA и предназначенная для разработки параллельных программ. В сочетании с развитой программной платформой архитектура CUDA позволяет программисту задействовать невероятную мощь графических процессоров для создания высокопроизводительных приложений, включая научные, инженерные и финансовые приложения.

Книга написана двумя старшими членами команды по разработке программной платформы CUDA. Новая технология представлена в ней с точки зрения программиста. Авторы рассматривают все аспекты разработки на CUDA, иллюстрируя изложение работающими примерами. После краткого введения в саму платформу и архитектуру CUDA, а также беглого обзора языка CUDA C, начинается подробное обсуждение различных функциональных возможностей CUDA и связанных с ними компромиссов. Вы узнаете, когда следует использовать то или иное средство и как писать программы, демонстрирующие поистине выдающуюся производительность.

Издание предназначено для программистов, а также будет полезно инженерам, научным работникам и студентам вузов.

УДК 004.3'144: 004.383.5CUDA
ББК 32.973.26-04

Все права защищены. Любая часть этой книги не может быть воспроизведена в какой бы то ни было форме и какими бы то ни было средствами без письменного разрешения владельцев авторских прав.

Материал, изложенный в данной книге, многократно проверен. Но поскольку вероятность технических ошибок все равно существует, издательство не может гарантировать абсолютную точность и правильность приводимых сведений. В связи с этим издательство не несет ответственности за возможные ошибки, связанные с использованием книги.

*Нашим семьям и друзьям, которые во всем поддерживали нас.
Нашим читателям, которые открывают перед нами будущее.
И учителям, которые научили наших читателей читать.*



Содержание

Предисловие	10
Вступление	12
Благодарности	14
Об авторах	15
Глава 1. Почему CUDA? Почему именно теперь?	16
1.1. О чем эта глава	16
1.2. Век параллельной обработки.....	16
1.2.1. Центральные процессоры.....	17
1.3. Развитие GPU-вычислений	18
1.3.1. Краткая история GPU	18
1.3.2. Ранние этапы GPU-вычислений	19
1.4. Технология CUDA.....	20
1.4.1. Что такое архитектура CUDA?	21
1.4.2. Использование архитектуры CUDA	21
1.5. Применение CUDA	22
1.5.1. Обработка медицинских изображений.....	22
1.5.2. Вычислительная гидродинамика.....	23
1.5.3. Науки об окружающей среде.....	24
1.6. Резюме	24
Глава 2. Приступая к работе	26
2.1. О чем эта глава	26
2.2. Среда разработки.....	26
2.2.1. Графические процессоры, поддерживающие архитектуру CUDA	26
2.2.2. Драйвер устройства NVIDIA	28
2.2.3. Комплект средств разработки CUDA Development Toolkit	28
2.2.4. Стандартный компилятор	30
Windows	30
Linux	30
Macintosh OS X	31
2.3. Резюме	31
Глава 3. Введение в CUDA C	32
3.1. О чем эта глава	32

3.2. Первая программа.....	32
3.2.1. Здравствуй, мир!	32
3.2.2. Вызов ядра	33
3.2.3. Передача параметров	34
3.3. Получение информации об устройстве	36
3.4. Использование свойств устройства	40
3.5. Резюме.....	42

Глава 4. Параллельное программирование на CUDA

4.1. О чём эта глава	43
4.2. Параллельное программирование в CUDA.....	43
4.2.1. Сложение векторов.....	43
Сложение векторов на CPU	44
Сложение векторов на GPU	46
4.2.2. Более интересный пример.....	50
Вычисление фрактала Джулия на CPU.....	51
Вычисление фрактала Джулия на GPU.....	53
4.3. Резюме.....	58

Глава 5. Взаимодействие нитей

5.1. О чём эта глава	59
5.2. Расщепление параллельных блоков.....	59
5.2.1. И снова о сложении векторов.....	60
Сложение векторов на GPU с использованием нитей	60
Сложение более длинных векторов на GPU	62
Сложение векторов произвольной длины на GPU	64
5.2.2. Создание эффекта волн на GPU с использованием нитей	67
5.3. Разделяемая память и синхронизация	72
5.3.1. Скалярное произведение.....	72
5.3.1. Оптимизация скалярного произведения (неправильная)	81
5.3.2. Растровое изображение в разделяемой памяти	83
5.4. Резюме.....	86

Глава 6. Константная память и события

6.1. О чём эта глава	88
6.2. Константная память	88
6.2.1. Введение в метод трассировки лучей	89
6.2.2. Трассировка лучей на GPU	89
6.2.3. Трассировка лучей с применением константной памяти	94
6.2.4. Производительность версии с константной памятью	96
6.3. Измерение производительности с помощью событий	97
6.3.1. Измерение производительности трассировщика лучей	99
6.4. Резюме.....	102

Глава 7. Текстурная память

7.1. О чём эта глава	104
7.2. Обзор текстурной памяти	104
7.3. Моделирование теплообмена.....	105

7.3.1. Простая модель теплообмена	105
7.3.2. Обновление температур	106
7.3.3. Анимация моделирования	108
7.3.4. Применение текстурной памяти.....	112
7.3.5. Использование двумерной текстурной памяти	116
7.4. Резюме.....	120
Глава 8. Интероперабельность с графикой.....	121
8.1. О чём эта глава	121
8.2. Взаимодействие с графикой.....	122
8.3. Анимация волн на GPU с применением интероперабельности с графикой	128
8.3.1. Структура GPUAnimBitmap	129
8.3.2. И снова об анимации волн на GPU	132
8.4. Моделирование теплообмена с использованием интероперабельности с графикой	134
8.5. Интероперабельность с DirectX.....	138
8.6. Резюме	138
Глава 9. Атомарные операции.....	140
9.1. О чём эта глава	140
9.2. Вычислительные возможности	140
9.2.1. Вычислительные возможности NVIDIA GPU	141
9.2.2. Компиляция программы для GPU с заданным минимальным уровнем вычислительных возможностей	142
9.3. Обзор атомарных операций	143
9.4. Вычисление гистограмм	145
9.4.1. Вычисление гистограммы на CPU	145
9.4.2. Вычисление гистограммы на GPU	147
Ядро вычисления гистограммы с применением атомарных операций с глобальной памятью	151
Ядро вычисления гистограммы с применением атомарных операций с глобальной и разделяемой памятью	153
9.5. Резюме	155
Глава 10. Потоки	156
10.1. О чём эта глава	156
10.2. Блокированная память CPU	156
10.3. Потоки CUDA.....	161
10.4. Использование одного потока CUDA	161
10.5. Использование нескольких потоков CUDA	166
10.6. Планирование задач на GPU	171
10.7. Эффективное использование потоков CUDA.....	173
10.8. Резюме	175
Глава 11. CUDA C на нескольких GPU	176
11.1. О чём эта глава	176
11.2. Нуль-копируемая память CPU	176

11.2.1. Вычисление скалярного произведения с применением нуль-копируемой памяти.....	177
11.2.2. Производительность нуля-копирования.....	183
11.3. Использование нескольких GPU.....	184
11.4. Переносимая закрепленная память	188
11.5. Резюме	193
Глава 12. Последние штрихи	194
12.1. О чём эта глава	194
12.2. Инструментальные средства CUDA	195
12.2.1. CUDA Toolkit	195
12.2.2. Библиотека CUFFT	195
12.2.3. Библиотека CUBLAS	196
12.2.4. Комплект NVIDIA GPU Computing SDK.....	196
12.2.5. Библиотека NVIDIA Performance Primitives	197
12.2.6. Отладка программ на языке CUDA C.....	197
CUDA-GDB	197
NVIDIA Parallel Nsight	198
12.2.7. CUDA Visual Profiler.....	198
12.3. Текстовые ресурсы	200
12.3.1. Programming Massively Parallel Processors: A Hands-On Approach.....	200
12.3.2. CUDA U	200
Материалы университетских курсов.....	201
Журнал DR. DOBB'S	201
12.3.3. Форумы NVIDIA	201
12.4. Программные ресурсы	202
12.4.1. Библиотека CUDA Data Parallel Primitives Library.....	202
12.4.2. CULAtools.....	202
12.4.3. Интерфейсы к другим языкам	203
12.5. Резюме	203
Приложение А. Еще об атомарных операциях	204
A.1. И снова скалярное произведение	204
A.1.1. Атомарные блокировки	206
A.1.2. Возвращаясь к скалярному произведению: атомарная блокировка	208
A.2. Реализация хеш-таблицы	211
A.2.1. Обзор хеш-таблиц	212
A.2.2. Реализация хеш-таблицы на CPU.....	214
A.2.3. Многонитевая реализация хеш-таблицы	218
A.2.4. Реализация хеш-таблицы на GPU.....	219
A.2.5. Производительность хеш-таблицы	225
A.3. Резюме.....	226
Предметный указатель	227



Предисловие

Недавние разработки ведущих производителей микросхем, таких как NVIDIA, со всей очевидностью показали, что будущие микропроцессоры и крупные высокопроизводительные вычислительные системы (HPC) будут гибридными (гетерогенными). В их основу будут положены компоненты двух основных типов в разных пропорциях:

- **мультиядерные и многоядерные центральные процессоры:** количество ядер будет и дальше возрастать из-за желания поместить все больше компонентов на один кристалл, не упираясь в барьер мощности, памяти и параллелизма на уровне команд;
- **специализированное оборудование и массивно-параллельные ускорители:** например, графические процессоры (GPU) от NVIDIA в последние годы превзошли стандартные CPU в производительности вычислений с плавающей точкой. Да и программировать их стало так же просто, как многоядерные GPU (если не проще).

Каким будет соотношение между этими компонентами в будущих проектах, пока не ясно, и со временем оно, скорее всего, будет меняться. Но не вызывает сомнений, что вычислительные системы нового поколения – от ноутбуков до суперкомпьютеров – будут состоять из гетерогенных компонентов. Именно такая система преодолела барьер в один петафлоп (10^{15} операций с плавающей точкой в секунду).

И тем не менее задачи, с которыми сталкиваются разработчики, вступающие в новый мир гибридных процессоров, все еще весьма сложны. Критическим частям программной инфраструктуры уже очень трудно поспевать за изменениями. В некоторых случаях производительность не масштабируется с увеличением числа ядер, потому что все большая доля времени тратится не на вычисления, а на перемещение данных. В других случаях разработка программ, оптимизированных для достижения максимального быстродействия, завершается спустя годы после выхода на рынок оборудования, и потому они оказываются устаревшими уже в момент поставки. А иногда, как в случае последних моделей GPU, программа вообще перестает работать, так как окружение изменилось слишком сильно.

Книга «Технология CUDA в примерах» посвящена рассмотрению одного из самых новаторских и эффективных решений задачи программирования массивно-параллельных процессоров, появившемуся в последние годы.

В ней вы на примерах познакомитесь с программированием на языке CUDA C, узнаете о конструкции графических процессоров NVIDIA и научитесь их эффективно использовать. Эта книга – введение в идеи параллельных вычислений, здесь

вы найдете и простые примеры, и технику отладки (поиск логических ошибок и разрешение проблем с производительностью), и более сложные темы, связанные с разработкой и использованием конкретных приложений. Все излагаемые идеи иллюстрируются примерами кода.

Эта книга – обязательное чтение для всех работающих с вычислительными системами, содержащими ускорители. Параллельные вычисления здесь рассматриваются достаточно глубоко, предлагаются подходы к решению различных возникающих задач. Особенно полезна она будет разработчикам приложений, авторам библиотек для численных расчетов, а также студентам, изучающим параллельные вычисления, и преподавателям.

Я получил большое удовольствие от чтения этой книги и почерпнул из нее кое-что новое для себя. Уверен, что и вам это предстоит.

*Джек Донгарра,
заслуженный профессор,
заслуженный исследователь
университета штата Теннесси,
национальная лаборатория Оук Ридж*



Вступление

В этой книге демонстрируется, как, задействовав мощь графического процессора (GPU), имеющегося в вашем компьютере, можно создавать высокопроизводительные приложения для самых разных целей. Хотя первоначально GPU предназначались для визуализации компьютерной графики на экране монитора (и по сей день используются для этой цели), позже они оказались востребованы и в других, столь же требовательных к вычислительным ресурсам областях науки, техники, финансовой математики и др. Программы для GPU, которые решают задачи, не связанные с графикой, мы будем называть программами *общего назначения* (general purpose). Для чтения этой книги необходим опыт работы с языками С или С++, но разбираться в компьютерной графике, к счастью, необязательно. Вообще не нужно! Программирование GPU просто открывает возможность по-новому – и с большой пользой – применить уже имеющиеся у вас навыки.

Чтобы использовать NVIDIA GPU для решения задач общего назначения, необходимо знать, что такое технология CUDA. Все NVIDIA GPU построены на базе *архитектуры CUDA*. Можно считать, что это та основа, на которой компания NVIDIA сконструировала GPU, умеющие решать как традиционные задачи рендеринга, *так и* задачи общего назначения. Для программирования GPU на основе CUDA мы будем использовать язык CUDA C. Как вы скоро увидите, CUDA C – это по существу язык С¹, в который введены расширения, позволяющие программировать массивно-параллельные компьютеры, каковыми являются NVIDIA GPU.

При написании книги мы ориентировались на опытных программистов на С или С++, не испытывающих затруднений при чтении и написании кода на языке С. Текст, изобилующий примерами, позволит вам быстро освоить разработанный компанией NVIDIA язык программирования CUDA C. Разумеется, мы не предполагаем, что вы уже разрабатывали крупномасштабные программные системы, писали компилятор языка С или ядро операционной системы либо досконально знаете все закоулки стандарта ANSI C. Но в то же время мы не тратим время на обзор синтаксиса языка С или таких хорошо известных библиотечных функций, как `malloc()` или `memcpuy()`, поскольку считаем, что с этими вопросами читатель уже знаком.

Вы встретите ряд приемов, которые можно было бы назвать общими парадигмами параллельного программирования, но мы не ставили себе целью написать учебник по основам параллельного программирования. Кроме того, хотя мы так

¹ Точнее, С+++. – Прим. перев.

или иначе демонстрируем почти все части CUDA API, эта книга не может служить заменой полному справочнику по CUDA API, и мы не углубляемся во все детали всех инструментальных средств, помогающих разрабатывать программы на CUDA C. Поэтому мы настоятельно рекомендуем читать эту книгу вместе со свободно доступной документацией, в частности *Руководством по программированию в среде NVIDIA CUDA* (NVIDIA CUDA Programming Guide) и *Настоящим по рекомендованным приемам разработки в среде NVIDIA CUDA* (NVIDIA CUDA Best Practices Guide). Но не стоит сразу бросаться скачивать эти документы; по ходу изложения мы расскажем, что и когда нужно делать.

Ну и хватит слов, мир программирования NVIDIA GPU на CUDA C ждет вас!

Благодарности

Говорят, что, для того чтобы выпустить в свет техническую книгу, нужно семь нянь. И книга «Технология CUDA в примерах» – не исключение. Авторы обязаны многим людям, и некоторых из них хотели бы поблагодарить на этих страницах.

Ян Бак (Ian Buck), отвечающий в NVIDIA за направление разработки ПО для программирования GPU, оказал неоценимую помощь на всех этапах работы над этой книгой – от поддержки самой идеи до консультации по различным деталям. Мы также обязаны Тому Маррею (Tim Murrey), нашему неизменно улыбчивому рецензенту; то, что эта книга технически точна и читабельна, – в основном его заслуга. Большое спасибо нашему художнику Дарвину Тату (Darwin Tat), который создал потрясающую обложку и рисунки в очень сжатые сроки. Наконец, мы очень признательны Джону Парку (John Park), который помог утрясти все юридические тонкости, связанные с изданием книги.

Без помощи со стороны издательства Addison-Wesley эта книга так и осталась бы в мечтах авторов. Питер Гордон (Peter Gordon), Ким Бёдигхаймер (Kim Boedigheimer) и Джуллия Нахил (Julie Nahil) выказали безмерное терпение и профессионализм, благодаря им публикация этой книги прошла без сучка и без задоринки. А стараниями выпускающего редактора Молли Шарп (Molly Sharp) и корректора Ким Уимпсетт (Kim Wimpsett) куча изобилующих ошибками документов превратилась в аккуратный томик, который вы сейчас читаете.

Часть приведенного к этой книге материала не удалось бы включить без содействия со стороны помощников. Точнее, Надим Мохаммад (Nadeem Mohammad) поведал о примерах применения CUDA, упомянутых в главе 1, а Натан Уайтхэд (Nathan Whitehead) щедро поделился кодом, включенным в различные примеры.

Невозможно не поблагодарить и тех, кто прочел первые черновики и прислал полезные отзывы, в том числе Женевьеву Брид (Genevieve Breed) и Курта Уолла (Kurt Wall). Многие работающие в NVIDIA программисты оказали бесценную техническую помощь, в частности Марк Хэйр groove (Mark Hairgrove) скрупулезно проштудировал книгу в поисках всех и всяческих ошибок – технических, типографских и грамматических. Стив Хайнс (Steve Hines), Николас Вилт (Nicholas Wilt) и Стивен Джоунс (Stephen Jones) консультировали нас по отдельным разделам CUDA, помогая пролить свет на нюансы, которые иначе авторы могли бы оставить без внимания. Благодарим также Рандима Фернандо (Randima Fernando), который помог запустить этот проект, и Майкла Шидловски (Michael Schidlowsky) за упоминание Джейсона в своей книге.

И конечно, раздел «Благодарности» немыслим без выражения горячей признательности семьям, которые поддерживали нас во время работы и без которых эта книга никогда не состоялась бы. А особенно мы хотели бы поблагодарить наших любящих родителей, Эдварда и Кэтлин Кэндрот и Стивена и Хелен Сандерс. Спасибо также нашим братьям Кеннету Кэндроту и Кори Сандерсу. Спасибо за неустанную заботу и поддержку.



Об авторах

Джейсон Сандерс – старший инженер-программист в группе CUDA Platform в компании NVIDIA. Работая в NVIDIA, он принимал участие в разработке первых версий системы CUDA и в работе над спецификацией OpenCL 1.0, промышленном стандарте гетерогенных вычислений. Джейсон получил степень магистра по компьютерным наукам в Калифорнийском университете в Беркли, где опубликовал результаты исследований по вычислениям на графических процессорах. Он также получил степень бакалавра по электротехнике и электронике в Принстонском университете. До поступления на работу в NVIDIA он работал в ATI Technologies, Apple и Novell. Когда Джейсон не пишет книги, он любит играть в футбол и заниматься фотографированием.

Эдвард Кэндрот – старший инженер-программист в группе CUDA Algorithms в компании NVIDIA. Вот уже больше 20 лет он занимается оптимизацией кода и повышением производительности различных программ, в том числе Photoshop и Mozilla. Кэндрот работал в Adobe, Microsoft и Google, а также консультировал различные компании, включая Apple и Autodesk. В свободное от кодирования время он любит играть в World of Warcraft или отведывать восхитительные блюда в Лас-Вегасе.



Глава 1. Почему CUDA?

Почему именно теперь?

Еще сравнительно недавно на параллельные вычисления смотрели как на «экзотику», находящуюся на периферии информатики. Но за последние несколько лет положение кардинально изменилось. Теперь практически каждый честолюбивый программист *вынужден* изучать параллельное программирование, если хочет добиться успеха в компьютерных дисциплинах. Быть может, вы взяли эту книгу, еще сомневаясь в важности той роли, которую параллельное программирование играет в компьютерном мире сегодня и будет играть в будущем. В этой вводной главе мы обсудим современные тенденции в области разработки оборудования, лежащие в основе ПО, которое предстоит создавать нам, программистам. И надеемся убедить вас в том, что «параллельно-вычислительная революция» *уже* произошла и что, изучая CUDA C, вы делаете правильный шаг, который позволит занять подобающее место среди тех, кто будет писать высокопроизводительные приложения для гетерогенных платформ, содержащих центральные и графические процессоры.

1.1. О чем эта глава

Прочитав эту главу, вы:

- узнаете о возрастающей роли параллельных вычислений;
- получите представление об истории вычислений с помощью GPU и технологии CUDA;
- познакомитесь с некоторыми успешными приложениями на базе CUDA.

1.2. Век параллельной обработки

За последние годы в компьютерной индустрии произошло немало событий, повлекших за собой масштабный сдвиг в сторону параллельных вычислений. В 2010 году почти все компьютеры потребительского класса будут оборудованы многоядерными центральными процессорами. С появлением дешевых двухъядерных нетбуков и рабочих станций с числом ядер от 8 до 16 параллельные вычисления перестают быть привилегией экзотических суперкомпьютеров и майн-фреймов. Более того, даже в мобильные телефоны и портативные аудиоплееры производители постепенно начинают встраивать средства параллельной обработки, стремясь наделить их такой функциональностью, которая была немыслима для устройств предыдущего поколения.

Разработчики ПО испытывают все более настоятельную потребность в освоении разнообразных платформ и технологий параллельных вычислений, чтобы предложить все более требовательным и знающим пользователям новаторские и функционально насыщенные решения. Время командной строки уходит, настает время многопоточных графических интерфейсов. Сотовые телефоны, умеющие только звонить, вымирают, им на смену идут телефоны, умеющие воспроизводить музыку, заходить в Интернет и поддерживать GPS-навигацию.

1.2.1. Центральные процессоры

В течение 30 лет одним из основных методов повышения производительности бытовых компьютеров было увеличение тактовой частоты процессора. В первых персональных компьютерах, появившихся в начале 1980-х годов, генератор тактовых импульсов внутри CPU работал на частоте 1 МГц или около того. Прошло 30 лет – и теперь тактовая частота процессоров в большинстве настольных компьютеров составляет от 1 до 4 ГГц, то есть примерно в 1000 быстрее своих прародителей. Хотя увеличение частоты тактового генератора – далеко не единственный способ повышения производительности вычислений, он всегда был наиболее надежным из всех.

Однако в последние годы производители оказались перед необходимостью искать замену этому традиционному источнику повышения быстродействия. Из-за фундаментальных ограничений при производстве интегральных схем уже невозможно рассчитывать на увеличение тактовой частоты процессора как средство получения дополнительной производительности от существующих архитектур. Ограничения на потребляемую мощность и на тепловыделение, а также быстро приближающийся физический предел размера транзистора заставляют исследователей и производителей искать решение в другом месте.

Тем временем суперкомпьютеры, оставаясь в стороне от мира потребительских компьютеров, на протяжении десятков лет добивались повышения производительности сходными методами. Производительность применяемых в них процессоров росла столь же быстрыми темпами, как в персональных компьютерах. Однако, помимо впечатляющего повышения быстродействия одного процессора, производители суперкомпьютеров нашли и другой источник роста общей производительности – увеличение *числа* процессоров. Самые быстрые современные суперкомпьютеры насчитывают десятки и сотни тысяч процессорных ядер, работающих согласованно. И, глядя на успехи суперкомпьютеров, естественно задаться вопросом: может быть, не гнаться за повышением производительности одного ядра, а поместить в персональный компьютер несколько таких ядер? При таком подходе мощность персональных компьютеров можно наращивать и дальше, не пытаясь любой ценой увеличить тактовую частоту.

В 2005 году, столкнувшись с ростом конкуренции на рынке и имея не так уж много вариантов выбора, ведущие производители CPU стали предлагать процессоры с двумя вычислительными ядрами вместо одного. В последующие

годы эта тенденция продолжилась выпуском CPU с тремя, четырьмя, шестью и восьмью ядрами. Эта так называемая *мультиядерная революция* знаменовала колossalный скачок в развитии рынка потребительских компьютеров.

Сегодня весьма затруднительно купить настольный компьютер с одним единственным процессорным ядром. Даже самые дешевые маломощные CPU содержат не менее двух ядер. Ведущие производители CPU уже анонсировали планы выпуска CPU с 12 и 16 ядрами, лишний раз подтвердив, что настало время параллельных вычислений.

1.3. Развитие GPU-вычислений

По сравнению с традиционным конвейером обработки данных в центральном процессоре, выполнение вычислений общего характера в графическом процессоре (GPU) – идея новая. Да и сама концепция GPU появилась сравнительно недавно. Однако мысль о том, чтобы производить вычисления в графическом процессоре, не так нова, как может показаться.

1.3.1. Краткая история GPU

Мы уже говорили об эволюции центральных процессоров в плане увеличения тактовой частоты и числа ядер. А тем временем в области обработки графики произошла настоящая революция. В конце 1980 – начале 1990-х годов рост популярности графических операционных систем типа Microsoft Windows создал рынок для процессоров нового типа. В начале 1990-х годов пользователи начали покупать ускорители двумерной графики для своих ПК. Эти устройства позволяли аппаратно выполнять операции с растровыми изображениями, делая работу с графической операционной системой более комфортной.

Примерно в то же время – на протяжении всех 1980-х годов – компания Silicon Graphics, работавшая в области профессионального оборудования и ПО, стремилась вывести трехмерную графику на различные рынки, в том числе приложения для правительства и министерства обороны, визуализация при решении научно-технических задач, а также инструменты для создания впечатляющих кинематографических эффектов. В 1992 году Silicon Graphics раскрыла программный интерфейс к своему оборудованию, выпустив библиотеку OpenGL. Silicon Graphics рассчитывала, что OpenGL станет стандартным, платформенно независимым методом написания трехмерных графических приложений. Как и в случае параллельной обработки и новых CPU, приход новых технологий в потребительские приложения – всего лишь вопрос времени.

К середине 1990-х годов спрос на потребительские приложения с трехмерной графикой резко увеличился, что подготовило условия для двух весьма существенных направлений разработки. Во-первых, выход на рынок игр шутеров от первого лица (First Person Shooter, FPS), таких как Doom, Duke Nukem 3D и Quake, знаменовал начало гонки за создание все более и более реалистичных трехмерных сцен для игр на ПК. Хотя в конечном итоге 3D-графика проникнет практически

во все компьютерные игры, популярность только нарождающихся «стрелялок» от первого лица существенно ускорила внедрение 3D-графики в компьютеры потребительского класса. В то же время такие компании, как NVIDIA, ATI Technologies и 3dfx Interactive, начали выпускать доступные по цене графические ускорители, способные заинтересовать широкую публику. Все это закрепило за 3D-графикой место на рынке перспективных технологий.

Выпуск компанией NVIDIA карты GeForce 256 еще больше расширил возможности графического оборудования для потребительских компьютеров. Впервые вычисление геометрических преобразований и освещения сцены стало возможно производить непосредственно в графическом процессоре, что позволило создавать еще более визуально привлекательные приложения. Поскольку обработка преобразований и освещения уже входили неотъемлемой частью в графический конвейер OpenGL, выход GeForce 256 ознаменовал начало этапа реализации все большего и большего числа компонентов графического конвейера аппаратно.

А с точки зрения параллельных вычислений, выпуск в 2001 году серии GeForce 3 представляет, пожалуй, самый важный прорыв в технологии производства GPU. Это была первая микросхема, в которой был реализован тогда еще новый стандарт Microsoft DirectX 8.0. Этот стандарт требовал, чтобы совместимое оборудование включало возможность программируемой обработки вершин и пикселов (шейдинга). Впервые разработчики получили средства для частичного контроля над тем, какие именно вычисления будут проводиться на GPU.

1.3.2. Ранние этапы GPU-вычислений

Появление на рынке GPU с программируемым конвейером привлекло внимание многих исследователей к возможности использования графического оборудования не только для рендеринга изображений средствами OpenGL или DirectX. В те первые годы вычисления с помощью GPU были чрезвычайно запутанными. Поскольку единственным способом взаимодействия с GPU оставались графические API типа OpenGL и DirectX, любая попытка запрограммировать для GPU произвольные вычисления была подвержена ограничениям, налагаемым графическим API. Поэтому исследователи старались представить задачу общего характера как традиционный рендеринг.

По существу, GPU начала 2000-х годов предназначались для вычисления цвета каждого пикселя на экране с помощью программируемых арифметических устройств, *пиксельных шейдеров* (pixel shader). В общем случае пиксельный шейдер получает на входе координаты (x, y) точки на экране и некоторую дополнительную информацию, а на выходе должен выдать конечный цвет этой точки. В качестве дополнительной информации могут выступать входные цвета, текстурные координаты или иные атрибуты, передаваемые шейдеру на этапе его выполнения. Но поскольку арифметические действия, производимые над входными цветами и текстурами, полностью контролируются программистом, то, как заметили исследователи, в качестве «цветов» могли выступать *любые* данные.

Если на вход подавались числовые данные, содержащие не цвета, то ничто не мешало программисту написать шейдер, выполняющий с ними произвольные вычисления. GPU возвращал результат как окончательный «цвет» пикселя, хотя на деле цветом оказывался результат запрограммированных вычислений над входными данными. Результат можно было считать в программу, а GPU было безразлично, как именно он интерпретируется. Таким образом, программист «обманом» заставлял GPU проделать вычисления, не имеющие никакого отношения к рендерингу, подавая их под личиной стандартной задачи рендеринга. Трюк остроумный, но уж больно неестественный.

Поскольку скорость выполнения арифметических операций на GPU была очень высока, результаты первых экспериментов прочили GPU-вычислениям блестящее будущее. Однако модель программирования была слишком ограничивающей для формирования критической массы разработчиков. Имели место жесткие ограничения на ресурсы, поскольку программа могла получать входные данные только в виде горстки цветов и текстурных блоков. Существовали серьезные ограничения на то, как и в какое место памяти можно записывать результаты, поэтому алгоритмы, для которых требовалась возможность записывать в произвольные ячейки памяти (с разбросом, scatter) на GPU не могли быть запущены. Кроме того, было почти невозможно предсказать, как конкретный GPU поведет себя в отношении чисел с плавающей точкой (и будет ли вообще их обрабатывать), поэтому для большинства научных расчетов GPU оказывался непригоден. Наконец, в процессе разработки неизбежно случается, что программа выдает неправильные результаты, не завершается или попросту подвешивает компьютер, но никакого сколько-нибудь приемлемого способа отладки кода, исполняемого GPU, не существовало.

Более того, всякий человек, которого эти ограничения не напугали и который все-таки хотел использовать GPU для выполнения вычислений общего назначения, должен был выучить OpenGL или DirectX, так как никакого другого способа взаимодействия с GPU не было. А это означает, что не только данные следует хранить в виде графических текстур и вычисления над ними выполнять путем вызова функций OpenGL или DirectX, но и сами вычисления записывать на специальных языках графического программирования – *шейдерных языках*. Необходимость работать в жестких рамках ограничений на ресурсы и способы программирования да при этом еще изучать компьютерную графику и шейдерные языки – и все только для того, чтобы в будущем воспользоваться вычислительной мощью GPU, – оказалась слишком серьезным препятствием на пути к широкому признанию технологии.

1.4. Технология CUDA

Лишь спустя пять лет после выпуска серии GeForce 3 наступил расцвет GPU-вычислений. В ноябре 2006 года NVIDIA торжественно объявила о выпуске первого в истории GPU с поддержкой стандарта DirectX 10, GeForce 8800 GTX. Он был построен на архитектуре CUDA. Она включала несколько новых компонен-

тов, предназначенных исключительно для GPU-вычислений и призванных снять многие ограничения, которые препятствовали полноценному применению прежних графических процессоров для вычислений общего назначения.

1.4.1. Что такое архитектура CUDA?

В отличие от предыдущих поколений GPU, в которых вычислительные ресурсы подразделялись на вершинные и пиксельные шейдеры, в архитектуру CUDA включен унифицированный шейдерный конвейер, позволяющий программе, выполняющей вычисления общего назначения, задействовать любое арифметически-логическое устройство (АЛУ, ALU), входящее в микросхему. Поскольку NVIDIA рассчитывала, что новое семейство графических процессоров будет использоваться для вычислений общего назначения, то АЛУ были сконструированы с учетом требований IEEE к арифметическим операциям над числами с плавающей точкой одинарной точности; кроме того, был разработан набор команд, ориентированный на вычисления общего назначения, а не только на графику. Наконец, исполняющим устройствам GPU был разрешен произвольный доступ к памяти для чтения и записи, а также доступ к программно-управляемому кэшу, получившему название *разделяемая память* (shared memory). Все эти средства были добавлены в архитектуру CUDA с целью создать GPU, который отличноправлялся бы с вычислениями общего назначения, а не только с традиционными задачами компьютерной графики.

1.4.2. Использование архитектуры CUDA

Однако усилия NVIDIA, направленные на то, чтобы предложить потребителям продукт, одинаково хорошо приспособленный для вычислений общего назначения и для обработки графики, не могли ограничиться разработкой оборудования, построенного на базе архитектуры CUDA. Сколько бы новых возможностей для вычислений ни включала NVIDIA в свои микросхемы, все равно единственным способом доступа к ним оставался OpenGL или DirectX. И, значит, пользователи по-прежнему должны были маскировать вычисления под графические задачи и оформлять их на шейдерном языке типа GLSL, входящего в OpenGL, или Microsoft HLSL.

Чтобы охватить максимальное количество разработчиков, NVIDIA взяла стандартный язык C и дополнила его некоторыми новыми ключевыми словами, позволяющими задействовать специальные средства, присущие архитектуре CUDA. Через несколько месяцев после выпуска GeForce 8800 GTX открыла доступ к компилятору нового языка CUDA C. Он стал первым языком, специально разработанным компанией по производству GPU с целью упростить программирование GPU для вычислений общего назначения.

Помимо создания языка для программирования GPU, NVIDIA предлагает специализированный драйвер, позволяющий использовать возможности массивно-параллельных вычислений в архитектуре CUDA. Теперь пользователям нет

нужды изучать программные интерфейсы OpenGL или DirectX или представлять свои задачи в виде задач компьютерной графики.

1.5. Применение CUDA

Со времени дебюта в 2007 году на языке CUDA C было написано немало приложений в различных отраслях промышленности. Во многих случаях за счет этого удалось добиться повышения производительности на несколько порядков. Кроме того, приложения, работающие на графических процессорах NVIDIA, демонстрируют большую производительность в расчете на доллар вложенных средств и на ватт потребленной энергии по сравнению с реализациями, построенными на базе одних лишь центральных процессоров. Ниже описаны несколько направлений успешного применения языка CUDA C и архитектуры CUDA.

1.5.1. Обработка медицинских изображений

За последние 20 лет резко возросло количество женщин, страдающих раком груди. Но благодаря неустанным усилиям многих людей в последние годы интенсифицировались исследования, направленные на предотвращение и лечение этого страшного заболевания. Конечная цель состоит в том, чтобы диагностировать рак груди на ранней стадии, дабы избежать губительных побочных эффектов облучения и химиотерапии, постоянных напоминаний, которые оставляет хирургическое вмешательство, и летальных исходов в случаях, когда лечение не помогает. Поэтому исследователи прилагают все силы, чтобы отыскать быстрые, точные способы с минимальным вмешательством для диагностики начальных симптомов рака груди.

У маммограммы, одного из лучших современных методов ранней диагностики рака груди, есть несколько существенных ограничений. Необходимо получить два или более изображений, причем пленку должен проявить и интерпретировать опытный врач, способный распознать потенциальную опухоль. Кроме того, частые рентгеновские исследования сопряжены с риском облучения пациентки. После тщательного изучения врачам нередко требуется изображение конкретного участка – и даже биопсия, – чтобы исключить возможность рака. Такие положительные результаты приводят к дорогостоящим дополнительным исследованиям и вызывают у пациента ненужный стресс в ожидании окончательного заключения.

Ультразвуковые исследования безопаснее рентгеновских, поэтому врачи часто назначают их в сочетании с маммографией при диагностике и лечении рака груди. Однако у традиционного УЗИ груди тоже есть свои ограничения. Попытки решить эту проблему привели к образованию компании TechniScan Medical Systems. TechniScan разработала многообещающую методику трехмерного ультразвукового сканирования, но ее решение не было внедрено в практику по очень простой причине: нехватка вычислительных мощностей. Проще говоря, вычисления, необходимые для преобразования собранных в результате УЗИ данных в трехмерное

изображение, занимают слишком много времени, поэтому признаны чрезмерно дорогими для клинического применения.

Появление первого GPU компании NVIDIA, основанного на архитектуре CUDA, вкупе с языком программирования CUDA C стало той платформой, на которой TechniScan смогла претворить мечты своих основателей в реальность. В системе ультразвукового сканирования Svara для получения изображения груди пациентки применяются ультразвуковые волны. В системе используются два процессора NVIDIA Tesla C1060, обрабатывающие 35 Гб данных, собранных за 15 минут сканирования. Благодаря вычислительной мощности Tesla C1060 уже через 20 минут врач может рассматривать детализированное трехмерное изображение груди пациентки. TechniScan рассчитывает на широкое внедрение системы Svara, начиная с 2010 года.

1.5.2. Вычислительная гидродинамика

В течение многих лет проектирование эффективных винтов и лопаток оставалось черной магией. Невероятно сложное движение воздуха и жидкости, обтекающих эти устройства, невозможно исследовать на простых моделях, а точные модели оказываются слишком ресурсоемкими с вычислительной точки зрения. Лишь самые мощные суперкомпьютеры могли предложить ресурсы, достаточные для расчета численных моделей, требуемого для разработки и проверки конструкции. Поскольку лишь немногие организации имеют доступ к таким машинам, проектирование подобных механизмов находится в застое.

Кэмбриджский университет, продолжая великие традиции, заложенные Чарльзом Бэббиджем, является полигоном для активных исследований в области параллельных вычислений. Д-р Грэхем Пуллан и д-р Тобиас Брэндвик из «многоядерной группы» правильно оценили потенциал архитектуры CUDA для беспрецедентного ускорения гидродинамических расчетов. Первоначальная прикидка показала, что персональная рабочая станция, оборудованная GPU, уже способна достичь приемлемой производительности. Впоследствии небольшой кластер, собранный из GPU, с легкостью обставлял куда более дорогие суперкомпьютеры, подтвердив предположение о том, что GPU компании NVIDIA отлично подходят для решения интересующих их задач.

Для исследователей из Кэмбриджа колossalный выигрыш в производительности, полученный за счет использования CUDA C, оказался больше, чем простое дополнение к ресурсам их суперкомпьютера. Наличие многочисленных дешевых вычислительных устройств позволило ученым проводить эксперименты и быстро получать их результаты. А когда результат численного эксперимента доступен через несколько секунд, возникает обратная связь, приводящая в конечном итоге к прорыву. В результате применение дешевых GPU-кластеров принципиально изменило подход к проведению исследований. Почти интерактивное моделирование открыло новые возможности для новаторских идей в области, которая раньше страдала от застоя.

1.5.3. Науки об окружающей среде

Естественным следствием быстрой индустриализации глобальной экономики является увеличивающая потребность в потребительских товарах, которые не загрязняют окружающую среду. Озабоченность в связи с изменением климата, неуклонно растущие цены на топливо и увеличение концентрации загрязняющих веществ в воде и воздухе – все это остро поставило вопрос о косвенном ущербе от промышленных выбросов. Моющие и чистящие средства давно уже стали необходимыми, но потенциально вредными потребительскими продуктами ежедневного пользования. Поэтому многие ученые начали искать, как сократить пагубное влияние моющих средств на окружающую среду, не снижая их эффективности. Но получить что-то из ничего – задача не из простых.

Основными компонентами чистящих средств являются поверхностно-активные вещества (ПАВ). Именно их молекулы определяют чистящую способность и текстуру стиральных порошков и шампуней, но они же оказывают разрушающее воздействие на окружающую среду. Эти молекулы сцепляются с грязью, а затем соединяются с водой, так что ПАВ смывается вместе с грязью. Традиционное измерение чистящей способности нового вещества требует длительных лабораторных исследований различных комбинаций материалов и загрязнений. Неудивительно, что процедура оказывается медленной и дорогой.

Университет Темпл работает совместно с промышленным гигантом, компанией Procter & Gamble, над моделированием взаимодействия молекул ПАВ с грязью, водой и другими материалами. Внедрение компьютерных моделей позволило не только ускорить традиционный подход, но и расширить диапазон исследований, включив многочисленные вариации окружающих условий, – дело, совершенно немыслимое в прошлом. Исследователи из университета Темпл воспользовались GPU-ускоренной программой моделирования Highly Optimized Object Oriented Many-particle Dynamics (HOOMD – высокооптимизированная объектно-ориентированная многочастичная динамика), разработанной в лаборатории Эймса при министерстве энергетики. Распределив моделирование между двумя GPU NVIDIA Tesla, они сумели достичь производительности, эквивалентной суперкомпьютеру Cray XT3 с 128 процессорными ядрами или IBM BlueGene/L с 1024 процессорами. Увеличив количество Tesla GPU, они уже могут моделировать взаимодействия ПАВ в 16 раз быстрее, чем на старых платформах. Поскольку архитектура CUDA помогла сократить время исчерпывающего моделирования с нескольких недель до нескольких часов, то в ближайшем будущем должны появиться новые продукты, одновременно более эффективные и менее вредные для окружающей среды.

1.6. Резюме

Компьютерная индустрия стоит на пороге революции, связанной с массовым переходом к параллельным вычислениям, а язык CUDA C, разработанный компанией NVIDIA, пока что является одним из самых успешных языков для парал-

лельных вычислений. Прочитав эту книгу, вы научитесь писать код на CUDA C. Мы расскажем о специализированных расширениях языка C и программных интерфейсах (API), которые NVIDIA создала для программирования GPU. *Не* предполагается знакомство со стандартами OpenGL или DirectX, а также какие-либо знания в области компьютерной графики.

Мы не останавливаемся на основах программирования на C, поэтому не можем порекомендовать эту книгу людям, совсем незнакомым с программированием. Некоторые знания в области параллельного программирования были бы полезны, но мы *не предполагаем*, что вы когда-либо занимались параллельным программированием по-настоящему. Все необходимые термины и концепции, относящиеся к параллельному программированию, объясняются прямо в тексте. Не исключено даже, что вы столкнетесь с ситуацией, когда допущения, сделанные благодаря знанию традиционного параллельного программирования, оказываются неверными в применении к программированию GPU. Так что в действительности единственным обязательным условием для чтения этой книги является наличие какого-то опыта программирования на языке C или C++.

В следующей главе мы расскажем, как подготовить свой компьютер для GPU-вычислений – установить все необходимые аппаратные и программные компоненты. После этого можно будет приступить к изучению CUDA C. Если вы уже имеете опыт работы с CUDA C или уверены, что система настроена для разработки на CUDA C, то можете сразу перейти к главе 3.



Глава 2. Приступая к работе

Надеемся, что глава 1 пробудила в вас желание поскорее приступить к изучению CUDA C. Поскольку подход, принятый в этой книге, – обучение на примерах, то вам понадобится среда разработки. Конечно, можно было бы постоять на обочине и понаблюдать за происходящим, но думается, что вы получите больше удовольствия и дольше сохраните интерес, если вмешаетесь и, не откладывая в долгий ящик, начнете экспериментировать с CUDA C. Исходя из этого предположения, мы в этой главе расскажем о том, какие аппаратные и программные компоненты понадобятся, чтобы приступить к работе. Спешим порадовать – все программное обеспечение бесплатно, так что деньги можете потратить на то, что вам по душе.

2.1. О чем эта глава

Прочитав эту главу, вы:

- скачаете все необходимые для работы с этой книгой программные компоненты;
- настроите среду для компиляции кода на CUDA C.

2.2. Среда разработки

Прежде чем отправляться в путешествие, необходимо настроить среду, в которой вы будете вести разработку на CUDA C. Для этого понадобятся:

- графический процессор, поддерживающий архитектуру CUDA;
- драйвер устройства NVIDIA;
- комплект средств разработки CUDA (CUDA Toolkit);
- стандартный компилятор языка C.

И сейчас мы подробнее обсудим каждое из этих условий.

2.2.1. Графические процессоры, поддерживающие архитектуру CUDA

К счастью, найти графический процессор, построенный на базе архитектуры CUDA, совсем несложно, потому что таковыми являются все NVIDIA GPU, начиная с выпущенной в 2006 году карты GeForce 8800 GTX. NVIDIA регулярно выпускает новые GPU на базе архитектуры CUDA, поэтому приведенный ниже

список подходящих GPU наверняка неполон. Но так или иначе, все перечисленные в нем GPU поддерживают CUDA.

Полный список имеется на сайте NVIDIA по адресу www.nvidia.com/cuda, хотя можно без опаски считать, что любой из недавно выпущенных (начиная с 2007 года) GPU с графической памятью объемом не менее 256 Мб годится для разработки и исполнения кода на языке CUDA C.

GeForce GTX 480	GeForce 8300 mGPU	Quadro FX 5600
GeForce GTX 470	GeForce 8200 mGPU	Quadro FX 4800
GeForce GTX 295	GeForce 8100 mGPU	Quadro FX 4800 for Mac
GeForce GTX 285	Tesla S2090	Quadro FX 4700 X2
GeForce GTX 285 for Mac	Tesla M2090	Quadro FX 4600
GeForce GTX 280	Tesla S2070	Quadro FX 3800
GeForce GTX 275	Tesla M2070	Quadro FX 3700
GeForce GTX 260	Tesla C2070	Quadro FX 1800
GeForce GTS 250	Tesla S2050	Quadro FX 1800
GeForce GT 220	Tesla M2050	Quadro FX 580
GeForce G210	Tesla C2050	Quadro FX 570
GeForce GTS 150	Tesla S1070	Quadro FX 470
GeForce GT 130	Tesla S1060	Quadro FX 380
GeForce GT 120	Tesla S870	Quadro FX 370
GeForce G100	Tesla C870	Quadro FX 370 Low Profile
GeForce 9800 GX2	Tesla D870	Quadro CX
GeForce 9800 GTX+	Продукты Quadro Mobile	Quadro NVS 450
GeForce 9800 GTX		Quadro NVS 420
GeForce 9800 GT	Quadro FX 3700M	Quadro NVS 295
GeForce 9600 GSO	Quadro FX 3600M	Quadro NVS 290
GeForce 9600 GT	Quadro FX 2700M	Quadro Plex 2100 D4
GeForce 9500 GT	Quadro FX 1700M	Quadro Plex 2100 D2
GeForce 9400GT	Quadro FX 1600M	Quadro Plex 2100 S4
GeForce 8800 Ultra	Quadro FX 770M	Quadro Plex 1000 Model IV
GeForce 8800 GTX	Quadro FX 570M	Продукты GeForce mobile
GeForce 8800 GTS	Quadro FX 370M	
GeForce 8800 GT	Quadro FX 360M	GeForce GTX 280M
GeForce 8800 GS	Quadro NVS 320M	GeForce GTX 260M
GeForce 8600 GTS	Quadro NVS 160M	GeForce GTS 260M
GeForce 8600 GT	Quadro NVS 150M	GeForce GTS 250M
GeForce 8500 GT	Quadro NVS 140M	GeForce GTS 160M
GeForce 8400 GS	Quadro NVS 135M	GeForce GTS 150M
GeForce 9400 mGPU	Quadro NVS 130M	GeForce GT 240M
GeForce 9300 mGPU	Quadro FX 5800	GeForce GT 230M
GeForce GT 130M	GeForce 9700M GTS	GeForce 9200M GS
GeForce G210M	GeForce 9700M GT	GeForce 9100M G
GeForce G110M	GeForce 9650M GS	GeForce 8800M GTS

GeForce G105M	GeForce 9600M GT	GeForce 8700M GT
GeForce G102M	GeForce 9600M GS	GeForce 8600M GT
GeForce 9800M GTX	GeForce 9500M GS	GeForce 8600M GS
GeForce 9800M GT	GeForce 9500M G	GeForce 8400M GT
GeForce 9800M GTS	GeForce 9300M GS	GeForce 8400M GS
GeForce 9800M GS	GeForce 9300M G	

2.2.2. Драйвер устройства NVIDIA

NVIDIA поставляет системное ПО, которое позволяет программам взаимодействовать с оборудованием, поддерживающим CUDA. Если NVIDIA GPU был установлен правильно, то, скорее всего, это ПО уже находится на вашей машине. Но всегда полезно проверить, не появились ли более свежие версии драйверов, поэтому мы рекомендуем зайти на сайт www.nvidia.com/cuda и щелкнуть по ссылке *Download Drivers* (Скачать драйверы). Выберите вариант, соответствующий вашей графической карте и операционной системе, в которой вы намереваетесь вести разработку.

Выполнив все инструкции по установке, вы получите систему с новейшим системным ПО компании NVIDIA.

2.2.3. Комплект средств разработки *CUDA Development Toolkit*

Имея GPU с поддержкой CUDA и драйвер устройства NVIDIA, вы уже можете исполнять откомпилированный код на CUDA. Это означает, что, скачав какое-нибудь приложение на CUDA, вы сможете выполнить его на своем графическом процессоре. Однако мы предполагаем, что вы хотите не только запускать чужой код, иначе зачем было покупать эту книгу? Если вы собираетесь разрабатывать код для GPU NVIDIA на языке CUDA C, то понадобится дополнительное программное обеспечение. Впрочем, как мы и обещали, оно не будет стоить вам ни копейки.

Подробнее об этом мы будем говорить в следующей главе, но уже сейчас отметим, что поскольку приложения, написанные на CUDA C, исполняются на двух разных процессорах, то понадобятся два компилятора. Один будет компилировать код для GPU, второй – код для CPU. NVIDIA предлагает только компилятор кода, предназначенного для GPU. Как и драйвер устройства, комплект средств разработки *CUDA Toolkit* можно скачать со страницы по адресу <http://developer.nvidia.com/object/gpucomputing.html>. Она показана на рис. 2.1.

Вас снова попросят выбрать платформу: 32- или 64-разрядную версию Windows XP, Windows Vista, Windows 7, Linux или Mac OS. Из предлагаемых продуктов необходимо скачать только CUDA Toolkit – без него вы не сможете собрать приведенные в этой книге примеры. Кроме того, мы рекомендуем, хотя это и не обязательно, скачать примеры, содержащиеся в комплекте GPU Computing SDK;

DEVELOPER ZONE

Search Developer Zone

Last Updated: 03 / 30 / 2010

Quick Links

- Home
- News
- Developer Newsletter
- Newsletter Sign-Up
- CUDA Newsletter Sign-Up
- Drivers
- Registered Developer Login
- Become a Registered Developer
- Events Calendar

Parallel Nsight™

Graphics

- DirectX
- OpenGL
- 3D Vision™
- Documentation

GPU Computing

- Downloads
- CUDA
- DirectCompute
- OpenGL
- Free GPU Computing Seminars

Tegra

NVIDIA® Application Acceleration Tools

- SceniX™
- ComplexX™
- OptiX™
- PhysX™
- Cg Toolkit

Forums

- Developer Forums
- GPU Computing Forums

NOTE: The NVIDIA Developer Forums and the GPU Computing Forums require separate logins. We will fix this in the near future when the two forums are merged. Thank you for your patience!

Contact

Legal Information

Site Feedback

CUDA 3.0 Downloads

[Click here to view all CUDA Toolkit releases](#)

Download Quick Links [Windows] [Linux] [MacOS]

Release Highlights

- Support for the new Fermi architecture, with:
 - Native 64-bit GPU support
 - Multiple Copy Engine support
 - ECC reporting
 - Concurrent Kernel Execution
 - Fermi HW debugging support in cuda-gdb
 - Fermi HW profiling support for CUDA C and OpenCL in Visual Profiler
- C++ Class Inheritance and Template Inheritance support for increased programmer productivity
- A new unified interoperability API for Direct3D and OpenGL, with support for:
 - OpenGL texture interop
 - Direct3D 11 interop support
- CUDA Driver / Runtime Buffer Interoperability, which allows applications using the CUDA Driver API to also use libraries implemented using the CUDA C Runtime such as CUFFT and CUBLAS.
- CUBLAS now supports all BLAS1, 2, and 3 routines including those for single and double precision complex numbers
- Up to 100x performance improvement while debugging applications with cuda-gdb
- cuda-gdb hardware debugging support for applications that use the CUDA Driver API
- cuda-gdb support for JIT-compiled kernels
- New CUDA Memory Checker reports misalignment and out of bounds errors, available as a stand-alone utility and debugging mode within cuda-gdb
- CUDA Toolkit libraries are now versioned, enabling applications to require a specific version, support multiple versions explicitly, etc.
- CUDA C/C++ kernels are now compiled to standard ELF format
- Support for device emulation mode has been packaged in a separate version of the CUDA C Runtime (CUDART), and is deprecated in this release. Now that more sophisticated hardware debugging tools are available and more are on the way, NVIDIA will be focusing on supporting these tools instead of the legacy device emulation functionality.
- On Windows, use the new Parallel Nsight development environment for Visual Studio, with integrated GPU debugging and profiling tools (was code-named "Nexus"). Please see [www.nvidia.com/nsight](#) for details.
- On Linux, use cuda-gdb and cuda-memcheck, and check out the solutions from Allinea and TotalView that will be available soon.
- Support for all the OpenCL features in the latest R195 production driver package:
 - Double Precision
 - Graphics Interoperability with OpenCL, Direct3D9, Direct3D10, and Direct3D11 for high performance visualization
 - OpenCL Attribute Capability, so you can target optimizations for GPU architectures (`cl_nv_device_attribute_query`)
 - Ability to control compiler optimization settings via support for pragma unroll in OpenCL kernels and an extension that allows programmers to set compiler flags. (`cl_nv_compiler_options`)
 - OpenCL Images support, for better/faster image filtering
 - 32-bit global and local atomics for fast, convenient data manipulation
 - Byte Addressable Stores, for faster video/image processing and compression algorithms
 - Support for the latest OpenCL spec revision 1.0.48 and latest official Khronos OpenCL headers as of 2010-02-17

For more information on general purpose computing features of the Fermi architecture, see: [www.nvidia.com/fermi](#).

Please review the release notes for additional important information about this release.

Note: The developer driver packages below provide baseline support for the widest number of NVIDIA products in the smallest number of installers. More recent production driver packages for end users are available at [www.nvidia.com/drivers](#).

[Windows] [Linux] [MacOS]

Рис. 2.1. Страница загрузки CUDA

среди них вы найдете десятки полезных примеров. Комплект GPU Computing SDK не рассматривается в этой книге, но включенные в него примеры удачно дополняют излагаемый материал, а когда изучаешь новый стиль программирования, то чем больше примеров, тем лучше. Имейте в виду, что хотя почти весь приведенный в этой книге код должен работать на платформах Linux, Windows и Mac OS, мы тестировали приложения только в Linux и в Windows. Пользователи Mac OS X ведут жизнь, полную опасностей, и запускают неподдерживаемые примеры программ.

2.2.4. Стандартный компилятор

Как мы уже сказали, вам понадобятся два компилятора: для кода, исполняемого на GPU и CPU. Если вы скачали и установили комплекс CUDA Toolkit, как описано в предыдущем разделе, то компилятор для GPU у вас уже есть. Осталось только решить вопрос с компилятором для CPU, и можно переходить к интересным вещам.

Windows

На платформах Microsoft Windows, включая Windows XP, Windows Vista, Windows Server 2008 и Windows 7, мы рекомендуем пользоваться компилятором Microsoft Visual Studio C. NVIDIA поддерживает семейства продуктов Visual Studio 2005 и Visual Studio 2008⁴. Когда Microsoft выпустит новые версии Visual Studio, NVIDIA, скорее всего, начнет их поддерживать, прекращая вместе с тем поддержку предыдущих версий. У многих разработчиков на C и C++ уже стоит Visual Studio 2005 или Visual Studio 2008; если вы из их числа, можете смело пропустить этот подраздел.

Если у вас нет поддерживаемой версии Visual Studio и вы не готовы покупать ее, то Microsoft предлагает на своем сайте бесплатные издания Visual Studio 2008 Express. Хотя для разработки коммерческого ПО они, как правило, не годятся, но для того, чтобы получить представление о работе с CUDA C на платформах Windows, не тратя денег на приобретение лицензии, это как раз то, что надо. Поэтому, если вам нужна Visual Studio 2008, отправляйтесь на сайт www.microsoft.com/visualstudio/!

Linux

В большинство дистрибутивов Linux, как правило, уже входит та или иная версия компилятора GNU C (gcc). Что касается CUDA 3.0, то поддерживаемые версии gcc включены в следующие дистрибутивы Linux:

- Red Hat Enterprise Linux 4.8;
- Red Hat Enterprise Linux 5.3;
- OpenSUSE 11.1;
- SUSE Linux Enterprise Desktop 11;

⁴ На данный момент также поддерживается Visual Studio 2010.

- Ubuntu 9.04;
- Fedora 10.

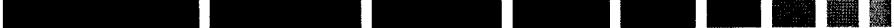
Если вы «упертый линуксоид», то, наверное, знаете, что большинство программных пакетов для Linux работают не только на «поддерживаемых» платформах. Комплект CUDA Toolkit – не исключение, так что если в этом списке нет вашего любимого дистрибутива, то все равно стоит попробовать. Совместимость определяется главным образом версией ядра, gcc и glibc.

Macintosh OS X

Собираясь вести разработку на Mac OS X, проверьте, что установлена как минимум версия Mac OS X 10.5.7. Подойдет, в частности, и версия 10.6, Mac OS X «Snow Leopard». Затем нужно будет установить компилятор gcc, скачав и поставив пакет Apple Xcode. Это ПО бесплатно предоставляется членам сообщества Apple Developer Connection (ADC), а скачать его можно со страницы по адресу <http://developer.apple.com/tools/Xcode>. Код, приведенный в этой книге, разрабатывался и тестировался на платформах Linux и Windows, но должен без изменений работать и в системах Mac OS X.

2.3. Резюме

Если вы сделали все, о чем написано в этой главе, то теперь готовы приступить к разработке на CUDA C. Быть может, вы даже поэкспериментировали с какими-то примерами из комплекта GPU Computing SDK, который скачали с сайта NVIDIA. Если так, аплодируем вашей готовности к встрече с неизведанным! Если нет, тоже не страшно. Все необходимое вы найдете в этой книге. Как бы то ни было, вам, наверное, не терпится написать первую программу на CUDA C, так не будем откладывать.



Глава 3. Введение в CUDA C

Если вы прочитали главу 1, то, надеемся, прониклись нашей убежденностью в том, что графические процессоры обладают колossalной вычислительной мощью и нужен лишь программист, который ее обуздает. А если вы не пропустили главу 2, то теперь имеете среду, настроенную для компиляции и запуска кода, написанного на языке CUDA C. Если же вы пропустили первые главы – быть может, потому что просто проглядываете примеры или листаете книгу в магазине и открыли ее на случайной странице или умираете от желания поскорее начать, – тоже не страшно (мы никому не скажем). В любом случае, если вы готовы к встрече с первыми примерами кода, начнем.

3.1. О чем эта глава

Прочитав эту главу, вы:

- напишете свой первый код на CUDA C;
- узнаете, в чем разница между кодом для CPU и для *устройства*;
- научитесь запускать код для устройства из программы для CPU;
- узнаете о том, как можно использовать память устройств, поддерживающих CUDA;
- научитесь получать от системы информацию о наличии устройств, поддерживающих CUDA.

3.2. Первая программа

Поскольку мы собираемся изучать язык CUDA C на примерах, то и начнем с примера. Чтобы не отступать от традиций написания книг по программированию, в качестве первого примера мы возьмем программу «Здравствуй, мир!».

3.2.1. Здравствуй, мир!

```
#include "../common/book.h"

int main( void ) {
    printf( "Здравствуй, мир!\n" );
    return 0;
}
```

Наверняка вы сейчас подумали, что авторы вас попросту надули. Это ведь обычный С, не так ли? А существует ли вообще CUDA С? Ответ на оба вопроса утвердительный, и эта книга – не ловкое мошенничество. Пример программы «Здравствуй, мир!» приведен только для того, чтобы показать: между CUDA С и «стандартным С, к которому вы привыкли, нет фундаментальных различий.

Своей простотой пример обязан тому факту, что работает целиком на основном процессоре – *CPU* (*host*). В этой книге мы будем называть GPU вкупе с его памятью *устройством* (*device*)¹. Этот пример так сильно напоминает код, который мы писали раньше, потому что вообще игнорирует какие-либо вычислительные устройства, кроме CPU.

Чтобы избавить вас от горького ощущения, что вы потратили деньги на дорогостоящее собрание банальностей, мы будем постепенно расширять этот пример. Для начала взглянем на нечто, способное использовать GPU (*устройство*) для исполнения кода. Функция, исполняемая на устройстве, обычно называется *ядром* (*kernel*).

3.2.2. Вызов ядра

Сейчас на основе предыдущего примера мы напишем код, который уже не так напоминает простенькую программу «Здравствуй, мир!».

```
#include <stdio.h>

__global__ void kernel( void ) {
}

int main( void ) {
    kernel<<<1,1>>>();
    printf( "Здравствуй, мир!\n" );
    return 0;
}
```

В этой программе есть два добавления к первоначальному примеру:

- пустая функция *kernel()* с квалификатором *_global_*;
- вызов этой пустой функции, украшенный синтаксисом *<<<1,1>>>*.

В предыдущем разделе мы видели, что по умолчанию код обрабатывается «стандартным компилятором С». Так, в Linux код для CPU будет компилировать GNU gcc, а в Windows – Microsoft Visual C. Инструментальные средства NVIDIA просто передают код компилятору для CPU, и все работает так, будто никакой CUDA в мире не существует.

Теперь же мы видим, что CUDA С дополняет стандартный язык С квалификатором *_global_*. Тем самым мы говорим компилятору, что эта функция должна

¹ В англоязычной литературе для обозначения CPU используется термин *host*, однако в книге будет использоваться термин *CPU*.

компилироваться для исполнения устройством, а не CPU. В данном случае nvcc передает функцию `kernel()` компилятору, обрабатывающему код для устройства, а функцию `main()` – компилятору для CPU, как в предыдущем примере.

Но что это за таинственное обращение к `kernel()` и почему мы должны уродовать стандартный синтаксис С угловыми скобками и набором чисел? Готовьтесь, именно здесь и творится волшебство.

Мы видели, что CUDA C нуждается в каком-то синтаксическом способе пометить, что функция должна исполняться устройством. Ничего особенного тут нет; это просто признак, позволяющий отправлять код для CPU одному компилятору, а код для устройства – другому. Настоящий трюк в том, чтобы вызвать код для устройства из кода, исполняемого CPU. Одно из достоинств CUDA C заключается в том, что он обеспечивает языковую интеграцию, благодаря которой вызовы функций для устройства выглядят очень похоже на вызовы функций для CPU. Позже мы обсудим, что реально происходит за кулисами, а пока достаточно знать, что компилятор и исполняющая среда CUDA берут на себя заботу о вызове кода для устройства из программы, исполняемой CPU.

Итак, странный вызов служит для обращения к коду, исполняемому устройством, но к чему все-таки эти угловые скобки и числа? Угловыми скобками обозначаются аргументы, которые мы собираемся передать исполняющей среде. Это не аргументы функции, исполняемой устройством, а параметры, влияющие на то, как исполняющая среда будет запускать код для устройства. Разговор об этих параметрах мы отложим до следующей главы. Аргументы же самой функции передаются в круглых скобках – как обычно.

3.2.3. Передача параметров

Мы обещали рассказать о том, как передаются параметры ядру, и теперь пришло время исполнить обещание. Рассмотрим следующую модификацию программы «Здравствуй, мир!»:

```
#include <stdio.h>
#include "../common/book.h"

__global__ void add( int a, int b, int *c ) {
    *c = a + b;
}

int main( void ) {
    int c;
    int *dev_c;
    HANDLE_ERROR( cudaMalloc( (void**)&dev_c, sizeof(int) ) );
    add<<<1>>>( 2, 7, dev_c );
    HANDLE_ERROR( cudaMemcpy( &c,
                           dev_c,
```

```
        sizeof(int),  
        cudaMemcpyDeviceToHost ) );  
    printf( "2 + 7 = %d\n", c );  
    cudaFree( dev_c );  
  
    return 0;  
}
```

Новых строк здесь много, но новых концепций всего две:

- ❑ мы можем передавать параметры ядру, как любой другой функции на С;
- ❑ мы должны выделить память, если хотим, чтобы устройство сделало нечто полезное, например вернуло данные CPU.

Ничего особенного в передаче параметров ядру нет. Несмотря на угловые скобки, вызов ядра выглядит и работает как вызов любой другой написанной на стандартном С функции. Исполняющая среда берет на себя заботу о передаче параметров от CPU устройству.

Более интересно выделение памяти с помощью функции `cudaMalloc()`. Она очень похожа на стандартную функцию `malloc()`, но говорит исполняющей среде CUDA, что память должна быть выделена на устройстве. Первый аргумент – это указатель на указатель, в котором будет возвращен адрес выделенной области памяти, а второй – размер этой области. Если не считать того, что указатель на выделенную область не возвращается функцией в виде значения, то поведение ничем не отличается от `malloc()`, даже тип возвращаемого значения `void*` совпадает. Конструкция `HANDLE_ERROR()`, окружающая обращения, – это служебный макрос, определенный в коде, прилагаемом к книге. Когда функция возвращает ошибку, он печатает сообщение и завершает приложение с кодом `EXIT_FAILURE`. Хотя никто не запрещает вам использовать этот макрос и в своих программах, скорее всего, такой стратегии обработки ошибок в промышленном коде будет недостаточно.

И тут возникает тонкий, но важный момент. Своей простотой и мощью язык CUDA C во многом обязан стиранию грани между кодом для CPU и для устройства. Однако программист не должен разыменовывать указатель, возвращаемый `cudaMalloc()`, в коде, исполняемом CPU. Этот указатель можно передавать другим функциям, выполнять с ним арифметические операции и даже приводить к другому типу. Но ни читать, ни писать в эту область памяти нельзя.

К сожалению, компилятор не может защитить от такой ошибки. Он с радостью позволит разыменовать указатель на память устройства в коде, исполняемом CPU, потому что синтаксически этот указатель ничем не отличается от любого другого. (Формулируем правила использования указателей на память устройства:

Разрешается передавать указатели на память, выделенную `cudaMalloc()`, функциям, исполняемым устройством.

Разрешается использовать указатели на память, выделенную `cudaMalloc()`, для чтения и записи в эту память в коде, который исполняется устройством.

Разрешается передавать указатели на память, выделенную `cudaMalloc()`, функциям, исполняемым CPU.

Не разрешается использовать указатели на память, выделенную cudaMalloc(), для чтения и записи в эту память в коде, который исполняется CPU.

Если вы читаете внимательно, то, наверное, предвидите и следующее правило: нельзя использовать стандартную функцию `free()` для освобождения памяти, выделенной функцией `cudaMalloc()`. Чтобы освободить память, выделенную `cudaMalloc()`, следует обращаться к функции `cudaFree()`, которая ведет себя точно так же, как `free()`.

Мы видели, как выделять и освобождать память устройства, но при этом столкнулись с печальным фактом: модифицировать эту память в коде, исполняемом CPU, нельзя. В следующих двух строчках примера иллюстрируются два наиболее употребительных способа доступа к памяти устройства: использование возвращенных указателей в коде, исполняемом самим устройством, и обращение к функции `cudaMemcpy()`.

Внутри кода, исполняемого устройством, указатели используются точно так же, как в обычном коде на C, который исполняется CPU. Оператор `*c = a + b` не таит в себе никаких подвохов. Мы складываем параметры `a` и `b` и сохраняем результат в области памяти, на которую указывает `c`. Это настолько просто, что даже неинтересно.

Мы перечислили способы правильного и неправильного использования указателей на память устройства. Все то же самое относится и к указателям на память CPU. Передавать такие указатели между исполняемыми устройством функциями разрешается, но попытка обратиться к памяти CPU, адресуемой указателем, из кода, исполняемого устройством, закончится катастрофой. Итак, указатели на память CPU можно использовать для доступа к памяти из кода, исполняемого CPU, а указатели на память устройства – из кода, исполняемого устройством.

Но к памяти устройства можно обратиться также с помощью вызова функции `cudaMemcpy()` из кода, исполняемого CPU. Она похожа на стандартную функцию `memcpy()`, но принимает дополнительный параметр, который говорит о том, какой из двух указателей адресует память устройства. В нашем примере последний параметр `cudaMemcpy()` равен `cudaMemcpyDeviceToHost`, то есть начальный указатель адресует память устройства, а конечный – память CPU.

Вряд ли вы удивитесь, узнав, что константа `cudaMemcpyHostToDevice` описывает противоположную ситуацию, когда исходные данные находятся в памяти CPU, а конечный адрес – в памяти устройства. Наконец, можно сказать, что *оба* указателя относятся к памяти устройства, для этого нужно передать константу `cudaMemcpyDeviceToDevice`. Если же и начальный, и конечный указатели адресуют память CPU, то копирование производится стандартной функцией `memcpy()`.

3.3. Получение информации об устройстве

Поскольку мы хотели бы выделять память и выполнять код на нашем устройстве, то программе было бы полезно знать, сколько у устройства памяти и что оно умеет делать. К тому же часто в компьютере установлено несколько устройств

с поддержкой CUDA. В таких ситуациях крайне желательно отличать один GPU от другого.

Например, во многие материнские платы уже интегрирован графический процессор NVIDIA. Если производитель или пользователь поставит в такой компьютер еще и дискретный графический процессор, то окажется, что процессоров с поддержкой CUDA два. А некоторые продукты NVIDIA, в частности карта GeForce GTX 295, изначально оснащены двумя GPU. Если в компьютере находится такая карта, то она тоже будет распознаваться как два процессора с поддержкой CUDA.

Прежде чем приступить к написанию кода для устройства, хорошо бы иметь механизм, позволяющий узнать, сколько устройств присутствует и какие возможности поддерживает каждое из них. К счастью, для получения подобной информации есть очень простой интерфейс. Первым делом нужно узнать, сколько в системе есть устройств, построенных на базе архитектуры CUDA. Они и смогут исполнять ядра, написанные на CUDA C. Узнать число CUDA-устройств позволяет функция `cudaGetDeviceCount()`. Нет нужды говорить, что мы претендуем на получение приза за лучшее имя функции.

```
int count;
HANDLE_ERROR( cudaGetDeviceCount( &count ) );
```

После возврата из `cudaGetDeviceCount()` мы можем перебрать имеющиеся устройства и запросить информацию о каждом из них. Исполняющая среда CUDA возвращает свойства устройства в структуре типа `cudaDeviceProp`. Какие свойства мы можем получить? В версии CUDA 3.0¹ структура `cudaDeviceProp` объявлена следующим образом:

```
struct cudaDeviceProp {
    char name[256];
    size_t totalGlobalMem;
    size_t sharedMemPerBlock;
    int regsPerBlock;
    int warpSize;
    size_t memPitch;
    int maxThreadsPerBlock;
    int maxThreadsDim[3];
    int maxGridSize[3];
    size_t totalConstMem;
    int major;
    int minor;
    int clockRate;
    size_t textureAlignment;
    int deviceOverlap;
    int multiProcessorCount;
```

¹ На момент издания перевода книги доступна уже версия CUDA 4.0 и в ней в эту структуру добавлен ряд новых полей.

```

int kernelExecTimeoutEnabled;
int integrated;
int canMapHostMemory;
int computeMode;
int maxTexture1D;
int maxTexture2D[2];
int maxTexture3D[3];
int maxTexture2DArray[3];
int concurrentKernels;
}

```

Некоторые поля не нуждаются в пояснениях, для остальных в табл. 3.1 приведено краткое описание.

Таблица 3.1. Свойства CUDA-устройства

Свойство устройства	Описание
char name[256]	ASCII-строка, идентифицирующая устройство (например, «GeForce GTX 280»)
size_t totalGlobalMem	Объем глобальной памяти устройства в байтах
size_t sharedMemPerBlock	Максимальный размер разделяемой памяти в одном блоке (в байтах)
int regsPerBlock	Количество 32-разрядных регистров в одном блоке
int warpSize	Количество нитей в варпе
size_t memPitch	Максимальный шаг (pitch), разрешенный для копирования в памяти (в байтах)
int maxThreadsPerBlock	Максимальное количество нитей в одном блоке
int maxThreadsDim[3]	Максимальное количество нитей вдоль каждого измерения блока
int maxGridSize[3]	Максимальное количество блоков вдоль каждого измерения сетки
size_t totalConstMem	Объем имеющейся константной памяти
int major	Старшая часть уровня вычислительных возможностей устройства
int minor	Младшая часть уровня вычислительных возможностей устройства
size_t textureAlignment	Требования, предъявляемые устройством к выравниванию текстур
int deviceOverlap	Булевский флаг, показывающий, может ли устройство одновременно выполнять функцию cudaMemcpy() и ядро
int multiProcessorCount	Количество мультипроцессоров в устройстве
int kernelExecTimeoutEnabled	Булевский флаг, показывающий, существует ли ограничение на время исполнения ядер устройством
int integrated	Булевский флаг, показывающий, является ли устройство интегрированным GPU (то есть частью набора микросхем, а не дискретным GPU)
int canMapHostMemory	Булевский флаг, показывающий, может ли устройство отображать память CPU на адресное пространство CUDA-устройства

Таблица 3.1. Свойства CUDA-устройства (окончание)

Свойство устройства	Описание
<code>int computeMode</code>	Режим вычислений для данного устройства: по умолчанию, исключительный, запрещено
<code>int maxTexture1D</code>	Максимальный поддерживаемый размер одномерных текстур
<code>int maxTexture2D[2]</code>	Максимальный поддерживаемый размер (для каждого измерения) для массива двухмерных текстур
<code>int maxTexture3D[3]</code>	Максимальный поддерживаемый размер (по каждому измерению) для трехмерных текстур
<code>int maxTexture2DArray[3]</code>	Максимальное количество измерений для массивов двухмерных текстур
<code>int concurrentKernels</code>	Булевский флаг, показывающий, поддерживает ли устройство одновременное исполнение нескольких ядер в одном контексте

Мы не хотели бы проваливаться в кроличью нору слишком глубоко или слишком быстро, поэтому пока отложим обсуждение этих свойств. На самом деле в представленном выше списке некоторые важные подробности опущены, поэтому имеет смысл поискать дополнительную информацию в руководстве *NVIDIA CUDA Programming Guide*. Когда вы начнете сами писать приложения, эти свойства окажутся весьма полезными. Ну а пока просто покажем, как опросить каждое устройство и распечатать его свойства. Опрос устройств мог бы выглядеть примерно так:

```
#include "../common/book.h"

int main( void ) {
    cudaDeviceProp prop;

    int count;
    HANDLE_ERROR( cudaGetDeviceCount( &count ) );
    for (int i=0; i< count; i++) {
        HANDLE_ERROR( cudaGetDeviceProperties( &prop, i ) );

        // Сделать что-то со свойствами устройства
    }
}
```

А зная, какие поля имеются в структуре, мы можем конкретизировать комментарий «Сделать что-то...» и написать что-то не столь тривиальное:

```
#include "../common/book.h"

int main( void ) {
    cudaDeviceProp prop;

    int count;
    HANDLE_ERROR( cudaGetDeviceCount( &count ) );
```

```

for (int i=0; i< count; i++) {
    HANDLE_ERROR( cudaGetDeviceProperties( &prop, i ) );
    printf( " --- Общая информация об устройстве %d ---\n", i );
    printf( "Имя: %s\n", prop.name );
    printf( "Вычислительные возможности: %d.%d\n", prop.major, prop.minor );
    printf( "Тактовая частота: %d\n", prop.clockRate );
    printf( "Перекрытие копирования: " );
    if (prop.deviceOverlap)
        printf( "Разрешено\n" );
    else
        printf( "Запрещено\n" );
    printf( "Тайм-аут выполнения ядра : " );
    if (prop.kernelExecTimeoutEnabled)
        printf( "Включен\n" );
    else
        printf( "Выключен\n" );

    printf( " --- Информация о памяти для устройства %d ---\n", i );
    printf( "Всего глобальной памяти: %ld\n", prop.totalGlobalMem );
    printf( "Всего константной памяти: %ld\n", prop.totalConstMem );
    printf( "Максимальный шаг: %ld\n", prop.memPitch );
    printf( "Выравнивание текстур: %ld\n", prop.textureAlignment );

    printf( " --- Информация о мультипроцессорах для устройства %d ---\n", i );
    printf( "Количество мультипроцессоров: %d\n",
           prop.multiProcessorCount );
    printf( "Разделяемая память на один МП: %ld\n", prop.sharedMemPerBlock );
    printf( "Регистров на один МП: %d\n", prop.regsPerBlock );
    printf( "Нитей в варпе: %d\n", prop.warpSize );
    printf( "Макс. количество нитей в блоке: %d\n",
           prop.maxThreadsPerBlock );
    printf( "Макс. количество нитей по измерениям: (%d, %d, %d)\n",
           prop.maxThreadsDim[0], prop.maxThreadsDim[1],
           prop.maxThreadsDim[2] );
    printf( "Максимальные размеры сетки: (%d, %d, %d)\n",
           prop.maxGridSize[0], prop.maxGridSize[1],
           prop.maxGridSize[2] );
    printf( "\n" );
}
}

```

3.4. Использование свойств устройства

Ну а помимо написания приложения, которое красиво распечатывает все свойства устройства с поддержкой CUDA, для чего еще можно их использовать? Поскольку мы как разработчики программы хотели бы, чтобы она работала

максимально быстро, интересно было бы выбрать для исполнения программы CPU с наибольшим числом мультипроцессоров. А если ядро должно тесно взаимодействовать с CPU, то хорошо бы исполнять программу на интегрированном GPU, который разделяет системную память с CPU. То и другое можно узнать, вызвав функцию `cudaGetDeviceProperties()`.

Предположим, что наша программа зависит от поддержки чисел с плавающей точкой двойной точности. Заглянув в приложение A руководства *NVIDIA CUDA Programming Guide*, мы узнаем, что арифметику двойной точности для чисел с плавающей точкой поддерживают устройства с уровнем вычислительных возможностей не ниже 1.3. Поэтому, чтобы наша программа заработала, необходимо найти хотя бы одно такое устройство.

С помощью функций `cudaGetDeviceCount()` и `cudaGetDeviceProperties()` мы могли бы перебрать все устройства и найти те, для которых старшая часть уровня больше 1 или старшая часть равна единице, а дополнительная – не меньше 3. Но поскольку эту относительно простую процедуру выполнять скучно, исполняющая среда CUDA позволяет ее автоматизировать. Сначала запишем в структуру `cudaDeviceProp` свойства, которыми должно обладать устройство:

```
cudaDeviceProp prop;
memset( &prop, 0, sizeof( cudaDeviceProp ) );
prop.major = 1;
prop.minor = 3;
```

Затем передадим эту структуру функции `cudaChooseDevice()`, поручив исполняющей среде CUDA найти устройство, удовлетворяющее заданным ограничениям. Эта функция вернет идентификатор устройства, который мы затем сможем передать функции `cudaSetDevice()`. Начиная с этого момента, все операции, относящиеся к устройству, будут производиться на устройстве, обнаруженному функцией `cudaChooseDevice()`.

```
#include "../common/book.h"

int main( void ) {
    cudaDeviceProp prop;
    int dev;

    HANDLE_ERROR( cudaGetDevice( &dev ) );
    printf( "Ид текущего CUDA-устройства: %d\n", dev );

    memset( &prop, 0, sizeof( cudaDeviceProp ) );
    prop.major = 1;
    prop.minor = 3;
    HANDLE_ERROR( cudaChooseDevice( &dev, &prop ) );
    printf( "Ид CUDA-устройства, ближайшего к уровню 1.3: %d\n", dev );
    HANDLE_ERROR( cudaSetDevice( dev ) );
}
```

Системы с несколькими GPU получают все большее распространение. Например, многие наборы микросхем для материнских плат производства NVIDIA содержат интегрированный GPU с поддержкой CUDA. Если в компьютер с такой материнской платой добавить еще и дискретный GPU, то получится платформа с несколькими GPU. Кроме того, разработанная NVIDIA технология SLI позволяет устанавливать в один компьютер сразу несколько дискретных GPU. В таком случае приложение может, руководствуясь теми или иными соображениями, предпочесть один GPU другому. Если приложению необходима поддержка процессором определенных возможностей или оно хотело бы работать на самом быстром из имеющихся GPU, то не забывайте об описанном выше API, поскольку нет никакой гарантии, что исполняющая среда по умолчанию выберет именно тот GPU, который лучше всего подходит вашему приложению.

3.5. Резюме

Наконец-то мы что-то написали на CUDA C, и это оказалось совсем на так страшно, как вы, возможно, ожидали. В принципе, CUDA C – это стандартный язык C, пополненный кое-какими средствами, позволяющими указать, какой код должен исполняться устройством, а какой – CPU. Добавив ключевое слово `_global_`, мы говорим компилятору, что функция должна исполняться GPU. Мы также узнали, что для доступа к памяти GPU в CUDA C имеются функции, похожие на стандартные `malloc()`, `memcpy()` и `free()`. Их аналоги в CUDA – `cudaMalloc()`, `cudaMemcpy()` и `cudaFree()` – позволяют выделить память устройства, скопировать данные между устройством и CPU и освободить память устройства, когда в ней отпала необходимость.

Дальше мы увидим более интересные примеры эффективного использования устройства в качестве массивно-параллельного сопроцессора. Но уже сейчас вы должны усвоить, что ничего сверхъестественного в языке CUDA C нет. В следующей главе мы покажем, как просто выполнить параллельный код на GPU.



Глава 4. Параллельное программирование на CUDA C

В предыдущей главе мы видели, как просто написать код, исполняемый графическим процессором. Мы даже научились складывать два числа, правда, пока только 2 и 7. Признаем, что этот пример мало кого может впечатлить. Но надеемся, вы убедились, что хотя бы поначалу в CUDA C нет ничего сложного, и горите желанием учиться дальше. Перспективы использования GPU-вычислений в основном связаны с массивно-параллельной структурой многих задач. Поэтому мы посвятим эту главу вопросу об исполнении на GPU параллельного кода, написанного на CUDA C.

4.1. О чем эта глава

Прочитав эту главу, вы:

- узнаете один из основных способов организации параллелизма в CUDA;
- напишете свою первую параллельную программу на CUDA C.

4.2. Параллельное программирование в CUDA

Выше мы видели, как просто выполнить стандартную С-функцию на устройстве. Достаточно всего лишь добавить квалификатор `_global_` и вызвать ее, применяя специальный синтаксис с угловыми скобками. Чрезвычайно просто, но и чрезвычайно неэффективно, поскольку инженеры NVIDIA оптимизировали графические процессоры для выполнения сотен вычислений параллельно. Однако до сих пор мы запустили ядро, выполняемое последовательно. В этой главе мы увидим, как запустить ядро, которое будет производить вычисления параллельно.

4.2.1. Сложение векторов

Мы рассмотрим простой пример, на котором продемонстрируем использование нитей¹ в программе на CUDA C. Предположим, что есть два списка чисел, мы

¹ В русскоязычной технической литературе слово «thread» принято переводить как «поток», соответственно, «multithreaded» – «многопоточный». Однако при этом возникает конфликт со словом «stream», которое также переводится как «поток». Поскольку в этой книге оба слова встречаются в одном контексте, то принято решение переводить «thread», как «нить». Собственно, это и есть изначальный смысл этого слова, который обыгрывается в смежных терминах. Например, образование, более мелкое, чем нить, называется, *fiber* (волокно), а более крупное (группа нитей) – *warp* (варп, канат). – *Прим. перев.*



Хотим сложить соответственные элементы из каждого списка и сохранить результат в третьем списке. Весь процесс показан на рис. 4.1. Если вы чуть-чуть знакомы с линейной алгеброй, то легко распознаете в этой операции сложение двух векторов.

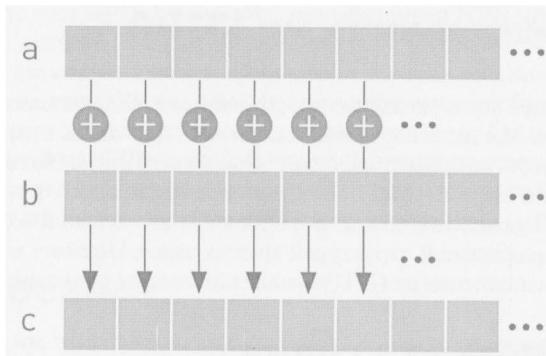


Рис. 4.1. Сложение двух векторов

Сложение векторов на CPU

Сначала покажем, как реализуется сложение векторов в традиционной программе на C:

```
#include "../common/book.h"

#define N 10

void add( int *a, int *b, int *c ) {
    int tid = 0;      // это CPU с номером 0, поэтому начинаем с нуля
    while (tid < N) {
        c[tid] = a[tid] + b[tid];
        tid += 1;    // у нас всего один CPU, поэтому увеличиваем на 1
    }
}

int main( void ) {
    int a[N], b[N], c[N];

    // заполняем массивы 'a' и 'b' на CPU
    for (int i=0; i<N; i++) {
        a[i] = -i;
        b[i] = i * i;
    }

    add( a, b, c );

    // выводим результат
```

```

for (int i=0; i<N; i++) {
    printf( "%d + %d = %d\n", a[i], b[i], c[i] );
}
return 0;
}
  
```

Большая часть кода не нуждается в пояснениях, но функцию add() мы все же рассмотрим и объясним, зачем мы ее так усложнили.

```

void add( int *a, int *b, int *c ) {
    int tid = 0;      // это CPU с номером 0, поэтому начинаем с нуля
    while (tid < N) {
        c[tid] = a[tid] + b[tid];
        tid += 1;     // у нас всего один CPU, поэтому увеличиваем на 1
    }
}
  
```

Сумма вычисляется в цикле while, где индекс tid пробегает значения от 0 до N-1. Мы складываем соответственные элементы массивов a[] и b[] и помещаем результат в элемент массива c[]. Обычно такой код записывают проще:

```

void add( int *a, int *b, int *c ) {
    for (i=0; i < N; i++) {
        c[i] = a[i] + b[i];
    }
}
  
```

А наш более запутанный вариант рассчитан для показа возможности распараллеливания кода в системе с несколькими CPU или процессорными ядрами. Например, при наличии двухъядерного процессора можно было бы увеличивать индекс на 2, при этом первое ядро инициализировало бы tid значением 0, а второе – значением 1. Тогда первое ядро складывало бы элементы с четными индексами, а второе – с нечетными. И каждое из двух ядер выполняло бы такой код:

Ядро 1	Ядро 2
<pre> void add(int *a, int *b, int *c) { int tid = 0; while (tid < N) { c[tid] = a[tid] + b[tid]; tid += 2; } } </pre>	<pre> void add(int *a, int *b, int *c) { int tid = 1; while (tid < N) { c[tid] = a[tid] + b[tid]; tid += 2; } } </pre>

Разумеется, чтобы в действительности проделать это на CPU, пришлось бы написать гораздо больше. Нужно было бы добавить инфраструктурный код для создания рабочих нитей, исполняющих функцию add(), а также допустить, что

нити и вправь выполняются параллельно. Правда, это допущение об алгоритме планирования верно не всегда.

Сложение векторов на GPU

То же самое сложение можно реализовать на GPU, представив add() в виде функции устройства. Это будет выглядеть примерно так, как код из предыдущей главы. Но прежде чем перейти к коду, исполняемому устройством, рассмотрим функцию main(). Хотя ее реализация для GPU отличается от версии для CPU, ничего существенно нового здесь нет:

```
#include "../common/book.h"

#define N 10

int main( void ) {
    int a[N], b[N], c[N];
    int *dev_a, *dev_b, *dev_c;

    // выделить память на GPU
    HANDLE_ERROR( cudaMalloc( (void**)&dev_a, N * sizeof(int) ) );
    HANDLE_ERROR( cudaMalloc( (void**)&dev_b, N * sizeof(int) ) );
    HANDLE_ERROR( cudaMalloc( (void**)&dev_c, N * sizeof(int) ) );

    // заполняем массивы 'a' и 'b' на CPU
    for (int i=0; i<N; i++) {
        a[i] = -i;
        b[i] = i * i;
    }

    // копируем массивы 'a' и 'b' на GPU
    HANDLE_ERROR( cudaMemcpy( dev_a, a, N * sizeof(int),
                           cudaMemcpyHostToDevice ) );
    HANDLE_ERROR( cudaMemcpy( dev_b, b, N * sizeof(int),
                           cudaMemcpyHostToDevice ) );

    add<<<N,1>>>( dev_a, dev_b, dev_c );

    // копируем массив 'c' с GPU на CPU
    HANDLE_ERROR( cudaMemcpy( c, dev_c, N * sizeof(int),
                           cudaMemcpyDeviceToHost ) );

    // выводим результат
    for (int i=0; i<N; i++) {
        printf( "%d + %d = %d\n", a[i], b[i], c[i] );
    }

    // освобождаем память, выделенную на GPU
    cudaFree( dev_a );
```

```

    cudaFree( dev_b );
    cudaFree( dev_c );

    return 0;
}

```

Еще раз обращаем ваше внимание на общие приемы.

- ❑ Мы выделяем три массива в памяти устройства, обращаясь к функции `cudaMalloc()`: два массива, `dev_a` и `dev_a`, будут содержать исходные данные, а третий, `dev_c`, – результат.
- ❑ Поскольку мы заботимся об окружающей среде, то перед уходом прибираем за собой, вызывая функцию `cudaFree()`.
- ❑ С помощью функции `cudaMemcpy()` мы копируем входные данные в память устройства, передавая четвертым параметром константу `cudaMemcpyHostToDevice`, а результат копируем назад в память CPU, передавая константу `cudaMemcpyDeviceToHost`.
- ❑ Мы вызываем функцию `add()`, исполняемую устройством, из функции `main()`, исполняемой CPU. При этом используется синтаксис с тремя угловыми скобками.

По ходу дела может возникнуть вопрос: почему входные массивы заполняются на CPU. Нет никакой причины, *обязывающей* нас поступать именно так. На самом деле программа даже будет работать быстрее, если заполнять массивы на GPU. Но мы хотели всего лишь показать, как реализуется конкретная операция – сложение векторов – на графическом процессоре. Поэтому просто представьте, что это небольшой фрагмент более крупного приложения, в котором входные массивы `a[]` и `b[]` генерируются каким-то алгоритмом или загружаются с диска. Короче говоря, будет достаточно притвориться, что данные появились из ниоткуда, а наша задача – что-то сделать с ними.

Пойдем дальше. Наша функция `add()` похожа на ту, что была написана для CPU:

```

__global__ void add( int *a, int *b, int *c ) {
    int tid = blockIdx.x;           // обработать данные, находящиеся по этому индексу
    if (tid < N)
        c[tid] = a[tid] + b[tid];
}

```

И снова мы видим общий прием.

- ❑ Мы написали функцию `add()`, которая исполняется устройством. Для этого мы взяли соответствующий код на C и добавили перед именем функции квалификатор `_global_`.

До сих пор в этом примере не было ничего нового, разве что он умеет складывать не только числа 2 и 7. Однако стоит отметить две детали: параметры в угловых скобках и код, содержащийся в ядре. То и другое – новые концепции.

До этого момента мы запускали ядра так:

```
kernel<<<1,1>>>( param1, param2, ... );
```

Но в этом примере в угловых скобках находится число, отличное от 1:

```
add<<<N,1>>>( dev _ a, dev _ b, dev _ c );
```

Ну и что?

Напомним, что мы еще не объяснили смысл двух чисел в угловых скобках, ограничившись туманным замечанием о том, что это параметры, которые говорят исполняющей среде, как запускать ядро. Так вот, первый из этих параметров задает количество параллельных блоков, в которых устройство должно выполнять наше ядро. В данном случае в качестве этого параметра мы передали значение N.

Например, если при запуске ядра мы пишем `kernel<<<2,1>>>()`, то можете считать, что исполняющая среда создает два ядра и исполняет их параллельно. Каждый параллельно выполняемый экземпляр называется **блоком (block)**. Запись `kernel<<<256,1>>>()` означает, что GPU будет выполнять 256 блоков. Никогда еще параллельное программирование не было таким простым делом.

Но тогда возникает законный вопрос: GPU исполняет N копий кода ядра, но как внутри кода определить, в каком блоке он исполняется? И тут мы подходим ко второй особенности нашего примера – самому коду ядра. Точнее, переменной `blockIdx.x`:

```
_global_ void add( int *a, int *b, int *c ) {
    int tid = blockIdx.x;           // обработать данные, находящиеся по этому индексу
    if (tid < N)
        c[tid] = a[tid] + b[tid];
}
```

На первый взгляд кажется, что компилятор должен сообщить об ошибке, потому что мы присваиваем значение переменной `tid`, но нигде не объявляем переменную `blockIdx`. Однако объявлять переменную `blockIdx` и не нужно; это одна из встроенных переменных, предоставляемых исполняющей средой CUDA. Более того, мы используем ее в полном соответствии с ее именем. Она содержит индекс того блока устройства, в котором исполняется текущий код.

Но почему, спросите вы, не просто `blockIdx.x`? Почему `blockIdx.x`? А потому что CUDA C позволяет определять двумерную группу блоков. Для двумерных задач, например операций над матрицами или обработки изображений, часто удобно использовать двумерную индексацию, чтобы не переходить каждый раз от линейных индексов к прямоугольным. Не переживайте, если вы не сталкивались с такими задачами; просто поверьте, что иногда двумерная индексация удобнее одномерной. Но никто не заставляет вас ей пользоваться. Не хотите – не надо: мы не обидимся.

При запуске ядра мы задали N в качестве числа параллельных блоков. Набор параллельных блоков называется **сеткой (grid)**. Таким образом, мы говорим ис-

полняющей среде, что хотим получить одномерную сетку из N блоков (скалярные значения считаются одномерными). Значение blockIdx.x для нитей в этих блоках изменяется от 0 до $N-1$. Представьте себе четыре блока, каждый из которых исполняет один и тот же код, но значения blockIdx.x в них разные. Вот как выглядит фактически исполняемый в этих блоках код после того, как исполняющая среда подставит вместо blockIdx.x индекс соответствующего блока:

Блок 1	Блок 2
<pre>__global__ void add(int *a, int *b, int *c) { int tid = 0; if (tid < N) c[tid] = a[tid] + b[tid]; }</pre>	<pre>__global__ void add(int *a, int *b, int *c) { int tid = 1; if (tid < N) c[tid] = a[tid] + b[tid]; }</pre>
Блок 3	Блок 4
<pre>__global__ void add(int *a, int *b, int *c) { int tid = 3; if (tid < N) c[tid] = a[tid] + b[tid]; }</pre>	<pre>__global__ void add(int *a, int *b, int *c) { int tid = 4; if (tid < N) c[tid] = a[tid] + b[tid]; }</pre>

Вспомните пример для CPU, с которого мы начали. Тогда для сложения двух векторов нам пришлось обходить элементы с индексами от 0 до $N-1$. Но поскольку исполняющая среда уже запустила ядро в N блоках и в каждом блоке представлен ровно один индекс, то практически вся работа уже сделана. А так как мы ленивы, то это не может не радовать. У нас остается время написать что-нибудь в свой блог, например о том, насколько мы ленивы.

И последний вопрос: зачем мы проверяем, что переменная tid меньше N ? Она ведь обязана быть меньше N , потому что ядро запускалось так, что это условие непременно выполняется. Но стремление лениться имеет оборотную сторону – навязчивую боязнь того, что кто-то нарушит предположения, сделанные нами о поведении программы. А если предположения нарушены, то код перестает работать. А это извещения об ошибках, сидение допоздна в компании отладчика и вообще куча ненужной работы, стоящей между нами и нашим блогом. Если мы не проверим, что tid меньше N и из-за этого заберемся в чужую память, ничего хорошего не жди. На самом деле это, скорее всего, приведет к насильственному завершению ядра, потому что в GPU встроены изощренные схемы управления памятью, которые завершают процессы, нарушающие правила работы с памятью.

Если вы столкнетесь с подобной проблемой, то один из макросов `HANDLE_ERROR()`, которые мы так щедро рассыпали по всему коду, обнаружит ошибку и из-

вестит о том, что произошло. Как и в традиционном программировании на С, урок заключается в том, что если функция возвращает код ошибки, то тому есть причина. Хотя всегда хочется эти коды проигнорировать, мы очень хотим избавить вас от многочасовых страданий, которые мы претерпели, и потому настоятельно рекомендуем проверять исход любой операции, которая теоретически может завершиться неудачно. Часто бывает, что такие ошибки не прекращают выполнение программы, но почти наверняка вызывают непредсказуемые и крайне неблагоприятные последствия где-то в другом месте.

Итак, мы имеем код, параллельно исполняемый графическим процессором. Быть может, вы слышали, что это трудно или что для программирования задач общего назначения на GPU нужно разбираться в компьютерной графике. Надеемся, что теперь вы понимаете, насколько CUDA C упрощает написание параллельного кода для GPU. В этом примере мы складывали векторы длины 10. Если вам интересно посмотреть, как создается массивно-параллельное приложение, замените 10 в директиве `#define N 10` на 10 000 или 50 000, чтобы запустить десятки тысяч параллельно исполняемых блоков. Но имейте в виду: ни одна из размерностей массива блоков не должна превышать 65 535. Это ограничение налагается оборудованием, так что попытка запустить больше блоков приведет к ошибке. В следующей главе мы покажем, как обойти это ограничение.

4.2.2. Более интересный пример

Мы не хотим сказать, что сложение векторов – безделка, но следующий пример удовлетворит тех, кто жаждет видеть впечатляющие примеры параллельного программирования на CUDA C.

Мы покажем, как нарисовать часть фрактала Джюлиа. Для тех, кто не знает, сообщим, что фрактал Джюлиа – это граница некоторого класса функций комплексных переменных. Это, конечно, не идет ни в какое сравнение с такими серьезными задачами, как сложение векторов или умножение матриц, однако почти для всех значений параметров функции граница образует фрактал – одну из самых захватывающих и красивых математических диковинок.

Вычисления, которые нужно проделать для генерации фрактала, очень просты. По сути дела, речь идет о простой рекуррентной формуле для точек на комплексной плоскости. Точка *не является* частью фрактала, если в процессе рекуррентных вычислений по этой формуле получается расходящаяся последовательность. Иными словами, если последовательность значений, образующихся в результате повторного применения формулы, стремится к бесконечности, то точка *не принадлежит* фракталу. Наоборот, если последовательность остается ограниченной, то точка *принадлежит* фракталу.

С точки зрения вычислений, рекуррентная формула очень проста; она приведена в формуле 4.1.

$$Z_{n+1} = Z_n^2 + C. \quad (4.1)$$

Вычисление по формуле 4.1 сводится к возведению текущего значения в квадрат и прибавлению константы.

Вычисление фрактала Джулия на CPU

Теперь рассмотрим код, который будет вычислять и визуализировать фрактал Джулия. Поскольку эта программа сложнее предыдущих, то мы разобьем ее на части. А в конце главы приведем весь код целиком.

```
int main( void ) {
    CPUBitmap bitmap( DIM, DIM );
    unsigned char *ptr = bitmap.get_ptr();

    kernel( ptr );

    bitmap.display_and_exit();
}
```

Главная функция на удивление проста. Она создает растровое изображение нужного размера, пользуясь библиотекой вспомогательных функций. Затем указатель на изображение передается функции `kernel`.

```
void kernel( unsigned char *ptr ){
    for (int y=0; y<DIM; y++) {
        for (int x=0; x<DIM; x++) {
            int offset = x + y * DIM;

            int juliaValue = julia( x, y );
            ptr[offset*4 + 0] = 255 * juliaValue;
            ptr[offset*4 + 1] = 0;
            ptr[offset*4 + 2] = 0;
            ptr[offset*4 + 3] = 255;
        }
    }
}
```

Вычислительное ядро просто обходит все точки, которые мы собираемся визуализировать, и для каждой вызывает функцию `julia()`, которая определяет, принадлежит ли эта точка фракталу Джулия. Функция `julia()` возвращает 1, если точка принадлежит фракталу, и 0 в противном случае. Мы задаем для точки красный цвет, если `julia()` вернула 1, и черный – если 0. Цвета выбраны произвольно, можете задать такие, которые отвечают вашему эстетическому чувству.

```
int julia( int x, int y ) {
    const float scale = 1.5;
    float jx = scale * (float)(DIM/2 - x)/(DIM/2);
    float jy = scale * (float)(DIM/2 - y)/(DIM/2);

    cuComplex c(-0.8, 0.156);
    cuComplex a(jx, jy);

    int i = 0;
```

```

for (i=0; i<200; i++) {
    a = a * a + c;
    if (a.magnitude2() > 1000)
        return 0;
}
return 1;
}

```

В этой функции происходит вся основная работа. Сначала мы преобразуем координаты пикселя в координаты на комплексной плоскости. Чтобы расположить начало координат комплексной плоскости в центре изображения, мы производим сдвиг на $\text{DIM}/2$. Затем, чтобы изображение находилось в диапазоне от -1.0 до 1.0 , мы масштабируем его с коэффициентом $\text{DIM}/2$. Таким образом, точка изображения с координатами (x, y) преобразуется в точку комплексной плоскости с координатами $((\text{DIM}/2 - x)/(\text{DIM}/2), ((\text{DIM}/2 - y)/(\text{DIM}/2))$.

Далее, чтобы обеспечить возможность увеличения или уменьшения, мы вводим коэффициент масштабирования. В этой версии «зашито» значение 1.5 , но, изменяя его, можно варьировать размер картинки. Если чувствуете в себе силы, можете задавать эту величину в командной строке.

Получив точку на комплексной плоскости, мы должны определить, принадлежит ли она фракталу Джулиа. Как вы помните, для этого нужно произвести вычисления по формуле $Z_{n+1} = Z_n^2 + C$. Поскольку C – произвольная комплексная константа, мы решили взять значение $-0.8 + 0.156i$, поскольку при этом получается интересная картинка. Если хотите полюбоваться на другие варианты фрактала Джулиа, поиграйте с этой константой.

В данном примере мы вычисляем 200 итераций формулы. После каждой итерации проверяется, не превзошла ли абсолютная величина результата пороговое значение (мы выбрали 1000). Если да, то мы считаем, что последовательность расходится, и возвращаем 0, означающий, что точка *не принадлежит* фракталу. С другой стороны, если после 200 итераций абсолютная величина не превысила 1000, то мы считаем, что точка принадлежит фракталу, и возвращаем значение 1.

Поскольку все вычисления производятся с комплексными числами, то мы определяем структуру для хранения комплексного числа.

```

struct cuComplex {
    float r;
    float i;
    cuComplex( float a, float b ) : r(a), i(b) {}
    float magnitude2( void ) { return r * r + i * i; }
    cuComplex operator*(const cuComplex& a) {
        return cuComplex(r*a.r - i*a.i, i*a.r + r*a.i);
    }
    cuComplex operator+(const cuComplex& a) {
        return cuComplex(r+a.r, i+a.i);
    }
};

```

Этот класс представляет комплексное число в виде вещественной части g и мнимой части i с одинарной точностью. В нем также определены операторы сложения и умножения комплексных чисел. (Если вы совсем незнакомы с комплексными числами, найдите какое-нибудь краткое введение в сети.) Наконец, мы определяем метод, возвращающий абсолютную величину комплексного числа.

Вычисление фрактала Джулия на GPU

В полном соответствии с отмеченной ранее тенденцией реализация для графического процессора очень похожа на реализацию для CPU.

```
int main( void ) {
    CPUBitmap bitmap( DIM, DIM );
    unsigned char *dev_bitmap;

    HANDLE_ERROR( cudaMalloc( (void**)&dev_bitmap,
                           bitmap.image_size() ) );

    dim3 grid(DIM,DIM);
    kernel<<<grid,1>>>( dev_bitmap );

    HANDLE_ERROR( cudaMemcpy( bitmap.get_ptr(),
                           dev_bitmap,
                           bitmap.image_size(),
                           cudaMemcpyDeviceToHost ) );

    bitmap.display_and_exit();

    cudaFree( dev_bitmap );
}
```

Эта версия `main()` выглядит намного сложнее версии для CPU, но общая структура точно такая же. Как и в версии для CPU, мы создаем растровое изображение размером $DIM \times DIM$ с помощью библиотечной функции. Но поскольку вычисления будут производиться на GPU, то мы еще объявляем указатель `dev_bitmap` на область памяти устройства, где будут храниться данные. А чтобы выделить эту область, нужно вызвать функцию `cudaMalloc()`.

Далее мы запускаем функцию `kernel()` точно так же, как в версии для CPU, хотя теперь она помечена квалификатором `__global__` и, значит, выполняется на GPU. Как и раньше, мы передаем функции `kernel()` полученный в предыдущей строке указатель на область для сохранения результата. Разница лишь в том, что память теперь принадлежит графическому процессору, а не CPU.

Самое существенное отличие от первоначальной версии заключается в том, что мы задаем число параллельных блоков, которые будут исполнять функцию `kernel()`. Поскольку каждую точку можно обсчитывать независимо от остальных, то мы просто выделяем для каждой точки свой экземпляр функции. Выше мы отмечали, что для некоторых задач наиболее естественно выглядит двумерная ин-

дексация. Неудивительно, что обсчет точек на двумерной комплексной плоскости как раз относится к таким задачам. Поэтому в следующей строке мы определяем двумерную сетку блоков:

```
dim3 grid(DIM, DIM);
```

`dim3` – это не стандартный тип С, не пугайтесь, что с вами случился провал в памяти. Просто в заголовочных файлах исполняющей среды CUDA определены типы, инкапсулирующие многомерные кортежи. В частности, тип `dim3` представляет трехмерный вектор, с помощью которого мы описываем сетку блоков. Но почему трехмерный, когда мы четко сказали, что сетка будет *двумерной*?

Если честно, то потому, что исполняющая среда CUDA ожидает получить значение типа `dim3`. И хотя трехмерные сетки в настоящее время не поддерживаются, CUDA все равно хочет видеть значение типа `dim3`, в котором последняя компонента равна 1. Если мы при инициализации указываем только две компоненты, как в предложении `dim3 grid(DIM, DIM)`, то исполняющая среда CUDA автоматически подставит вместо третьей значение 1, поэтому все будет работать правильно. Вполне возможно, что в будущем NVIDIA станет поддерживать и трехмерные сетки, но пока мы должны придерживаться API запуска ядра, потому что в борьбе программиста с API всегда побеждает API.

Далее мы передаем переменную `grid` типа `dim3` исполняющей среде CUDA:

```
kernel<<<grid, 1>>>( dev _ bitmap );
```

И наконец, так как результаты, сформированные в результате выполнения `kernel()`, находятся в памяти устройства, мы должны скопировать их в память CPU. Мы уже знаем, что для этого следует вызывать функцию `cudaMemcpy()`, передав в последнем аргументе константу `cudaMemcpyDeviceToHost`.

```
HANDLE_ERROR( cudaMemcpy( bitmap.get_ptr(),
                         dev_bitmap,
                         bitmap.image_size(),
                         cudaMemcpyDeviceToHost ) );
```

Еще одно отличие новой реализации от первоначальной – сама функция `kernel()`.

```
_global_ void kernel( unsigned char *ptr ) {
    // отобразить blockIdx на позицию пикселя
    int x = blockIdx.x;
    int y = blockIdx.y;
    int offset = x + y * gridDim.x;

    // вычислить значение в этой позиции
    int juliaValue = julia( x, y );
    ptr[offset*4 + 0] = 255 * juliaValue;
```

```

ptr[offset*4 + 1] = 0;
ptr[offset*4 + 2] = 0;
ptr[offset*4 + 3] = 255;
}

```

Прежде всего мы должны объявить функцию `kernel()` как `__global__`, чтобы она исполнялась устройством, а не CPU. И, в отличие от версии для CPU, нам больше не нужны вложенные циклы `for` для порождения индексов пикселей, передаваемых функции `julia()`. Как и в задаче о сложении векторов, исполняющая среда CUDA предоставляет нам эти индексы в переменной `blockIdx`. Это работает, потому что сетка блоков имеет столько же размерностей, сколько изображение, так что мы получаем по одному блоку для каждой пары целых чисел (x, y) в диапазоне от $(0, 0)$ до $(\text{DIM}-1, \text{DIM}-1)$.

Единственное, что нам еще нужно, – это линейное смещение от начала буфера `ptr`. Оно вычисляется с помощью еще одной встроенной переменной, `gridDim`. Эта переменная одинакова во всех блоках и содержит размерности запущенной сетки. В нашем случае она будет содержать значение (DIM, DIM) . Поэтому, чтобы получить смещение от начала буфера в диапазоне от 0 до $(\text{DIM} \times \text{DIM} - 1)$, мы умножаем индекс строки на ее ширину и прибавляем индекс столбца:

```
int offset = x + y * gridDim.x;
```

И напоследок приведем код, который определяет, принадлежит точка фрактalu Джюлии или нет. Как и в предыдущих примерах, он почти не отличается от версии для CPU.

```

__device__ int julia( int x, int y ) {
    const float scale = 1.5;
    float jx = scale * (float)(DIM/2 - x)/(DIM/2);
    float jy = scale * (float)(DIM/2 - y)/(DIM/2);

    cuComplex c(-0.8, 0.156);
    cuComplex a(jx, jy);

    int i = 0;
    for (i=0; i<200; i++) {
        a = a * a + c;
        if (a.magnitude2() > 1000)
            return 0;
    }

    return 1;
}

```

Мы снова определяем структуру `cuComplex`, где, в частности, присутствует метод для создания комплексного числа, компоненты которого представлены числами с плавающей точкой одинарной точности. Там же определены операторы сло-

жения и умножения комплексных чисел и функция для вычисления абсолютной величины комплексного числа.

```
struct cuComplex {
    float r;
    float i;

    cuComplex( float a, float b ) : r(a), i(b) {}
    __device__ float magnitude2( void ) {
        return r * r + i * i;
    }

    __device__ cuComplex operator*(const cuComplex& a) {
        return cuComplex(r*a.r - i*a.i, i*a.r + r*a.i);
    }

    __device__ cuComplex operator+(const cuComplex& a) {
        return cuComplex(r+a.r, i+a.i);
    }
};
```

Обратите внимание, что в CUDA C применяются те же языковые конструкции, что в версии для CPU. Единственное отличие – квалификатор `__device__`, который показывает, что код исполняется устройством, а не CPU. Напомним, что функции, объявленные с квалификатором `__device__`, можно вызывать из других функций с квалификатором `__device__` или `__global__`.

Мы слишком часто прерывали код комментариями, поэтому ниже приведен полный текст от начала до конца:

```
#include "../common/book.h"
#include "../common/cpu_bitmap.h"

#define DIM 1000

struct cuComplex {
    float r;
    float i;

    cuComplex( float a, float b ) : r(a), i(b) {}
    __device__ float magnitude2( void ) {
        return r * r + i * i;
    }

    __device__ cuComplex operator*(const cuComplex& a) {
        return cuComplex(r*a.r - i*a.i, i*a.r + r*a.i);
    }

    __device__ cuComplex operator+(const cuComplex& a) {
        return cuComplex(r+a.r, i+a.i);
    }
};
```

```
device__ int julia( int x, int y ) {
    const float scale = 1.5;
    float jx = scale * (float)(DIM/2 - x)/(DIM/2);
    float jy = scale * (float)(DIM/2 - y)/(DIM/2);

    cuComplex c(-0.8, 0.156);
    cuComplex a(jx, jy);

    int i = 0;
    for (i=0; i<200; i++) {
        a = a * a + c;
        if (a.magnitude2() > 1000)
            return 0;
    }

    return 1;
}

_global__ void kernel( unsigned char *ptr ) {
    // отобразить blockIdx на позицию пикселя
    int x = blockIdx.x;
    int y = blockIdx.y;
    int offset = x + y * gridDim.x;

    // вычислить значение в этой позиции
    int juliaValue = julia( x, y );
    ptr[offset*4 + 0] = 255 * juliaValue;
    ptr[offset*4 + 1] = 0;
    ptr[offset*4 + 2] = 0;
    ptr[offset*4 + 3] = 255;
}

int main( void ) {
    CPUBitmap bitmap( DIM, DIM );
    unsigned char *dev_bitmap;

    HANDLE_ERROR( cudaMalloc( (void**)&dev_bitmap,
                           bitmap.image_size() ) );

    dim3 grid(DIM,DIM);
    kernel<<<grid,1>>>( dev_bitmap );

    HANDLE_ERROR( cudaMemcpy( bitmap.get_ptr(), dev_bitmap,
                           bitmap.image_size(),
                           cudaMemcpyDeviceToHost ) );
    bitmap.display_and_exit();

    HANDLE_ERROR( cudaFree( dev_bitmap ) );
}
```

Запустив это приложение, вы увидите изображение фрактала Джулиа. Дабы убедить вас в том, что этот раздел заслуживает названия «Забавный пример», на рис. 4.2 приведен снимок экрана.

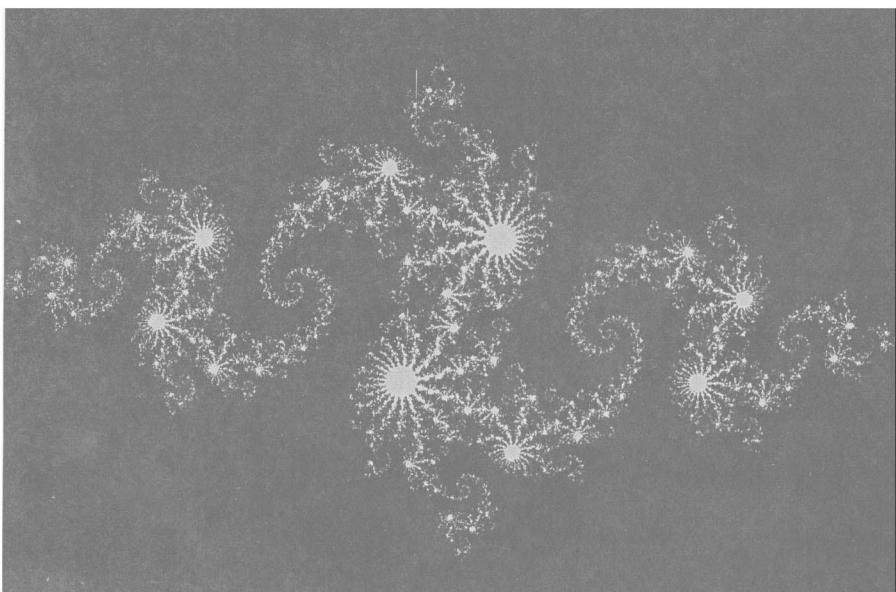


Рис. 4.2. Снимок экрана с изображением фрактала Джулия

4.3. Резюме

Примите поздравления, теперь вы умеете писать, компилировать и запускать массивно-параллельные программы, работающие на графическом процессоре! Можете похвастаться друзьям. И если они все еще думают, что GPU-вычисления – что-то экзотическое и трудное для освоения, то будут потрясены. А то, с какой легкостью вы достигли этого результата, останется нашим секретом. Но если вы готовы поделиться секретом с друзьями, порекомендуйте им купить эту книгу.

Пока что мы видели, как заставить исполняющую среду CUDA запустить несколько экземпляров программы, которые будут параллельно исполняться в *блоках*. Набор таких блоков мы назвали *сеткой*. Само название подразумевает, что сетка может быть одномерной или двумерной. Любой экземпляр ядра может узнать, в каком блоке он исполняется, опросив встроенную переменную `blockIdx`. А размерность сетки он может получить из встроенной переменной `gridDim`. Обе эти переменные оказались полезны нашему ядру – с их помощью оно вычисляло индекс элемента данных, за который отвечает блок.



Глава 5. Взаимодействие нитей

Итак, мы написали первую программу на языке CUDA C и узнали, как пишется параллельный код, исполняемый GPU. Для начала неплохо. Но, пожалуй, одним из важнейших в параллельном программировании является вопрос о том, как параллельно исполняемые элементы взаимодействуют между собой при решении задачи. Редко встречаются такие задачи, когда каждый процессор может вычислить свой результат и завершить выполнение, не заботясь о том, что в это время делают другие. Даже в относительно простых алгоритмах параллельно исполняемые экземпляры кода должны взаимодействовать и кооперироваться. Но пока мы еще ни слова не сказали о механизмах такого взаимодействия в языке CUDA C. К счастью, решение есть, и мы рассмотрим его в этой главе.

5.1. О чем эта глава

Прочитав эту главу, вы:

- узнаете о *нитях* в языке CUDA C;
- узнаете, как нити взаимодействуют друг с другом;
- узнаете о механизмах синхронизации параллельно работающих нитей.

5.2. Расщепление параллельных блоков

В предыдущей главе мы видели, как на GPU исполняется параллельный код. Для этого мы сообщали исполняющей среде CUDA о том, сколько нужно запустить параллельных экземпляров ядра. Эти параллельные экземпляры мы назвали *блоками*.

Исполняющая среда CUDA позволяет расщепить блоки на *нити*. Напомним, что при запуске нескольких параллельных блоков мы указывали их количество в первом аргументе внутри угловых скобок. Так, в задаче о сложении векторов мы запускали по одному блоку для каждого элемента вектора размером N :

```
add<<<N, 1>>>( dev_a, dev_b, dev_c );
```

Второй параметр в угловых скобках говорит, сколько нитей создать для каждого блока. До сих пор мы запускали блоки всего с одной нитью. В предыдущем примере ситуация выглядела так:

N блоков \times 1 нить/блок = N параллельных нитей.

Но можно было бы вместо этого запустить $N/2$ блоков с двумя нитями в блоке, $N/4$ блоков с четырьмя нитями каждый и т. д. Вернемся к задаче о сложении векторов, вооружившись новой информацией о возможностях CUDA C.

5.2.1. И снова о сложении векторов

Мы собираемся решить ту же задачу, что и в предыдущей главе, то есть, имея два вектора на входе, получить третий вектор, содержащий их сумму. Но на этот раз вместо блоков мы воспользуемся нитями.

Возможно, вам не понятно, какие преимущества дает использование нитей вместо блоков. Пока никаких достойных обсуждения преимуществ не видно. Однако параллельные нити внутри блока могут делать такие вещи, которые параллельным блокам не под силу. Так что наберитесь терпения и позвольте нам переписать программу из предыдущей главы, заменив блоки нитями.

Сложение векторов на GPU с использованием нитей

Начнем с двух заметных изменений. Раньше мы запускали N параллельных блоков с одной нитью в каждом:

```
add<<<N, 1>>>( dev_a, dev_b, dev_c );
```

а теперь запустим N нитей в одном блоке:

```
add<<<1, N>>>( dev_a, dev_b, dev_c );
```

Изменить придется только способ индексации данных. Раньше входные и выходные данные индексировались в ядре с помощью номера блока:

```
int tid = blockIdx.x;
```

Теперь же, поскольку блок всего один, индексировать нужно с помощью номера нити:

```
int tid = threadIdx.x;
```

И это все, что нужно сделать для перехода от реализации с параллельными блоками к реализации с параллельными нитями. Для полноты приведем исходный текст целиком, выделив измененные строки полужирным шрифтом:

```
#include "../common/book.h"

#define N 10

__global__ void add( int *a, int *b, int *c ) {
    int tid = threadIdx.x;
```

```
if (tid < N)
    c[tid] = a[tid] + b[tid];
}

int main( void ) {
    int a[N], b[N], c[N];
    int *dev_a, *dev_b, *dev_c;

    // выделить память на GPU
    HANDLE_ERROR( cudaMalloc( (void**)&dev_a, N * sizeof(int) ) );
    HANDLE_ERROR( cudaMalloc( (void**)&dev_b, N * sizeof(int) ) );
    HANDLE_ERROR( cudaMalloc( (void**)&dev_c, N * sizeof(int) ) );

    // заполнить массивы 'a' и 'b' на CPU
    for (int i=0; i<N; i++) {
        a[i] = i;
        b[i] = i * i;
    }

    // скопировать массивы 'a' и 'b' на GPU
    HANDLE_ERROR( cudaMemcpy( dev_a,
                           a,
                           N * sizeof(int),
                           cudaMemcpyHostToDevice ) );
    HANDLE_ERROR( cudaMemcpy( dev_b,
                           b,
                           N * sizeof(int),
                           cudaMemcpyHostToDevice ) );

    add<<<1,N>>>( dev_a, dev_b, dev_c );

    // скопировать массив 'c' с GPU на CPU
    HANDLE_ERROR( cudaMemcpy( c,
                           dev_c,
                           N * sizeof(int),
                           cudaMemcpyDeviceToHost ) );

    // вывести результат
    for (int i=0; i<N; i++) {
        printf( "%d + %d = %d\n", a[i], b[i], c[i] );
    }

    // освободить память, выделенную на GPU
    cudaFree( dev_a );
    cudaFree( dev_b );
    cudaFree( dev_c );

    return 0;
}
```

Все просто, не правда ли? В следующем разделе мы поговорим об ограничении, присущем решению, которое основано на одних лишь потоках. А затем расскажем, зачем вообще нужно расщеплять блоки на более мелкие параллельные компоненты.

Сложение более длинных векторов на GPU

В предыдущей главе мы упомянули об аппаратном ограничении на количество блоков в одной сетке – 65 535 по каждому измерению. Существует также ограничение и на количество нитей в одном блоке. Точнее, оно не должно превышать величины в поле `maxThreadsPerBlock` в структуре, описывающей свойства устройства (см. главу 3). Для многих современных графических процессоров количество нитей в блоке не должно превышать 512. И как же воспользоваться подходом на основе нитей для сложения векторов длиннее 512? Для этого понадобится комбинация нитей и блоков.

Как и раньше, мы должны будем внести два изменения: способ вычисления индекса внутри ядра и запуск самого ядра. Когда имеются несколько блоков и нитей, индексация напоминает стандартный метод перехода от двумерного индекса к линейному.

```
int tid = threadIdx.x + blockIdx.x * blockDim.x;
```

Здесь мы пользуемся еще одной встроенной переменной, `blockDim`. Она одинакова во всех блоках и содержит количество нитей вдоль каждой размерности блока. Поскольку мы работаем с одномерным блоком, то интерес представляет только величина `blockDim.x`. Напомним, что имеется похожая переменная `gridDim`, в которой хранится количество блоков вдоль каждого измерения сетки. Правда, переменная `gridDim` двумерная, а `blockDim` – трехмерная. Иначе говоря, исполняющая среда CUDA позволяет запускать двумерную сетку блоков, каждый из которых представляет собой трехмерный массив нитей. Да уж, целая куча измерений, и маловероятно, что все они вам понадобятся, но если что – они к вашим услугам.

Вычисление индекса в линейном массиве по приведенной выше формуле интуитивно очевидно. Если вам так не кажется, то представьте себе, что набор блоков, включающих нити, расположен в пространстве как двумерный массив пикселей. Эта структура показана на рис. 5.1.

Если нитям соответствуют столбцы, а блокам – строки, то для получения индекса конкретной нити нужно умножить номер блока на количество нитей в одном блоке и прибавить к произведению номер нити в блоке. Точно так же в задаче о фрактале Джгулия мы вычисляли смещение от начала буфера, зная двумерный индекс пикселя:

```
int offset = x + y * DIM;
```

А теперь предположим, что `DIM` – размерность блока (измеренная в нитях), `y` – номер блока, а `x` – номер нити в блоке. Вот и получается формула для вычисления линейного индекса нити:

Блок 0	Нить 0	Нить 1	Нить 2	Нить 3
Блок 1	Нить 0	Нить 1	Нить 2	Нить 3
Блок 2	Нить 0	Нить 1	Нить 2	Нить 3
Блок 3	Нить 0	Нить 1	Нить 2	Нить 3

Рис. 5.1. Двумерная организация блоков и нитей

```
tid = threadIdx.x + blockIdx.x * blockDim.x.
```

Второе изменение следует внести в код самого ядра. Нам, как и раньше, нужно запустить N параллельных нитей, только теперь мы хотим распределить их по нескольким блокам, чтобы не выйти за рамки ограничения 512 нитей на блок. Одно из возможных решений – произвольно выбрать фиксированное количество нитей в блоке, например 128. Тогда для получения N нитей нужно будет запустить $N/128$ блоков.

Беда в том, что результатом целочисленного деления $N/128$ является целое число. В частности, если N равно 127, то $N/128$ равно нулю, а запустив 0 нитей, мы ничего не вычислим. Да и вообще, если N не кратно 128, то мы запустим слишком мало нитей. Это плохо. Хотелось бы, чтобы частное от деления округлялось с избытком.

Для решения этой задачи без привлечения функции `ceil()` существует стандартный прием: вычислять не $N/128$, а $(N+127)/128$. Можете поверить нам на слово, что эта формула дает наименьшее целое число, большее или равное частному от деления N на 128. Или потратите минутку на то, чтобы убедиться в этом самостоятельно. Поскольку мы решили, что в блоке будет 128 нитей, то запускать ядро нужно следующим образом:

```
add<<< (N+127)/128, 128 >>>( dev_a, dev_b, dev_c );
```

После этого изменения будет запущено заведомо достаточное количество нитей и даже *слишком много*, если N не кратно 128. Но эта проблема легко решается, и наше ядро уже позаботилось об этом. Нам нужно лишь проверить, что индекс нити находится в диапазоне от 0 до N , – только в этом случае его можно использовать для доступа к массивам входных и выходных данных:

```
if (tid < N)
    c[tid] = a[tid] + b[tid];
```

Теперь если индекс оказывается за концом массива – а так будет всегда, когда длина массива не кратна 128, – мы не производим никаких вычислений. И, что еще важнее, не пытаемся читать или записывать данные в область памяти вне массивов.

Сложение векторов произвольной длины на GPU

Мы не были до конца откровенны, когда первый раз обсуждали запуск параллельных блоков на GPU. Помимо ограничения на количество нитей, существует также аппаратное ограничение на количество блоков (хотя и гораздо более либеральное). Как мы уже упоминали, количество блоков вдоль любого измерения сетки не должно превышать 65 535.

Но тогда в последней реализации сложения векторов возникает проблема. Если запускать $N/128$ блоков, то мы столкнемся с ошибкой, когда число элементов в каждом векторе окажется больше $65\,535 * 128 = 8\,388\,480$. Вроде бы большое число, но при нынешних объемах памяти от 1 до 4 Гб наиболее мощные графические процессоры легко справляются с хранением данных, совокупный размер которых на несколько порядков больше, чем 8 миллионов элементов.

К счастью, есть очень простое решение. Сначала изменим ядро.

```
__global__ void add( int *a, int *b, int *c ) {
    int tid = threadIdx.x + blockIdx.x * blockDim.x;
    while (tid < N) {
        c[tid] = a[tid] + b[tid];
        tid += blockDim.x * gridDim.x;
    }
}
```

Выглядит совсем как *первоначальная* версия программы сложения векторов! Действительно, давайте сравним с реализацией для CPU из предыдущей главы:

```
void add( int *a, int *b, int *c ) {
    int tid = 0;      // это CPU с номером 0, поэтому начинаем с нуля
    while (tid < N) {
        c[tid] = a[tid] + b[tid];
        tid += 1;     // у нас всего один CPU, поэтому увеличиваем на 1
    }
}
```

Здесь мы тоже обходили данные в цикле `while`. Вспомните, мы говорили, что в версии для нескольких CPU или процессорных ядер индекс массива следовало бы увеличивать не на 1, а на количество процессоров. Тот же принцип мы сейчас применим в версии для GPU.

В реализации для GPU мы можем рассматривать параллельно работающие нити как процессоры. Хотя в реальном GPU процессорных блоков может быть меньше (или больше), мы считаем, что каждая нить логически исполняется параллельно, а планирование фактического выполнения поручаем оборудованию.

Отделение распараллеливания от истинного способа аппаратного исполнения – бремя, которое CUDA снимает с плеч программиста. И это действительно мания небесная, если учесть, что в современных процессорах NVIDIA количество АЛУ варьируется от 8 до 480!

Теперь, разобравшись с принципом, лежащим в основе реализации, мы должны только понять, как вычислить начальный индекс каждой параллельно исполняемой нити и на сколько его увеличивать в цикле. Мы хотим, чтобы каждая нить начинала работу с разных индексов, поэтому нужно линеаризовать пару (номер блока, номер нити), как показано в разделе «Сложение более длинных векторов на GPU». Начальный индекс для каждой нити будем вычислять по следующей формуле:

```
int tid = threadIdx.x + blockIdx.x * blockDim.x;
```

После того как нить закончит работать с текущим индексом, мы должны увеличить его на общее число нитей, запущенных в сетке. Эта величина равна произведению числа блоков на число нитей в одном блоке, то есть `blockDim.x * gridDim.x`. Таким образом, приращение вычисляется так:

```
tid += blockDim.x * gridDim.x;
```

Мы почти у цели! Осталось только подправить код запуска. Если помните, мы начали этот разговор с того, что вызов `add<<<(N+127)/128, 128>>>(dev_a, dev_b, dev_c)` завершится ошибкой, если $(N+127)/128$ больше 65 535. Чтобы гарантированно не запускать слишком много блоков, нужно лишь заранее выбрать разумное их количество. Ну и чтобы не вводить лишних чисел, остановимся на 128: 128 блоков по 128 нитей в каждом.

```
add<<<128, 128>>>( dev _ a, dev _ b, dev _ c );
```

Можете взять любые другие значения, лишь бы они не превышали вышеупомянутых ограничений. Позже мы обсудим влияние того или иного выбора на производительность, но пока будем считать, что 128 блоков по 128 нитей достаточно. Теперь мы можем складывать векторы произвольной длины и ограничены лишь объемом памяти, доступной GPU. Ниже приведен полный исходный текст программы:

```
#include "../common/book.h"

#define N (33 * 1024)

__global__ void add( int *a, int *b, int *c ) {
    int tid = threadIdx.x + blockIdx.x * blockDim.x;
    while (tid < N) {
        c[tid] = a[tid] + b[tid];
        tid += blockDim.x * gridDim.x;
    }
}
```

```
}

int main( void ) {
    int a[N], b[N], c[N];
    int *dev_a, *dev_b, *dev_c;

    // выделить память на GPU
    HANDLE_ERROR( cudaMalloc( (void**)&dev_a, N * sizeof(int) ) );
    HANDLE_ERROR( cudaMalloc( (void**)&dev_b, N * sizeof(int) ) );
    HANDLE_ERROR( cudaMalloc( (void**)&dev_c, N * sizeof(int) ) );

    // заполнить массивы 'a' и 'b' на CPU
    for (int i=0; i<N; i++) {
        a[i] = i;
        b[i] = i * i;
    }

    // скопировать массивы 'a' и 'b' на GPU
    HANDLE_ERROR( cudaMemcpy( dev_a,
                            a,
                            N * sizeof(int),
                            cudaMemcpyHostToDevice ) );
    HANDLE_ERROR( cudaMemcpy( dev_b,
                            b,
                            N * sizeof(int),
                            cudaMemcpyHostToDevice ) );

    add<<<128,128>>>( dev_a, dev_b, dev_c );

    // скопировать массив 'c' с GPU на CPU
    HANDLE_ERROR( cudaMemcpy( c,
                            dev_c,
                            N * sizeof(int),
                            cudaMemcpyDeviceToHost ) );

    // проверить, что GPU сделал все, что мы просили
    bool success = true;
    for (int i=0; i<N; i++) {
        if ((a[i] + b[i]) != c[i]) {
            printf( "Ошибка: %d + %d != %d\n", a[i], b[i], c[i] );
            success = false;
        }
    }
    if (success) printf( "Мы сделали это!\n" );

    // освободить память, выделенную на GPU
    cudaFree( dev_a );
```

```
    cudaFree( dev_b );
    cudaFree( dev_c );

    return 0;
}
```

5.2.2. Создание эффекта волн на GPU с использованием нитей

Как и в предыдущей главе, мы вознаградим вас за терпение, представив более интересный пример, в котором демонстрируются некоторые из обсуждавшихся приемов. Мы снова воспользуемся вычислительной мощью GPU для программной генерации картинок. Но чтобы было интереснее, мы их еще и анимируем. Не пугайтесь, весь не относящийся к делу код анимации мы вынесли во вспомогательные функции, так что владения компьютерной графикой или техникой анимации от вас не потребуется.

```
struct DataBlock {
    unsigned char *dev_bitmap;
    CPUAnimBitmap *bitmap;
};

// освободить выделенную память устройства
void cleanup( DataBlock *d ) {
    cudaFree( d->dev_bitmap );
}

int main( void ) {
    DataBlock data;
    CPUAnimBitmap bitmap( DIM, DIM, &data );
    data.bitmap = &bitmap;

    HANDLE_ERROR( cudaMalloc( (void**)&data.dev_bitmap,
                           bitmap.image_size() ) );

    bitmap.anim_and_exit( (void (*)(void*, int))generate_frame,
                          (void (*)(void*))cleanup );
}
```

Наиболее сложная часть `main()` скрыта во вспомогательном классе `CPUAnimBitmap`. Как и раньше, мы выделяем память с помощью `cudaMalloc()`, исполняем на устройстве код, в котором используется выделенная память, а затем освобождаем память, вызывая `cudaFree()`. Эта последовательность теперь должна казаться вам рутинной.

В этом примере мы немного усложнили способ выполнения среднего шага, «исполнение на устройстве кода, в котором используется выделенная память».

Мы передаем методу `anim_and_exit()` указатель на функцию `generate_frame()`. Класс будет вызывать ее всякий раз, как понадобится сгенерировать новый кадр анимации.

```
void generate_frame( DataBlock *d, int ticks ) {
    dim3 blocks(DIM/16,DIM/16);
    dim3 threads(16,16);
    kernel<<<blocks,threads>>>( d->dev_bitmap, ticks );

    HANDLE_ERROR( cudaMemcpy( d->bitmap->get_ptr(),
                           d->dev_bitmap,
                           d->bitmap->image_size(),
                           cudaMemcpyDeviceToHost ) );
}
```

В четырех строчках этой функции заключены важные концепции CUDA C. Во-первых, мы объявляем двумерные переменные `blocks` и `threads`. Понятно, что `blocks` – это число параллельных блоков в запускаемой сетке, а `threads` – число нитей в одном блоке. Так как мы собираемся генерировать изображение, то используем двумерную индексацию, так что у каждой нити будет уникальный индекс (x, y), которому легко поставить в соответствие пиксель в создаваемом изображении. Мы решили, что блоки будут содержать массив нитей размером 16×16 . Если размер изображения составляет $DIM \times DIM$ пикселей, то, для того чтобы на каждый пиксель приходилась одна нить, нужно будет запустить $DIM/16 \times DIM/16$ блоков. На рис. 5.2 показано, как выглядит такая конфигурация блоков и нитей для изображения шириной 48 пиксель и высотой 32 пикселя – до смешного маленького.

Если вам доводилось заниматься многопоточным программированием для CPU, то, наверное, вас удивляет, зачем так много нитей. Ведь чтобы выполнить анимацию высокой четкости на экране размером 1920×1080 , будет создано более двух миллионов нитей. На GPU создание и планирование такого гигантского количества нитей – обычное дело, тогда как на CPU никто об этом и помышлять бы не стал. Поскольку управление нитями на CPU осуществляется программно, он просто не может справиться с таким количеством нитей, как GPU. Но раз мы в состоянии создать по одной нити для каждого обрабатываемого элемента, то параллельное программирование для GPU оказывается гораздо проще, чем для CPU.

Объявив переменные, в которых будут храниться параметры параллельного исполнения, мы просто запускаем ядро, вычисляющее значения пикселей.

```
kernel<<< blocks,threads>>>( d->dev _ bitmap, ticks );
```

Ядру необходимо знать две вещи, которые мы передаем в виде параметров. Во-первых, нужен указатель на область памяти устройства, в котором будут храниться вычисленные пиксели. Эта память выделена в функции `main()`, и указатель на нее помещен в глобальную переменную. Однако переменная «глобальная» лишь в коде, исполняемом CPU, поэтому ее необходимо передавать в параметре, что-

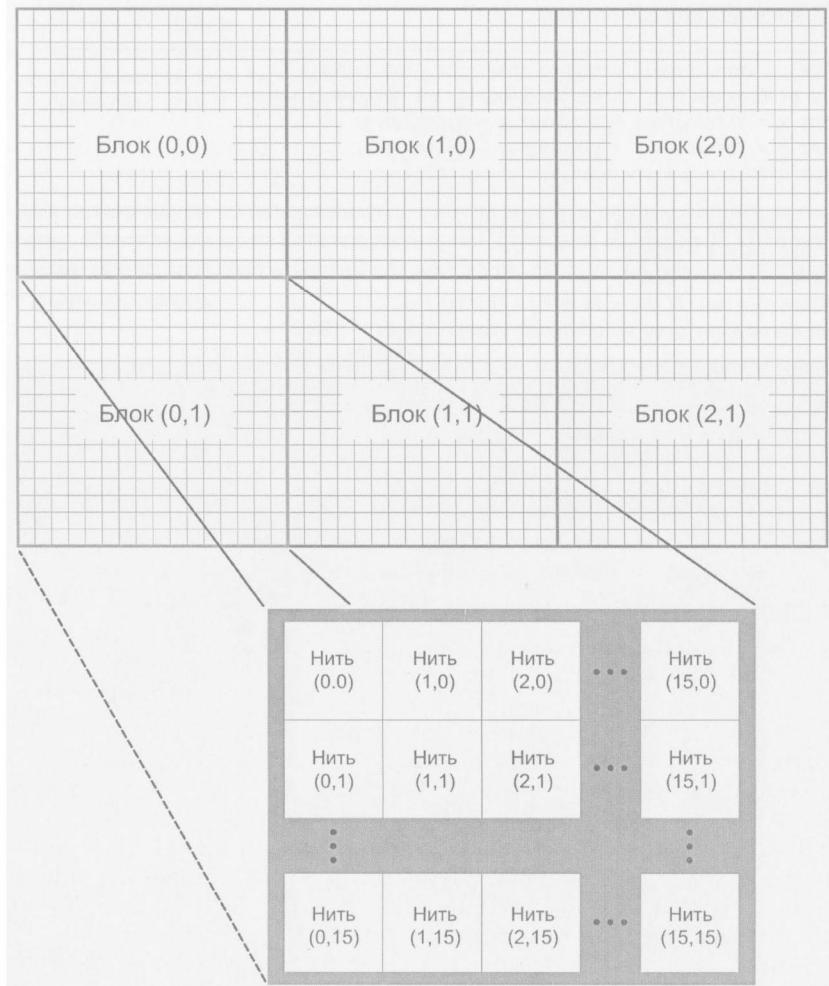


Рис. 5.2. Двумерная иерархия блоков и нитей для обработки изображения размером 48 × 32 пикселя – по одной нити на пиксель

бы исполняющая среда CUDA могла сделать ее доступной в коде, исполняемом устройством.

Во-вторых, ядро должно знать текущее время анимации, чтобы генерировать правильный кадр. Это время, ticks, передается функции `generate_frame()` из инфраструктурного кода в классе `CPUAnimBitmap`, поэтому мы можем переправить его ядру без изменения.

И наконец, код самого ядра:

```
global__ void kernel( unsigned char *ptr, int ticks ) {
    // отображаем пару threadIdx/BlockIdx на позицию пикселя
    int x = threadIdx.x + blockIdx.x * blockDim.x;
    int y = threadIdx.y + blockIdx.y * blockDim.y;
    int offset = x + y * blockDim.x * gridDim.x;

    // вычисляем значение в этой позиции
    float fx = x - DIM/2;
    float fy = y - DIM/2;
    float d = sqrtf( fx * fx + fy * fy );
    unsigned char grey = (unsigned char) ( 128.0f + 127.0f *
                                         cos(d/10.0f - ticks/7.0f) /
                                         (d/10.0f + 1.0f));

    ptr[offset*4 + 0] = grey;
    ptr[offset*4 + 1] = grey;
    ptr[offset*4 + 2] = grey;
    ptr[offset*4 + 3] = 255;
}
```

Первые три строчки ядра – самые важные:

```
int x = threadIdx.x + blockIdx.x * blockDim.x;
int y = threadIdx.y + blockIdx.y * blockDim.y;
int offset = x + y * blockDim.x * gridDim.x;
```

Здесь каждая нить определяет свой индекс в блоке, а также индекс своего блока в сетке и преобразует их в уникальный индекс (x, y) пикселя в изображении. Таким образом, когда нить с индексом (3, 5) в блоке (12, 8) начинает выполнение, она знает, что слева от нее есть 12 полных блоков, а сверху – 8 полных блоков. Внутри же блока слева от нити с индексом (3, 5) находятся три нити, а над ней – пять нитей. Поскольку всего в каждом блоке 16 нитей, то получается, что слева от рассматриваемой нити имеются 3 нити + 12 блоков * 16 нитей/блок = 195 нитей, а над ней – 5 нитей + 8 блоков * 16 нитей/блок = 128 нитей.

Это и есть вычисление x и y в первых двух строчках; именно так мы отображаем индексы нити и блока на координаты пикселя в изображении. Далее мы просто ли-неаризуем эти значения x и y и получаем смещение пикселя от начала выходного буфера. Все это аналогично вычислениям, проделанным в разделах «Сложение более длинных векторов на GPU» и «Сложение векторов произвольной длины на GPU».

```
int offset = x + y * blockDim.x * gridDim.x;
```

Поскольку мы знаем пиксель (x, y), который должна обсчитывать нить, и время, когда она должна вычислять его значение, то можем вычислить произвольную функцию от (x, y, t) и поместить результат в выходной буфер. В данном случае функция порождает синусоидальную «рябь».

```
float fx = x - DIM/2;  
float fy = y - DIM/2;  
float d = sqrtf( fx * fx + fy * fy );  
unsigned char grey = (unsigned char)( 128.0f + 127.0f *  
                           cos(d/10.0f - ticks/7.0f) /  
                           (d/10.0f + 1.0f));
```

Мы рекомендуем не задерживаться слишком долго на вычислении переменной `grey`. По сути дела, это простая двумерная функция от времени, которая при анимации порождает симпатичный эффект волн. На рис. 5.3 показан снимок одного кадра анимации.

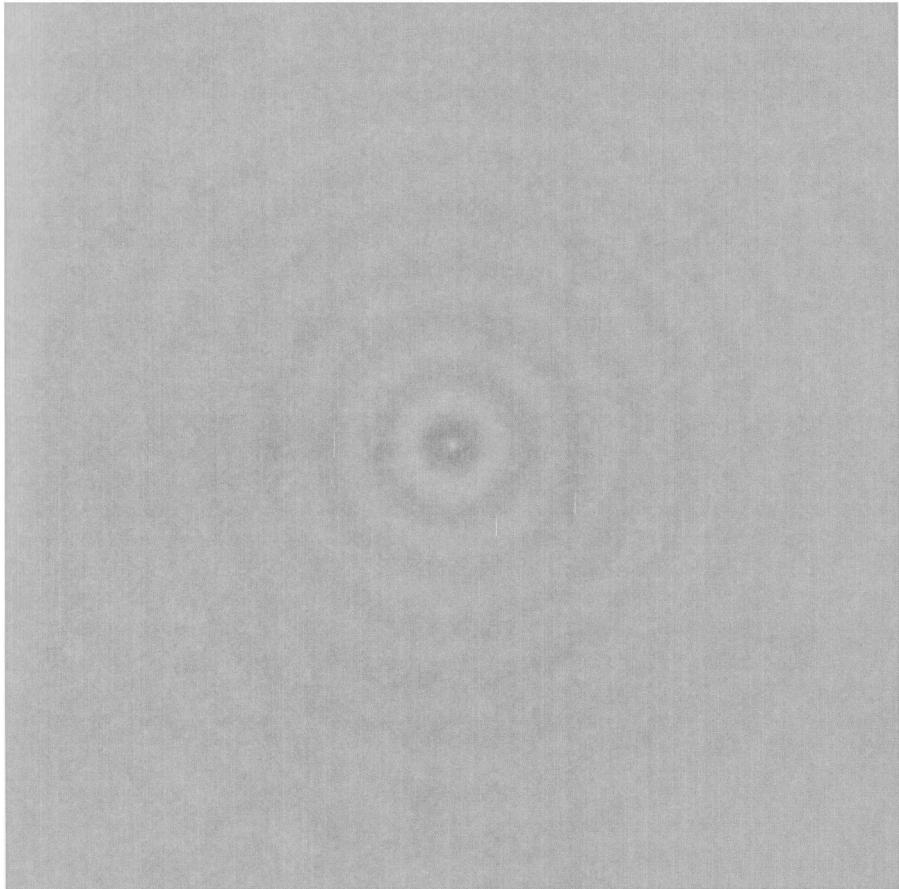


Рис. 5.3. Снимок одного кадра анимации волн на GPU

5.3. Разделяемая память и синхронизация

До сих пор единственным побудительным мотивом расщепления блоков на нити было стремление обойти аппаратные ограничения на количество одновременно работающих блоков. Это довольно слабая мотивация, потому что исполняющая среда CUDA могла бы и сама справиться с этой задачей. Но есть и другие причины.

Язык CUDA C предоставляет в распоряжение программы так называемую *разделяемую память*. С этой областью памяти связано еще одно расширение языка C, близкое к ключевым словам `_device_` и `_global_`. Если в объявление любой переменной добавить ключевое слово `_shared_`, то эта переменная будет размещена в разделяемой памяти. Но зачем это нужно?

Рады, что вы спросили. Компилятор CUDA C обрабатывает переменные в разделяемой памяти иначе, чем обычные переменные. Он создает копию такой переменной в каждом блоке, запускаемом на GPU. Все нити, работающие в одном блоке, разделяют эту переменную, но не могут ни увидеть, ни модифицировать ее копии, видимые в других блоках. Это создаст прекрасный механизм взаимодействия и кооперации нитей, находящихся в одном блоке. Кроме того, буферы разделяемой памяти физически находятся в самом GPU, а не в DRAM (динамическое запоминающее устройство с произвольной выборкой) вне кристалла. Это означает, что время задержки при доступе к разделяемой памяти существенно меньше, чем при доступе к обычным буферам, то есть разделяемая память играет роль внутриблочного программно управляемого кэша.

Перспективы взаимодействия между нитями должны вас вдохновлять. Нас они тоже вдохновляют. Но ничто в жизни не дается бесплатно, и межнитевое взаимодействие – не исключение. Если мы хотим, чтобы нити взаимодействовали, то необходим и механизм синхронизации. Например, если нить A записывает какое-то значение в разделяемую память, а нить B должна что-то с этим значением сделать, то нить B не должна приступать к работе, пока нить A не сделала свое дело. Без синхронизации возникла бы «гонка» (*race condition*), и правильность результата оказалась бы зависящей от недетерминированных особенностей работы оборудования. Рассмотрим пример, в котором эти средства используются.

5.3.1. Скалярное произведение

Поздравляем! Мы успешно справились со сложением векторов и теперь займемся их скалярным произведением (иногда эту операцию называют также *внутренним произведением*). Вкратце напомним, что такое скалярное произведение, на случай, если вы незнакомы с векторной алгеброй (или с годами подзабыли ее). Вычисление состоит из двух шагов. Сначала мы перемножаем соответственные элементы двух входных векторов. Это очень похоже на сложение векторов, только место операции сложения занимает операция умножения. Но, вместо того чтобы

сохранять результаты в третьем, выходном векторе, мы суммируем их, получая при этом скалярную величину.

Например, умножение двух векторов из четырех элементов производится по формуле 5.1.

$$(x_1 x_2 x_3 x_4) \cdot (y_1 y_2 y_3 y_4) = x_1 y_1 + x_2 y_2 + x_3 y_3 + x_4 y_4. \quad (5.1)$$

Наверное, алгоритм, который мы собираемся применить, вам уже понятен. Первый шаг выполняется точно так же, как при сложении векторов. Каждая нить перемножает пару соответственных элементов, а затем переходит к следующей паре. Поскольку в результате должна быть выдана сумма попарных произведений, то каждая нить сохраняет сумму произведений тех пар, которые обрабатывала. После обработки очередной пары индекс увеличивается на величину, равную общему числу нитей, так что ни один элемент не будет пропущен и ни один не будет обработан дважды. Вот как выглядит код первого шага вычисления скалярного произведения:

```
#include "../common/book.h"

#define imin(a,b) (a<b?a:b)

const int N = 33 * 1024;
const int threadsPerBlock = 256;

__global__ void dot( float *a, float *b, float *c ) {
    __shared__ float cache[threadsPerBlock];
    int tid = threadIdx.x + blockIdx.x * blockDim.x;
    int cacheIndex = threadIdx.x;
    float temp = 0;
    while (tid < N) {
        temp += a[tid] * b[tid];
        tid += blockDim.x * gridDim.x;
    }

    // сохранить значение в кэше
    cache[cacheIndex] = temp;
}
```

Как видите, мы объявили буфер в разделяемой памяти с именем `cache`. В этом буфере будет храниться частичная сумма, вычисленная данной нитью. Вскоре мы увидим, *почему* мы так поступили, а пока рассмотрим сам механизм. Объявить, что переменная должна находиться в разделяемой памяти, очень просто; точно так же в стандартном С вы говорите, что переменная должна быть `static` или `volatile`:

```
__shared__ float cache[threadsPerBlock];
```

Мы объявляем массив размером `threadsPerBlock`, так чтобы у каждой нити в блоке было свое место для хранения временного результата. Напомним, что при выделении глобальной памяти мы принимали во внимание все нити, исполняю-

щие ядро, то есть размер области вычислялся как произведение `threadsPerBlock` на общее число блоков. Но, поскольку компилятор сам создает копии разделяемых переменных для каждого блока, то нам нужно выделить лишь столько памяти, чтобы у каждой нити в блоке был свой элемент.

Выделив разделяемую память, мы вычисляем индексы элементов данных — так же, как и раньше:

```
int tid = threadIdx.x + blockIdx.x * blockDim.x;
int cacheIndex = threadIdx.x;
```

Порядок вычисления переменной `tid` уже нам знаком; требуется объединить индексы блока и нити так, чтобы получилось глобальное смещение от начала входного массива. В качестве смещения от начала кэша в разделяемой памяти используется просто индекс нити. Еще раз повторим, что индекс блока в этом смещении не участвует, потому что у каждого блока имеется своя частная копия разделяемой памяти.

Область памяти, выделенная под кэш, инициализируется, чтобы впоследствии можно было просуммировать все элементы массива, не опасаясь, что там находятся бессмысленные данные. Может случиться, что будут использованы не все элементы массива; так произойдет, если длина входных векторов не кратна числу нитей в блоке. В подобном случае в последнем блоке окажутся нити, которым нечего делать, поэтому они ничего и не запишут в кэш.

Каждый поток вычисляет частичную сумму произведений соответственных элементов векторов `a` и `b`. Дойдя до конца массива, нить сохраняет вычисленную сумму в кэше.

```
float temp = 0;
while (tid < N) {
    temp += a[tid] * b[tid];
    tid += blockDim.x * gridDim.x;
}

// сохранить значение в кэше
cache[cacheIndex] = temp;
```

Теперь мы должны просуммировать все значения, находящиеся в массиве `cache`. Для этого необходимо, чтобы какие-то нити прочитали эти значения. Однако, как отмечалось выше, это опасная ситуация. Необходимо как-то гарантировать, что все операции записи в разделяемый массив `cache[]` завершатся до того, как кто-то попытается его прочитать. К счастью, такой механизм существует:

```
// синхронизировать нити в этом блоке
__syncthreads();
```

Этот вызов гарантирует, что все нити в блоке закончат выполнять команды, предшествующие `__syncthreads()`, до того как аппаратура разрешит выполнить сле-

дующую команду в какой-нибудь нити. Как раз то, что надо! Теперь мы знаем, что когда первая нить выполнит первую команду после обращения к `_syncthreads()`, все остальные нити также дошли до `_syncthreads()`.

Теперь, когда временный кэш гарантированно заполнен, мы можем просуммировать находящиеся в нем значения. Процедура, в результате которой из входного массива получается массив меньшего размера, называется *редукцией*. Редукция так часто встречается в параллельном программировании, что для нее было придумано специальное название.

Наивный способ выполнить редукцию заключается в том, чтобы выделить одну нить, которая обойдет массив разделяемой памяти и вычислит его сумму. Это займет время, пропорциональное длине массива. Но поскольку в нашем распоряжении сотни нитей, то редукцию можно произвести параллельно за время, пропорциональное логарифму длины массива. Поначалу приведенный ниже код может показаться запутанным, поэтому рассмотрим его по частям.

Идея в том, что каждая нить складывает два значения, находящихся в массиве `cache[]`, и помещает результат снова в `cache[]`. Поскольку при этом два элемента объединяются в один, то по завершении данного шага количество элементов в массиве уменьшится вдвое. На следующем шаге та же операция применяется к оставшейся половине. Продолжая в том же духе, мы за $\log_2(\text{threadsPerBlock})$ шагов вычислим сумму всех элементов массива `cache[]`. В нашем случае блок содержит 256 нитей, поэтому для редукции 256 элементов массива `cache[]` в единственное значение, сумму, потребуются 8 итераций.

Код выглядит следующим образом:

```
// для успешной редукции threadsPerBlock должно быть степенью 2
// из-за следующего кода
int i = blockDim.x/2;
while (i != 0) {
    if (cacheIndex < i)
        cache[cacheIndex] += cache[cacheIndex + i];
    __syncthreads();
    i /= 2;
}
```

На первом шаге `i` будет равно половине `threadsPerBlock`. Мы хотим, чтобы работали лишь нити с индексами, меньшими этого значения, поэтому складываем элементы массива `cache[]`, только если индекс нити меньше `i`, то есть сложение производится при условии `if(cacheIndex < i)`. Каждая нить берет из массива `cache[]` два элемента – индекс первого совпадает с ее индексом, а индекс второго смещен относительно ее индекса на `i`, – вычисляет их сумму и помещает результат обратно в `cache[]`.

Предположим, что в кэше было восемь элементов, тогда `i` равно 4. На рис. 5.4 показан один шаг редукции. После того как шаг завершен, мы сталкиваемся с той же ситуацией, что при вычислении попарных произведений. Прежде чем читать значения, только что помещенные в `cache[]`, необходимо убедиться, что

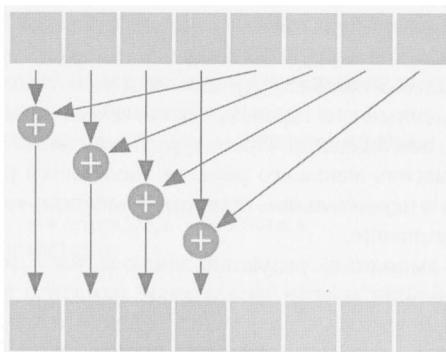


Рис. 5.4. Один шаг суммирования путем редукции

все нити, которые должны были что-то записать в `cache[]`, сделали это. Вызов `_syncthreads()` после присваивания гарантирует, что это условие выполнено.

По выходе из цикла `while` в каждом блоке остается всего одно число. Оно находится в первом элементе массива `cache[]` и является суммой попарных произведений, вычисленных нитями из этого блока. Мы сохраняем это единственное значение в глобальной памяти и завершаем ядро:

```
if (cacheIndex == 0)
    c[blockIdx.x] = cache[0];
}
```

Почему запись в глобальную память производит только нить, для которой `cacheIndex == 0`? Так ведь число только одно, поэтому для его сохранения хватит и одной нити. Конечно, мы могли бы выполнить эту операцию в каждой нити, и программа работала бы правильно, но это лишь создает ненужный трафик между процессором и памятью. Для простоты мы выбрали нить с индексом 0, хотя могли бы взять любую другую. Наконец, поскольку каждый блок записывает ровно одно значение в глобальный массив `c[]`, ничто не мешает записать его в элемент с индексом `blockIdx.x`.

Теперь у нас имеется массив `c[]`, в каждом элементе которого находится частичная сумма, вычисленная одним из параллельных блоков. На последнем шаге вычисления скалярного произведения мы должны просуммировать эти элементы. Но хотя скалярное произведение еще не вычислено до конца, мы в этот момент выходим из ядра и возвращаем управление CPU. Почему же мы так поступаем?

Выше мы сказали, что операция, подобная скалярному произведению, называется *редукцией*. Смысл названия в том, что на выходе получается меньше элементов данных, чем на входе. В случае скалярного произведения на выходе получается ровно одно значение независимо от размера входных векторов. Оказывается, что массивно-параллельная машина, какой является GPU, будет просто вп-

стую тратить ресурсы, выполняя последний шаг редукции, потому что размер набора данных в этот момент слишком мал; трудновато занять 480 арифметических устройств сложением 32 чисел!

Поэтому мы возвращаем управление CPU, оставляя ему последний шаг вычисления – суммирование элементов массива $c[]$. В более серьезном приложении GPU мог бы начать вычисление следующего скалярного произведения или заняться еще чем-нибудь полезным. Но в этом примере работа GPU закончена.

При рассмотрении данного примера мы нарушили традицию и сразу перешли к ядру. Надеемся, что у вас не возникло трудностей с пониманием той части тела `main()`, которая предшествует вызову ядра, ведь она так мало отличается от того, что мы видели раньше.

```
const int blocksPerGrid =
    imin( 32, (N+threadsPerBlock-1) / threadsPerBlock );

int main( void ) {
    float *a, *b, c, *partial_c;
    float *dev_a, *dev_b, *dev_partial_c;

    // выделить память на CPU
    a = (float *)malloc( N*sizeof(float) );
    b = (float *)malloc( N*sizeof(float) );
    partial_c = (float *)malloc( blocksPerGrid*sizeof(float) );

    // выделить память на GPU
    HANDLE_ERROR( cudaMalloc( (void**)&dev_a,
        N*sizeof(float) ) );
    HANDLE_ERROR( cudaMalloc( (void**)&dev_b,
        N*sizeof(float) ) );
    HANDLE_ERROR( cudaMalloc( (void**)&dev_partial_c,
        blocksPerGrid*sizeof(float) ) );

    // заполнить память CPU данными
    for (int i=0; i< N; i++) {
        a[i] = i;
        b[i] = i*2;
    }

    // скопировать массивы 'a' и 'b' в память GPU
    HANDLE_ERROR( cudaMemcpy( dev_a, a, N*sizeof(float),
        cudaMemcpyHostToDevice ) );
    HANDLE_ERROR( cudaMemcpy( dev_b, b, N*sizeof(float),
        cudaMemcpyHostToDevice ) );

    dot<<<blocksPerGrid,threadsPerBlock>>> ( dev_a,
        dev_b,
        dev_partial_c );
```

Чтобы вы не сбежали от скуки, мы кратенько опишем, что происходит в этом коде.

1. Выделить память CPU и устройства для массивов – двух входных и одного выходного.
2. Заполнить входные массивы $a[]$ и $b[]$, затем скопировать их в память устройства с помощью функции `cudaMemcpy()`.
3. Вызвать ядро, вычисляющее скалярное произведение, задав фиксированное количество нитей в блоке и блоков в сетке.

Несмотря на то что большая часть этого кода уже стала общим местом, мы все же остановимся на вычислении числа запускаемых блоков. Мы говорили о том, что скалярное произведение – частный случай редукции и что каждый запущенный блок вычисляет частичную сумму. Длина списка частичных сумм должна быть, с одной стороны, достаточно мала, чтобы не перегружать CPU, а с другой – достаточно велика, чтобы занять работой даже самый быстрый GPU. Мы выбрали 32 блока, хотя это тот случай, когда при другом выборе производительность может как повыситься, так и снизиться – в зависимости от соотношения скоростей CPU и GPU.

Но что, если входной список очень короткий и 32 блока по 256 нитей в каждом – слишком много? Если имеется N элементов данных, то для вычисления скалярного произведения понадобится только N нитей. Таким образом, мы должны выбрать наименьшее число, кратное `threadsPerBlock`, которое было бы больше или равно N . С подобной ситуацией мы уже встречались при сложении векторов. В данном случае искомая величина вычисляется по формуле $(N + (\text{threadsPerBlock} - 1)) / \text{threadsPerBlock}$. Это весьма распространенный прием при работе с целыми числами, поэтому имеет смысл запомнить его, даже если большую часть времени вы пишете программы, не имеющие отношения к CUDA C.

Итак, количество запускаемых блоков должно быть равно минимуму из двух чисел: 32 и $(N + (\text{threadsPerBlock} - 1)) / \text{threadsPerBlock}$.

```
const int blocksPerGrid =
    imin( 32, (N+threadsPerBlock-1) / threadsPerBlock );
```

Теперь должно быть понятно, откуда взялся код в функции `main()`. После завершения ядра нам еще предстоит просуммировать частичные результаты. Но сначала необходимо скопировать эти результаты в память CPU. Что мы и делаем:

```
// скопировать массив 'c' из памяти GPU в память CPU
HANDLE_ERROR( cudaMemcpy( partial_c, dev_partial_c,
                           blocksPerGrid*sizeof(float),
                           cudaMemcpyDeviceToHost ) );

// завершить вычисление на CPU
c = 0;
for (int i=0; i<blocksPerGrid; i++) {
    c += partial_c[i];
}
```

И наконец, мы проверяем правильность вычислений и освобождаем память, выделенную на GPU и CPU. Проверить результат просто, потому что мы заполнили входные массивы вполне определенными данными. Если помните, в массив a[] были записаны целые числа от 0 до N-1, а b[] – это просто 2*a[].

```
// заполнить память CPU данными
for (int i=0; i<N; i++) {
    a[i] = i;
    b[i] = i*2;
}
```

Наше скалярное произведение равно удвоенной сумме квадратов целых чисел от 0 до N-1. Читатели, которые любят дискретную математику (а кто ее не любит?!), возможно, захотят отвлечься и вывести формулу этой суммы. Ну а для тех, кому это неинтересно или не терпится, мы приведем и готовую формулу, и оставшуюся часть тела main():

```
#define sum_squares(x) (x*(x+1)*(2*x+1)/6)
printf( "Значение, вычисленное GPU %.6g = %.6g?\n", c,
        2 * sum_squares( (float)(N - 1) ) );

// освободить память GPU
cudaFree( dev_a );
cudaFree( dev_b );
cudaFree( dev_partial_c );

// освободить память CPU
delete [] a;
delete [] b;
delete [] partial_c;
}
```

Если вас утомили наши пояснения, то ниже приведен полный исходный текст, не прерываемый комментариями.

```
#include "../common/book.h"

#define imin(a,b) (a<b?a:b)

const int N = 33 * 1024;
const int threadsPerBlock = 256;
const int blocksPerGrid =
    imin( 32, (N+threadsPerBlock-1) / threadsPerBlock );

__global__ void dot( float *a, float *b, float *c ) {
    __shared__ float cache[threadsPerBlock];
    int tid = threadIdx.x + blockIdx.x * blockDim.x;
    int cacheIndex = threadIdx.x;
```

```
float temp = 0;
while (tid < N) {
    temp += a[tid] * b[tid];
    tid += blockDim.x * gridDim.x;
}

// сохранить значение в кэше
cache[cacheIndex] = temp;

// синхронизировать нити в этом блоке
__syncthreads();

// для успешной редукции threadsPerBlock должно быть степенью 2
// из-за следующего кода
int i = blockDim.x/2;
while (i != 0) {
    if (cacheIndex < i)
        cache[cacheIndex] += cache[cacheIndex + i];
    __syncthreads();
    i /= 2;
}

if (cacheIndex == 0)
    c[blockIdx.x] = cache[0];
}

int main( void ) {
    float *a, *b, c, *partial_c;
    float *dev_a, *dev_b, *dev_partial_c;

    a = (float *)malloc( N*sizeof(float) );
    b = (float *)malloc( N*sizeof(float) );
    partial_c = (float *)malloc( blocksPerGrid*sizeof(float) );

    // выделить память на GPU
    HANDLE_ERROR( cudaMalloc( (void**)&dev_a,
                           N*sizeof(float) ) );
    HANDLE_ERROR( cudaMalloc( (void**)&dev_b,
                           N*sizeof(float) ) );
    HANDLE_ERROR( cudaMalloc( (void**)&dev_partial_c,
                           blocksPerGrid*sizeof(float) ) );

    // заполнить память CPU данными
    for (int i=0; i< N; i++) {
        a[i] = i;
        b[i] = i*2;
    }

    // скопировать массивы 'a' и 'b' в память GPU
    HANDLE_ERROR( cudaMemcpy( dev_a, a, N*sizeof(float),
```

```
        cudaMemcpyHostToDevice ) );  
HANDLE_ERROR( cudaMemcpy( dev_b, b, N*sizeof(float),  
                           cudaMemcpyHostToDevice ) );  
  
dot<<<blocksPerGrid,threadsPerBlock>>> ( dev_a,  
                                              dev_b,  
                                              dev_partial_c );  
  
// скопировать массив 'c' из памяти GPU в память CPU  
HANDLE_ERROR( cudaMemcpy( partial_c, dev_partial_c,  
                           blocksPerGrid*sizeof(float),  
                           cudaMemcpyDeviceToHost ) );  
  
// завершить вычисление на CPU  
c = 0;  
for (int i=0; i<blocksPerGrid; i++) {  
    c += partial_c[i];  
}  
  
#define sum_squares(x) (x*(x+1)*(2*x+1)/6)  
printf( "Значение, вычисленное GPU %.6g = %.6g?\n", c,  
       2 * sum_squares( (float)(N - 1) ) );  
  
// освободить память GPU  
cudaFree( dev_a );  
cudaFree( dev_b );  
cudaFree( dev_partial_c );  
  
// освободить память CPU  
delete [] a;  
delete [] b;  
delete [] partial_c;  
}
```

5.3.1. Оптимизация скалярного произведения (неправильная)

Мы лишь мимоходом упомянули второй вызов `_syncthreads()` в задаче о вычислении скалярного произведения. А теперь присмотримся к нему внимательнее и попробуем кое-что улучшить. Как вы помните, второй вызов `_syncthreads()` нужен потому, что мы обновили массив `cache[]` в разделяемой памяти и хотим, чтобы изменения были видны всем нитям на следующей итерации цикла.

```
int i = blockDim.x/2;  
while (i != 0) {  
    if (cacheIndex < i)  
        cache[cacheIndex] += cache[cacheIndex + i];
```

```

    __syncthreads();
    i /= 2;
}

```

Обратите внимание, что массив `cache[]` изменяется, только если `cacheIndex` меньше `i`. Так как `cacheIndex` – не что иное, как `threadIdx.x`, то это означает, что лишь *часть* нитей изменяют элементы в разделяемом кэше. Поскольку смысл `__syncthreads` в том, чтобы изменения гарантированно произошли до последующей обработки, то вроде бы скорость работы должна увеличиться, если ждать только те нити, которые действительно пишут в разделяемую память. Чтобы реализовать это изменение, достаточно перенести синхронизацию внутрь предложения `if`:

```

int i = blockDim.x/2;
while (i != 0) {
    if (cacheIndex < i) {
        cache[cacheIndex] += cache[cacheIndex + i];
        __syncthreads();
    }
    i /= 2;
}

```

Увы, эта героическая попытка оптимизации не работает. Больше того, ситуация стала только хуже. GPU вообще перестает отвечать, и мы вынуждены принудительно завершать программу. Но как столь невинное на первый взгляд изменение могло привести к подобной катастрофе? Чтобы ответить на этот вопрос, поставьте себя на место нити, выполняющей последовательность команд. Все нити выполняют одни и те же команды, но с разными данными. Но что случится, если некоторая команда выполняется условно, например в предложении `if`? Тогда ее будет выполнять не каждая нить, правда? Пусть, к примеру, ядро содержит такой фрагмент кода, где некоторую переменную изменяют лишь нити с нечетными индексами:

```

int myVar = 0;
if( threadIdx.x % 2 )
    myVar = threadIdx.x;

```

Строку, выделенную полужирным шрифтом, будут обновлять только нити с нечетными индексами, потому что четные индексы не удовлетворяют условию `if(threadIdx.x % 2)`. Пока нечетные нити выполняют эту команду, четные не делают ничего. Ситуация, когда одни нити выполняют некоторую команду, а другие – нет, называется *расхождением нитей* (*thread divergence*). При нормальных обстоятельствах наличие расходящихся ветвей просто приводит к тому, что некоторые нити пропускают, пока другие выполняют команды, находящиеся внутри ветви.

Но в случае `__syncthreads()` результат оказывается трагическим. Архитектура CUDA гарантирует, что *ни одна* нить не начнет выполнять команду, следующую за

`__syncthreads()`, пока все нити в этом блоке не выполнят `__syncthreads()`. К несчастью, если вызов `__syncthreads()` находится в расходящейся ветви, то некоторые нити никогда не дойдут до него. Но так как ни одна команда после `__syncthreads()` не может быть выполнена, пока все нити не выполнят `__syncthreads()`, то аппаратура будет ждать эти нити. И ждать. И ждать. Вечно.

Именно такая ситуация возникла в задаче о скалярном произведении, когда мы перенесли вызов `__syncthreads()` внутрь блока `if`. Ни одна нить, для которой `cacheIndex` больше или равен `i`, не выполнит `__syncthreads()`. И в результате процессор зависнет в ожидании того, что никогда не произойдет.

```
if (cacheIndex < i) {  
    cache[cacheIndex] += cache[cacheIndex + i];  
    __syncthreads();  
}
```

Мораль сей басни такова: `__syncthreads()` – единственное средство, позволяющее гарантировать правильную работу массивно-параллельного приложения. Но обращаться с ним нужно с осторожностью, не забывая о возможных нежелательных последствиях.

5.3.2. Растровое изображение в разделяемой памяти

Мы рассмотрели примеры, в которых использовались разделяемая память и функция `__syncthreads()`, гарантирующая, что данные готовы для продолжения обработки. Может возникнуть искушение рискнуть и во имя быстродействия опустить `__syncthreads()`. Сейчас мы рассмотрим пример графического приложения, в котором без `__syncthreads()` результаты оказываются неверными. Мы приведем два снимка экрана: ожидаемый и получающийся, когда вызов `__syncthreads()` опущен. Второй вам не понравится.

Тело функции `main()` такое же, как в задаче о фрактале Джулиа, хотя на этот раз мы запускаем несколько нитей в одном блоке.

```
#include "cuda.h"  
#include "../common/book.h"  
#include "../common/cpu_bitmap.h"  
  
#define DIM 1024  
#define PI 3.1415926535897932f  
  
int main( void ) {  
    CPUBitmap bitmap( DIM, DIM );  
    unsigned char *dev_bitmap;  
  
    HANDLE_ERROR( cudaMalloc( (void**)&dev_bitmap,  
                            bitmap.image_size() ) );
```

```

dim3 grids(DIM/16,DIM/16);
dim3 threads(16,16);
kernel<<<grids,threads>>>( dev_bitmap );

HANDLE_ERROR( cudaMemcpy( bitmap.get_ptr(), dev_bitmap,
                         bitmap.image_size(),
                         cudaMemcpyDeviceToHost ) );
bitmap.display_and_exit();

cudaFree( dev_bitmap );
}

```

Как и в примере фрактала Джулиа, каждая нить вычисляет значение пикселя в одной точке. Сначала нить вычисляет координаты (x, y) «своей» точки результирующего изображения. Делается это так же, как при вычислении переменной `tid` в задаче о сложении векторов, только на этот раз измерения два.

```

__global__ void kernel( unsigned char *ptr ) {
    // отобразить пару threadIdx/blockIdx на позицию пикселя
    int x = threadIdx.x + blockIdx.x * blockDim.x;
    int y = threadIdx.y + blockIdx.y * blockDim.y;
    int offset = x + y * blockDim.x * gridDim.x;

```

Поскольку для кэширования результатов вычисления мы будем использовать буфер в разделяемой памяти, объявим его так, чтобы каждой нити в нашем блоке 16×16 соответствовал отдельный элемент.

```
_shared__ float shared[16][16];
```

Затем каждая нить вычисляет значение, помещаемое в буфер.

```

// вычислить значение в этой позиции
const float period = 128.0f;
shared[threadIdx.x][threadIdx.y] =
    255 * (sinf(x*2.0f*PI / period) + 1.0f) *
    (sinf(y*2.0f*PI / period) + 1.0f) / 4.0f;

```

И напоследок запишем эти значения в пиксель, поменяв местами `x` и `y`:

```

ptr[offset*4 + 0] = 0;
ptr[offset*4 + 1] = shared[15-threadIdx.x][15-threadIdx.y];
ptr[offset*4 + 2] = 0;
ptr[offset*4 + 3] = 255;
}

```

Признаем, вычисления не слишком осмысленны. Это просто сетка, состоящая из зеленых круглых пятен. После компиляции и запуска ядра получается изображение, показанное на рис. 5.5.

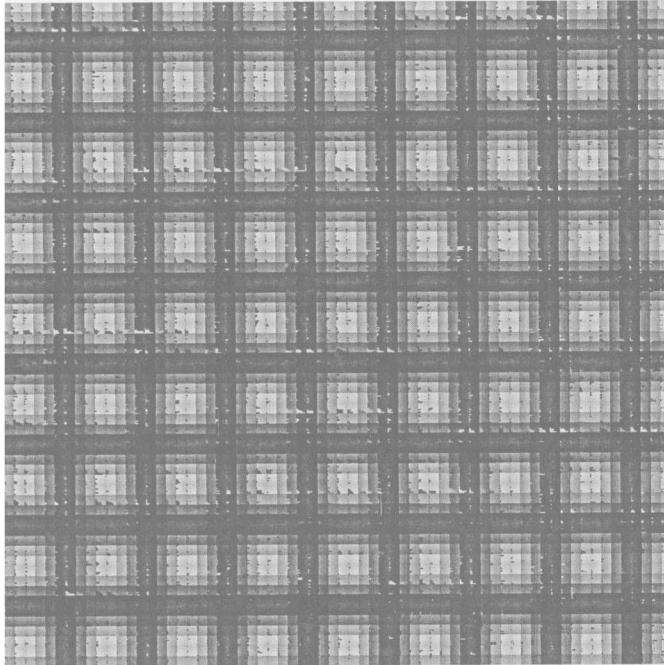


Рис. 5.5. Снимок экрана, полученный в результате работы программы без синхронизации

В чем дело? Как вы, наверное, догадались, мы опустили важную точку синхронизации. Вполне может случиться, что к моменту записи в пиксель значения, хранящегося в массиве `shared[][]`, нить, которая должна была поместить значение в данный элемент массива, еще не закончила работать. И гарантировать, что такого не произойдет, можно только одним способом – вызвав `__syncthreads()`. А результат налицо – искажения на картинке.

Это, конечно, не конец света, но программа могла бы производить и более важные вычисления.

Так что необходимо добавить точку синхронизации между записью в разделяемую память и последующим чтением из нее.

```
shared[threadIdx.x][threadIdx.y] =
    255 * (sinf(x*2.0f*PI/ period) + 1.0f) *
        (sinf(y*2.0f*PI/ period) + 1.0f) / 4.0f;

__syncthreads();

ptr[offset*4 + 0] = 0;
ptr[offset*4 + 1] = shared[15-threadIdx.x][15-threadIdx.y];
```

```
ptr[offset*4 + 2] = 0;  
ptr[offset*4 + 3] = 255;  
}
```

Теперь, когда вызов `__syncthreads()` на месте, мы получаем куда более предсказуемый (и эстетически приятный) результат, показанный на рис. 5.6.

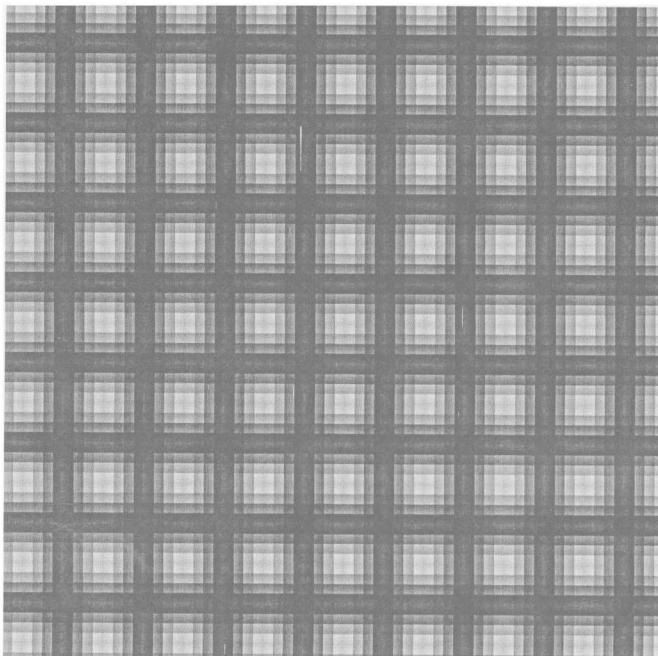


Рис. 5.6. Снимок экрана,
полученный после добавления синхронизации

5.4. Резюме

Теперь мы знаем, что блоки можно разбить на более мелкие единицы параллельного выполнения – *нити*. Мы переработали пример сложения векторов из предыдущей главы, показав, как можно складывать векторы произвольной длины. Мы также привели пример *редукции* и на нем продемонстрировали использование разделяемой памяти и синхронизации. На самом деле это пример сотрудничества GPU и CPU для достижения результата. Наконец, мы показали, какие опасности подстерегают того, кто пренебрежает синхронизацией.

Вы многое узнали об основах языка CUDA C, о тех аспектах, в которых он похож на стандартный C, и о тех, в которых он существенно отличается. Теперь самое время поразмыслить над своими задачами и прикинуть, какие из них можно было бы с успехом распараллелить за счет применения CUDA C. В дальнейшем мы познакомимся со средствами, позволяющими реализовать на GPU параллельные задачи, а также с некоторыми более сложными API, предлагаемыми архитектурой CUDA.



Глава 6. Константная память и события

Надеемся, что теперь вы лучше понимаете, как писать код, исполняемый GPU. Вы знаете, как запускать параллельные блоки, исполняющие ядро, и как расщеплять блоки на параллельные нити. Вы также видели, как осуществляются взаимодействие между нитями и их синхронизация. Но книга-то еще не закончилась, и, значит, в языке CUDA C есть и другие полезные средства.

Два таких средства мы и рассмотрим в этой главе. Точнее, есть несколько способов использовать специальные области памяти GPU для ускорения приложения. В данной главе мы поговорим об одной из таких областей – *константной памяти*. Кроме того, раз уж речь пойдет о повышении производительности приложения, то надо бы научиться измерять ее; для этого предназначены *события* CUDA. На основе результатов измерений вы сможете количественно оценить выигрыш (или проигрыш!), достигнутый в результате той или иной модификации.

6.1. О чём эта глава

Прочитав эту главу, вы:

- узнаете об использовании константной памяти в CUDA C;
- узнаете о характеристиках производительности константной памяти;
- узнаете о том, как измерять производительность с помощью событий CUDA.

6.2. Константная память

Выше мы говорили, что современные GPU обладают колоссальной вычислительной мощностью. Именно превосходство графических процессоров над CPU в плане скорости вычислений и привлекло интерес к GPU как к средству для выполнения вычислений общего назначения. При наличии сотен арифметических устройств узким местом GPU часто становится не скорость вычислений, а скорость обмена с памятью. В графическом процессоре так много АЛУ, что иногда мы просто не успеваем подавать им данные в темпе, достаточном для поддержания высокой скорости вычислений. Поэтому имеет смысл рассмотреть средства, с помощью которых можно уменьшить количество операций передачи данных между процессором и памятью.

До сих пор мы видели программы, в которых использовалась глобальная и разделяемая память. Но язык CUDA C поддерживает еще один вид памяти – *конс-*

тантную. Как следует из самого названия, константная память служит для хранения данных, которые не изменяются в процессе исполнения ядра. Оборудование NVIDIA предоставляет 64 Кб константной памяти, работа с которой организована иначе, чем со стандартной глобальной памятью. В некоторых случаях использование константной памяти вместо глобальной позволяет уменьшить трафик между памятью и процессором.

6.2.1. Введение в метод трассировки лучей

Один из способов использования константной памяти мы рассмотрим на примере простого приложения для *трассировки лучей*. Но сначала расскажем об основах этого метода. Если вы уже знакомы с трассировкой лучей, то можете сразу перейти к разделу «Трассировка лучей на GPU».

Говоря по-простому, трассировка лучей – это способ построения двумерного образа сцены, содержащей трехмерные объекты. Но не для этого ли изначально и предназначались GPU? Насколько это отличается от того, что делают OpenGL или DirectX, когда вы играете в свою любимую игру? Да, GPU действительно решают ту же задачу, но применяют для этого технику *растеризации*. О растеризации написано много хороших книг, поэтому не станем здесь подробно разъяснять различия. Достаточно сказать, что это совершенно разные методы решения одной и той же задачи.

Так каким же образом трассировка лучей позволяет получить образ трехмерной сцены? Идея проста. Выбираем на сцене точку, в которую помещаем воображаемую камеру. В этой упрощенной цифровой камере имеется датчик света, поэтому, чтобы получить изображение, мы должны решить, какой свет попадает в этот датчик. Каждый пиксель результирующего изображения будет иметь цвет и яркость луча света, падающего на датчик в камере.

Поскольку свет, падающий на точку датчика, может исходить из любого места на сцене, проще решать противоположную задачу. То есть, вместо того чтобы пытаться понять, какой луч освещает данный пиксель, мы проведем воображаемый луч из пикселя на сцену. Тогда каждый пиксель играет роль глаза, «смотрящего» на сцену. На рис. 6.1 показаны лучи, исходящие из каждого пикселя.

Чтобы вычислить, какой цвет «видит» каждый пиксель, мы проводим луч из этого пикселя через сцену, пока не упремся в какой-нибудь объект. Тогда мы говорим, что пиксель видит этот объект, и можем назначить ему цвет, исходя из цвета объекта. Вычисления, необходимые для трассировки луча, в основном сводятся к нахождению пересечений этого луча с объектами на сцене.

В более сложных моделях трассировки блестящие объекты могут отражать, а полупрозрачные – преломлять лучи света. В результате образуются вторичные лучи, третичные лучи и т. д. На самом деле одна из привлекательных особенностей метода трассировки лучей как раз и состоит в том, что очень просто реализовать базовую схему, а затем включать в алгоритм трассировки различные усложнения для получения более реалистичного изображения.

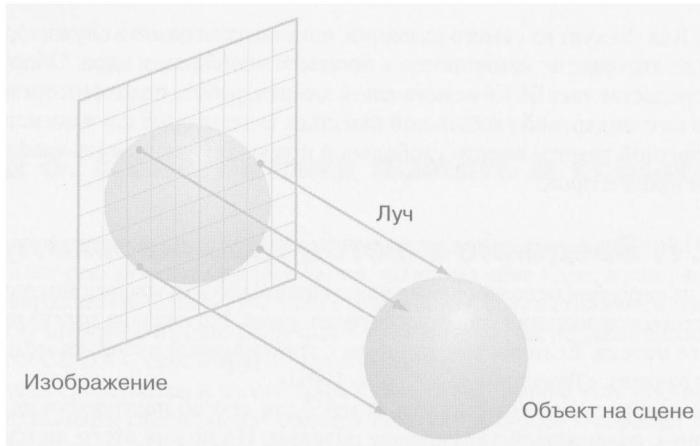


Рис. 6.1. Схема простой трассировки лучей

6.2.2. Трассировка лучей на GPU

Поскольку API типа OpenGL и DirectX не предназначены для трассировки лучей, то мы реализуем простейший трассировщик целиком на CUDA C. Мы не будем ничего усложнять, а сосредоточимся на вопросе о работе с константной памятью, поэтому если вы ожидаете увидеть основу для полноценного промышленного механизма рендеринга, то будете разочарованы. Наш трассировщик будет поддерживать только сцены, состоящие из сфер, а камера будет расположена на оси z и направлена в сторону начала координат. Кроме того, мы не поддерживаем освещение сцены, чтобы не связываться с вторичными лучами. Вместо того чтобы вычислять эффекты освещения, мы назначим каждой сфере один и тот же цвет и будем выбирать его оттенок с помощью предварительно заданной функции.

Тогда что же *делает* трассировщик? Он проводит луч из каждого пикселя и смотрит, какие сферы этот луч пересекает. Кроме того, он вычисляет расстояние до каждого пересечения. Если луч пересекает несколько сфер, то видимой считается только ближайшая. По сути дела, наш «трассировщик лучей» всего лишь скрывает поверхности, которые камера не видит.

Сфера моделируется с помощью структуры данных, содержащей ее центр (x , y , z), радиус и цвет (r , b , g).

```
#define INF 2e10f

struct Sphere {
    float r,b,g;
    float radius;
    float x,y,z;
__device__ float hit( float ox, float oy, float *n ) {
    float dx = ox - x;
    float dy = oy - y;
```

```

    if (dx*dx + dy*dy < radius*radius) {
        float dz = sqrtf( radius*radius - dx*dx - dy*dy );
        *n = dz / sqrtf( radius * radius );
        return dz + z;
    }
    return -INF;
}
};

```

В этой структуре имеется также метод `hit(float ox, float oy, float *n)`. Он определяет, пересекает ли данную сферу луч, проведенный из пикселя (`ox, oy`). Если да, то метод вычисляет расстояние от камеры до точки пересечения. Это нужно знать для того, чтобы выбрать ближайшую сферу в случае, когда луч пересекает несколько сфер, поскольку именно эта сфера будет видна.

Функция `main()` устроена примерно так же, как в ранее рассмотренных примерах генерации изображений.

```

#include "cuda.h"
#include "../common/book.h"
#include "../common/cpu_bitmap.h"

#define rnd( x ) (x * rand() / RAND_MAX)
#define SPHERES 20

Sphere *s;

int main( void ) {
    // запомнить время начала
    cudaEvent_t start, stop;
    HANDLE_ERROR( cudaEventCreate( &start ) );
    HANDLE_ERROR( cudaEventCreate( &stop ) );
    HANDLE_ERROR( cudaEventRecord( start, 0 ) );

    CPUBitmap bitmap( DIM, DIM );
    unsigned char *dev_bitmap;

    // выделить на GPU память для выходного изображения
    HANDLE_ERROR( cudaMalloc( (void**)&dev_bitmap,
                            bitmap.image_size( ) ) );
    // выделить память для набора сфер
    HANDLE_ERROR( cudaMalloc( (void**)&s,
                            sizeof(Sphere) * SPHERES ) );

```

Мы выделяем память для входных данных – массива сфер, составляющих сцену. Поскольку эти данные нужны GPU, но генерируются CPU, то необходимо выделить память и на GPU, и на CPU, обратившись соответственно к `cudaMalloc()` и `malloc()`. Кроме того, мы выделяем память для растрового изображения, которое будем заполнять данными о пикселях по мере трассировки сфер.

После того как вся необходимая память выделена, мы случайным образом генерируем координаты центров, цвета и радиусы сфер.

```
// выделить временную память, инициализировать ее, скопировать
// на GPU, а затем освободить
Sphere *temp_s = (Sphere*)malloc( sizeof(Sphere) * SPHERES );
for (int i=0; i<SPHERES; i++) {
    temp_s[i].r = rnd( 1.0f );
    temp_s[i].g = rnd( 1.0f );
    temp_s[i].b = rnd( 1.0f );
    temp_s[i].x = rnd( 1000.0f ) - 500;
    temp_s[i].y = rnd( 1000.0f ) - 500;
    temp_s[i].z = rnd( 1000.0f ) - 500;
    temp_s[i].radius = rnd( 100.0f ) + 20;
}
```

Сейчас программа генерирует массив из 20 случайных сфер, но эта константа задана в директиве `#define` и может быть легко изменена.

Затем массив сфер копируется в память GPU с помощью `cudaMemcpy()`, после чего временный буфер освобождается.

```
HANDLE_ERROR( cudaMemcpy( s, temp_s,
                         sizeof(Sphere) * SPHERES,
                         cudaMemcpyHostToDevice ) );
free( temp_s );
```

Теперь входные данные находятся в памяти GPU, память для выходных данных выделена, и мы можем запускать ядро.

```
// сгенерировать растровое изображение сфер
dim3 grids(DIM/16,DIM/16);
dim3 threads(16,16);
kernel<<<grids,threads>>>( dev_bitmap );
```

Само ядро мы рассмотрим чуть позже, а пока примите на веру, что оно действительно трассирует сцену и генерирует цвета пикселей. Напоследок мы копируем вычисленные данные в память CPU и отображаем их. И, разумеется, освобождаем всю выделенную память.

```
// скопировать изображение в память CPU для вывода на экран
HANDLE_ERROR( cudaMemcpy( bitmap.get_ptr(), dev_bitmap,
                         bitmap.image_size(),
                         cudaMemcpyDeviceToHost ) );
bitmap.display_and_exit();

// освободить память
cudaFree( dev_bitmap );
cudaFree( s );
}
```

Все это вам уже привычно и знакомо. Однако как все-таки производится трассировка луча? Поскольку мы выбрали очень простую модель трассировки, то в коде ядра разобраться совсем несложно. Каждая нить генерирует один пиксель выходного изображения, поэтому мы, как обычно, начинаем с вычисления координат (x, y) пикселя, соответствующего нити, и его смещения от начала буфера. Кроме того, мы сдвигаем точку (x, y) на $DIM/2$, чтобы ось z проходила через центр изображения.

```
__global__ void kernel( unsigned char *ptr ) {
    // отобразить пару threadIdx/BlockIdx на позицию пикселя
    int x = threadIdx.x + blockIdx.x * blockDim.x;
    int y = threadIdx.y + blockIdx.y * blockDim.y;
    int offset = x + y * blockDim.x * gridDim.x;
    float ox = (x - DIM/2);
    float oy = (y - DIM/2);
```

Так как для каждого луча нужно проверить пересечение со всеми сферами, то мы обходим массив сфер и производим необходимые вычисления:

```
float r=0, g=0, b=0;
float maxz = -INF;
for(int i=0; i<SPHERES; i++) {
    float n;
    float t = s[i].hit( ox, oy, &n );
    if (t > maxz) {
        float fscale = n;
        r = s[i].r * fscale;
        g = s[i].g * fscale;
        b = s[i].b * fscale;
        maxz = t;
    }
}
```

Очевидно, что самые интересные вычисления находятся внутри цикла `for`. Мы перебираем все сферы и для каждой вызываем метод `hit()`, чтобы узнать, «видит» ли луч эту сферу. Если да, то мы смотрим, расположена ли данная точка пересечения ближе к камере, чем последняя увиденная. Если да, то мы запоминаем расстояние до новой ближайшей сферы. Кроме того, запоминается цвет этой сферы, чтобы по выходе из цикла нить знала цвет сферы, ближайшей к камере. Поскольку именно этот цвет «видит» луч, проведенный из нашего пикселя, мы считаем его цветом самого пикселя и сохраняем значение в буфере результирующего изображения.

После того как пересечение со всеми сферами проверено, мы можем записать текущий цвет в выходное изображение:

```
ptr[offset*4 + 0] = (int)(r * 255);
ptr[offset*4 + 1] = (int)(g * 255);
```

```
ptr[offset*4 + 2] = (int)(b * 255);
ptr[offset*4 + 3] = 255;
}
```

Отметим, что если луч не пересекает ни одну сферу, то в качестве цвета будут записаны начальные значения переменных *r*, *b* и *g*. В данном случае мы инициализировали эти переменные нулями, то есть получится черный фон. Если хотите, чтобы фон был другого цвета, измените начальные значения. На рис. 6.2 показан пример картинки, получающейся при визуализации 20 сфер на черном фоне.

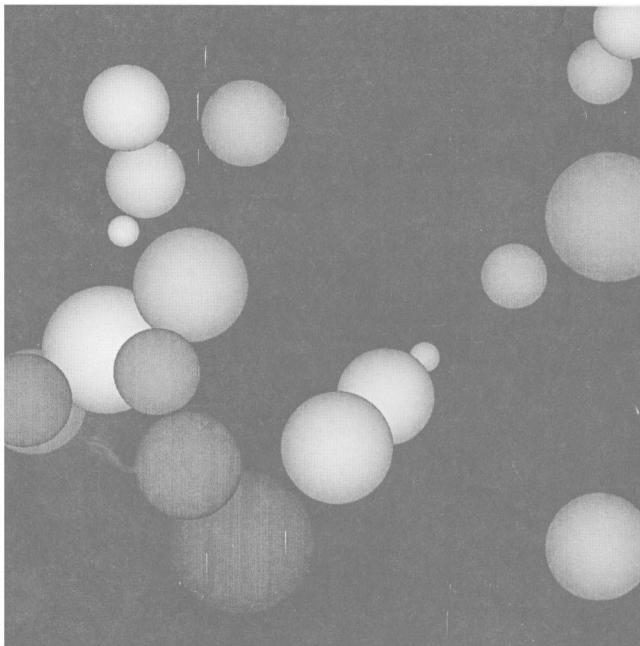


Рис. 6.2. Снимок экрана для программы трассировки лучей

Мы генерировали позиции, цвета и радиусы сфер случайным образом, так что не впадайте в панику, если у вас получится другая картинка.

6.2.3. Трассировка лучей с применением константной памяти

Вы, конечно, обратили внимание, что в предыдущем примере константная память нигде не упоминалась. Теперь улучшим программу, воспользовавшись преимуществами константной памяти. Поскольку константную память нельзя моди-

Константная память

фицировать, то использовать ее для хранения выходного изображения, конечно, невозможно. А коль скоро в этом примере входными данными является только массив сфер, то сомнений в том, что именно поместить в константную память, не возникает.

Для объявления константной памяти применяется такой же механизм, как для разделяемой памяти. Мы просто добавляем в начало объявления

```
Sphere *s;
```

модификатор `__constant__`:

```
__constant__ Sphere s[SPHERES];
```

Отметим, что в первой версии примера мы объявили указатель, а затем с помощью `cudaMalloc()` выделили для него область памяти GPU. При переходе на константную память мы изменили объявление так, что память выделяется статически. Теперь вызывать `cudaMalloc()` и `cudaFree()` для массива сфер не нужно, зато размер этого массива должен быть известен на этапе компиляции. Во многих приложениях это приемлемая плата за более высокое быстродействие константной памяти. Но прежде чем говорить о достигнутом выигрыше, посмотрим, как изменилась наша функция `main()`:

```
int main( void ) {
    CPUBitmap bitmap( DIM, DIM );
    unsigned char *dev_bitmap;

    // выделить на GPU память для выходного изображения
    HANDLE_ERROR( cudaMalloc( (void**)&dev_bitmap,
                            bitmap.image_size() ) );

    // выделить временную память, инициализировать ее, скопировать
    // в константную память GPU, а затем освободить
    Sphere *temp_s = (Sphere*)malloc( sizeof(Sphere) * SPHERES );
    for( int i=0; i<SPHERES; i++ ) {
        temp_s[i].r = rnd( 1.0f );
        temp_s[i].g = rnd( 1.0f );
        temp_s[i].b = rnd( 1.0f );
        temp_s[i].x = rnd( 1000.0f ) - 500;
        temp_s[i].y = rnd( 1000.0f ) - 500;
        temp_s[i].z = rnd( 1000.0f ) - 500;
        temp_s[i].radius = rnd( 100.0f ) + 20;
    }
    HANDLE_ERROR( cudaMemcpyToSymbol( s, temp_s,
                                    sizeof(Sphere) * SPHERES ) );
    free( temp_s );

    // сгенерировать растровое изображение сфер
    dim3 grids(DIM/16,DIM/16);
```

```

dim3 threads(16,16);
kernel<<< grids, threads>>>( dev_bitmap );

// скопировать изображение в память CPU для вывода на экран
HANDLE_ERROR( cudaMemcpy( bitmap.get_ptr(), dev_bitmap,
                         bitmap.image_size(),
                         cudaMemcpyDeviceToHost ) );
bitmap.display_and_exit();

// освободить память
cudaFree( dev_bitmap );
}

```

Изменений, по сравнению с первоначальной версией `main()`, немного. Как уже было сказано, теперь нет нужды вызывать `cudaMalloc()` для выделения памяти под массив сфер. Вторая модификация выделена полужирным шрифтом:

```
HANDLE_ERROR( cudaMemcpyToSymbol( s, temp_s,
                                 sizeof(Sphere) * SPHERES ) );


---


```

Этот специальный вариант `cudaMemcpy()` применяется при копировании из памяти CPU в константную память GPU. Единственное различие между `cudaMemcpyToSymbol()` и `cudaMemcpy()` с параметром `cudaMemcpyHostToDevice` заключается в том, что `cudaMemcpyToSymbol()` копирует в константную память, а `cudaMemcpy()` – в глобальную.

Если не считать модификатора `__constant__` и двух изменений `main()`, то версии с константной памятью и без нее идентичны.

6.2.4. Производительность версии с константной памятью

Включив в объявление модификатор `__constant__`, мы запрещаем всякую запись в область памяти. Но в обмен на такое самоограничение мы ожидаем что-то получить. Как отмечалось выше, чтение из константной памяти может уменьшить количество операций передачи между памятью и процессором по сравнению с чтением из глобальной памяти. Тому есть две причины:

- ❑ одно чтение из константной памяти может быть передано (broadcast) другим «близким» нитям, что позволяет сэкономить до 15 операций чтения;
- ❑ константная память кэшируется, поэтому последующие операции чтения из того же адреса не порождают дополнительного трафика.

Что понимается под словом **близкие**? Чтобы ответить на этот вопрос, нам придется рассказать о понятии *warp* (warp). Для тех читателей, которые лучше знакомы с фильмом «Стар трек», чем с текстильной промышленностью, скажем, что слово *warp* в этом контексте не имеет никакого отношения к скорости перемещения в космосе. В текстильной промышленности этим словом называется груп-

на спрятанных вместе *нитей* (основа), из которых ткется ткань. В архитектуре CUDA *варпом* (дословно – канат) называется группа из 32 «спрятанных» нитей, которые исполняются синхронно. Каждая нить, принадлежащая одному варпу, в любой момент времени исполняет одну и ту же команду над разными данными.

При чтении из константной памяти оборудование NVIDIA может транслировать прочитанные данные на целый полуварп. Полуварпом называется группа из 16 нитей – половина 32-нитевого варпа. Если каждая нить в полуварпе запрашивает данные из одного и того же адреса в константной памяти, то GPU выдает только один запрос на чтение, а затем передает прочитанные данные каждой нити. Если программа часто читает данные из константной памяти, то трафик сокращается в 16 раз по сравнению с чтением из глобальной памяти (примерно 6%).

Но экономия не ограничивается 94%-ным сокращением трафика между процессором и памятью! Поскольку мы пообещали, что память не будет изменяться, оборудование может кэшировать константные данные в GPU. Иными словами, после того как данные по некоторому адресу в константной памяти один раз прочитаны, нити из других полуварпов, читающие тот же адрес, обнаружат данные в кэше, поэтому повторного обращения к памяти не произойдет.

В примере трассировки лучей каждая запущенная нить читает параметры сферы, чтобы можно было проверить, пересекает ли ее луч. После перемещения данных о сферах в константную память аппаратуре достаточно отправить лишь один запрос на чтение из памяти. Затем данные кэшируются, и все остальные нити не порождают трафика по одной из двух причин:

- нить находится в том же полуварпе и уже получила данные в результате широковещательной трансляции;
- нить нашла данные в кэше константной памяти.

К сожалению, при работе с константной памятью есть и негативный эффект. Трансляция на полуварп – это палка о двух концах. Она может здорово повысить производительность, если все 16 нитей читают один и тот же адрес. Но если нити обращаются к разным адресам, то программа еле «шевелится».

Платой за трансляцию одной операции чтения на 16 нитей является то, что в каждый момент времени все 16 нитей могут помещаться на шину только один запрос чтения. Значит, если 16 нитям в полуварпе нужно прочитать разные данные из константной памяти, то эти нити выстраиваются в очередь, и на отправку запроса на чтение уходит в 16 раз больше времени. Если бы нити читали из глобальной памяти, то все запросы можно было бы отправить одновременно. В описанной ситуации чтение из константной памяти, скорее всего, окажется медленнее, чем из глобальной.

6.3. Измерение производительности с помощью событий

Понимая, что возможны как позитивные, так и негативные последствия, вы все же решаете изменить трассировщик лучей и перейти на константную память. Как узнать, что при этом произошло с производительностью программы? Одна из

простейших метрик требует ответа на вопрос: какая версия завершается быстрее? Можно было бы воспользоваться таймером CPU или операционной системы, но это повлечет за собой задержки и вариативность, обусловленные природой источника (планирование потоков ОС, доступность высокоточных таймеров CPU и т. д.). Кроме того, пока GPU исполняет ядро, на CPU могут асинхронно производиться другие вычисления. Измерить длительность этих вычислений можно только с помощью механизмов, встроенных в CPU или в операционную систему. А для изменения времени, которое затрачивает GPU, мы воспользуемся API событий CUDA.

В CUDA *событием* (event) называется временная метка GPU, запомненная в определенный пользователем момент времени. Поскольку GPU ставит эту метку сам, то мы обходим все проблемы, связанные с измерением времени GPU с помощью таймеров CPU. API несложен – для получения временной метки нужно проделать два шага: создать событие и зарегистрировать событие. Например, в начале некоторой последовательности предложений программы мы просим исполняющую среду отметить текущее время. Для этого сначала создается, а потом регистрируется событие:

```
cudaEvent_t start;
cudaEventCreate(&start);
cudaEventRecord( start, 0 );
```

Обратите внимание, что при регистрации события начала функции cudaEventRecord() передается второй аргумент. В примере выше он равен 0. Точный смысл этого аргумента нам сейчас не важен, поэтому не станем раскрывать эту тайну, пока не разворшим следующий муравейник. Если вас терзает неумное любопытство, то сообщим, что вернемся к этому вопросу при обсуждении *потоков* (stream).

Чтобы измерить время выполнения блока кода, нужно создать события начала и окончания. Мы просим исполняющую среду CUDA зарегистрировать время начала, затем выполнить какие-то действия на GPU, а по завершении зарегистрировать время окончания:

```
cudaEvent_t start, stop;
cudaEventCreate(&start);
cudaEventCreate(&stop);
cudaEventRecord( start, 0 );

// что-то сделать на GPU

cudaEventRecord( stop, 0 );
```

К сожалению, при таком хронометраже кода, исполняемого GPU, еще остается одна проблема. Чтобы исправить ее, нужна всего одна строчка, но не обойтись без пояснений. Штука в том, что некоторые обращения к исполняющей среде CUDA на самом деле выполняются *асинхронно*. Например, при запуске ядра

трассировщика лучей GPU начинает исполнять наш код, но CPU продолжает выполнять следующую строку программы, не дожидаясь окончания работы GPU. С точки зрения производительности это прекрасно, потому что вычисления на GPU и CPU могут производиться одновременно, но вот хронометраж осложняется.

Обращения к функции `cudaEventRecord()` удобно представлять себе как запросы на получение текущего времени, которые ставятся в очередь. Это означает, что регистрация события произойдет только после того, как GPU покончит со всеми заданиями, поставленными в очередь раньше `cudaEventRecord()`. Если речь идет об использовании события окончания для измерения правильного времени, то это как раз то, что нужно. Однако мы не можем безопасно *прочитать* значение события, пока GPU не закончит все предыдущие задания. К счастью, имеется простой способ заставить GPU синхронизироваться с событием – вызов функции `cudaEventSynchronize()`:

```
cudaEvent_t start, stop;
cudaEventCreate(&start);
cudaEventCreate(&stop);
cudaEventRecord( start, 0 );

// сделать что-то на GPU

cudaEventRecord( stop, 0 );
cudaEventSynchronize( stop );
```

Тем самым мы попросили исполняющую среду не приступать к следующей команде, пока GPU не достигнет события окончания. После того как функция `cudaEventSynchronize()` вернет управление, есть уверенность, что вся работа до события окончания уже завершена, поэтому мы можем без опаски прочитать зарегистрированную временную метку. Стоит отметить, что поскольку события CUDA реализованы в самом GPU, то они непригодны для хронометража смешанного кода, исполняемого как GPU, так и CPU. Иными словами, при попытке использовать события CUDA для чего-нибудь, кроме выполнения ядра и копирования с участием памяти устройства, результаты будут ненадежны.

6.3.1. Измерение производительности трассировщика лучей

Для хронометража нашего трассировщика лучей нужно создать события начала и окончания, как было только что описано. Ниже приведена хронометрируемая версия трассировщика без использования константной памяти.

```
int main( void ) {
    // запомнить время начала
    cudaEvent_t start, stop;
    HANDLE_ERROR( cudaEventCreate( &start ) );
    HANDLE_ERROR( cudaEventCreate( &stop ) );
```

```
HANDLE_ERROR( cudaEventRecord( start, 0 ) );  
  
CPUBitmap bitmap( DIM, DIM );  
unsigned char *dev_bitmap;  
  
// выделить на GPU память для выходного изображения  
HANDLE_ERROR( cudaMalloc( (void**)&dev_bitmap,  
                           bitmap.image_size() ) );  
  
// выделить память для набора сфер  
HANDLE_ERROR( cudaMalloc( (void**)&s,  
                           sizeof(Sphere) * SPHERES ) );  
  
// выделить временную память, инициализировать ее, скопировать  
// на GPU, а затем освободить  
Sphere *temp_s = (Sphere*)malloc( sizeof(Sphere) * SPHERES );  
for (int i=0; i<SPHERES; i++) {  
    temp_s[i].r = rnd( 1.0f );  
    temp_s[i].g = rnd( 1.0f );  
    temp_s[i].b = rnd( 1.0f );  
    temp_s[i].x = rnd( 1000.0f ) - 500;  
    temp_s[i].y = rnd( 1000.0f ) - 500;  
    temp_s[i].z = rnd( 1000.0f ) - 500;  
    temp_s[i].radius = rnd( 100.0f ) + 20;  
}  
HANDLE_ERROR( cudaMemcpy( s, temp_s,  
                           sizeof(Sphere) * SPHERES,  
                           cudaMemcpyHostToDevice ) );  
free( temp_s );  
  
// сгенерировать растровое изображение сфер  
dim3 grids(DIM/16,DIM/16);  
dim3 threads(16,16);  
kernel<<<grids,threads>>>( s, dev_bitmap );  
  
// скопировать изображение в память CPU для вывода на экран  
HANDLE_ERROR( cudaMemcpy( bitmap.get_ptr(), dev_bitmap,  
                           bitmap.image_size(),  
                           cudaMemcpyDeviceToHost ) );  
  
// получить время окончания и вывести результаты хронометража  
HANDLE_ERROR( cudaEventRecord( stop, 0 ) );  
HANDLE_ERROR( cudaEventSynchronize( stop ) );  
  
float elapsedTime;  
HANDLE_ERROR( cudaEventElapsedTime( &elapsedTime,  
                                    start, stop ) );  
printf( "Время генерации: %3.1f ms\n", elapsedTime );  
HANDLE_ERROR( cudaEventDestroy( start ) );
```

```
HANDLE_ERROR( cudaEventDestroy( stop ) );

// вывести изображение
bitmap.display_and_exit();

// освободить память
cudaFree( dev_bitmap );
cudaFree( s );
}
```

Обратите внимание на две новые функции: `cudaEventElapsedTime()` и `cudaEventDestroy()`. Функция `cudaEventElapsedTime()` просто вычисляет время в миллисекундах между двумя ранее зарегистрированными событиями. Результат возвращается в виде числа с плавающей точкой, адрес которого передан в первом аргументе.

Когда событие, созданное функцией `cudaEventCreate()`, перестает быть нужным, его следует уничтожить с помощью функции `cudaEventDestroy()`. Причина точно такая же, как при вызове `free()` для освобождения памяти, выделенной `malloc()`, поэтому не будем объяснять, почему так важно, чтобы каждый вызов `cudaEventCreate()` сопровождался вызовом `cudaEventDestroy()`.

Аналогично можно добавить хронометраж в версию трассировщика, в которой используется константная память.

```
int main( void ) {
    // запомнить время начала
    cudaEvent_t start, stop;
    HANDLE_ERROR( cudaEventCreate( &start ) );
    HANDLE_ERROR( cudaEventCreate( &stop ) );
    HANDLE_ERROR( cudaEventRecord( start, 0 ) );

    CPUBitmap bitmap( DIM, DIM );
    unsigned char *dev_bitmap;

    // выделить на GPU память для выходного изображения
    HANDLE_ERROR( cudaMalloc( (void**)&dev_bitmap,
                           bitmap.image_size() ) );

    // выделить временную память, инициализировать ее, скопировать
    // в константную память GPU, а затем освободить
    Sphere *temp_s = (Sphere*)malloc( sizeof(Sphere) * SPHERES );
    for( int i=0; i<SPHERES; i++ ) {
        temp_s[i].r = rnd( 1.0f );
        temp_s[i].g = rnd( 1.0f );
        temp_s[i].b = rnd( 1.0f );
        temp_s[i].x = rnd( 1000.0f ) - 500;
        temp_s[i].y = rnd( 1000.0f ) - 500;
        temp_s[i].z = rnd( 1000.0f ) - 500;
        temp_s[i].radius = rnd( 100.0f ) + 20;
    }
}
```

```
HANDLE_ERROR( cudaMemcpyToSymbol( s, temp_s,
                                  sizeof(Sphere) * SPHERES ) );
free( temp_s );

// сгенерировать растровое изображение сфер
dim3 grids(DIM/16,DIM/16);
dim3 threads(16,16);
kernel<<<grids,threads>>>( dev_bitmap );

// скопировать изображение в память CPU для вывода на экран
HANDLE_ERROR( cudaMemcpy( bitmap.get_ptr(), dev_bitmap,
                        bitmap.image_size(),
                        cudaMemcpyDeviceToHost ) );

// получить время окончания и вывести результаты хронометража
HANDLE_ERROR( cudaEventRecord( stop, 0 ) );
HANDLE_ERROR( cudaEventSynchronize( stop ) );

float elapsedTime;
HANDLE_ERROR( cudaEventElapsedTime( &elapsedTime,
                                    start, stop ) );
printf( "Время генерации: %3.1f ms\n", elapsedTime );
HANDLE_ERROR( cudaEventDestroy( start ) );
HANDLE_ERROR( cudaEventDestroy( stop ) );

// вывести изображение
bitmap.display_and_exit();

// освободить память
cudaFree( dev_bitmap );
}
```

Теперь, имея две версии трассировщика, мы можем сравнить время их выполнения на GPU. Это даст общее представление о том, повысилась производительность в результате перехода на константную память или наоборот. В данном случае производительность возросла – и очень заметно. Наши эксперименты на GeForce GTX 280 показали, что трассировщик с константной памятью работает примерно на 50% быстрее трассировщика с глобальной памятью. На других GPU цифры могут отличаться, но в любом случае трассировщик с константной памятью должен работать не медленнее, чем без нее.

6.4. Резюме

Помимо глобальной и разделяемой памяти, рассмотренной в предыдущих главах, оборудование NVIDIA предоставляет и другие виды памяти. Константная память обременена дополнительными ограничениями по сравнению с глобальной, но в некоторых случаях, согласившись ограничить себя, мы можем выиграть

Резюме

в производительности. Точнее, выигрыш появляется, когда все потоки в одном warpе должны обращаться к одним и тем же данным только для чтения. Если обращения к константной памяти устроены именно так, то трафик между памятью и процессором уменьшается за счет трансляции результатов операции чтения на полуварп и за счет кэша памяти на кристалле. Доступ к памяти является узким местом во многих алгоритмах, поэтому наличие механизма, позволяющего улучшить ситуацию в этом отношении, может оказаться чрезвычайно полезным.

Мы также научились применять события CUDA для регистрации временных меток в конкретные моменты времени выполнения программы на GPU. Мы видели, как синхронизировать CPU и GPU по таким событиям и как вычислить промежуток времени между двумя событиями. Попутно мы сравнили время работы двух версий трассировщика лучей и пришли к выводу, что в этом случае использование константной памяти дает ощутимый выигрыш.



Глава 7. Текстурная память

При рассмотрении константной памяти мы видели, что при определенных условиях использование специальных видов памяти может заметно ускорить работу приложения. Мы также научились измерять выигрыш и тем самым получать информацию для принятия обоснованного решения. В этой главе мы покажем, как выделить и использовать *текстурную память*. Это еще один вид памяти, предназначенный только для чтения и позволяющей повысить производительность и сократить трафик между процессором и памятью при определенных способах доступа. Хотя изначально текстурная память предназначалась для традиционных графических программ, ее можно с успехом применять и в некоторых вычислительных приложениях GPU.

7.1. О чём эта глава

Прочитав эту главу, вы:

- узнаете о характеристиках производительности текстурной памяти;
- научитесь использовать одномерную текстурную память в CUDA C;
- научитесь использовать двумерную текстурную память в CUDA C.

7.2. Обзор текстурной памяти

Если вы прочли введение к этой главе, то уже знаете секрет: в программах на CUDA C можно использовать еще один вид доступной только для чтения памяти. Читатели, знакомые с работой графической аппаратуры, вряд ли этому удивятся. Однако же хитроумный механизм текстурной памяти GPU применим и к вычислениям общего назначения. Хотя NVIDIA проектировала текстурные устройства для поддержки классических конвейеров рендеринга OpenGL и DirectX, у текстурной памяти есть некоторые свойства, делающие ее чрезвычайно полезной для вычислений вообще.

Как и константная память, текстурная память кэшируется на кристалле, поэтому в некоторых случаях позволяет уменьшить количество обращений к внешнему DRAM. Если быть точным, текстурные кэши предназначены для графических приложений, в которых доступ к памяти характеризуется высокой *пространственной локальностью*. В вычислительном приложении общего назначения это означает, что нить с большой вероятностью будет обращаться к адресам, расположенным «рядом» с адресами, к которым обращаются близкие нити (см. рис. 7.1).

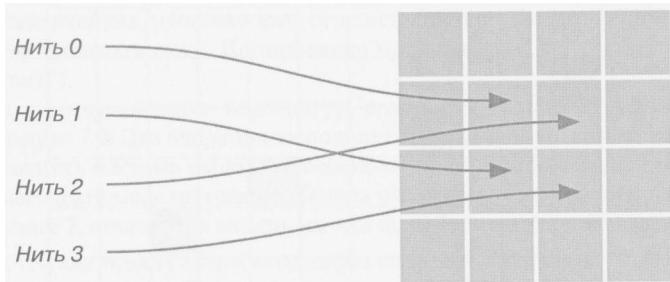


Рис. 7.1. Соответствие между нитями и двумерной областью памяти

С точки зрения арифметики, четыре показанных на рисунке адреса не являются соседними, поэтому не будут кэшироваться вместе при использовании стандартной схемы кэширования, применяемой в CPU. Но текстурные кэши GPU специально разработаны для ускорения доступа в таких ситуациях, поэтому применение текстурной памяти вместо глобальной может дать выигрыш. А, как мы увидим ниже, такой характер доступа к памяти вовсе не является чем-то исключительным.

7.3. Моделирование теплообмена

Физические модели – одни из самых вычислительно сложных задач. В них надо находить компромисс между точностью и вычислительной сложностью. Поэтому в последние годы – во многом благодаря революции в параллельных вычислениях, сделавшей возможным увеличение точности, – компьютерное моделирование стало играть столь важную роль. Многие физические модели легко распараллеливаются, и в этом разделе мы рассмотрим одну очень простую модель такого рода.

7.3.1. Простая модель теплообмена

Чтобы продемонстрировать ситуацию, в которой можно эффективно задействовать текстурную память, мы построим простую двумерную модель теплообмена. Для начала предположим, что имеется прямоугольная комната, которую мы разобьем на одинаковые ячейки. Внутри сетки случайным образом расположены «нагреватели» разной температуры. На рис. 7.2 показан пример такой сетки.

Имея прямоугольную сетку и зная конфигурацию нагревателей, мы хотим построить модель, которая позволит узнать, как изменяется температура каждой ячейки со временем. Для простоты будем считать, что ячейки с нагревателями сохраняют постоянную температуру. На каждом шаге моделирования тепло «переносится» из каждой ячейки в соседние. Если соседняя ячейка теплее данной, то данная ячейка нагревается. Наоборот, если соседняя ячейка холоднее данной, то данная ячейка остывает. На рис. 7.3 изображена качественная картина теплообмена.

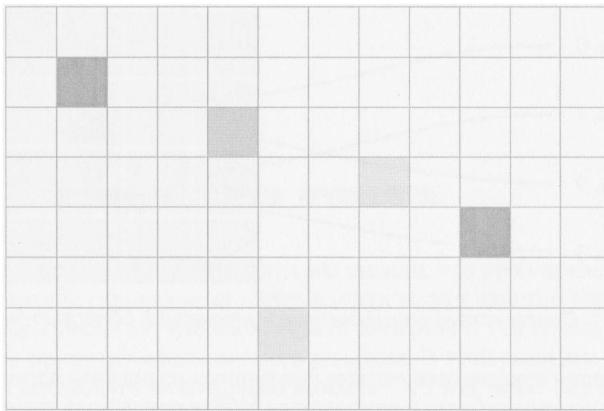


Рис. 7.2. Комната с «нагревателями» разной температуры

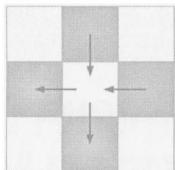


Рис. 7.3. Перенос тепла из теплых ячеек в холодные

В нашей модели теплообмена новая температура ячейки вычисляется как сумма разностей между ее текущей температурой и температурами ее соседей, то есть по формуле 7.1.

$$T_{\text{new}} = T_{\text{old}} + \sum_{\text{по соседям}} k(T_{\text{neighbor}} - T_{\text{old}}) \quad (7.1)$$

В этой формуле постоянная k описывает скорость переноса тепла. При больших значениях k система быстро перейдет в равновесное состояние, когда температура каждой ячейки перестанет изменяться, а при малых градиентах температуры будет сохраняться дольше. Так как мы рассматриваем только четырех соседей (верхний, нижний, левый и правый), а k и T_{old} постоянны, то эту формулу можно переписать следующим образом:

$$T_{\text{new}} = T_{\text{old}} + k \cdot (T_{\text{top}} + T_{\text{bottom}} + T_{\text{left}} + T_{\text{right}} - 4 \cdot T_{\text{old}}) \quad (7.2)$$

Как и в задаче о трассировке лучей из предыдущей главы, эта модель не претендует на близость к практически применимым моделям (на самом деле она не является даже отдаленным приближением к физической реальности). Мы до предела все упростили, чтобы сосредоточиться на применяемой технике программирования. Имея это в виду, посмотрим все же, как по формуле 7.2 производятся вычисления на GPU.

7.3.2. Обновление температур

Детали каждого шага мы обсудим чуть ниже, а в целом процедура обновления состояния выглядит следующим образом:

1. Имея сетку входных температур, скопировать в нее ячейки с нагревателями. Тем самым мы сотрем ранее вычисленные значения температуры


```
// отобразить пару threadIdx/BlockIdx на позицию пикселя
int x = threadIdx.x + blockIdx.x * blockDim.x;
int y = threadIdx.y + blockIdx.y * blockDim.y;
int offset = x + y * blockDim.x * gridDim.x;

int left = offset - 1;
int right = offset + 1;
if (x == 0) left++;
if (x == DIM-1) right--;

int top = offset - DIM;
int bottom = offset + DIM;
if (y == 0) top += DIM;
if (y == DIM-1) bottom -= DIM;

outSrc[offset] = inSrc[offset] + SPEED * (inSrc[top] +
    inSrc[bottom] + inSrc[left] + inSrc[right] -
    inSrc[offset]*4);
}
```

Начало точно такое же, как в задачах генерации изображений. Но вместо цвета пикселя нить вычисляет температуру ячейки. И тем не менее индексы threadIdx и blockIdx преобразуются все в те же x, y и offset. Вы, наверное, можете отбара-банить эти строки даже во сне (хотя надеемся, что делать этого вам не придется).

Далее мы вычисляем смещения всех четырех соседних ячеек, чтобы затем прочесть значения температуры в них. Эти значения понадобятся для вычисле-ния новой температуры в текущей ячейке. Единственная сложность – граничные ячейки, для которых индексы следует подкорректировать, чтобы не выйти за границы сетки. И в последней, выделенной строке производится вычисление по формуле 7.2.

7.3.3. Анимация моделирования

В оставшейся части программы мы инициализируем сетку, а затем отображаем анимированную карту распределения тепла. Рассмотрим этот код.

```
#include "cuda.h"
#include "../common/book.h"
#include "../common/cpu_anim.h"

#define DIM 1024
#define PI 3.1415926535897932f
#define MAX_TEMP 1.0f
#define MIN_TEMP 0.0001f
#define SPEED 0.25f

// глобальные данные, необходимые функции обновления
struct DataBlock {
```

Моделирование теплообмена

```

unsigned char *output_bitmap;
float      *dev_inSrc;
float      *dev_outSrc;
float      *dev_constSrc;
CPUAnimBitmap *bitmap;
cudaEvent_t start, stop;
float      totalTime;
float      frames;

};

void anim_gpu( DataBlock *d, int ticks ) {
    HANDLE_ERROR( cudaEventRecord( d->start, 0 ) );
    dim3 blocks(DIM/16,DIM/16);
    dim3 threads(16,16);
    CPUAnimBitmap *bitmap = d->bitmap;

    for (int i=0; i<90; i++) {
        copy_const_kernel<<<blocks,threads>>>( d->dev_inSrc,
                                                d->dev_constSrc );
        blend_kernel<<<blocks,threads>>>( d->dev_outSrc,
                                                d->dev_inSrc );
        swap( d->dev_inSrc, d->dev_outSrc );
    }
    float_to_color<<<blocks,threads>>>( d->output_bitmap,
                                            d->dev_inSrc );

    HANDLE_ERROR( cudaMemcpy( bitmap->get_ptr(),
                            d->output_bitmap,
                            bitmap->image_size(),
                            cudaMemcpyDeviceToHost ) );
    HANDLE_ERROR( cudaEventRecord( d->stop, 0 ) );
    HANDLE_ERROR( cudaEventSynchronize( d->stop ) );
    float elapsedTime;
    HANDLE_ERROR( cudaEventElapsedTime( &elapsedTime,
                                       d->start, d->stop ) );
    d->totalTime += elapsedTime;
    ++d->frames;
    printf( "Среднее время на один кадр: %3.1f ms\n",
            d->totalTime/d->frames );
}

void anim_exit( DataBlock *d ) {
    cudaFree( d->dev_inSrc );
    cudaFree( d->dev_outSrc );
    cudaFree( d->dev_constSrc );

    HANDLE_ERROR( cudaEventDestroy( d->start ) );
    HANDLE_ERROR( cudaEventDestroy( d->stop ) );
}

```

Мы включили в код хронометраж на базе событий, как в задаче о трассировке лучей из предыдущей главы. Назначение хронометража то же, что и раньше. Поскольку мы собираемся ускорить первоначальную реализацию, то сразу позаботились о механизме измерения производительности, чтобы убедиться в успехе своего начинания.

Каркас анимации вызывает функцию `anim_gpu()` в каждом кадре. В качестве аргументов ей передается указатель на структуру `DataBlock` и количество уже прошедших тактов анимации. Как и в примерах анимации выше, мы используем блоки из 256 нитей, организованные в сетку 16×16 . На каждой итерации цикла `for` в функции `anim_gpu()` вычисляется один временной шаг моделирования, описанный в начале раздела 7.3.2 «Обновление температур». Поскольку структура `DataBlock` содержит буфер постоянных температур нагревателей, а также результат последнего шага моделирования, то она инкапсулирует все состояние анимации и, следовательно, `anim_gpu()` вообще не нуждается в переменной `ticks`.

Обратите внимание, что в одном кадре мы выполняем 90 шагов моделирования. Это не какое-то волшебное число, а величина, которая была экспериментально выбрана в качестве разумного компромисса между необходимостью выгружать растр на каждом шаге и вычислением слишком большого числа шагов в одном кадре, что приводит к прерывистой анимации. Если вас больше интересует наблюдение за результатом каждого шага моделирования, чем анимация в реальном времени, то можете изменить тело функции так, чтобы в каждом кадре вычислялся ровно один шаг.

После вычисления 90 шагов моделирования с момента окончания предыдущего кадра функция `anim_gpu()` готова скопировать растровое изображение текущего кадра назад в память CPU. Поскольку цикл `for` оставил входной и выходной буфера переставленными, мы передаем следующему ядру входной буфер, который в действительности содержит результат, получившийся после 90-го шага моделирования. Мы преобразуем температуры в цвета с помощью ядра `float_to_color()`, а затем копируем результирующее изображение в память CPU, вызывая функцию `cudaMemcpy()` с параметром `cudaMemcpyDeviceToHost`. Наконец, чтобы подготовиться к следующей последовательности шагов моделирования, мы обмениваем входной буфер с выходным, поскольку последний будет выступать в роли входного на следующем шаге моделирования.

```
int main( void ) {
    DataBlock data;
    CPUAnimBitmap bitmap( DIM, DIM, &data );
    data.bitmap = &bitmap;
    data.totalTime = 0;
    data.frames = 0;
    HANDLE_ERROR( cudaEventCreate( &data.start ) );
    HANDLE_ERROR( cudaEventCreate( &data.stop ) );
    HANDLE_ERROR( cudaMalloc( (void**)&data.output_bitmap,
                           bitmap.image_size() ) );
```

Моделирование теплообмена

```

// предполагаем, что размер float равен 4 байтам (то есть rgba)
HANDLE_ERROR( cudaMalloc( (void**)&data.dev_inSrc,
                           bitmap.image_size() ) );
HANDLE_ERROR( cudaMalloc( (void**)&data.dev_outSrc,
                           bitmap.image_size() ) );
HANDLE_ERROR( cudaMalloc( (void**)&data.dev_constSrc,
                           bitmap.image_size() ) );

float *temp = (float*)malloc( bitmap.image_size() );
for (int i=0; i<DIM*DIM; i++) {
    temp[i] = 0;
    int x = i % DIM;
    int y = i / DIM;
    if ((x>300) && (x<600) && (y>310) && (y<601))
        temp[i] = MAX_TEMP;
}
temp[DIM*100+100] = (MAX_TEMP + MIN_TEMP)/2;
temp[DIM*700+100] = MIN_TEMP;
temp[DIM*300+300] = MIN_TEMP;
temp[DIM*200+700] = MIN_TEMP;
for (int y=800; y<900; y++) {
    for (int x=400; x<500; x++) {
        temp[x+y*DIM] = MIN_TEMP;
    }
}
HANDLE_ERROR( cudaMemcpy( data.dev_constSrc, temp,
                           bitmap.image_size(),
                           cudaMemcpyHostToDevice ) );

for (int y=800; y<DIM; y++) {
    for (int x=0; x<200; x++) {
        temp[x+y*DIM] = MAX_TEMP;
    }
}
HANDLE_ERROR( cudaMemcpy( data.dev_inSrc, temp,
                           bitmap.image_size(),
                           cudaMemcpyHostToDevice ) );

free( temp );

bitmap.anim_and_exit( (void (*)(void*,int))anim_gpu,
                      (void (*)(void*))anim_exit );
}

```

На рис. 7.4 показано, как может выглядеть картина теплообмена. На рисунке видны некоторые нагреватели, которые выглядят как островки размером в пиксель, нарушающие непрерывность распределения температур.

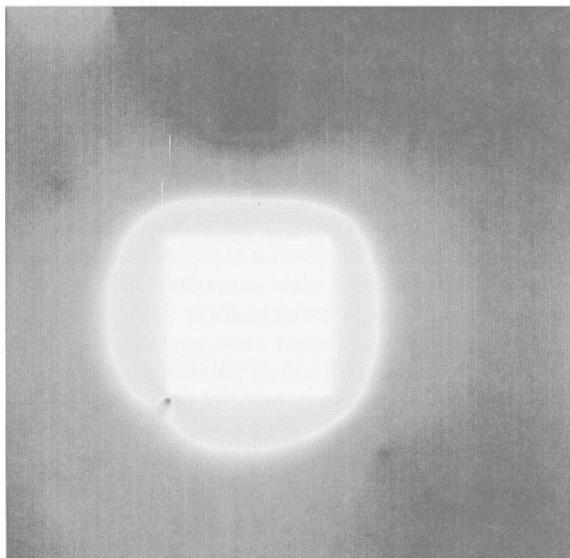


Рис. 7.4. Снимок экрана, полученный в процессе моделирования теплообмена

7.3.4. Применение текстурной памяти

Доступ к памяти в алгоритме пересчета температур на каждом шаге характеризуется *пространственной локальностью*. Выше мы говорили, что именно для ускорения работы в этом случае и была придумана текстурная память. Но, прежде чем воспользоваться текстурной памятью, надо понять механизм ее работы.

Сначала мы должны объявить входные данные как ссылки на текстуры. Мы используем текстуры с плавающей точкой, потому что именно в таком виде представлены данные о температуре.

```
// эти данные находятся в памяти GPU
texture<float> texConstSrc;
texture<float> texIn;
texture<float> texOut;
```

Следующая особенность заключается в том, что после выделения памяти GPU для всех трех буферов мы должны *привязать* эти ссылки к буферам в памяти с помощью функции `cudaBindTexture()`¹. Тем самым мы сообщаем исполняющей среде CUDA две вещи:

¹ Обратите внимание, что в версии CUDA 3.2 и выше механизм работы с текстурной памятью был несколько изменен.

- что собираемся использовать указанный буфер в качестве текстуры;
- что собираемся использовать указанную ссылку в качестве «имени» текстуры.

В программе моделирования теплообмена после трех операций выделения памяти мы привязываем ко всем трем областям ранее объявленные ссылки (`texConstSrc`, `texIn` и `texOut`).

```

HANDLE_ERROR( cudaMalloc( (void**)&data.dev_inSrc,
                          imageSize ) );
HANDLE_ERROR( cudaMalloc( (void**)&data.dev_outSrc,
                          imageSize ) );
HANDLE_ERROR( cudaMalloc( (void**)&data.dev_constSrc,
                          imageSize ) );

HANDLE_ERROR( cudaBindTexture( NULL, texConstSrc,
                               data.dev_constSrc,
                               imageSize ) );
HANDLE_ERROR( cudaBindTexture( NULL, texIn,
                               data.dev_inSrc,
                               imageSize ) );
HANDLE_ERROR( cudaBindTexture( NULL, texOut,
                               data.dev_outSrc,
                               imageSize ) );

```

В этот момент текстуры полностью настроены, и мы готовы запустить ядро. Однако при чтении данных из текстур внутри ядра необходимо вызывать специальные функции, которые говорят GPU, что запросы должны проходить через текстурный блок, а не через стандартную глобальную память. Поэтому для чтения из буферов обычные квадратные скобки уже не годятся; нужно модифицировать `blend_kernel()`, так чтобы для чтения из памяти вызывалась функция `tex1Dfetch()`.

Между глобальной и текстурной памятью имеется и еще одно различие, вынуждающее проделать дополнительное изменение. Хотя `tex1Dfetch()` выглядит как вызов функции, на самом деле ее код генерирует компилятор. А поскольку ссылки на текстуры должны быть объявлены глобально в области видимости файла, то мы больше не можем передавать входной и выходной буфера в виде параметров `blend_kernel()`, потому что еще на этапе компиляции компилятор должен знать, к каким текстурам будет обращаться `tex1Dfetch()`. Вместо того чтобы передавать указатели на входной и выходной буфера, мы теперь передаем `blend_kernel()` булевский флаг `dstOut`, который показывает, какой буфер входной, а какой – выходной. Ниже изменения в коде `blend_kernel()` выделены полужирным шрифтом:

```

__global__ void blend_kernel( float *dst,
                             bool dstOut ) {
    . . . // отобразить пару threadIdx/BlockIdx на позицию пикселя
    int x = threadIdx.x + blockIdx.x * blockDim.x;
    int y = threadIdx.y + blockIdx.y * blockDim.y;

```

```

int offset = x + y * blockDim.x * gridDim.x;

int left = offset - 1;
int right = offset + 1;
if (x == 0) left++;
if (x == DIM-1) right--;

int top = offset - DIM;
int bottom = offset + DIM;
if (y == 0) top += DIM;
if (y == DIM-1) bottom -= DIM;

float t, l, c, r, b;
if (dstOut) {
    t = tex1Dfetch(texIn,top);
    l = tex1Dfetch(texIn,left);
    c = tex1Dfetch(texIn,offset);
    r = tex1Dfetch(texIn,right);
    b = tex1Dfetch(texIn,bottom);
} else {
    t = tex1Dfetch(texOut,top);
    l = tex1Dfetch(texOut,left);
    c = tex1Dfetch(texOut,offset);
    r = tex1Dfetch(texOut,right);
    b = tex1Dfetch(texOut,bottom);
}
dst[offset] = c + SPEED * (t + b + r + l - 4 * c);
}

```

Ядро `copy_const_kernel()` читает из буфера, в котором хранятся позиции и температуры нагревателей, поэтому необходимо проделать аналогичное изменение, чтобы чтение производилось из текстурной, а не глобальной памяти:

```

__global__ void copy_const_kernel( float *iptr ) {
    // отобразить пару threadIdx/BlockIdx на позицию пикселя
    int x = threadIdx.x + blockIdx.x * blockDim.x;
    int y = threadIdx.y + blockIdx.y * blockDim.y;
    int offset = x + y * blockDim.x * gridDim.x;

    float c = tex1Dfetch(texConstSrc,offset);
    if (c != 0)
        iptr[offset] = c;
}

```

Так как функция `blend_kernel()` теперь принимает флаг, описывающий назначение буферов, то необходимо внести соответствующее исправление в функцию `anim_gpu()`. Вместо того чтобы менять местами сами буфера, мы меняем значение флага на противоположное – `dstOut = !dstOut` – после каждой серии вызовов:

```

void anim_gpu( DataBlock *d, int ticks ) {
    HANDLE_ERROR( cudaEventRecord( d->start, 0 ) );
    dim3 blocks(DIM/16,DIM/16);
    dim3 threads(16,16);
    CPUAnimBitmap *bitmap = d->bitmap;

    // так как текстура глобальная и привязанная, то мы используем
    // флаг, который указывает, какой буфер является на данной итерации
    // входным, а какой - выходным
    volatile bool dstOut = true;
    for (int i=0; i<90; i++) {
        float *in, *out;
        if (dstOut) {
            in = d->dev_inSrc;
            out = d->dev_outSrc;
        } else {
            out = d->dev_inSrc;
            in = d->dev_outSrc;
        }
        copy_const_kernel<<<blocks,threads>>>( in );
        blend_kernel<<<blocks,threads>>>( out, dstOut );
        dstOut = !dstOut;
    }
    float_to_color<<<blocks,threads>>>( d->output_bitmap,
                                              d->dev_inSrc );

    HANDLE_ERROR( cudaMemcpy( bitmap->get_ptr(),
                           d->output_bitmap,
                           bitmap->image_size(),
                           cudaMemcpyDeviceToHost ) );

    HANDLE_ERROR( cudaEventRecord( d->stop, 0 ) );
    HANDLE_ERROR( cudaEventSynchronize( d->stop ) );
    float elapsedTime;
    HANDLE_ERROR( cudaEventElapsedTime( &elapsedTime,
                                       d->start, d->stop ) );
    d->totalTime += elapsedTime;
    ++d->frames;
    printf( "Среднее время на кадр: %3.1f ms\n",
           d->totalTime/d->frames );
}

```

И последнее изменение касается очистки в конце работы приложения. Мы должны не только освобождать глобальные буфера, но и отвязать текстуры:

```

// очистить память, выделенную на GPU
void anim_exit( DataBlock *d ) {
    cudaUnbindTexture( texIn );
}

```

```

cudaUnbindTexture( texOut );
cudaUnbindTexture( texConstSrc );
cudaFree( d->dev_inSrc );
cudaFree( d->dev_outSrc );
cudaFree( d->dev_constSrc );

HANDLE_ERROR( cudaEventDestroy( d->start ) );
HANDLE_ERROR( cudaEventDestroy( d->stop ) );
}

```

7.3.5. Использование двумерной текстурной памяти

В начале этой книги мы отметили, что некоторые задачи по природе своей двумерны, поэтому иногда удобно использовать двумерные блоки и сетки. То же самое верно и в отношении текстурной памяти. Есть много случаев, когда полезно иметь двумерную область памяти; для человека, знакомого с многомерными массивами в стандартном С, это неудивительно. Посмотрим, как можно модифицировать программу расчета теплообмена с применением двумерных текстур.

Во-первых, изменим объявление ссылок на текстуры. Если не указано противное, ссылки на текстуры по умолчанию одномерны, поэтому добавим аргумент 2, задающий размерность.

```

texture<float,2> texConstSrc;
texture<float,2> texIn;
texture<float,2> texOut;

```

Упрощение, связанное с переходом на двумерные текстуры, проявляется в функции `blend_kernel()`. Нам придется заменить обращения к `tex1Dfetch()` обращениями к `tex2D()`, зато больше не нужна переменная `offset`, которая применялась для вычисления линейных смещений `top`, `left`, `right` и `bottom`. Двумерную текстуру мы можем адресовать непосредственно с помощью координат `x` и `y`.

Далее. После перехода на `tex2D()` уже не нужно беспокоиться о выходе за границы массива. Если `x` или `y` меньше нуля, то `tex2D()` вернет значение, находящееся в позиции, у которой соответствующая координата равна нулю. Аналогично если какая-то координата больше ширины, то `tex2D()` вернет значение в позиции, соответствующей ширине. Отметим, что в нашем приложении это поведение идеально, но в других оно может быть нежелательно. После всех упрощений код ядра становится более ясным и компактным.

```

__global__ void blend_kernel( float *dst,
                             bool dstOut ) {
    // отобразить пару threadIdx/BlockIdx на позицию пикселя
    int x = threadIdx.x + blockIdx.x * blockDim.x;
    int y = threadIdx.y + blockIdx.y * blockDim.y;

```

```

int offset = x + y * blockDim.x * gridDim.x;

float t, l, c, r, b;
if (dstOut) {
    t = tex2D(texIn, x, y-1);
    l = tex2D(texIn, x-1, y);
    c = tex2D(texIn, x, y);
    r = tex2D(texIn, x+1, y);
    b = tex2D(texIn, x, y+1);
} else {
    t = tex2D(texOut, x, y-1);
    l = tex2D(texOut, x-1, y);
    c = tex2D(texOut, x, y);
    r = tex2D(texOut, x+1, y);
    b = tex2D(texOut, x, y+1);
}
dst[offset] = c + SPEED * (t + b + r + l - 4 * c);
}

```

Так как все предыдущие обращения к `tex1Dfetch()` нужно заменить на обращения к `tex2D()`, то мы вносим соответствующие изменения в ядро `copy_const_kernel()`. Как и в `blend_kernel()`, больше нет необходимости использовать смещение для адресации текстуры, это можно сделать с помощью самих координат `x` и `y`:

```

__global__ void copy_const_kernel( float *iptr ) {
    // отобразить пару threadIdx/BlockIdx на позицию пикселя
    int x = threadIdx.x + blockIdx.x * blockDim.x;
    int y = threadIdx.y + blockIdx.y * blockDim.y;
    int offset = x + y * blockDim.x * gridDim.x;

    float c = tex2D(texConstSrc, x, y);
    if (c != 0)
        iptr[offset] = c;
}

```

Последнее изменение лежит в том же русле, что и предыдущие. Точнее, в функции `main()` следует изменить привязку, сообщив исполняющей среде, что буфер следует рассматривать как двумерную, а не одномерную текстуру:

```

HANDLE_ERROR( cudaMalloc( (void**)&data.dev_inSrc,
                           imageSize ) );
HANDLE_ERROR( cudaMalloc( (void**)&data.dev_outSrc,
                           imageSize ) );
HANDLE_ERROR( cudaMalloc( (void**)&data.dev_constSrc,
                           imageSize ) );

cudaChannelFormatDesc desc = cudaCreateChannelDesc<float>();
HANDLE_ERROR( cudaBindTexture2D( NULL, texConstSrc,

```

```

        data.dev_constSrc,
        desc, DIM, DIM,
        sizeof(float) * DIM ) );
HANDLE_ERROR( cudaBindTexture2D( NULL, texIn,
        data.dev_inSrc,
        desc, DIM, DIM,
        sizeof(float) * DIM ) );
HANDLE_ERROR( cudaBindTexture2D( NULL, texOut,
        data.dev_outSrc,
        desc, DIM, DIM,
        sizeof(float) * DIM ) );

```

Как и в версиях без текстур и с одномерной текстурой, мы вначале выделяем память для входных массивов. От одномерной версии этот код отличается тем, что в случае привязки двумерных текстур исполняющая среда CUDA требует передачи `cudaChannelFormatDesc` при привязке текстур. В листинге выше имеется объявление дескриптора формата канала. В данном случае можно принять параметры по умолчанию и указать лишь, что требуется дескриптор с плавающей точкой. Затем с помощью функции `cudaBindTexture2D()` мы привязываем все три входных буфера как двумерные текстуры, передавая размеры текстуры (`DIM × DIM`) и дескриптор формата канала (`desc`). Больше ничего не изменяется.

```

int main( void ) {
    DataBlock data;
    CPUAnimBitmap bitmap( DIM, DIM, &data );
    data.bitmap = &bitmap;
    data.totalTime = 0;
    data.frames = 0;
    HANDLE_ERROR( cudaEventCreate( &data.start ) );
    HANDLE_ERROR( cudaEventCreate( &data.stop ) );

    int imageSize = bitmap.image_size();

    HANDLE_ERROR( cudaMalloc( (void**)&data.output_bitmap,
                            imageSize ) );
    // предполагаем, что размер float равен 4 байтам (т.е. rgba)
    HANDLE_ERROR( cudaMalloc( (void**)&data.dev_inSrc,
                            imageSize ) );
    HANDLE_ERROR( cudaMalloc( (void**)&data.dev_outSrc,
                            imageSize ) );
    HANDLE_ERROR( cudaMalloc( (void**)&data.dev_constSrc,
                            imageSize ) );

    cudaChannelFormatDesc desc = cudaCreateChannelDesc<float>();
    HANDLE_ERROR( cudaBindTexture2D( NULL, texConstSrc,
        data.dev_constSrc,
        desc, DIM, DIM,
        sizeof(float) * DIM ) );

```

```
HANDLE_ERROR( cudaBindTexture2D( NULL, texIn,
                                data.dev_inSrc,
                                desc, DIM, DIM,
                                sizeof(float) * DIM ) );
HANDLE_ERROR( cudaBindTexture2D( NULL, texOut,
                                data.dev_outSrc,
                                desc, DIM, DIM,
                                sizeof(float) * DIM ) );

// инициализировать константные данные
float *temp = (float*)malloc( imageSize );
for (int i=0; i<DIM*DIM; i++) {
    temp[i] = 0;
    int x = i % DIM;
    int y = i / DIM;
    if ((x>300) && (x<600) && (y>310) && (y<601))
        temp[i] = MAX_TEMP;
}
temp[DIM*100+100] = (MAX_TEMP + MIN_TEMP)/2;
temp[DIM*700+100] = MIN_TEMP;
temp[DIM*300+300] = MIN_TEMP;
temp[DIM*200+700] = MIN_TEMP;
for (int y=800; y<900; y++) {
    for (int x=400; x<500; x++) {
        temp[x+y*DIM] = MIN_TEMP;
    }
}
HANDLE_ERROR( cudaMemcpy( data.dev_constSrc, temp,
                        imageSize,
                        cudaMemcpyHostToDevice ) );
// инициализировать входные данные
for (int y=800; y<DIM; y++) {
    for (int x=0; x<200; x++) {
        temp[x+y*DIM] = MAX_TEMP;
    }
}
HANDLE_ERROR( cudaMemcpy( data.dev_inSrc, temp,
                        imageSize,
                        cudaMemcpyHostToDevice ) );
free( temp );
bitmap.anim_and_exit( (void (*)(void*,int))anim_gpu,
                      (void (*)(void*))anim_exit );
}
```

Хотя для привязки одномерных и двумерных текстур нужны разные функции, для отвязывания используется одна и та же функция `cudaUnbindTexture()`. Поэтому функция очистки остается без изменения:

```
// очистить память, выделенную на GPU
void anim_exit( DataBlock *d ) {
    cudaUnbindTexture( texIn );
    cudaUnbindTexture( texOut );
    cudaUnbindTexture( texConstSrc );
    cudaFree( d->dev_inSrc );
    cudaFree( d->dev_outSrc );
    cudaFree( d->dev_constSrc );

    HANDLE_ERROR( cudaEventDestroy( d->start ) );
    HANDLE_ERROR( cudaEventDestroy( d->stop ) );
}
```

У версии программы моделирования теплообмена с двумерными текстурами практически такие же характеристики производительности, как у версии с одномерными текстурами. Поэтому с точки зрения производительности решение о выборе одномерной или двумерной текстуры, скорее всего, несущественно. В нашем конкретном приложении применение двумерных текстур позволило немного упростить код, потому что сама предметная область двумерна. Но в общем случае так бывает не всегда, поэтому мы рекомендуем подходить к вопросу о выборе размерности текстуры с учетом особенностей конкретной задачи.

7.4. Резюме

Как и в случае константной памяти, преимущества текстурной памяти отчасти обусловлены кэшированием внутри процессора. Особенно это заметно в таких приложениях, как моделирование теплообмена, характеризующихся пространственной локальностью доступа к данным. Мы также видели, что можно использовать одномерные и двумерные текстуры с примерно одинаковыми характеристиками производительности. Как и в случае блока или сетки, размерность текстуры выбирается в основном из соображений удобства. Поскольку при работе с двумерными текстурами код становится немного чище, а выход за границы массива обрабатывается автоматически, то в задаче о теплообмене мы, пожалуй, остановились бы на двумерных текстурах. Но, как вы видели, оба варианта работают одинаково хорошо.

Текстурная память могла бы дать дополнительное повышение быстродействия, если бы мы воспользовались средствами автоматического преобразования упакованных данных в отдельные переменные или возможностью конвертации 8- и 16-разрядных целых в нормализованные числа с плавающей точкой. В задаче о теплообмене ни то, ни другое нам не понадобилось, но вы можете найти полезные применения этим механизмам!



Глава 8. Интероперабельность с графикой

Поскольку эта книга посвящена вычислениям общего назначения, то мы, как правило, игнорируем тот факт, что GPU содержат и ряд специальных компонентов. Своим успехом GPU обязаны умению решать сложные задачи рендеринга в реальном масштабе времени, позволяя прочим частям системы заниматься другими делами. Но тогда возникает естественный вопрос: можно ли использовать GPU для рендеринга и вычислений общего назначения *в одной и той же программе*? Что, если визуализируемое изображение зависит от результатов вычислений? Или если мы хотим применить алгоритмы обработки изображений к визуализированному кадру? Или подвергнуть его статистическому анализу?

К счастью, взаимодействие между вычислениями общего назначения и рендерингом не только возможно, но и легко реализуется теми средствами, о которых вы уже знаете. Приложения на языке CUDA естественно интегрируются с обеими наиболее популярными библиотеками рендеринга в реальном времени: OpenGL и DirectX. В этой главе мы поговорим о том, как это достигается.

Приведенные здесь примеры отличаются от примеров из предыдущих глав. В частности, предполагается, что вы достаточно уверенно владеете другими технологиями. Конкретно: мы включили довольно много кода, в котором используется OpenGL и GLUT, хотя подробно ни о той, ни о другой библиотеке не рассказываем. Есть немало отличных ресурсов для изучения графических API как в сети, так и в печатном виде, но эта тематика далеко выходит за рамки данной книги. Мы лишь хотели показать, как язык CUDA C и включенные в него механизмы позволяют осуществить интеграцию с графическими приложениями. Если вы не знакомы с OpenGL или DirectX, то вряд ли почерпнете что-то полезное из этой главы, так что можете сразу перейти к следующей.

8.1. О чем эта глава

Прочитав эту главу, вы:

- узнаете, что такое *интероперабельность с графикой* и зачем она может понадобиться;
- узнаете, как настроить устройство CUDA для поддержки интероперабельности с графикой;
- научитесь разделять данные между ядрами CUDA C и подсистемой рендеринга OpenGL.

8.2. Взаимодействие с графикой

Для демонстрации механизма взаимодействия между графикой и CUDA C мы напишем программу, работающую в два этапа. На первом этапе ядро на CUDA C генерирует данные изображения. На втором этапе эти данные передаются драйверу OpenGL для визуализации. При решении этой задачи мы воспользуемся многими из тех средств CUDA C, которые были рассмотрены в предыдущих главах, а также некоторыми вызовами OpenGL и GLUT.

В самом начале мы включаем заголовочные файлы GLUT и CUDA, в которых определены необходимые функции и перечисления. Кроме того, определяется размер окна, в котором приложение будет рисовать. При размере 512×512 пикселей изображение получится сравнительно небольшим.

```
#define GL_GLEXT_PROTOTYPES
#include "GL/glut.h"
#include "cuda.h"
#include "cuda_gl_interop.h"
#include "../common/book.h"
#include "../common/cpu_bitmap.h"

#define DIM 512
```

Далее мы объявляем две глобальные переменные, в которых будут храниться описатели данных, общих для OpenGL и CUDA. Как они используются, мы скоро покажем, а пока имейте в виду, что это разные описатели *одного и того же* буфера. Две переменные нужны потому, что в OpenGL и CUDA применяются разные «имена» одного буфера. Переменная `bufferObj` выступает в роли OpenGL-имени данных, а переменная `resource` – в роли CUDA C-имени.

```
GLuint bufferObj;
cudaGraphicsResource *resource;
```

Теперь перейдем к самому приложению. Прежде всего необходимо выбрать, на каком CUDA-устройстве оно будет работать. Во многих случаях эта процедура несложна, так как имеется всего один GPU с поддержкой CUDA. Однако все чаще можно встретить систему, содержащую несколько таких GPU, поэтому необходим способ, позволяющий выбрать конкретный процессор. К счастью, в исполняющую среду CUDA такое средство уже встроено.

```
int main( int argc, char **argv ) {
    cudaDeviceProp prop;
    int dev;

    memset( &prop, 0, sizeof( cudaDeviceProp ) );
    prop.major = 1;
```

```
prop.minor = 0;  
HANDLE_ERROR( cudaChooseDevice( &dev, &prop ) );
```

Вы, возможно, помните обсуждение функции `cudaChooseDevice()` в главе 3, но поскольку тогда это был второстепенный вопрос, то мы рассмотрим ее снова. Показанный выше код просит исполняющую среду выбрать любой GPU с уровнем *вычислительных возможностей* 1.0 или выше. Для этого сначала создается и обнуляется структура `cudaDeviceProp`, а затем в ней устанавливаются номера старшего и младшего уровней вычислительных возможностей (1 и 0 соответственно). Эта информация передается функции `cudaChooseDevice()`, которая просит исполняющую среду выбрать GPU, удовлетворяющий ограничениям, заданным в структуре `cudaDeviceProp`. В следующей главе мы подробнее обсудим, что такое *вычислительные возможности GPU*, а пока достаточно будет сказать, что эта характеристика определяет, какие функции GPU поддерживает. Для всех GPU с поддержкой CUDA уровень вычислительных возможностей не меньше 1.0, поэтому код, показанный выше, просто выберет произвольное устройство с поддержкой CUDA и вернет его идентификатор в переменной `dev`. Нет никакой гарантии, что это самый лучший или самый быстрый GPU, как и гарантии того, что при смене версии исполняющей среды CUDA будет выбираться один и тот же GPU.

Если результат процедуры выбора устройства столь ничтожен, то зачем вообще напрягаться – заполнять структуру `cudaDeviceProp` и вызывать функцию `cudaChooseDevice()` для получения идентификатора устройства? Да и незанимались мы раньше этими глупостями, так с чего вдруг теперь? Хорошие вопросы. А дело в том, что идентификатор устройства нужно знать, чтобы сообщить исполняющей среде CUDA, что мы собираемся использовать это устройство для CUDA и OpenGL. Для этого мы вызываем функцию `cudaGLSetGLDevice()`, передавая ей идентификатор устройства, полученный от `cudaChooseDevice()`:

```
HANDLE_ERROR( cudaGLSetGLDevice( dev ) );
```

Завершив инициализацию исполняющей среды CUDA, мы можем перейти к инициализации драйвера OpenGL, которая производится с помощью функций из библиотеки GL Utility Toolkit (GLUT). Последовательность обращений знакома любому пользователю GLUT:

```
// эти вызовы GLUT должны предшествовать всем остальным обращениям к GL  
glutInit( &argc, argv );  
glutInitDisplayMode( GLUT_DOUBLE | GLUT_RGBA );  
glutInitWindowSize( DIM, DIM );  
glutCreateWindow( "bitmap" );
```

Итак, мы подготовили исполняющую среду CUDA к взаимодействию с драйвером OpenGL, вызвав функцию `cudaGLSetGLDevice()`. Затем мы инициализирова-

ли библиотеку GLUT и создали окно с именем «bitmap», в которое будем выводить результаты. Теперь все готово к совместной работе с OpenGL!

Ключевым компонентом взаимодействия между ядром CUDA С и рендерингом OpenGL являются разделяемые буфера данных. Чтобы передать данные от OpenGL к CUDA и обратно, мы должны сначала создать буфер, который будет использоваться обоими API. Процедура начинается с создания пиксельного буфера в OpenGL и сохранения его идентификатора в глобальной переменной bufferObj типа GLuint:

```
glGenBuffers( 1, &bufferObj );
 glBindBuffer( GL_PIXEL_UNPACK_BUFFER_ARB, bufferObj );
 glBufferData( GL_PIXEL_UNPACK_BUFFER_ARB, DIM * DIM * 4,
 NULL, GL_DYNAMIC_DRAW_ARB );
```

Если вам никогда не доводилось создавать объект пиксельного буфера (PBO) в OpenGL, то скажем, что обычно эта операция выполняется в три приема. Сначала с помощью функции glGenBuffers() генерируется идентификатор буфера. Затем этот идентификатор связывается с пиксельным буфером с помощью функции glBindBuffer(). И наконец, мы просим драйвер OpenGL выделить для нас буфер, вызывая функцию glBufferData(). В данном случае мы запросили буфер для хранения $DIM \times DIM$ 32-разрядных чисел и указали константу GL_DYNAMIC_DRAW_ARB, которая говорит, что этот буфер будет модифицироваться программой. Поскольку никаких данных для предварительной загрузки в буфер у нас нет, то в качестве предпоследнего аргумента glBufferData() передается NULL.

Теперь для настройки интероперабельности с графикой осталось только сообщить исполняющей среде CUDA о том, что мы собираемся разделять буфер OpenGL, на который указывает переменная bufferObj, с CUDA. Для этого bufferObj регистрируется в CUDA как графический ресурс.

```
HANDLE_ERROR(
    cudaGraphicsGLRegisterBuffer( &resource,
                                bufferObj,
                                cudaGraphicsMapFlagsNone )
);
```

Чтобы уведомить исполняющую среду CUDA о нашем намерении использовать пиксельный буфер bufferObj как в OpenGL, так и в CUDA, мы вызываем функцию cudaGraphicsGLRegisterBuffer(). Исполняющая среда возвращает понятный CUDA описатель буфера в переменной resource. Этот описатель будет использоваться для ссылки на bufferObj в последующих обращениях к исполняющей среде CUDA.

Флаг cudaGraphicsMapFlagsNone говорит, что никакого специального поведения этого буфера мы не определяем. Если бы мы хотели сделать буфер доступным только для чтения, то могли бы задать флаг cudaGraphicsMapFlagsReadOnly.

Существует также флаг `cudaGraphicsMapFlagsWriteDiscard`, который говорит, что предыдущее содержимое буфера следует отбросить; при этом буфер, по существу, становится доступным только для записи. Эти необязательные флаги позволяют CUDA и драйверу OpenGL оптимизировать аппаратные установки для буферов с ограниченным доступом.

Вызов `glBufferData()` запрашивает у драйвера OpenGL буфер, достаточно большой для размещения $DIM \times DIM$ 32-разрядных значений. В последующих обращениях к OpenGL мы будем ссылаться на этот буфер по его идентификатору `bufferObj`, а при обращениях к исполняющей среде CUDA – по указателю `resource`. Так как мы предполагаем читать и писать в этот буфер из ядер, написанных на CUDA C, то одного идентификатора объекта недостаточно. Нужен еще фактический адрес в памяти устройства, который можно было бы передать ядру. Для этого мы просим исполняющую среду CUDA отобразить разделяемый ресурс в свое адресное пространство и дать нам указатель на отображенный ресурс.

```

uchar4* devPtr;
size_t size;
HANDLE_ERROR( cudaGraphicsMapResources( 1, &resource, NULL ) );
HANDLE_ERROR(
    cudaGraphicsResourceGetMappedPointer( (void**)&devPtr,
                                         &size,
                                         resource )
);

```

Впоследствии `devPtr` можно использовать как обычный указатель на область памяти устройства, но при этом хранящиеся в этой области данные могут выступать в роли источника пикселей для OpenGL. После всех этих предварительных ухищрений функция `main()` продолжает работу. Сначала мы запускаем ядро, передавая ему указатель на разделяемый буфер. Это ядро, код которого мы еще не приводили, генерирует данные визуализируемого изображения. Затем мы отключаем разделяемый ресурс. Это необходимо делать перед рендерингом, потому что таким образом обеспечивается синхронизация между CUDA и графическими частями программы. Точнее, мы гарантируем, что все операции, начатые CUDA перед обращением к `cudaGraphicsUnmapResources()`, завершаются еще до первого вызова графического API.

И напоследок мы регистрируем в GLUT функции обработки для клавиатуры и дисплея (`key_func` и `draw_func`) и передаем управление циклу рендеринга GLUT, вызывая `glutMainLoop()`.

```

dim3 grids(DIM/16,DIM/16);
dim3 threads(16,16);
kernel<<<grids,threads>>>( devPtr );

HANDLE_ERROR( cudaGraphicsUnmapResources( 1, &resource, NULL ) );

// настроить GLUT и запустить главный цикл

```

```

glutKeyboardFunc( key_func );
glutDisplayFunc( draw_func );
glutMainLoop();
}

```

Осталось рассмотреть только три функции, выделенные полужирным шрифтом в листинге выше: `kernel()`, `key_func()` и `draw_func()`.

Функция ядра принимает указатель на память устройства и генерирует данные изображения. Ниже показано ядро, устроенное так же, как в примере анимации волн из главы 5.

```

// основано на коде анимации волн, но в качестве типа данных
// используется uchar4 - тип, применяемый при взаимодействии
// с графикой
__global__ void kernel( uchar4 *ptr ) {
    // отобразить пару threadIdx/BlockIdx на позицию пикселя
    int x = threadIdx.x + blockDim.x * blockIdx.x;
    int y = threadIdx.y + blockDim.y * blockIdx.y;
    int offset = x + y * blockDim.x * gridDim.x;

    // вычислить значение, находящееся в этой позиции
    float fx = x/(float)DIM - 0.5f;
    float fy = y/(float)DIM - 0.5f;
    unsigned char green = 128 + 127 *
        sin( abs(fx*100) - abs(fy*100) );

    // обращаемся как к uchar4, а не unsigned char*
    ptr[offset].x = 0;
    ptr[offset].y = green;
    ptr[offset].z = 0;
    ptr[offset].w = 255;
}

```

Здесь мы встречаем многое уже знакомых приемов. Способ преобразования индексов блока и потока в координаты x , y и линейного смещения рассматривался неоднократно. Затем мы производим более-менее произвольные вычисления цвета пикселя в позиции (x, y) и сохраняем результат в памяти. Для программной генерации изображения на GPU мы снова используем CUDA C. Важно понимать, что впоследствии это изображение будет передано *непосредственно* OpenGL для рендеринга без вмешательства CPU. Напомним, что в примере из главы 5 мы тоже генерировали изображение на GPU, но затем копировали буфер в память CPU для отображения.

Но как все-таки визуализировать сгенерированный CUDA буфер с помощью OpenGL? Вспомните следующую строчку из кода инициализации в `main()`:

```
glBindBuffer( GL_PIXEL_UNPACK_BUFFER_ARB, bufferObj );
```

Здесь мы связали разделяемый буфер, назначив его источником пикселей для драйвера OpenGL, чтобы последующие вызовы `glDrawPixels()` использовали его как источник данных. И теперь вызова `glDrawPixels()` достаточно для визуализации изображения, сгенерированного ядром CUDA C. Стало быть, функции `draw_func()` остается сделать следующее:

```
static void draw_func( void ) {
    glDrawPixels( DIM, DIM, GL_RGB, GL_UNSIGNED_BYTE, 0 );
    glutSwapBuffers();
}
```

Возможно, вам встречались обращения к `glDrawPixels()`, в которых последним аргументом передается указатель на буфер. Драйвер OpenGL будет копировать данные из этого буфера, если никакой другой буфер не был назначен в качестве источника путем вызова `glBindBuffer()` с аргументом `GL_PIXEL_UNPACK_BUFFER_ARB`. Но поскольку наши данные уже находятся в памяти GPU и мы связали разделяемый буфер, назначив его источником, то последний параметр интерпретируется как смещение от начала буфера. Так как мы хотим визуализировать весь буфер, то смещение равно 0. Последний кусочек приложения не таит в себе ничего интересного; просто мы решили предоставить пользователю способ выйти из приложения. Для этого мы написали функцию обратного вызова `key_func()`, которая реагирует только на клавишу `Esc` и при ее нажатии производит очистку и выход.

```
static void key_func( unsigned char key, int x, int y ) {
    switch (key) {
        case 27:
            // очистка OpenGL и CUDA
            HANDLE_ERROR( cudaGraphicsUnregisterResource( resource ) );
            glBindBuffer( GL_PIXEL_UNPACK_BUFFER_ARB, 0 );
            glDeleteBuffers( 1, &bufferObj );
            exit(0);
    }
}
```

Эта программа рисует завораживающую картинку с чередованием черного и зеленого (оттенка NVIDIA) цветов, изображенную на рис. 8.1. Попробуйте с ее помощью загипнотизировать своих друзей (или врагов).

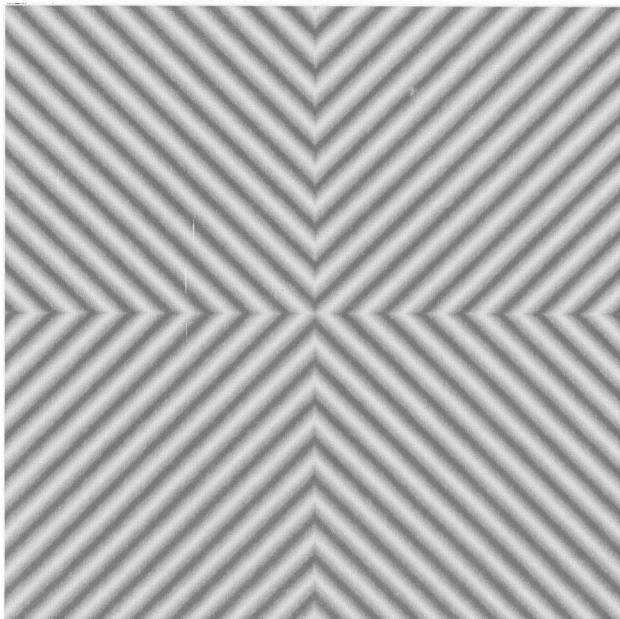


Рис. 8.1. Изображение, сгенерированное программой демонстрации взаимодействия с графикой

8.3. Анимация волн на GPU с применением интероперабельности с графикой

В разделе 8.2 «Взаимодействие с графикой» мы несколько раз упомянули программу анимации волн из главы 5. Если помните, эта программа создавала объект `CPUAnimBitmap` и передавала ему функцию, вызываемую при генерации каждого кадра.

```
int main( void ) {
    DataBlock data;
    CPUAnimBitmap bitmap( DIM, DIM, &data );
    data.bitmap = &bitmap;

    HANDLE_ERROR( cudaMalloc( (void**)&data.dev_bitmap,
                           bitmap.image_size() ) );

    bitmap.anim_and_exit( (void (*)(void*, int))generate_frame,
```

```

    (void (*)(void*))cleanup );
}

```

Мы намереваемся применить технику, рассмотренную в предыдущем разделе, к созданию структуры `GPUAnimBitmap`. Она служит тем же целям, что `CPUAnimBitmap`, но в новой версии CUDA и OpenGL будут взаимодействовать без вмешательства CPU. Когда мы все сделаем, функция `main()` окажется совсем простой:

```

int main( void ) {
    GPUAnimBitmap bitmap( DIM, DIM, NULL );

    bitmap.anim_and_exit(
        (void (*)(uchar4*,void*,int))generate_frame, NULL );
}

```

При работе со структурой `GPUAnimBitmap` будут использоваться вызовы тех же функций, что в разделе 8.2. Но теперь эти вызовы абстрагированы в самой структуре `GPUAnimBitmap`, поэтому код последующих примеров (и, возможно, ваших собственных приложений) станет чище.

8.3.1. Структура `GPUAnimBitmap`

Некоторые данные – члены `GPUAnimBitmap` уже знакомы по разделу 8.2 «Взаимодействие с графикой».

```

struct GPUAnimBitmap {
    GLuint bufferObj;
    cudaGraphicsResource *resource;
    int width, height;
    void *dataBlock;
    void (*fAnim)(uchar4*,void*,int);
    void (*animExit)(void*);
    void (*clickDrag)(void*,int,int,int,int);
    int dragStartX, dragStartY;
}

```

Мы знаем, что в OpenGL и в исполняющей среде CUDA используются разные имена для буфера в памяти GPU и что в зависимости от того, к чему мы обращаемся, нужно указывать то или иное имя. Поэтому в структуре хранятся оба имени: `bufferObj` для OpenGL и `resource` для исполняющей среды CUDA. Поскольку мы намереваемся визуализировать растровое изображение, для него должны быть заданы ширина и высота.

Чтобы клиенты `GPUAnimBitmap` могли зарегистрировать обратные вызовы, нам необходим указатель типа `void*` на произвольные пользовательские данные; мы храним его в поле `dataBlock`. Нашему классу эти данные безразличны, он просто передает указатель на них зарегистрированным функциям обратного вызова. Указатели на сами функции обратного вызова хранятся в полях `fAnim`, `animExit` и

`clickDrag`. Функция `fAnim()` вызывается при каждом обращении к `glutIdleFunc()` и отвечает за порождение данных изображения, визуализируемого в кадрах анимации. Функция `animExit()` вызывается при завершении анимации. Именно в ней пользователь должен реализовать код очистки, который нужно выполнить, когда анимация завершается. Наконец, необязательная функция `clickDrag()` реализует реакцию на события щелчка и буксировки мыши. Если клиент зарегистрирует эту функцию, то она будет вызываться после каждого события нажатия кнопки мыши, буксировки и отпускания кнопки. Позиция, в которой впервые была нажата кнопка мыши, сохраняется в полях (`dragStartX`, `dragStartY`), чтобы клиенту можно было передать начальную и конечную точки буксировки в момент отпускания кнопки мыши. С помощью этого механизма можно поразить своих друзей интерактивной анимацией.

Инициализация `GPUAnimBitmap` очень похожа на то, что мы видели в предыдущем примере. Сохранив аргументы в полях структуры, мы запрашиваем у исполняющей среды CUDA подходящее устройство:

```
GPUAnimBitmap( int w, int h, void *d ) {
    width = w;
    height = h;
    dataBlock = d;
    clickDrag = NULL;

    // первым делом находим CUDA-устройство и говорим, что
    // оно будет использоваться для взаимодействия с графикой
    cudaDeviceProp prop;
    int dev;
    memset( &prop, 0, sizeof( cudaDeviceProp ) );
    prop.major = 1;
    prop.minor = 0;
    HANDLE_ERROR( cudaChooseDevice( &dev, &prop ) );
```

Отыскав совместимое устройство, мы делаем очень важный вызов исполняющей среды CUDA, `cudaGLSetGLDevice()`, уведомляя ее о том, что будем использовать это устройство для взаимодействия с OpenGL:

```
cudaGLSetGLDevice( dev );
```

Поскольку мы пользуемся библиотекой GLUT для создания среды оконного рендеринга, то должны инициализировать GLUT. К сожалению, это неудобно для нас, поскольку функция `glutInit()` хочет получить аргументы командной строки для передачи оконной системе. Так как нам передавать нечего, то было бы разумно указать просто `NULL`. Увы, в некоторых версиях GLUT есть ошибка, из-за которой попытка передать `NULL` приводит к аварийному завершению программы. Поэтому мы обманываем GLUT, передавая ей фиктивный аргумент, и тогда все становится хорошо.

```
int c=1;
char *foo = "name";
glutInit( &c, &foo );
```

Инициализация GLUT продолжается так же, как в предыдущем примере. Мы создаем окно, в котором будем визуализировать изображение, и в качестве его заголовка указываем строку «bitmap». Если хотите, можете назвать его как-то иначе.

```
glutInitDisplayMode( GLUT_DOUBLE | GLUT_RGBA );
glutInitWindowSize( width, height );
glutCreateWindow( "bitmap" );
```

Далее мы запрашиваем у драйвера OpenGL идентификатор буфера, который сразу же связываем в режиме GL_PIXEL_UNPACK_BUFFER_ARB, чтобы будущие обращения к glDrawPixels() рисовали в нашем общем буфере:

```
glGenBuffers( 1, &bufferObj );
 glBindBuffer( GL_PIXEL_UNPACK_BUFFER_ARB, bufferObj );
```

И последнее, но отнюдь не самое маловажное действие – запрос к драйверу OpenGL о выделении области памяти GPU. Получив буфер, мы сообщаем о нем исполняющей среде CUDA, зарегистрировав bufferObj с помощью функции cudaGraphicsGLRegisterBuffer().

```
glBufferData( GL_PIXEL_UNPACK_BUFFER_ARB, width * height * 4,
               NULL, GL_DYNAMIC_DRAW_ARB );
HANDLE_ERROR(
    cudaGraphicsGLRegisterBuffer( &resource,
                                 bufferObj,
                                 cudaGraphicsMapFlagsNone ) );
}
```

Итак, структура GPUAnimBitmap настроена, и осталось только понять, как выполняется рендеринг. Основная работа производится внутри функции glutIdleFunction(). Она делает три вещи. Во-первых, отображает разделяемый буфер в адресное пространство GPU и получает указатель на него.

```
// статический метод, вызываемый из библиотеки GLUT
static void idle_func( void ) {
    static int ticks = 1;
    GPUAnimBitmap* bitmap = *(get_bitmap_ptr());
    uchar4* devPtr;
    size_t size;

    HANDLE_ERROR(
        cudaGraphicsMapResources( 1, &(bitmap->resource), NULL )
    );
```

```

HANDLE_ERROR(
    cudaGraphicsResourceGetMappedPointer( (void**)&devPtr,
                                         &size,
                                         bitmap->resource )
);

```

Во-вторых, вызывает заданную пользователем функцию fAnim(), которая предположительно запускает ядро на CUDA C для заполнения буфера по адресу devPtr данными изображения.

```
bitmap->fAnim( devPtr, bitmap->dataBlock, ticks++ );
```

И наконец, выключает отображение буфера на память GPU, что позволяет драйверу OpenGL использовать буфер для рендеринга. Рендеринг активируется обращением к функции glutPostRedisplay().

```

HANDLE_ERROR(
    cudaGraphicsUnmapResources( 1,
                                &(bitmap->resource),
                                NULL ) );
glutPostRedisplay();
}

```

Оставшаяся часть структуры GPUAnimBitmap содержит хотя и важный, но не относящийся к рассматриваемой теме инфраструктурный код. Если он вас интересует, не стесняйтесь, изучайте. Но если у вас нет на это времени или желания, то можно с успехом двигаться дальше, не отвлекаясь на остаток GPUAnimBitmap.

8.3.2. И снова об анимации волн на GPU

Теперь, имея версию CPUAnimBitmap для GPU, мы можем переделать программу анимации волн, так чтобы она работала целиком на GPU. Для начала включим файл gpu_anim.h, в котором находится определение GPUAnimBitmap. И ядро, мало чем отличающееся от рассмотренного в главе 5.

```

#include "../common/book.h"
#include "../common/gpu_anim.h"

#define DIM 1024

__global__ void kernel( uchar4 *ptr, int ticks ) {
    // отобразить пару threadIdx/BlockIdx на позицию пикселя
    int x = threadIdx.x + blockIdx.x * blockDim.x;
    int y = threadIdx.y + blockIdx.y * blockDim.y;
    int offset = x + y * blockDim.x * gridDim.x;

    // вычислить значение в этой позиции

```

```

float fx = x - DIM/2;
float fy = y - DIM/2;
float d = sqrtf( fx * fx + fy * fy );
unsigned char grey = (unsigned char)(128.0f + 127.0f *
    cos(d/10.0f -
        ticks/7.0f) /
    (d/10.0f + 1.0f));
ptr[offset].x = grey;
ptr[offset].y = grey;
ptr[offset].z = grey;
ptr[offset].w = 255;
}

```

Единственное изменение выделено полужирным шрифтом. Оно необходимо, потому что для взаимодействия с OpenGL разделяемые поверхности должны быть «дружественными к графике». Поскольку при рендеринге в реальном времени обычно используются массивы элементов, состоящих из четырех компонентов (красный/зеленый/синий/альфа-канал), то буфер больше не является массивом элементов типа `unsigned char`, как раньше. В нем должны храниться элементы типа `uchar4`. На самом деле мы и в главе 5 работали с буфером как с массивом четырехкомпонентных элементов, потому что доступ к элементу осуществлялся с помощью выражения `ptr[offset*4+k]`, где `k` – номер компоненты от 0 до 3. Но теперь четырехкомпонентная природа данных сделана явной за счет перехода к типу `uchar4`.

Поскольку у нас уже есть написанная на CUDA C функция `kernel()`, которая генерирует данные изображения, то остается только написать функцию анимации, которая будет вызываться из функции `idle_func()`, определенной в структуре `GPUAnimBitmap`. В нашем приложении на долю этой функции остается запуск ядра:

```

void generate_frame( uchar4 *pixels, void*, int ticks ) {
    dim3 grids(DIM/16,DIM/16);
    dim3 threads(16,16);
    kernel<<< grids, threads >>>( pixels, ticks );
}

```

Вот, собственно, и все, ведь трудная работа делается в структуре `GPUAnimBitmap`. Чтобы все закрутилось, мы должны только создать экземпляр `GPUAnimBitmap` и зарегистрировать функцию анимации `generate_frame()` в качестве функции обратного вызова.

```

int main( void ) {
    GPUAnimBitmap bitmap( DIM, DIM, NULL );
    bitmap.anim_and_exit(
        (void (*)(uchar4*,void*,int))generate_frame, NULL );
}

```

8.4. Моделирование теплообмена с использованием интероперабельности с графикой

А для чего мы все это делали? Если заглянуть в код структуры CPUAnimBitmap, которая использовалась в предыдущих примерах анимации, то обнаружится, что она работает почти так же, как код рендеринга, приведенный в разделе 8.2.

Почти.

Основное отличие CPUAnimBitmap от предыдущего примера кроется в вызове функции glDrawPixels().

```
glDrawPixels( bitmap->x,
               bitmap->y,
               GL_RGBA,
               GL_UNSIGNED_BYTE,
               bitmap->pixels );
```

В первом примере из этой главы мы упомянули о том, что вам, возможно, встречались обращения к glDrawPixels() с указателем на буфер в качестве последнего аргумента. Ну а если не встречались, то встретятся теперь. Этот вызов производится из функции Draw() в структуре CPUAnimBitmap и запускает копирование буфера bitmap->pixels из памяти CPU в память GPU. Для этого CPU в каждом кадре должен приостановить текущую работу и начать копирование на GPU, что требует синхронизации между CPU и GPU и вносит дополнительную задержку, необходимую для инициализации и завершения передачи по шине PCI Express. Поскольку функция glDrawPixels() ожидает получить в последнем аргументе указатель на память CPU, то получается, что после генерации одного кадра изображения в ядре программа из главы 5 должна была скопировать этот кадр из памяти GPU в память CPU с помощью функции cudaMemcpy().

```
void generate_frame( DataBlock *d, int ticks ) {
    dim3 grids(DIM/16,DIM/16);
    dim3 threads(16,16);
    kernel<<<grids,threads>>>( d->dev_bitmap, ticks );

    HANDLE_ERROR( cudaMemcpy( d->bitmap->get_ptr(),
                            d->dev_bitmap,
                            d->bitmap->image_size(),
                            cudaMemcpyDeviceToHost ) );
}
```

Все вместе это означает, что первоначальная версия программы анимации волн на GPU была, мягко говоря, дурацкой. Мы использовали CUDA C для генерации данных для каждого кадра, но по завершении вычислений копировали заполненный буфер на CPU, который затем копировал его *обратно* на GPU для

отображения. Следовательно, от максимальной производительности нас отделяли совершенно лишние операции передачи данных между CPU и устройством. Теперь давайте вернемся к счетной задаче анимации и посмотрим, удастся ли повысить производительность за счет интероперабельности с графикой.

Вспомните, что в программе моделирования теплообмена из предыдущей главы тоже использовалась структура CPUAnimBitmap для отображения результатов моделирования. Теперь мы модифицируем эту программу, применив новую структуру GPUAnimBitmap, и поглядим, как это отразится на производительности. Как и в примере анимации волн, GPUAnimBitmap почти без усилий можно подставить вместо CPUAnimBitmap, нужно лишь поменять тип unsigned char на uchar4. Соответственно, следует изменить сигнатуру функции анимации.

```
void anim_gpu( uchar4* outputBitmap, DataBlock *d, int ticks ) {
    HANDLE_ERROR( cudaEventRecord( d->start, 0 ) );
    dim3 blocks(DIM/16, DIM/16);
    dim3 threads(16, 16);

    // так как текстура глобальная и привязанная, то мы используем
    // флаг, который указывает, какой буфер является на данной итерации
    // входным, а какой - выходным
    volatile bool dstOut = true;
    for (int i=0; i<90; i++) {
        float *in, *out;
        if (dstOut) {
            in = d->dev_inSrc;
            out = d->dev_outSrc;
        } else {
            out = d->dev_inSrc;
            in = d->dev_outSrc;
        }
        copy_const_kernel<<<blocks, threads>>>( in );
        blend_kernel<<<blocks, threads>>>( out, dstOut );
        dstOut = !dstOut;
    }
    float_to_color<<<blocks, threads>>>( outputBitmap,
                                              d->dev_inSrc );
}

HANDLE_ERROR( cudaEventRecord( d->stop, 0 ) );
HANDLE_ERROR( cudaEventSynchronize( d->stop ) );
float elapsedTime;
HANDLE_ERROR( cudaEventElapsedTime( &elapsedTime,
                                    d->start, d->stop ) );
d->totalTime += elapsedTime;
++d->frames;
printf( "Среднее время на кадр: %.3f ms\n",
        d->totalTime/d->frames );
}
```

Поскольку ядро `float_to_color()` – единственная функция, в которой используется `outputBitmap`, только она и нуждается в модификации в результате перехода на тип `uchar4`. В предыдущей главе мы считали ее обычной вспомогательной функцией, и таковой она и останется. Однако мы решили перегрузить ее, и в файле `book.h` приведены два варианта – для `unsigned char` и для `uchar4`. Как легко заметить, различия между этими вариантами такие же, как между функцией `kernel()` в версиях анимации ряби для CPU и для GPU. Для краткости большая часть ядер `float_to_color()` опущена, но если вам интересны детали, загляните в файл `book.h`.

```
__global__ void float_to_color( unsigned char *optr,
                               const float *outSrc ) {
    // преобразовать число с плавающей точкой в
    // четырехкомпонентный цвет
    optr[offset*4 + 0] = value( m1, m2, h+120 );
    optr[offset*4 + 1] = value( m1, m2, h );
    optr[offset*4 + 2] = value( m1, m2, h -120 );
    optr[offset*4 + 3] = 255;
}

__global__ void float_to_color( uchar4 *optr,
                               const float *outSrc ) {
    // преобразовать число с плавающей точкой в
    // четырехкомпонентный цвет
    optr[offset].x = value( m1, m2, h+120 );
    optr[offset].y = value( m1, m2, h );
    optr[offset].z = value( m1, m2, h -120 );
    optr[offset].w = 255;
}
```

Помимо этих изменений, единственное существенное различие – замена структуры `CPUAnimBitmap` структурой `GPUAnimBitmap` при выполнении анимации.

```
int main( void ) {
    DataBlock data;
    GPUAnimBitmap bitmap( DIM, DIM, &data );
    data.totalTime = 0;
    data.frames = 0;
    HANDLE_ERROR( cudaEventCreate( &data.start ) );
    HANDLE_ERROR( cudaEventCreate( &data.stop ) );

    int imageSize = bitmap.image_size();

    // предполагаем, что размер float равен 4 байтам (то есть rgba)
    HANDLE_ERROR( cudaMalloc( (void**)&data.dev_inSrc,
                           imageSize ) );
    HANDLE_ERROR( cudaMalloc( (void**)&data.dev_outSrc,
                           imageSize ) );
    HANDLE_ERROR( cudaMalloc( (void**)&data.dev_constSrc,
```


Хотя взглянуть на оставшуюся часть улучшенной программы моделирования теплообмена, быть может, и поучительно, но она не настолько отличается от версии из предыдущей главы, чтобы описывать ее и дальше. Важно ответить на вопрос: как мы полностью перенесли работу программы на GPU. Отказ от копирования каждого кадра в память CPU для отображения, по идеи, должен бы заметно улучшить ситуацию.

Но как это выражается количественно? Раньше на обработку одного кадра уходило примерно 25,3 мс на тестовом компьютере с картой GeForce GTX 285. После того как мы стали использовать интероперабельность с графикой, это время упало на 15% до 21,6 мс на кадр. Стало быть, цикл рендеринга стал на 15% быстрее, и больше не требуется отвлекать CPU каждый раз, как мы хотим отобразить кадр. Не так плохо для одного дня работы!

8.5. Интероперабельность с DirectX

В примерах выше речь шла только о взаимодействии с системой рендеринга OpenGL, но взаимодействие с DirectX мало чем отличается. Мы по-прежнему используем тип `cudaGraphicsResource` для ссылки на буферы, разделяемые DirectX и CUDA, и должны вызывать функции `cudaGraphicsMapResources()` и `cudaGraphicsResourceGetMappedPointer()` для получения понятных CUDA указателей на разделяемые ресурсы.

Как правило, вызовы OpenGL с легкостью транслируются в вызовы DirectX. Например, вместо `cudaGLSetGLDevice()` мы вызываем `cudaD3D9SetDirect3DDevice()`, когда хотим настроить устройство CUDA для взаимодействия с Direct3D 9.0. Аналогично `cudaD3D10SetDirect3DDevice()` настраивает устройство для взаимодействия с Direct3D 10, а `cudaD3D11SetDirect3DDevice()` – для взаимодействия с Direct3D 11.

Детали интероперабельности с DirectX вряд ли вызовут удивление у того, кто проработал примеры, относящиеся к OpenGL. Но если вы хотите попробовать, то предлагаем перенести примеры из этой главы на DirectX. Для начала рекомендуем познакомиться с API, обратившись к руководству *NVIDIA CUDA Programming Guide*, и изучить примеры взаимодействия с DirectX, имеющиеся в комплекте GPU Computing SDK.

8.6. Резюме

Хотя большая часть этой книги посвящена использованию GPU для параллельных вычислений общего назначения, нельзя забывать и о том, что GPU прекрасно справляется с ролью механизма рендеринга. Многие приложения не могут обойтись без компьютерной графики. В области рендеринга GPU – большой мастер, и единственное, что мешало нам использовать эти его возможности, – непонимание того, как заставить исполняющую среду CUDA и графический драйвер работать совместно. Теперь мы знаем, как это делается, и можем обойтись без

посредничества CPU в визуализации результатов вычислений. Попутно мы ускорили цикл рендеринга и предоставили CPU возможность заниматься другими вычислениями. Но даже если никаких других вычислений в нашей программе нет, система в целом будет быстрее реагировать на события и уделять больше времени другим приложениям.

Мы не рассмотрели много других способов организовать интероперабельность. Мы сосредоточились на использовании ядра для записи в пиксельный буфер, отображаемый затем в окне. Но данные изображения можно использовать и как текстуру, применимую к любой поверхности на сцене. Помимо модификации объектов пиксельных буферов, CUDA и графическая подсистема могут разделять объекты вершинных буферов. Это, в частности, позволяет писать на CUDA C ядра, которые обнаруживают столкновения между объектами или вычисляют карты смещения вершин, применяемые для рендеринга объектов или поверхностей, взаимодействующих с пользователем или со своим окружением. Если вас интересует компьютерная графика, то API интероперабельности CUDA C с графикой откроет вам целый мир новых возможностей!



Глава 9. Атомарные операции

В первой половине книги мы видели много случаев, когда задача, с трудом решаемая в однопоточном приложении, становится совсем простой при реализации на CUDA C. Так, благодаря невидимому присутствию исполняющей среды CUDA нам больше нет нужды организовывать циклы `for` для попиксельного обновления кадров анимации или результатов моделирования теплообмена. Стоит вызвать из программы на стороне CPU функцию с квалификатором `_global_`, как создаются тысячи параллельно исполняемых блоков и нитей, идентифицируемых своими индексами.

С другой стороны, бывает и так, что задача, которая в однопоточном приложении тривиальна, вызывает серьезные сложности при попытке реализации в массивно-параллельной архитектуре. В этой главе мы рассмотрим некоторые ситуации, когда необходимо использовать специальные примитивы для безопасного решения задачи, которая в традиционном однопоточном приложении не представляет вообще никакой проблемы.

9.1. О чём эта глава

Прочитав эту главу, вы:

- ❑ узнаете о *вычислительных возможностях* различных NVIDIA GPU;
- ❑ узнаете о том, что такое атомарные операции и зачем они могут понадобиться;
- ❑ узнаете о том, как выполнять арифметические действия с помощью атомарных операций в ядрах, написанных на CUDA C.

9.2. Вычислительные возможности

Во всех ранее рассмотренных темах использовались только возможности, которыми обладает любой GPU с поддержкой CUDA. Например, любой GPU, построенный на базе архитектуры CUDA, может запускать ядра, обращаться к глобальной памяти и читать из константной или текстурной памяти. Но аналогично тому, как разные модели CPU могут обладать разными наборами команд (к примеру, MMX, SSE или SSE2), так и графические процессоры с поддержкой CUDA неодинаковы по своим возможностям. NVIDIA называет эту характеристику GPU *уровнем вычислительных возможностей* (*compute capability*).

9.2.1. Вычислительные возможности NVIDIA GPU

На момент подготовки книги к печати NVIDIA GPU поддерживали уровни вычислительных возможностей 1.0, 1.1, 1.2, 1.3, 2.0. Более высокий уровень включает возможности всех предыдущих, то есть образуется иерархия в виде «луковицы» или «матрешки» (уж как вам больше нравится). Так, GPU с уровнем вычислительных возможностей 1.2 поддерживает все возможности уровней 1.0 и 1.1. В руководстве *NVIDIA CUDA Programming Guide* приведен актуальный перечень всех GPU с поддержкой CUDA с указанием их вычислительных возможностей. В табл. 9.1 перечислены NVIDIA GPU на момент работы над этой книгой.

Таблица 9.1. GPU с поддержкой CUDA и их вычислительные возможности

GPU	Уровень вычислительных возможностей
GeForce GTX 480, GTX 470 2.0	2.0
GeForce GTX 295	1.3
GeForce GTX 285, GTX 280	1.3
GeForce GTX 260	1.3
GeForce 9800 GX2	1.1
GeForce GTS 250, GTS 150, 9800 GTX, 9800 GTX+, 8800 GTS 512	1.1
GeForce 8800 Ultra, 8800 GTX	1.0
GeForce 9800 GT, 8800 GT, GTX 280M, 9800M GTX	1.1
GeForce GT 130, 9600 GSO, 8800 GS, 8800M GTX, GTX 260M, 9800M GT	1.1
GeForce 8800 GTS	1.0
GeForce 9600 GT, 8800M GTS, 9800M GTS	1.1
GeForce 9700M GT	1.1
GeForce GT 120, 9500 GT, 8600 GTS, 8600 GT, 9700M GT, 9650M GS, 9600M GT, 9600M GS, 9500M GS, 8700M GT, 8600M GT, 8600M GS	1.1
GeForce G100, 8500 GT, 8400 GS, 8400M GT, 9500M G, 9300M G, 8400M GS, 9400 mGPU, 9300 mGPU, 8300 mGPU, 8200 mGPU, 8100 mGPU	1.1
GeForce 9300M GS, 9200M GS, 9100M G, 8400M G	1.1
Tesla S2070, S2050, C2070, C2050	2.0
Tesla S1070, C1060	1.3
Tesla S870 , D870, C870	1.0
Quadro Plex 2200 D2	1.3
Quadro Plex 2100 D4	1.1
Quadro Plex 2100 Model S4	1.0
Quadro Plex 1000 Model IV	1.0
Quadro FX 5800	1.3
Quadro FX 4800	1.3
Quadro FX 4700 X2	1.1
Quadro FX 3700M	1.1
Quadro FX 5600	1.0

Таблица 9.1. GPU с поддержкой CUDA и их вычислительные возможности (окончание)

GPU	Уровень вычислительных возможностей
Quadro FX 3700	1.1
Quadro FX 3600M	1.1
Quadro FX 4600	1.0
Quadro FX 2700M	1.1
Quadro FX 1700, FX 570, NVS 320M, FX 1700M, FX 1600M, FX 770M, FX 570M	1.1
Quadro FX 370, NVS 290, NVS 140M, NVS 135M, FX 360M	1.1
Quadro FX 370M, NVS 130M	1.1

NVIDIA постоянно выпускает новые графические процессоры, так что эта таблица несомненно устареет к моменту выхода книги из печати. Но у компании NVIDIA есть веб-сайт, а в нем – раздел CUDA Zone. Среди прочего в этом разделе публикуется актуальный перечень поддерживаемых устройств с поддержкой CUDA. Мы рекомендуем заглянуть в этот список, перед тем как решиться на какой-нибудь отчаянный поступок, не найдя своего новейшего GPU в табл. 9.1. А можете просто запустить программу из главы 3, которая распечатает вычислительные возможности всех обнаруженных в системе CUDA-устройств.

Поскольку эта глава посвящена атомарным операциям, особый интерес для нас представляет возможность выполнять атомарные операции с памятью. Прежде чем рассказывать о том, что такое атомарные операции и какое нам до них дело, отметим, что атомарные операции с глобальной памятью поддерживаются только GPU с уровнем вычислительных возможностей не ниже 1.1. А для атомарных операций с *разделяемой* памятью нужен GPU с уровнем вычислительных возможностей не ниже 1.2. Поскольку более высокий уровень включает все возможности более низкого уровня, то GPU версии 1.2 поддерживают атомарные операции как с глобальной, так и с *разделяемой* памятью. Как и GPU версии 1.3.

Если уровень вычислительных возможностей вашего GPU равен 1.0, то есть атомарные операции с глобальной памятью не поддерживаются, значит, самое время провести модернизацию! Если вы не хотите раскошелиться на новый процессор с поддержкой атомарных операций, то можете продолжать читать о том, что они собой представляют и когда необходимы. Но если невозможность прогнать примеры надрывает вам сердце, то можете спокойно перейти к следующей главе.

9.2.2. Компиляция программы для GPU с заданным минимальным уровнем вычислительных возможностей

Предположим, вы написали код, для работы которого требуется некий минимальный уровень вычислительных возможностей. К примеру, прочитали эту главу и принялись за разработку приложения, опирающегося на поддержку атомарных

операций с глобальной памятью. Поскольку вы внимательно проштудировали текст книги, то знаете, что для атомарных операций с глобальной памятью необходимы вычислительные возможности уровня не ниже 1.1. При компиляции программы необходимо сообщить компилятору, что ядро не может работать на оборудовании с уровнем вычислительных возможностей ниже 1.1. А уведомив об этом компилятор, мы уже вправе прибегнуть и к другим оптимизациям, возможным только на GPU, поддерживающем уровень 1.1 или выше. Чтобы уведомить компилятор, достаточно при запуске nvcc добавить в командную строку всего один флаг:

```
nvcc -arch=sm_11
```

Аналогично при компиляции ядра, требующего поддержки атомарных операций с разделяемой памятью, нужно уведомить компилятор о том, что минимально необходимый уровень равен 1.2:

```
nvcc -arch=sm_12
```

9.3. Обзор атомарных операций

При разработке традиционных однопоточных приложений у программиста обычно не возникает необходимости в атомарных операциях. Если эти слова относятся к вам, не волнуйтесь – мы объясним, что такое атомарные операции и зачем они нужны в многопоточных приложениях. Чтобы стало понятнее, в чем смысл атомарной операции, рассмотрим один из первых операторов, с которым вы встретились при изучении языка C или C++ – оператор инкремента:

```
x++;
```

В стандартном C это одно выражение, после вычисления которого значение x станет на единицу больше, чем было. Но какие операции при этом происходят? Чтобы прибавить единицу к x , мы должны сначала узнать, каково значение x сейчас. Прочитав значение x , мы можем его изменить. А затем должны записать новое значение назад в x .

Таким образом, вычисление инкремента производится в три приема:

1. Прочитать значение x .
2. Прибавить 1 к значению, прочитанному на шаге 1.
3. Записать результат в x .

Иногда эту последовательность операций называют *чтение–модификация–запись*, поскольку на шаге 2 может производиться любая операция, изменяющая прочитанное значение x .

А теперь рассмотрим ситуацию, когда инкрементировать x пытаются две нити: A и B. Чтобы инкрементировать x , обе нити должны выполнить все три описанные выше операции. Пусть вначале значение x было равно 7. В идеале хотелось бы, чтобы нити A и B выполняли шаги так, как показано в табл. 9.2.

Таблица 9.2. Две нити, увеличивающие значение x

Шаг	Пример
1. Нить А читает значение x	А читает 7 из x
2. Нить А прибавляет 1 к прочитанному значению	А вычисляет 8
3. Нить А записывает результат обратно в x	$x <- 8$
4. Нить В читает значение x	В читает 8 из x
5. Нить В прибавляет 1 к прочитанному значению	В вычисляет 9
6. Нить В записывает результат обратно в x	$x <- 9$

Поскольку в начале значение x было равно 7, а затем его увеличивали две нити, то мы ожидаем, что по завершении в x окажется значение 9. Если вычисления производились в той последовательности, что показана выше, то такой результат и получится. К сожалению, есть много других вариантов последовательности, при которых получается неправильное значение. Рассмотрим, например, порядок выполнения операций, показанный в табл. 9.3, когда нити А и В чередуются.

Таблица 9.3. Две нити, увеличивающие значение x, с чередованием операций

Шаг	Пример
Нить А читает значение x	А читает 7 из x
Нить В читает значение x	В читает 7 из x
Нить А прибавляет 1 к прочитанному значению	А вычисляет 8
Нить В прибавляет 1 к прочитанному значению	В вычисляет 8
Нить А записывает результат обратно в x	$x <- 8$
Нить В записывает результат обратно в x	$x <- 8$

Таким образом, при неблагоприятном планировании нитей результат вычислений оказывается неверным. Есть много способов упорядочить эти шесть операций; иногда получается правильный результат, иногда – нет. При переходе от однопоточной к многопоточной версии приложения мы сталкиваемся с непредсказуемостью результата, когда несколько нитей читают или изменяют разделяемые значения.

В рассмотренном примере нам необходимо как-то гарантировать, что последовательность *чтение–модификация–запись* не будет прерываться другой нитью. Точнее, никакая другая нить не должна ни читать, ни изменять значение x, пока мы не завершим операцию. Поскольку выполнение операции не может быть разбито на части в результате вмешательства другой нити, мы называем такую операцию *атомарной*. Язык CUDA C поддерживает несколько атомарных операций с памятью, которые выполняются безопасно даже при наличии тысяч нитей, конкурирующих за получение доступа. Теперь приведем пример задачи, в которой атомарные операции необходимы для получения правильного результата.

9.4. Вычисление гистограмм

Часто бывает необходимо вычислить *гистограмму* для некоторого набора данных. Если раньше вам не приходилось иметь дело с гистограммами, ничего страшного. Смысл в следующем: если имеется набор данных, содержащий множество элементов, то гистограмма описывает частоту появления в нем каждого элемента. Например, гистограмма распределения букв во фразе *Programming with CUDA C* выглядит, как показано на рис. 9.1.

Задача о вычислении гистограммы встречается в информатике на удивление часто – при обработке изображений, сжатии данных, в алгоритмах машинного зрения, машинного обучения, кодирования звука и во многих других случаях. В примерах ниже мы займемся вычислением гистограмм.

2	2	1	2	1	2	2	1	1	1	2	1	1	1
A	C	D	G	H	I	M	N	O	P	R	T	U	W

Рис. 9.1. Частотная гистограмма распределения букв в строке *Programming with CUDA C*

9.4.1. Вычисление гистограммы на CPU

Поскольку не все читатели знакомы с вычислением гистограмм, мы начнем с программы, работающей на CPU. И заодно покажем, что в однопоточном приложении вычисление гистограммы не вызывает особых сложностей. На вход приложения поступает большой поток данных. В реальной программе это могут быть, например, цвета пикселей или цифровые отсчеты звукового сигнала, а мы ограничимся потоком случайно сгенерированных байтов. Этот поток будет генерировать служебная функция `big_random_block()`. В нашем примере она порождает 100 Мб случайных данных.

```
#include "../common/book.h"

#define SIZE (100*1024*1024)

int main( void ) {
    unsigned char *buffer = (unsigned char*)big_random_block( SIZE );
```

Поскольку 8-разрядный байт может принимать 256 различных значений (от 0x00 до 0xFF), то в нашей гистограмме должно быть 256 *столбиков* (bins), в которых мы будем подсчитывать, сколько раз встретилось каждое значение. Создадим массив из 256 элементов и инициализируем его нулями.

```
unsigned int histo[256];
for (int i=0; i<256; i++)
    histo[i] = 0;
```

После того как гистограмма создана и все столбики обнулены, необходимо сформировать таблицу частот появления каждого значения в буфере данных `buffer[]`. Идея проста: встретив некоторое значение `z` в массиве `buffer[]`, мы должны увеличить на 1 значение в столбике гистограммы с индексом `z`. Таким образом мы подсчитываем количество вхождений каждого значения.

Если текущее значение – это `buffer[i]`, то увеличивать нужно счетчик, хранящийся в столбике с номером `buffer[i]`. Поскольку этому столбику соответствует элемент массива `histo[buffer[i]]`, то увеличить счетчик на единицу можно в одной строке кода:

```
histo[buffer[i]]++;
```

И это делается для каждого элемента массива `buffer[]` в простом цикле `for`:

```
for (int i=0; i<SIZE; i++)  
    histo[buffer[i]]++;
```

По завершении цикла гистограмма входных данных готова. В настоящей программе она была бы подана на вход следующего этапа вычисления. Ну а нам больше ничего не нужно, поэтому мы завершаем приложение, предварительно проверив, что суммирование всех столбиков дает ожидаемый результат:

```
long histoCount = 0;  
for (int i=0; i<256; i++) {  
    histoCount += histo[i];  
}  
printf( "Сумма гистограммы: %ld\n", histoCount );
```

Если вы были внимательны, то поняли, что эта сумма всегда будет одинакова, каким бы случайным ни был входной массив. Значение в каждом столбике показывает, сколько раз встретился соответствующий элемент данных, поэтому сумма по всем столбикам должна быть равна общему числу просмотренных элементов, то есть в нашем случае `SIZE`.

Нет нужды говорить (но мы все же скажем), что в конце нужно почистить за собой и вернуть управление.

```
free( buffer );  
return 0;  
}
```

На нашем лабораторном компьютере с процессором Core 2 Duo на построение гистограммы распределения значений в массиве размером 100 Мб было затрачено 0,416 с. С этой величиной мы и будем сравнивать производительность версии для GPU.

9.4.2. Вычисление гистограммы на GPU

Мы хотели бы адаптировать программу вычисления гистограммы для работы на GPU. Если входной массив достаточно велик, то мы сэкономим много времени, если поручим разным нитям обрабатывать разные части буфера. Само по себе это нетрудно, ведь нечто подобное мы уже проделывали неоднократно. Проблема же состоит в том, что несколько нитей могут инкрементировать один и тот же столбик результирующей гистограммы одновременно. И чтобы избежать ситуации, описанной в разделе 9.3 «Обзор атомарных операций», необходимо использовать атомарные операции инкремента.

Функция `main()` очень похожа на версию для CPU, хотя необходима стандартная обвязка, чтобы скопировать входные данные в память GPU и затем получить от GPU результаты. Начало функции выглядит точно так же, как для CPU:

```
int main( void ) {
    unsigned char *buffer = (unsigned char*)big_random_block( SIZE );
```

Нас интересует хронометраж программы, поэтому инициализируем события так же, как делали это раньше.

```
cudaEvent_t start, stop;
HANDLE_ERROR( cudaEventCreate( &start ) );
HANDLE_ERROR( cudaEventCreate( &stop ) );
HANDLE_ERROR( cudaEventRecord( start, 0 ) );
```

Покончив с инициализацией входных данных и событий, перейдем к работе с GPU. Необходимо выделить место для входных данных и результирующей гистограммы. Выделив память для входного буфера, скопируем в нее массив, сгенерированный функцией `big_random_block()`. А выделив место для гистограммы, обнулим ее, как и в версии для CPU.

```
// выделить память GPU для хранения данных
unsigned char *dev_buffer;
unsigned int *dev_histo;
HANDLE_ERROR( cudaMalloc( (void**)&dev_buffer, SIZE ) );
HANDLE_ERROR( cudaMemcpy( dev_buffer, buffer, SIZE,
                        cudaMemcpyHostToDevice ) );

HANDLE_ERROR( cudaMalloc( (void**)&dev_histo,
                        256 * sizeof( int ) ) );
HANDLE_ERROR( cudaMemset( dev_histo, 0,
                        256 * sizeof( int ) ) );
```

Обратите внимание на не встречающуюся ранее функцию `cudaMemset()`. Она похожа на стандартную функцию `memset()` и ведет себя почти так же. Разница заключается в том, что `cudaMemset()` возвращает код ошибки, а `memset()` – нет. Этот

код извещает вызывающую программу о том, была ли какая-нибудь ошибка при установке значений в памяти GPU. Ну и кроме того, `cudaMemset()` работает с памятью GPU, а `memset()` – с памятью CPU.

Инициализировав входные и выходные буферы, мы готовы приступить к вычислению гистограммы. Чуть ниже мы покажем, как подготовить и запустить соответствующее ядро. А пока предположим, что гистограмма уже вычислена на GPU. Теперь мы должны скопировать ее в память CPU, поэтому выделим память для массива из 256 элементов и скопируем данные из памяти устройства в память CPU.

```
unsigned int histo[256];
HANDLE_ERROR( cudaMemcpy( histo, dev_histo,
                         256 * sizeof( int ),
                         cudaMemcpyDeviceToHost ) );
```

В этот момент вычисление гистограммы завершено, поэтому мы можем остановить часы и вывести истекшее время. Код хронометража ничем не отличается от того, что мы видели в предыдущих главах.

```
// получить время завершения и вывести результаты хронометража
HANDLE_ERROR( cudaEventRecord( stop, 0 ) );
HANDLE_ERROR( cudaEventSynchronize( stop ) );
float elapsedTime;
HANDLE_ERROR( cudaEventElapsedTime( &elapsedTime,
                                    start, stop ) );
printf( "Время генерации: %3.1f ms\n", elapsedTime );
```

Теперь гистограмму можно было бы передать на следующий этап алгоритма, но, поскольку мы не собираемся ее как-то использовать, осталось лишь проверить, что результаты вычисления на GPU и на CPU совпадают. Сначала проверяем, что сумма гистограммы такая, как мы ожидаем. Тут никаких отличий от версии для CPU нет.

```
long histoCount = 0;
for (int i=0; i<256; i++) {
    histoCount += histo[i];
}
printf( "Сумма гистограммы: %ld\n", histoCount );
```

Но чтобы не оставалось никаких сомнений, мы вычислим ту же самую гистограмму на CPU. Проще всего было бы выделить память еще для одного массива, вычислить гистограмму по исходным данным, как было сделано в разделе 9.4.1, и убедиться, что в каждом столбике обеих гистограмм находятся одинаковые значения. Однако, вместо того чтобы выделять новый массив, мы начнем с гистограммы, вычисленной на GPU, и произведем «обратные вычисления». Это означает, что мы не будем инициализировать гистограмму нулями и увеличивать столбики,

а возьмем имеющуюся и будем *уменьшать* значение в столбике, соответствующем встреченному во входных данных элементу. Таким образом, можно утверждать, что CPU вычисляет ту же самую гистограмму, что GPU, в том и только в том случае, когда в конце значения во всех столбиках окажутся равными нулю. В некотором смысле мы вычисляем разность между двумя гистограммами. Код очень похож на тот, что мы писали для CPU, только вместо оператора инкремента применяется оператор декремента.

```
// проверить, что получились те же значения, что в версии для CPU
for (int i=0; i<SIZE; i++)
    histo[buffer[i]]--;
for (int i=0; i<256; i++) {
    if (histo[i] != 0)
        printf( "Ошибка в столбике %d!\n", i );
}
```

Как обычно, напоследок мы уничтожаем события CUDA и освобождаем память GPU и CPU.

```
HANDLE_ERROR( cudaEventDestroy( start ) );
HANDLE_ERROR( cudaEventDestroy( stop ) );
cudaFree( dev_histo );
cudaFree( dev_buffer );
free( buffer );
return 0;
}
```

Мы обещали рассмотреть само ядро после того, как обсудим окружение. Запуск ядра немного осложняется из-за необходимости позаботиться о производительности. Поскольку гистограмма содержит 256 столбиков, то будет и удобно, и оптимально запустить 256 нитей в одном блоке. Но вот что касается количества блоков, тут вариантов много. В нашем случае объем входных данных составляет 100 Мб, или 104 857 600 байтов. Можно было бы запустить один блок и поручить каждой нити анализ 409 600 элементов. Или запустить 409 600 блоков, тогда каждая нить будет анализировать ровно один элемент.

Как вы, наверное, догадываетесь, оптимальное решение находится между двумя крайностями. Эксперименты показали, что максимум производительности достигается, когда количество запущенных блоков вдвое больше числа мультипроцессоров в GPU. Например, на карте GeForce GTX 280 есть 30 мультипроцессоров, поэтому время вычисления гистограммы будет наименьшим при запуске 60 параллельных блоков.

В главе 3 мы обсуждали, как опрашивать различные свойства оборудования, на котором работает программа. Если мы собираемся динамически задавать количество запускаемых блоков, то надо будет поинтересоваться одним из свойств устройства. Ниже показан соответствующий код. Хотя вы еще не видели реализацию ядра, понять, что делается, не составит труда.

```

cudaDeviceProp prop;
HANDLE_ERROR( cudaGetDeviceProperties( &prop, 0 ) );
int blocks = prop.multiProcessorCount;
histo_kernel<<<blocks*2,256>>>( dev_buffer, SIZE, dev_histo );

```

Поскольку рассмотрение функции main() оказалось несколько фрагментарным, приведем ее код от начала до конца.

```

int main( void ) {
    unsigned char *buffer =
        (unsigned char*)big_random_block( SIZE );

    cudaEvent_t start, stop;
    HANDLE_ERROR( cudaEventCreate( &start ) );
    HANDLE_ERROR( cudaEventCreate( &stop ) );
    HANDLE_ERROR( cudaEventRecord( start, 0 ) );

    // выделить память GPU для хранения данных
    unsigned char *dev_buffer;
    unsigned int *dev_histo;
    HANDLE_ERROR( cudaMalloc( (void**)&dev_buffer, SIZE ) );
    HANDLE_ERROR( cudaMemcpy( dev_buffer, buffer, SIZE,
                           cudaMemcpyHostToDevice ) );

    HANDLE_ERROR( cudaMalloc( (void**)&dev_histo,
                           256 * sizeof( int ) ) );
    HANDLE_ERROR( cudaMemset( dev_histo, 0,
                           256 * sizeof( int ) ) );

    cudaDeviceProp prop;
    HANDLE_ERROR( cudaGetDeviceProperties( &prop, 0 ) );
    int blocks = prop.multiProcessorCount;
    histo_kernel<<<blocks*2,256>>>( dev_buffer, SIZE, dev_histo );

    unsigned int histo[256];
    HANDLE_ERROR( cudaMemcpy( histo, dev_histo,
                           256 * sizeof( int ),
                           cudaMemcpyDeviceToHost ) );

    // получить время завершения и вывести результаты хронометража
    HANDLE_ERROR( cudaEventRecord( stop, 0 ) );
    HANDLE_ERROR( cudaEventSynchronize( stop ) );
    float elapsedTime;
    HANDLE_ERROR( cudaEventElapsedTime( &elapsedTime,
                                       start, stop ) );
    printf( "Время генерации: %3.1f ms\n", elapsedTime );

    long histoCount = 0;
    for (int i=0; i<256; i++) {

```

```

histoCount += histo[i];
}
printf( "Сумма гистограммы: %ld\n", histoCount );

// проверить, что получились те же значения, что в версии для CPU
for (int i=0; i<SIZE; i++)
    histo[buffer[i]]--;
for (int i=0; i<256; i++) {
    if (histo[i] != 0)
        printf( "Ошибка в столбике %d!\n", i );
}

HANDLE_ERROR( cudaEventDestroy( start ) );
HANDLE_ERROR( cudaEventDestroy( stop ) );
cudaFree( dev_histo );
cudaFree( dev_buffer );
free( buffer );
return 0;
}

```

Ядро вычисления гистограммы с применением атомарных операций с глобальной памятью

А теперь самое интересное – исполняемый GPU код вычисления гистограммы! Ядру необходимо передать указатель на массив входных данных, его длину и указатель на результирующую гистограмму. Первым делом ядро вычисляет смещение от начала входного массива. Для каждой нити смещение находится в диапазоне от 0 до количества нитей минус единица. Шаг перемещения по массиву будет равен числу запущенных нитей. Надеемся, что вы помните этот прием; мы уже так поступали при сложении векторов произвольной длины, когда только начинали изучать нити.

```

#include "../common/book.h"

#define SIZE (100*1024*1024)

__global__ void histo_kernel( unsigned char *buffer,
                             long size,
                             unsigned int *histo ) {
    int i = threadIdx.x + blockIdx.x * blockDim.x;
    int stride = blockDim.x * gridDim.x;

```

После того как начальное смещение и шаг данной нити известны, программа приступает к обходу входного массива, попутно увеличивая значения в соответствующем столбике гистограммы.

```

while (i < size) {
    atomicAdd( &(histo[buffer[i]]), 1 );

```

```
i += stride;
}
}
```

В выделенной строке показан пример атомарной операции в CUDA C. Обращение `atomicAdd(addr, y)` порождает следующую атомарную последовательность операций: чтение значения, хранящегося по адресу `addr`, прибавление `y` к этому значению и сохранение результата в ячейке памяти с адресом `addr`. Оборудование гарантирует, что никакая другая нить не сможет прочитать или изменить значение в ячейке с адресом `addr`, пока вся последовательность не завершена. Тем самым обеспечивается предсказуемость результата. В нашем примере речь идет об адресе столбика гистограммы, соответствующего текущему байту. Как мы видели в версии для CPU, байту `buffer[i]` соответствует столбик `histo[buffer[i]]`. Поскольку атомарной операции необходим адрес этого столбика, то в первом аргументе мы передаем `&(histo[buffer[i]])`. А так как требуется увеличить хранящееся в нем значение на единицу, то второй аргумент равен 1.

А теперь, когда дым рассеялся, оказывается, что вычисление диаграммы на GPU очень похоже на версию для CPU.

```
#include "../common/book.h"

#define SIZE (100*1024*1024)

__global__ void histo_kernel( unsigned char *buffer,
                             long size,
                             unsigned int *histo ) {
    int i = threadIdx.x + blockIdx.x * blockDim.x;
    int stride = blockDim.x * gridDim.x;
    while (i < size) {
        atomicAdd( &(histo[buffer[i]]), 1 );
        i += stride;
    }
}
```

Однако радоваться рано. Если выполнить эту программу на GeForce GTX 285, то обнаружится, что на построение гистограммы для 100 Мб данных уходит 1,752 с. Ужас! Более чем в четыре раза медленнее, чем версия для CPU! Но именно для этого мы и измеряем эталонную производительность. Было бы стыдно согласиться на такую медленную реализацию только потому, что она работает на GPU.

Поскольку в ядре почти ничего не делается, то, скорее всего, проблема вызвана атомарными операциями с глобальной памятью. Действительно, когда тысячи нитей состязаются за доступ к горстке ячеек памяти, возникает сильная конкуренция. Чтобы гарантировать атомарность операций инкремента, оборудование должно сериализовать операции доступа к одной ячейке памяти. В результате выстраивается длинная очередь ожидающих операций, и весь потенциальный выигрыш съедается. Чтобы восстановить приемлемую производительность, необходимо улучшить алгоритм.

Ядро вычисления гистограммы с применением атомарных операций с глобальной и разделяемой памятью

По иронии судьбы, несмотря на то что снижение производительности обусловлено атомарными операциями, для преодоления этой проблемы количество таких операций нужно *увеличить*, а не уменьшить. Корень проблемы не в использовании атомарных операций как таковом, а в том, что тысячи нитей конкурируют за доступ к относительному небольшому количеству ячеек памяти. Решение же в том, чтобы разбить вычисление гистограммы на два этапа.

На первом этапе каждый параллельный блок будет вычислять отдельную гистограмму распределения тех данных, которые анализируют входящие в него нити. Поскольку блоки работают независимо, то эти гистограммы можно вычислять в разделяемой памяти, сэкономив на пересылке данных в DRAM. Но это не освобождает нас от необходимости выполнять операции атомарно, так как нити, принадлежащие одному блоку, все равно могут анализировать элементы массива с одинаковыми значениями. Однако теперь за доступ к 256 адресам конкурируют только 256 нитей, а не тысячи, как в первоначальной глобальной версии.

Итак, на первой фазе мы выделяем и обнуляем буфер в разделяемой памяти, где будет храниться промежуточная гистограмма блока. Напомним (см. главу 5), что поскольку на следующем шаге нам придется читать и модифицировать этот буфер, то каждая нить должна вызвать функцию `_syncthreads()`, перед тем как продолжить выполнение.

```
__global__ void histo_kernel( unsigned char *buffer,
                           long size,
                           unsigned int *histo ) {
    __shared__ unsigned int temp[256];
    temp[threadIdx.x] = 0;
    __syncthreads();
```

Следующий шаг после обнуления гистограммы очень похож на то, что мы делали в первоначальной версии. Единственное отличие – в том, что теперь используется буфер `temp[]` в разделяемой памяти, а не буфер `histo[]` в глобальной памяти, и что по завершении вычисления нужно вызвать функцию `_syncthreads()`, чтобы дождаться завершения всех операций.

```
int i = threadIdx.x + blockDim.x * gridDim.x;
int offset = blockDim.x * gridDim.x;
while (i < size) {
    atomicAdd( &temp[buffer[i]], 1);
    i += offset;
}
__syncthreads();
```

На последнем шаге модифицированной программы мы должны объединить временные гистограммы всех блоков в глобальном буфере `histo[]`. Предположим, что входные данные были разбиты на две половины, которые анализировали две нити, вычисляющие отдельные гистограммы. Если нить A встретила значение 0xFC 20 раз, а нить B встретила то же значение 5 раз, значит, байт 0xFC встречался во входных данных 25 раз. И вообще, каждый столбик окончательной гистограммы равен сумме соответствующих столбиков в гистограммах, вычисленных нитями A и B. Эта логика применима к любому числу нитей, поэтому объединение гистограмм отдельных блоков в окончательную гистограмму сводится к суммированию соответственных столбиков временных гистограмм. По изложенным выше причинам сложение следует выполнять атомарно:

```
    atomicAdd( &(histo[threadIdx.x]), temp[threadIdx.x] );
}
```

Поскольку у нас есть 256 нитей и 256 столбиков, каждая нить атомарно добавляет одно значения к итогу, хранящемуся в окончательной гистограмме. Если бы количество нитей и столбиков разнилось, то этот этап был бы сложнее. Отметим, что порядок объединения временных гистограмм в окончательную не определен, но поскольку сложение коммутативно, то результат всегда будет один и тот же при условии, что сложение производится атомарно.

На этом второй этап вычисления гистограммы завершается. Вот полный текст ядра:

```
__global__ void histo_kernel( unsigned char *buffer,
                             long size,
                             unsigned int *histo ) {
    __shared__ unsigned int temp[256];
    temp[threadIdx.x] = 0;
    __syncthreads();

    int i = threadIdx.x + blockIdx.x * blockDim.x;
    int offset = blockDim.x * gridDim.x;
    while (i < size) {
        atomicAdd( &temp[buffer[i]], 1 );
        i += offset;
    }

    __syncthreads();
    atomicAdd( &(histo[threadIdx.x]), temp[threadIdx.x] );
}
```

Эта версия программы работает намного быстрее предыдущей. На процессоре GeForce GTX 285 вычисление завершается за 0,057 с. Это не только гораздо лучше, чем при использовании атомарных операций с одной лишь глобальной памятью, но и заметно быстрее версии для CPU (0,057 вместо 0,416 с). Мы получили семи-

Резюме

кратное увеличение скорости по сравнению с версией для CPU. Таким образом, несмотря на разочарование, постигшее нас при первой попытке реализовать вычисление гистограммы на GPU, версию, в которой используется как глобальная, так и разделяемая память, следует считать успехом.

9.5. Резюме

Хотя мы часто и много говорили о том, как просто писать параллельные программы на языке CUDA C, при этом игнорировались случаи, когда массивно-параллельная архитектура типа той, что реализована в GPU, усложняет жизнь программиста. Иногда, для того чтобы справиться с ситуацией, в которой десятки тысяч нитей одновременно модифицируют один набор адресов, – типичной для массивно-параллельного компьютера, – приходится прилагать неординарные усилия. К счастью, у нас есть аппаратно поддержанные атомарные операции, упрощающие решение этой задачи.

Однако, как вы только что видели на примере вычисления гистограммы, бывает, что применение атомарных операций порождает проблемы с производительностью, решить которые можно только путем частичной переработки алгоритма. В данном случае мы перешли к двухэтапному алгоритму, уменьшившему конкуренцию за доступ к ячейкам глобальной памяти. Вообще говоря, стратегия, направленная на снижение конкуренции, обычно приносит плоды, и о ней всегда следует помнить при использовании атомарных операций в собственных приложениях.



Глава 10. Потоки

В этой книге мы неоднократно видели, как массивно-параллельная архитектура GPU может радикально повысить производительность по сравнению с аналогичной программой, исполняемой CPU. Но есть и еще один вид параллелизма, реализованный в графических процессорах NVIDIA. Это аналог *распараллеливания по задачам*, встречающегося во многих многопоточных приложениях для CPU. Вместо того чтобы вычислять одну и ту же функцию для разных элементов данных, как в распараллеливании по данным, технология распараллеливания по задачам подразумевает параллельное выполнение двух или более совершенно разных задач.

В контексте параллелизма *задачей* может быть все, что угодно. Например, одна задача может перерисовывать GUI, а другая – загружать по сети обновление программы. Эти задачи работают параллельно, хотя не имеют между собой ничего общего. В настоящее время механизм распараллеливания по задачам в GPU не такой гибкий, как в процессорах общего назначения, но тем не менее он позволяет добиться еще большего быстродействия от приложения, реализованного на GPU. В этой главе мы рассмотрим потоки CUDA (*stream*) и то, как они позволяют одновременно выполнять на GPU различные операции.

10.1. О чём эта глава

Прочитав эту главу, вы:

- узнаете о выделении заблокированных страниц памяти CPU;
- узнаете, что такое *потоки* CUDA;
- научитесь использовать потоки CUDA для ускорения работы приложений.

10.2. Блокированная память CPU

Во всех примерах из девяти предыдущих глав мы выделяли память GPU с помощью функции `cudaMalloc()`. А для выделения памяти CPU применяли стандартную библиотечную функцию `malloc()`. Однако исполняющая среда CUDA располагает собственным средством для выделения памяти CPU – функцией `cudaHostAlloc()`. Но зачем она нужна, если `malloc()` верой и правдой служит вам со дней написания первой программы на C?

А дело в том, что между памятью, выделенной `malloc()` и `cudaHostAlloc()`, есть существенная разница. Библиотечная функция `malloc()` выделяет стандартную страниценную память, тогда как `cudaHostAlloc()` – буфер, состоящий из *блокированных* страниц памяти. Иногда такие страницы называются *закрепленными* (*pinned*),

page-locked). Они обладают одним важным свойством: операционная система гарантирует, что закрепленная страница не будет выгружена на диск, а останется в физической памяти. Отсюда следует, что ОС может безопасно разрешить приложению доступ к такой странице по ее физическому адресу, потому что она никогда не будет вытеснена или перемещена в память.

Зная физический адрес буфера, GPU может воспользоваться механизмом прямого доступа к памяти (DMA) для копирования доступа в память CPU или из нее. Но копирование с помощью DMA происходит без участия CPU, а значит, CPU в принципе может в это время выгрузить буфер на диск или переместить его в память, обновив таблицы страниц операционной системы. Поэтому для копирования с помощью DMA абсолютно необходимо, чтобы страница была закреплена. На самом деле даже при попытке копирования обычной (страничной) памяти драйвер CUDA все равно применяет DMA для переноса буфера в память GPU. Таким образом, копирование производится дважды: сначала из страничного буфера в заблокированный временный буфер, а потом из заблокированного буфера в память GPU.

Поэтому при копировании из страничной памяти скорость копирования будет гарантированно ограничена *меньшей* из двух величин: скоростью передачи по шине PCI Express и скоростью внешней системной шины (Front Side Bus, FSB). В некоторых системах пропускная способность этих двух шин существенно различается, поэтому при копировании данных между GPU и CPU через заблокированную память можно добиться примерно двукратного повышения производительности по сравнению с обычной страничной памятью. Но даже в мире, где скорости внешней шины и шины PCI Express одинаковы, для буферов в страничной памяти все равно характерны издержки, связанные с дополнительным копированием, опосредуемым CPU.

Однако не спешите заменять все вхождения `malloc` на `cudaHostAlloc()`. Закрепленная память – палка о двух концах. Используя ее, вы лишаете себя всех преимуществ виртуальной памяти. Точнее, при работе такого приложения компьютер должен располагать физической памятью для всех заблокированных страниц, потому что выгрузить их на диск невозможно. А это значит, что память закончится гораздо быстрее, чем при обращении к стандартной функции `malloc()`. И, стало быть, приложение не только может начать сбить на машинах с небольшим объемом физической памяти, но и будет негативно влиять на производительность других программ, работающих на той же машине.

Нет, мы вовсе не хотим отговаривать вас от использования `cudaHostAlloc()`, просто нужно понимать все последствия заблокированных буферов. Мы рекомендуем использовать их только в качестве источника или места назначения при обращениях к `cudaMemcpy()` и освобождать сразу, как только в них отпадает необходимость, не дожидаясь завершения приложения, когда освобождается вся память. Работать с `cudaHostAlloc()` не сложнее, чем со всеми остальными ранее рассмотренными функциями, но все же рассмотрим пример выделения закрепленной памяти и заодно продемонстрируем выигрыш в производительности по сравнению с обычной страничной памятью.

Приложение будет очень простым, его единственная цель – протестировать производительность `cudaMemcpy()` для страничной и закрепленной памяти. Мы собираемся выделить буферы одинакового размера в памяти GPU и CPU, потом несколько раз скопировать данные из одного буфера в другой. Пользователь программы сможет задать направление копирования: «*up*» (из памяти CPU в память устройства) или «*down*» (из памяти устройства в память CPU). Для точного хронометража мы заведем события CUDA, соответствующие моментам начала и завершения последовательности операций копирования. Скорее всего, вы помните, как это делается, но на всякий случай приведем фрагмент, который освежит вашу память.

```
float cuda_malloc_test( int size, bool up ) {
    cudaEvent_t start, stop;
    int         *a, *dev_a;
    float       elapsedTime;

    HANDLE_ERROR( cudaEventCreate( &start ) );
    HANDLE_ERROR( cudaEventCreate( &stop ) );

    a = (int*)malloc( size * sizeof( *a ) );
    HANDLE_NULL( a );
    HANDLE_ERROR( cudaMalloc( (void**)&dev_a,
                           size * sizeof( *dev_a ) ) );
```

Независимо от направления копирования мы сначала выделяем в памяти CPU и GPU буферы, достаточные для размещения `size` целых чисел. Затем производим 100 операций копирования в направлении, заданном аргументом `up`, после чего останавливаем таймер.

```
HANDLE_ERROR( cudaEventRecord( start, 0 ) );
for (int i=0; i<100; i++) {
    if (up)
        HANDLE_ERROR( cudaMemcpy( dev_a, a,
                               size * sizeof( *dev_a ),
                               cudaMemcpyHostToDevice ) );
    else
        HANDLE_ERROR( cudaMemcpy( a, dev_a,
                               size * sizeof( *dev_a ),
                               cudaMemcpyDeviceToHost ) );
}
HANDLE_ERROR( cudaEventRecord( stop, 0 ) );
HANDLE_ERROR( cudaEventSynchronize( stop ) );
HANDLE_ERROR( cudaEventElapsedTime( &elapsedTime,
                                   start, stop ) );
```

После 100 операций копирования мы освобождаем оба буфера и уничтожаем события хронометража.

```
    free( a );
    HANDLE_ERROR( cudaFree( dev_a ) );
    HANDLE_ERROR( cudaEventDestroy( start ) );
    HANDLE_ERROR( cudaEventDestroy( stop ) );

    return elapsedTime;
}
```

Обращаем ваше внимание на то, что функция `cuda_malloc_test()` выделяет страничную память CPU, обращаясь к библиотечной функции `malloc()`. В версии с закрепленной памятью используется функция `cudaHostAlloc()`.

```
float cuda_host_alloc_test( int size, bool up ) {
    cudaEvent_t    start, stop;
    int           *a, *dev_a;
    float         elapsedTime;

    HANDLE_ERROR( cudaEventCreate( &start ) );
    HANDLE_ERROR( cudaEventCreate( &stop ) );

    HANDLE_ERROR( cudaHostAlloc( (void**)&a,
                                size * sizeof( *a ),
                                cudaHostAllocDefault ) );
    HANDLE_ERROR( cudaMalloc( void**&dev_a,
                            size * sizeof( *dev_a ) ) );

    HANDLE_ERROR( cudaEventRecord( start, 0 ) );
    for (int i=0; i<100; i++) {
        if (up)
            HANDLE_ERROR( cudaMemcpy( dev_a, a,
                                    size * sizeof( *a ),
                                    cudaMemcpyHostToDevice ) );
        else
            HANDLE_ERROR( cudaMemcpy( a, dev_a,
                                    size * sizeof( *a ),
                                    cudaMemcpyDeviceToHost ) );
    }
    HANDLE_ERROR( cudaEventRecord( stop, 0 ) );
    HANDLE_ERROR( cudaEventSynchronize( stop ) );
    HANDLE_ERROR( cudaEventElapsedTime( &elapsedTime,
                                      start, stop ) );

    HANDLE_ERROR( cudaFreeHost( a ) );
    HANDLE_ERROR( cudaFree( dev_a ) );
    HANDLE_ERROR( cudaEventDestroy( start ) );
    HANDLE_ERROR( cudaEventDestroy( stop ) );

    return elapsedTime;
}
```

Как видите, буфер, выделенный с помощью `cudaHostAlloc()`, используется точно так же, как буфер, выделенный `malloc()`. Но сигнатуры этих функций различаются: у `cudaHostAlloc()` есть дополнительный аргумент, в котором передаются флаги, позволяющие модифицировать поведение функции в части того, какую разновидность закрепленной памяти она выделяет. В данном случае он равен `cudaHostAllocDefault`. В следующей главе мы покажем, для чего предназначены другие флаги, а пока согласимся на выделение заблокированных страниц памяти, подразумеваемое по умолчанию. Для освобождения памяти, выделенной `cudaHostAlloc()`, применяется функция `cudaFreeHost()`. Как каждому вызову `malloc()` должен соответствовать вызов `free()`, так и каждому вызову `cudaHostAlloc()` должен соответствовать вызов `cudaFreeHost()`.

Тело функции `main()` не таит никаких сюрпризов:

```
#include "../common/book.h"

#define SIZE (10*1024*1024)

int main( void ) {
    float elapsedTime;
    float MB = (float)100*SIZE*sizeof(int)/1024/1024;
    elapsedTime = cuda_malloc_test( SIZE, true );
    printf( "Время при использовании cudaMalloc: %3.1f ms\n",
    elapsedTime );
    printf( "\tMб/с при копировании на GPU: %3.1f\n",
    MB/(elapsedTime/1000) );
```

Поскольку аргумент `up` функции `cuda_malloc_test()` равен `true`, то в приведенном выше фрагменте измеряется скорость копирования из памяти CPU в память устройства, то есть «вверх». Чтобы изменить скорость в противоположном направлении, выполняем те же самые вызовы, но во втором аргументе передаем `false`.

```
elapsedTime = cuda_malloc_test( SIZE, false );
printf( "Время при использовании cudaMalloc: %3.1f ms\n",
    elapsedTime );
printf( "\tMб/с при копировании на GPU: %3.1f\n",
    MB/(elapsedTime/1000) );
```

Те же тесты прогоним для замера производительности `cudaHostAlloc()`. Вызовем `cuda_host_alloc_test()` дважды, один раз передав в качестве аргумента `up` значение `true`, а другой раз – `false`.

```
elapsedTime = cuda_host_alloc_test( SIZE, true );
printf( "Время при использовании cudaHostAlloc: %3.1f ms\n",
    elapsedTime );
printf( "\tMб/с при копировании на GPU: %3.1f\n",
    MB/(elapsedTime/1000) );
```

```
elapsedTime = cuda_host_alloc_test( SIZE, false );
printf( "Время при использовании cudaHostAlloc: %3.1f ms\n",
        elapsedTime );
printf( "\tМб/с при копировании на CPU: %3.1f\n",
        MB/(elapsedTime/1000) );
}
```

Эксперименты с картой GeForce GTX 285 показали, что при использовании закрепленной памяти вместо страничной скорость копирования с CPU на GPU увелчилась с 2,77 Гб/с до 5,11 Гб/с. Скорость копирования с GPU на CPU увеличилась аналогично – с 2,43 Гб/с до 5,46 Гб/с. Таким образом, при работе большинства приложений, ограниченных пропускной способностью шины PCI Express, переход на закрепленную память даст заметное повышение производительности. Однако закрепленная память применяется не только ради увеличения производительности. В следующих разделах мы увидим, что иногда использовать ее просто необходимо.

10.3. Потоки CUDA

В главе 6 мы познакомились с событиями CUDA. Но при этом отложили на потом обсуждение второго аргумента функции `cudaEventRecord()`, упомянув лишь, что это *поток*, в который вставляется событие.

```
cudaEvent_t start;
cudaEventCreate(&start);
cudaEventRecord( start, 0 );
```

Потоки CUDA могут немало поспособствовать ускорению работы приложения. Под *потоком* понимается очередь операций GPU, выполняемых в определенном порядке. В очередь можно ставить такие операции, как запуск ядра, копирование памяти и даже регистрацию событий. Операции будут выполняться в том порядке, в котором добавлялись в поток. Можно считать, что каждый поток – это исполняемая на GPU *задача*, причем эти задачи могут исполняться параллельно. Сначала мы рассмотрим, как вообще используются потоки, а потом поговорим о том, как применить их для ускорения работы приложения.

10.4. Использование одного потока CUDA

Ниже мы увидим, что по-настоящему возможности потоков начинают проявляться, когда их несколько, но, чтобы проиллюстрировать порядок работы с ними, начнем с приложения, в котором есть всего один поток. Пусть имеется ядро, принимающее два входных буфера, *a* и *b*. Это ядро вычисляет результат, зависящий от данных в этих буферах, и помещает его в выходной буфер *c*. Подобную задачу мы

решали в программе сложения векторов, а сейчас вычислим среднее арифметическое трех значений в *a* и трех значений в *b*:

```
#include "../common/book.h"

#define N (1024*1024)
#define FULL_DATA_SIZE (N*20)

__global__ void kernel( int *a, int *b, int *c ) {
    int idx = threadIdx.x + blockIdx.x * blockDim.x;
    if (idx < N) {
        int idx1 = (idx + 1) % 256;
        int idx2 = (idx + 2) % 256;
        float as = (a[idx] + a[idx1] + a[idx2]) / 3.0f;
        float bs = (b[idx] + b[idx1] + b[idx2]) / 3.0f;
        c[idx] = (as + bs) / 2;
    }
}
```

Само ядро не так важно, поэтому не задерживайтесь на нем слишком долго, если не сразу поняли, что именно оно делает. Это просто своего рода заглушка, действительно важный код, относящийся к потокам, находится в функции *main()*.

```
int main( void ) {
    cudaDeviceProp prop;
    int whichDevice;
    HANDLE_ERROR( cudaGetDevice( &whichDevice ) );
    HANDLE_ERROR( cudaGetDeviceProperties( &prop, whichDevice ) );
    if (!prop.deviceOverlap) {
        printf( "Устройство не поддерживает перекрытие, поэтому "
               "потоки не приведут к ускорению\n" );
        return 0;
    }
}
```

Первым делом мы выбираем устройство и проверяем, поддерживает ли оно так называемое *перекрытие операций* (device overlap). GPU с поддержкой перекрытия операций может одновременно исполнять ядро на CUDA C и копирование между памятью устройства и CPU. Как и было обещано, для достижения такого перекрытия вычислений и передачи данных мы воспользуемся несколькими потоками, но сначала покажем, как создать и использовать один поток. Как и во всех примерах, имеющих целью измерить повышение (или снижение) производительности, мы начинаем с создания и регистрации события-таймера:

```
cudaEvent_t start, stop;
float elapsedTime;

// запустить таймер
```

```
HANDLE_ERROR( cudaEventCreate( &start ) );
HANDLE_ERROR( cudaEventCreate( &stop ) );
HANDLE_ERROR( cudaEventRecord( start, 0 ) );
```

Запустив таймер, мы создаем поток:

```
// инициализировать поток
cudaStream_t stream;
HANDLE_ERROR( cudaStreamCreate( &stream ) );
```

Вот, собственно, и все, что нужно для создания потока. Говорить тут не о чем, поэтому перейдем к выделению памяти для данных.

```
int *host_a, *host_b, *host_c;
int *dev_a, *dev_b, *dev_c;

// выделить память на GPU
HANDLE_ERROR( cudaMalloc( (void**)&dev_a,
    N * sizeof(int) ) );
HANDLE_ERROR( cudaMalloc( (void**)&dev_b,
    N * sizeof(int) ) );
HANDLE_ERROR( cudaMalloc( (void**)&dev_c,
    N * sizeof(int) ) );

// выделить закрепленную память, необходимую потоку
HANDLE_ERROR( cudaHostAlloc( (void**)&host_a,
    FULL_DATA_SIZE * sizeof(int),
    cudaHostAllocDefault ) );
HANDLE_ERROR( cudaHostAlloc( (void**)&host_b,
    FULL_DATA_SIZE * sizeof(int),
    cudaHostAllocDefault ) );
HANDLE_ERROR( cudaHostAlloc( (void**)&host_c,
    FULL_DATA_SIZE * sizeof(int),
    cudaHostAllocDefault ) );

for (int i=0; i<FULL_DATA_SIZE; i++) {
    host_a[i] = rand();
    host_b[i] = rand();
}
```

Мы выделили память для входных и выходных буферов на GPU и на CPU. Отметим, что на CPU выделена закрепленная память – с помощью функции `cudaHostAlloc()`. Тому есть очень веская причина, отнюдь не исчерпывающаяся желанием ускорить копирование. Детали мы обсудим чуть ниже, но сразу отметим, что будем использовать новую разновидность функции `cudaMemcpy()`, для которой *требуется*, чтобы память CPU была закреплена. Получив память для входных буферов, мы заполняем их случайными целыми числами, обращаясь к стандартной библиотечной функции `rand()`.

Итак, поток и события хронометража созданы, буферы на устройстве и CPU выделены, и мы наконец-то готовы что-нибудь посчитать! Раньше мы на этой стадии просто копировали оба входных буфера в память GPU, запускали ядро, а потом копировали выходной буфер в память CPU. Мы и сейчас будем придерживаться той же схемы, но с небольшими изменениями.

Во-первых, мы не станем копировать входные буфера на GPU целиком, а разобьем их на меньшие порции и для каждой порции выполним трехшаговую процедуру. Иначе говоря, мы скопируем на GPU какую-то часть входных буферов, выполним для этой части ядро, а затем скопируем получившуюся часть выходного буфера в память CPU. Причиной для такого образа действий может быть, например, то, что GPU располагает гораздо меньшим объемом памяти, чем CPU, поэтому вычисления приходится разбивать на части, так как весь входной буфер в память GPU не помещается. Код, выполняющий такую последовательность вычислений над порциями, мог бы выглядеть следующим образом:

```
// в цикле обходим данные, разбивая их на порции
for (int i=0; i<FULL_DATA_SIZE; i+= N) {
    // асинхронно копировать закрепленную память в устройство
    HANDLE_ERROR( cudaMemcpyAsync( dev_a, host_a+i,
        N * sizeof(int),
        cudaMemcpyHostToDevice,
        stream ) );
    HANDLE_ERROR( cudaMemcpyAsync( dev_b, host_b+i,
        N * sizeof(int),
        cudaMemcpyHostToDevice,
        stream ) );

    kernel<<<N/256,256,0,stream>>>( dev_a, dev_b, dev_c );

    // скопировать данные из памяти устройства в закрепленную память
    HANDLE_ERROR( cudaMemcpyAsync( host_c+i, dev_c,
        N * sizeof(int),
        cudaMemcpyDeviceToHost,
        stream ) );
}
```

Но в приведенном выше фрагменте есть два неожиданных отклонения от нормы. Во-первых, для копирования данных в память GPU и обратно вместо уже знакомой функции cudaMemcpy() применяется ранее не встречавшаяся функция cudaMemcpyAsync(). Между ними есть тонкое, но важное различие. Функция cudaMemcpy() ведет себя как стандартная библиотечная функция memcpy(). Точнее, она выполняется *синхронно*, то есть к моменту возврата из нее копирование полностью завершено и выходной буфер содержит все данные, которые планировалось в него поместить.

Напротив, функция cudaMemcpyAsync() работает асинхронно, что и отражено в ее имени. Обращение к cudaMemcpyAsync() просто помещает *запрос* на выполнение операции копирования в поток, определяемый аргументом stream. Когда

функция возвращает управление, нет никакой гарантии, что копирование хотя бы началось, не говоря уже о его завершении. Гарантируется лишь, что копирование будет выполнено до начала следующей операции, помещенной в тот же самый поток. Требуется, чтобы области памяти CPU, на которые направлены указатели, переданные `cudaMemcpyAsync()`, были выделены функцией `cudaHostAlloc()`. Иными словами, асинхронное копирование разрешено только в закрепленную память или из таковой.

Отметим, что обозначенное угловыми скобками ядро также принимает необязательный аргумент `stream`. Этот запуск ядра выполняется асинхронно, как и две предыдущие операции копирования в память GPU и последующая операция копирования в память CPU. Строго говоря, не исключено, что мы выйдем из цикла `encl` до того, как начнется копирование памяти или исполнение ядра. Гарантируется лишь, что первая операция копирования, помещенная в поток, завершится до начала второй операции копирования. А вторая операция копирования завершится до начала исполнения ядра, а исполнение ядра закончится до начала третьей операции копирования. То есть, как и было сказано в начале главы, поток ведет себя как упорядоченная очередь задач для GPU.

По выходе из цикла `for` в очереди еще могут находиться задания, которые GPU предстоит выполнить. Если нужна гарантия, что все вычисления на GPU и копирование памяти завершены, то необходимо синхронизировать его с CPU. То есть мы должны сказать CPU, чтобы он подождал что-либо делать, пока GPU не закончит работу. Для этого служит функция `cudaStreamSynchronize()`, а передается ей поток, который мы ждем:

```
HANDLE_ERROR( cudaStreamSynchronize( stream ) );
```

Так как после синхронизации потока с CPU все вычисления и копирование завершены, то мы можем остановить таймер, вывести данные о производительности и освободить все буферы.

```
HANDLE_ERROR( cudaEventRecord( stop, 0 ) );
HANDLE_ERROR( cudaEventSynchronize( stop ) );
HANDLE_ERROR( cudaEventElapsedTime( &elapsedTime,
                                    start, stop ) );
printf( "Затраченное время: %3.1f ms\n", elapsedTime );

// освободить потоки и память
HANDLE_ERROR( cudaFreeHost( host_a ) );
HANDLE_ERROR( cudaFreeHost( host_b ) );
HANDLE_ERROR( cudaFreeHost( host_c ) );
HANDLE_ERROR( cudaFree( dev_a ) );
HANDLE_ERROR( cudaFree( dev_b ) );
HANDLE_ERROR( cudaFree( dev_c ) );
```

И напоследок, перед тем как выйти из приложения, мы уничтожаем поток, в котором хранилась очередь операций:

```
HANDLE_ERROR( cudaStreamDestroy( stream ) );  
return 0;  
}
```

Честно говоря, в этом примере нет ничего, демонстрирующего полезность потоков. Разумеется, даже один поток может ускорить работу, если CPU есть что делать, пока GPU занят выполнением операций, помещенных в поток. Но и в том случае, когда CPU нечем заняться, потоки могут повысить производительность приложения, и в следующем разделе мы расскажем, как именно.

10.5. Использование нескольких потоков CUDA

Изменим программу из раздела 10.4 так, чтобы вместо одного потока использовалось два. В начале этой программы мы проверили, что устройство поддерживает *перекрытие*, и разбили вычисление на порции. Идея улучшенной версии проста и опирается на две вещи: вычисление по частям и перекрытие операций копирования с исполнением ядра. Мы хотим, чтобы поток 1 копировал входные буферы в память GPU, а поток 0 в это время исполнял свое ядро. Затем поток 1 будет выполнять свое ядро, а поток 0 – копировать свои результаты в память CPU. Далее поток 1 займется копированием своих результатов в память CPU, а поток 0 начнет выполнять свое ядро для следующей порции данных. В предположении, что операции копирования и исполнения ядра занимают примерно одинаковое время, временная диаграмма приложения будет выглядеть, как показано на рис. 10.1. Здесь считается, что GPU тратит на копирование памяти и на исполнение ядра одно и то же время, поэтому пустые прямоугольники представляют промежутки времени, когда один поток ожидает завершения операции, которая не может перекрываться с операцией из другого потока. Отметим также, что на этом и следующих рисунках мы сокращаем имя функции `cudaMemcpyAsync()` до «`тетсруд`».

На самом деле временная диаграмма работы может оказаться даже более благоприятной; некоторые из последних моделей NVIDIA GPU поддерживают одновременное исполнение ядра и двух операций копирования: одно – в память устройства, а другое – из памяти устройства¹. Так или иначе, на любом устройстве, поддерживающем перекрытие операций копирования памяти и исполнения ядра, общее время работы приложения при использовании нескольких потоков должно уменьшиться.

Несмотря на грандиозные планы по ускорению приложения, само вычислительное ядро никак не изменяется.

```
#include "../common/book.h"
```

¹ Fermi GPU также поддерживают одновременное выполнение нескольких ядер.

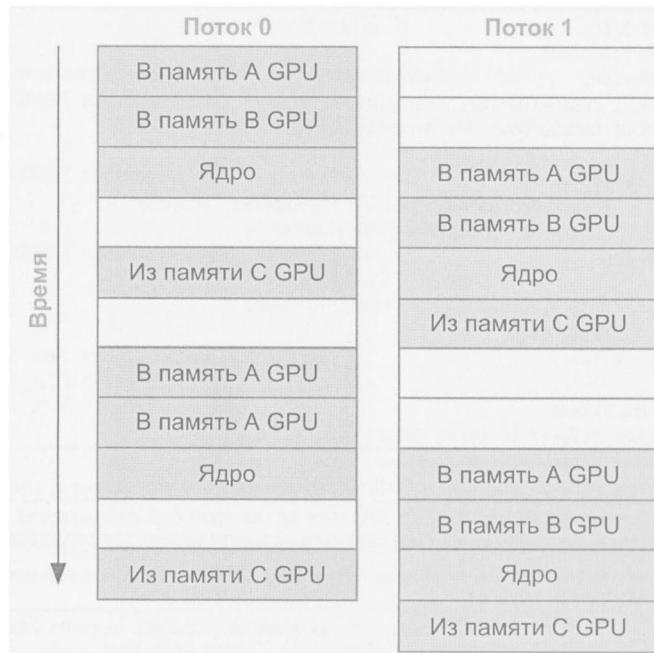


Рис. 10.1. Временная диаграмма желательной последовательности исполнения приложения при использовании двух независимых потоков

```
#define N (1024*1024)
#define FULL_DATA_SIZE (N*20)

__global__ void kernel( int *a, int *b, int *c ) {
    int idx = threadIdx.x + blockIdx.x * blockDim.x;
    if (idx < N) {
        int idx1 = (idx + 1) % 256;
        int idx2 = (idx + 2) % 256;
        float as = (a[idx] + a[idx1] + a[idx2]) / 3.0f;
        float bs = (b[idx] + b[idx1] + b[idx2]) / 3.0f;
        c[idx] = (as + bs) / 2
    }
}
```

Как и в версии с одним потоком, мы проверяем, что устройство поддерживает перекрытие вычислений с копированием памяти. Если это так, то поступаем так же, как раньше, – создаем события CUDA для хронометража.

```

int main( void ) {
    cudaDeviceProp prop;
    int whichDevice;
    HANDLE_ERROR( cudaGetDevice( &whichDevice ) );
    HANDLE_ERROR( cudaGetDeviceProperties( &prop, whichDevice ) );

    if (!prop.deviceOverlap) {
        printf( "Устройство не поддерживает перекрытие, поэтому "
                "потоки не приведут к ускорению\n" );
        return 0;
    }

    cudaEvent_t start, stop;
    float elapsedTime;

    // запустить таймер
    HANDLE_ERROR( cudaEventCreate( &start ) );
    HANDLE_ERROR( cudaEventCreate( &stop ) );
    HANDLE_ERROR( cudaEventRecord( start, 0 ) );

```

Далее создаем два потока так же, как в предыдущей версии создавали один:

```

// инициализировать потоки
cudaStream_t stream0, stream1;
HANDLE_ERROR( cudaStreamCreate( &stream0 ) );
HANDLE_ERROR( cudaStreamCreate( &stream1 ) );

```

Предполагается, что на стороне CPU по-прежнему есть два входных буфера и один выходной. Входные буфера, как и раньше, заполняются случайными данными. Однако поскольку для обработки данных мы собираемся использовать два потока, то на стороне GPU выделяются два одинаковых набора буферов, чтобы каждый поток мог обрабатывать свои порции независимо от другого.

```

int *host_a, *host_b, *host_c;
int *dev_a0, *dev_b0, *dev_c0; // буфера GPU для stream0
int *dev_a1, *dev_b1, *dev_c1; // буфера GPU для stream1

// выделить память на GPU
HANDLE_ERROR( cudaMalloc( (void**)&dev_a0,
                           N * sizeof(int) ) );
HANDLE_ERROR( cudaMalloc( (void**)&dev_b0,
                           N * sizeof(int) ) );
HANDLE_ERROR( cudaMalloc( (void**)&dev_c0,
                           N * sizeof(int) ) );
HANDLE_ERROR( cudaMalloc( (void**)&dev_a1,
                           N * sizeof(int) ) );
HANDLE_ERROR( cudaMalloc( (void**)&dev_b1,
                           N * sizeof(int) ) );
HANDLE_ERROR( cudaMalloc( (void**)&dev_c1,

```

```

    N * sizeof(int) ) );

// выделить закрепленную память, необходимую потокам
HANDLE_ERROR( cudaHostAlloc( (void**)&host_a,
                             FULL_DATA_SIZE * sizeof(int),
                             cudaHostAllocDefault ) );
HANDLE_ERROR( cudaHostAlloc( (void**)&host_b,
                             FULL_DATA_SIZE * sizeof(int),
                             cudaHostAllocDefault ) );
HANDLE_ERROR( cudaHostAlloc( (void**)&host_c,
                             FULL_DATA_SIZE * sizeof(int),
                             cudaHostAllocDefault ) );

for (int i=0; i<FULL_DATA_SIZE; i++) {
    host_a[i] = rand();
    host_b[i] = rand();
}

```

Далее мы в цикле обрабатываем порции данных, как и в первой версии. Но поскольку теперь есть два потока, на каждой итерации цикла выполняется вдвое больше действий, чем раньше. В потоке `stream0` мы ставим в очередь асинхронное копирование `a` и `b` в память GPU, исполнение ядра и обратное копирование с:

```

// в цикле обходим данные, разбивая их на порции
for (int i=0; i<FULL_DATA_SIZE; i+= N*2) {
    // асинхронно копировать закрепленную память в устройство
    HANDLE_ERROR( cudaMemcpyAsync( dev_a0, host_a+i,
                                  N * sizeof(int),
                                  cudaMemcpyHostToDevice,
                                  stream0 ) );
    HANDLE_ERROR( cudaMemcpyAsync( dev_b0, host_b+i,
                                  N * sizeof(int),
                                  cudaMemcpyHostToDevice,
                                  stream0 ) );

    kernel<<<N/256,256,0,stream0>>>( dev_a0, dev_b0, dev_c0 );

    // скопировать данные из памяти устройства в закрепленную память
    HANDLE_ERROR( cudaMemcpyAsync( host_c+i, dev_c0,
                                  N * sizeof(int),
                                  cudaMemcpyDeviceToHost,
                                  stream0 ) );

```

Поставив в очередь все операции для потока `stream0`, мы делаем то же самое, но уже для потока `stream1`.

```

// асинхронно копировать закрепленную память в устройство
HANDLE_ERROR( cudaMemcpyAsync( dev_a1, host_a+i+N,
                              N * sizeof(int),

```

```

        cudaMemcpyHostToDevice,
        stream1 ) );
HANDLE_ERROR( cudaMemcpyAsync( dev_b1, host_b+i+N,
N * sizeof(int),
cudaMemcpyHostToDevice,
stream1 ) );

kernel<<<N/256, 256, 0, stream1>>>( dev_a1, dev_b1, dev_c1 );

// скопировать данные из памяти устройства в закрепленную память
HANDLE_ERROR( cudaMemcpyAsync( host_c+i+N, dev_c1,
N * sizeof(int),
cudaMemcpyDeviceToHost,
stream1 ) );

```

Так мы и чередуем потоки в цикле, ставя в очередь каждого очередную порцию данных, пока все порции не будут распределены. По выходе из цикла мы синхронизируем GPU и CPU и только потом останавливаем таймер. Так как мы работали в двух потоках, то и синхронизировать нужно оба.

```
HANDLE_ERROR( cudaStreamSynchronize( stream0 ) );
HANDLE_ERROR( cudaStreamSynchronize( stream1 ) );
```

Конец функции `main()` ничем не отличается от версии с одним потоком. Мы останавливаем таймер, выводим затраченное время и прибираемся за собой. Конечно, теперь нужно уничтожить два потока и освободить вдвое больше буферов GPU, чем раньше, но в остальном все то же самое.

```

HANDLE_ERROR( cudaEventRecord( stop, 0 ) );
HANDLE_ERROR( cudaEventSynchronize( stop ) );
HANDLE_ERROR( cudaEventElapsedTime( &elapsedTime,
start, stop ) );
printf( "Затраченное время: %.1f ms\n", elapsedTime );

// освободить потоки и память
HANDLE_ERROR( cudaFreeHost( host_a ) );
HANDLE_ERROR( cudaFreeHost( host_b ) );
HANDLE_ERROR( cudaFreeHost( host_c ) );
HANDLE_ERROR( cudaFree( dev_a0 ) );
HANDLE_ERROR( cudaFree( dev_b0 ) );
HANDLE_ERROR( cudaFree( dev_c0 ) );
HANDLE_ERROR( cudaFree( dev_a1 ) );
HANDLE_ERROR( cudaFree( dev_b1 ) );
HANDLE_ERROR( cudaFree( dev_c1 ) );
HANDLE_ERROR( cudaStreamDestroy( stream0 ) );
HANDLE_ERROR( cudaStreamDestroy( stream1 ) );

return 0;
}
```

Мы замерили время работы обеих версий программы – с одним и двумя потоками – на GeForce GTX 285. Первая завершилась за 62 мс, вторая – за 61 мс.

О-хо-хо...

Что ж, для того и хронометрируем. Иногда все усилия, направленные на повышение производительности, не приводят ни к чему, кроме усложнения кода.

Но почему все-таки приложение не стало работать быстрее? Мы же так уверенно утверждали, что это должно случиться?! Не теряйте надежду, мы *действительно* можем ускорить работу, добавив второй поток, но чтобы пожать плоды перекрытия операций, нужно лучше понимать, как драйвер CUDA обрабатывает потоки. А для этого приглядимся внимательнее к драйверу CUDA и к принципам аппаратной архитектуры CUDA.

10.6. Планирование задач на GPU

Хотя логические потоки представляют собой независимые очереди операций, исполняемых GPU, на самом деле эта абстракция не вполне согласуется с механизмом обслуживания очередей графическим процессором. Будучи программистами, мы представляем себе потоки как упорядоченные последовательности операций копирования памяти и исполнения ядра. Однако оборудование ничего не знает о потоках. Имеются одна или несколько аппаратных подсистем для выполнения копирования памяти и подсистема для исполнения ядер. Эти подсистемы имеют независимые очереди команд, так что планирование задач выглядит, как показано на рис. 10.2. Стрелками показано, как операции, помещенные в очереди потоков, планируются для выполнения аппаратными подсистемами.

Таким образом, пользователь и оборудование по-разному видят работу очередей GPU, а обязанность сделать обе стороны счастливыми и довольными возлагается на драйвер CUDA. Прежде всего отметим, что существуют важные зависимости, определяемые порядком добавления операций в потоки. Например, на рис. 10.2 операция копирования памяти A в потоке 0 должна быть завершена до начала операции копирования B, которая, в свою очередь, должна завершиться до запуска ядра A. Но после того как эти операции помещены в очереди аппаратных подсистем копирования и исполнения ядер, эти зависимости потерялись, а проследить за тем, чтобы аппаратные подсистемы соблюдали межпотоковые зависимости, должен драйвер CUDA.

А что это означает для нас? Посмотрим, что в реальности происходит в примере из раздела 10.5 «Использование нескольких потоков CUDA». Заглянув в код, мы видим, что приложение выполняет `cudaMemcpyAsync()` для a, `cudaMemcpyAsync()` для b, запускает ядро, а потом с помощью `cudaMemcpyAsync()` копирует с обратно в память CPU. Сначала в очередь ставятся все операции в потоке 0, а потом все операции в потоке 1. Драйвер CUDA планирует выполнение этих операций оборудованием в том порядке, в котором они были заданы, сохраняя межпотоковые зависимости. Эти зависимости показаны на рис. 10.3, где стрелки, направленные от копирования к ядру, показывают, что операция копирования зависит от ядра и не может начаться, пока ядро не завершится.



Рис. 10.2. Отображение потоков CUDA на подсистемы GPU

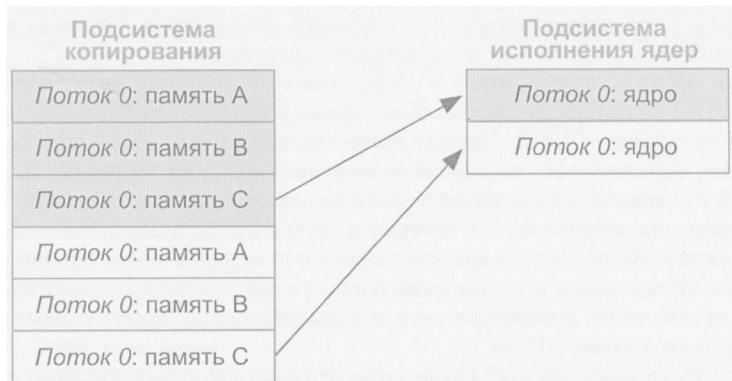


Рис. 10.3. Стрелки показывают зависимость вызовов `cudaMemcpyAsync()` от операций исполнения ядра в примере из раздела 10.5 «Использование нескольких потоков CUDA»

Теперь, разобравшись с тем, как работает планировщик GPU, мы можем взглянуть на временную диаграмму выполнения задач оборудованием (рис. 10.4). Поскольку обратное копирование буфера с в потоке 0 зависит от завершения ядра в том же потоке, совершенно независимые операции копирования а и в память

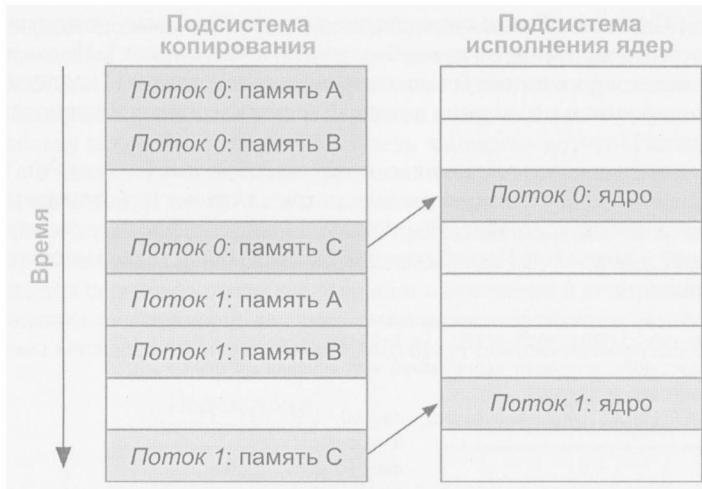


Рис. 10.4. Временная диаграмма работы программы из раздела 10.5 «Использование нескольких потоков CUDA»

GPU в потоке 1 оказываются заблокированы, потому что подсистемы GPU выполняют задачи в том порядке, в котором они были помещены в очередь. Вот из-за этой неэффективности версия приложения с двумя потоками и не показала никакого прироста производительности. Это прямое следствие нашего допущения, будто оборудование работает именно так, как подразумевает модель потокового программирования CUDA.

Мораль сей басни в том, что программист должен помочь системе обеспечить действительно параллельное выполнение независимых потоков. Памятуя о том, что аппаратные подсистемы копирования и исполнения ядер независимы, мы должны понимать, что от порядка постановки этих операций в очереди потоков зависит, как драйвер CUDA запланирует их выполнение. В следующем разделе мы покажем, как помочь оборудованию добиться перекрытия операций копирования и исполнения ядер.

10.7. Эффективное использование потоков CUDA

В предыдущем разделе мы видели, что если запланировать сразу все операции в одном потоке, то очень легко случайно заблокировать копирование или исполнение ядра в другом потоке. Чтобы избавиться от этого, достаточно ставить задачи в очередь не «вдоль» (depth-first), а «поперек» (breadth-first). Иными словами, вместо того чтобы сначала добавлять в очередь потока 0 операции копирования a, копирования b, исполнения ядра и копирования c, а только потом приступать к заполнению очереди потока 1, мы будем распределять задачи между потоками

вперемежку. Сначала добавим копирование а в поток 0, потом – копирование а в поток 1. Потом копирование б в поток 0 и копирование б в поток 1. Потом запланируем исполнение ядра в потоке 0 и исполнение ядра в потоке 1. И наконец, ставим обратное копирование с в очередь потока 0, а затем – обратное копирование с в очередь потока 1.

Чтобы понять, как это выглядит на практике, обратимся к коду. Мы изменили только порядок добавления операции в потоки, поэтому вся оптимизация сводится к копированию и вставке. Все остальное не меняется, то есть улучшения локализованы в цикле for. Новый порядок распределения задач между потоками выглядит так:

```
for (int i=0; i<FULL_DATA_SIZE; i+= N*2) {
    // поставить операции копирования а в очереди потоков
    // stream0 и stream1
    HANDLE_ERROR( cudaMemcpyAsync( dev_a0, host_a+i,
        N * sizeof(int),
        cudaMemcpyHostToDevice,
        stream0 ) );
    HANDLE_ERROR( cudaMemcpyAsync( dev_a1, host_a+i+N,
        N * sizeof(int),
        cudaMemcpyHostToDevice,
        stream1 ) );

    // поставить операции копирования б в очереди потоков
    // stream0 и stream1
    HANDLE_ERROR( cudaMemcpyAsync( dev_b0, host_b+i,
        N * sizeof(int),
        cudaMemcpyHostToDevice,
        stream0 ) );
    HANDLE_ERROR( cudaMemcpyAsync( dev_b1, host_b+i+N,
        N * sizeof(int),
        cudaMemcpyHostToDevice,
        stream1 ) );

    // поставить операции исполнения ядра в очереди потоков
    // stream0 и stream1
    kernel<<<N/256,256,0,stream0>>>( dev_a0, dev_b0, dev_c0 );
    kernel<<<N/256,256,0,stream1>>>( dev_a1, dev_b1, dev_c1 );

    // поставить операции копирования с из памяти устройства в
    // закрепленную память в очереди потоков stream0 и stream1
    HANDLE_ERROR( cudaMemcpyAsync( host_ci, dev_c0,
        N * sizeof(int),
        cudaMemcpyDeviceToHost,
        stream0 ) );
    HANDLE_ERROR( cudaMemcpyAsync( host_ci+N, dev_c1,
        N * sizeof(int),
        cudaMemcpyDeviceToHost,
        stream1 ) );
```

Если предположить, что времена операций копирования и исполнения ядра примерно одинаково, то временная диаграмма теперь будет выглядеть, как показано на рис. 10.5. Зависимости между подсистемами обозначены стрелками только для того, чтобы показать, что и при новом порядке планирования они удовлетворяются.

Поскольку операции ставились в очередь «поперек» потоков, то теперь копирование с в потоке 0 уже не блокирует начальное копирование а и в в потоке 1. А значит, GPU может копировать и исполнять ядра параллельно, что должно заметно ускорить работу приложения. Новая программа завершается за 48 мс, что на 21% быстрее первоначальной наивной реализации с двумя потоками. Если в приложении удается перекрыть почти все операции вычисления и копирования памяти, то такой подход даст примерно двукратное повышение производительности, так как подсистемы копирования и исполнения ядер будут постоянно загружены работой.

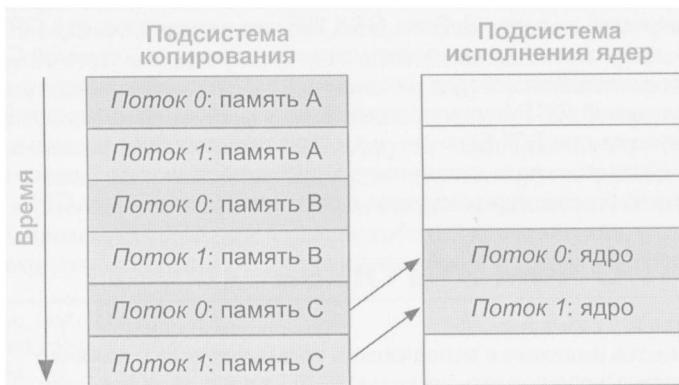


Рис. 10.5. Временная диаграмма работы улучшенной программы; стрелками показаны зависимости между подсистемами

10.8. Резюме

В этой главе мы видели, как реализовать распараллеливание по задачам в приложениях, написанных на CUDA C. Организовав два (или более) потоков CUDA, мы можем заставить GPU одновременно выполнять ядро и производить операцию копирования между памятью устройства и CPU. Но при этом следует проявлять осторожность. Во-первых, память CPU нужно выделять функцией `cudaHostAlloc()`, потому что операции копирования ставятся в очередь и асинхронно выполняются функцией `cudaMemcpyAsync()`, а для асинхронного копирования буфера должны быть закреплены в памяти. Во-вторых, нужно помнить, что от порядка добавления операций в очередь потоков зависит возможность перекрытия копирования и исполнения ядра. Общая рекомендация заключается в том, чтобы распределять работы между потоками «поперек», или по кругу. Если не знать, как аппаратура обслуживает очереди, этот совет может показаться противоречащим интуиции, поэтому помните о нем, когда начнете писать свои приложения.



Глава 11. CUDA С на нескольких GPU

Старая поговорка гласит: «Лучшие вычислений на GPU могут быть только вычисления на двух GPU». В последние годы системы, оборудованные несколькими графическими процессорами, получают все большее распространение. Конечно, системы с несколькими GPU похожи на системы с несколькими CPU в том смысле, что пока не стали общепринятыми, но тем не менее столкнуться с такой системой совсем несложно. Например, на карте GeForce GTX 295 уже установлены два GPU, а карта Tesla S1070 содержит аж четыре графических процессора с поддержкой CUDA. Системы на базе последних наборов микросхем NVIDIA содержат интегрированный GPU с поддержкой CUDA на материнской плате. Добавьте дискретный NVIDIA GPU в один из слотов PCI Express – и вы получите систему с несколькими GPU. Все эти сценарии вовсе не выглядят надуманными, поэтому имеет смысл научиться использовать во благо ресурсы системы, оснащенной несколькими GPU.

11.1. О чём эта глава

Прочитав эту главу, вы:

- научитесь выделять и использовать *нуль-копируемую память*;
- научитесь использовать несколько GPU в одном приложении;
- научитесь выделять и использовать переносимую закрепленную память.

11.2. Нуль-копируемая память CPU

В главе 10 мы рассматривали закрепление памяти CPU, гарантирующее, что буфер в физической памяти не будет выгружен на диск. Если помните, для выделения такой памяти вызывается функция `cudaHostAlloc()` с параметром `cudaHostAllocDefault`. Мы обещали показать и другие, более интересные способы выделения закрепленной памяти. Если это единственная причина, по которой вы продолжаете читать книгу, то радуйтесь – ожиданию пришел конец. Вместо `cudaHostAllocDefault` можно передать флаг `cudaHostAllocMapped`. Выделенная в результате область оказывается закрепленной в том же смысле, что и раньше, то есть не может быть ни выгружена, ни перемещена в физической памяти. Но, помимо использования для копирования между CPU и GPU, этот вид памяти CPU позволяет нарушить одно из правил, сформулированных в главе 3: к такой памяти разрешен прямой доступ из ядер, написанных на CUDA C. Поскольку не требуется копировать данные на GPU и обратно, такая память называется *нуль-копируемой* (*zero-copy*).

11.2.1. Вычисление скалярного произведения с применением нуль-копируемой памяти

Обычно GPU обращается только к памяти устройства, а CPU – только к памяти CPU. Но иногда бывает полезно нарушить эти правила. Чтобы понять, когда предпочтительно позволить GPU манипулировать памятью CPU, вернемся к нашему любимому примеру редукции: вычислению скалярного произведения векторов. Если вы читаете книгу подряд, то, наверное, помните первую попытку решить эту задачу. Мы скопировали оба входных вектора в память GPU, проделали вычисления, затем скопировали промежуточные результаты в память CPU и завершили вычисление уже на CPU.

В новой версии мы обойдемся без копирования входных векторов в память GPU, а вместо этого воспользуемся нуль-копируемой памятью, чтобы обращаться к данным напрямую из GPU. Программа будет устроена так же, как тест закрепленной памяти. Точнее, мы напишем две функции; одна, эталонная, будет работать со стандартной памятью CPU, а другая – выполнять редукцию на GPU с хранением входных и выходных буферов в нуль-копируемой памяти. Сначала приведем версию для вычисления скалярного произведения в стандартной памяти CPU. Начинаем, как обычно, с создания событий хронометража, выделения памяти для входных и выходных буферов и заполнения входных буферов данными.

```
float malloc_test( int size ) {
    cudaEvent_t start, stop;
    float *a, *b, c, *partial_c;
    float *dev_a, *dev_b, *dev_partial_c;
    float elapsedTime;

    HANDLE_ERROR( cudaEventCreate( &start ) );
    HANDLE_ERROR( cudaEventCreate( &stop ) );

    // выделить память на CPU
    a = (float*)malloc( size*sizeof(float) );
    b = (float*)malloc( size*sizeof(float) );
    partial_c = (float*)malloc( blocksPerGrid*sizeof(float) );

    // выделить память на GPU
    HANDLE_ERROR( cudaMalloc( (void**)&dev_a,
                           size*sizeof(float) ) );
    HANDLE_ERROR( cudaMalloc( (void**)&dev_b,
                           size*sizeof(float) ) );
    HANDLE_ERROR( cudaMalloc( (void**)&dev_partial_c,
                           blocksPerGrid*sizeof(float) ) );

    // заполнить память CPU данными
    for (int i=0; i<size; i++) {
```

```
a[i] = i;
b[i] = i*2;
}
```

Выделив память и заполнив ее данными, мы можем приступать к вычислению. Запускаем таймер, копируем входные буферы в память CPU, запускаем ядро вычисления скалярного произведения и копируем частичные результаты назад в память CPU.

```
HANDLE_ERROR( cudaEventRecord( start, 0 ) );
// скопировать массивы 'a' и 'b' в память GPU
HANDLE_ERROR( cudaMemcpy( dev_a, a, size*sizeof(float),
                        cudaMemcpyHostToDevice ) );
HANDLE_ERROR( cudaMemcpy( dev_b, b, size*sizeof(float),
                        cudaMemcpyHostToDevice ) );

dot<<<blocksPerGrid,threadsPerBlock>>>( size, dev_a, dev_b,
                                             dev_partial_c );

// скопировать массив 'c' с GPU на CPU
HANDLE_ERROR( cudaMemcpy( partial_c, dev_partial_c,
                        blocksPerGrid*sizeof(float),
                        cudaMemcpyDeviceToHost ) );
```

Теперь нужно закончить вычисления на CPU, как это было сделано в главе 5. Но сначала остановим таймер, потому что мы хотим измерять только время, затраченное GPU.

```
HANDLE_ERROR( cudaEventRecord( stop, 0 ) );
HANDLE_ERROR( cudaEventSynchronize( stop ) );
HANDLE_ERROR( cudaEventElapsedTime( &elapsedTime,
                                    start, stop ) );
```

Наконец, суммируем частичные результаты и освобождаем входные и выходные буферы.

```
// завершить вычисления на CPU
c = 0;
for (int i=0; i<blocksPerGrid; i++) {
    c += partial_c[i];
}

HANDLE_ERROR( cudaFree( dev_a ) );
HANDLE_ERROR( cudaFree( dev_b ) );
HANDLE_ERROR( cudaFree( dev_partial_c ) );

// освободить память на CPU
free( a );
```

```
    free( b );
    free( partial_c );

    // уничтожить события
    HANDLE_ERROR( cudaEventDestroy( start ) );
    HANDLE_ERROR( cudaEventDestroy( stop ) );

    printf( "Вычисленное значение: %f\n", c );

    return elapsedTime;
}
```

Версия с применением нуль-копируемой памяти будет очень похожа, если не считать выделения памяти. Как и раньше, начинаем с выделения памяти для входных и выходных буферов и заполнения входных буферов данными:

```
float cuda_host_alloc_test( int size ) {
    cudaEvent_t start, stop;
    float *a, *b, c, *partial_c;
    float *dev_a, *dev_b, *dev_partial_c;
    float elapsedTime;

    HANDLE_ERROR( cudaEventCreate( &start ) );
    HANDLE_ERROR( cudaEventCreate( &stop ) );

    // выделить память на CPU
    HANDLE_ERROR( cudaHostAlloc( (void**)&a,
                                size*sizeof(float),
                                cudaHostAllocWriteCombined | 
                                cudaHostAllocMapped ) );
    HANDLE_ERROR( cudaHostAlloc( (void**)&b,
                                size*sizeof(float),
                                cudaHostAllocWriteCombined | 
                                cudaHostAllocMapped ) );
    HANDLE_ERROR( cudaHostAlloc( (void**)&partial_c,
                                blocksPerGrid*sizeof(float),
                                cudaHostAllocMapped ) );

    // заполнить память CPU данными
    for (int i=0; i<size; i++) {
        a[i] = i;
        b[i] = i*2;
    }
```

Снова, как и в главе 10, мы встречаем функцию `cudaHostAlloc()`, но теперь заданы флаги, определяющие поведение, отличное от умалчиваемого. Флаг `cudaHostAllocMapped` говорит исполняющей среде, что мы собираемся обращаться к этому буферу из GPU. Иными словами, именно этот флаг делает выделенную

для буфера память *нуль-копируемой*. Для обоих входных буферов мы задаем также флаг cudaHostAllocWriteCombined. Он означает, что исполняющая среда должна установить для этого буфера режим объединенной записи (write-combined) относительно кэша CPU. Этот флаг не изменяет функциональность нашего приложения, но существенно повышает производительность при работе с буферами, которые читает только GPU. Однако память, работающая в режиме объединенной записи, может быть крайне неэффективна в случаях, когда ее читает также CPU, поэтому, принимая такое решение, внимательно проанализируйте, как приложение будет обращаться к памяти.

Поскольку память CPU выделена с флагом cudaHostAllocMapped, то GPU может обращаться к находящимся в ней буферам. Однако виртуальные адресные пространства GPU и CPU различаются, поэтому адреса буферов с точки зрения GPU и CPU различны. Функция cudaHostAlloc() возвращает указатель на область памяти в терминах CPU, а чтобы получить указатель на ту же область в терминах GPU, нужно вызывать функцию cudaHostGetDevicePointer(). Эти указатели передаются ядру, и затем GPU использует их для чтения и записи в память CPU.

```
HANDLE_ERROR( cudaHostGetDevicePointer( &dev_a, a, 0 ) );
HANDLE_ERROR( cudaHostGetDevicePointer( &dev_b, b, 0 ) );
HANDLE_ERROR( cudaHostGetDevicePointer( &dev_partial_c,
partial_c, 0 ) );
```

Имея указатели устройства, мы можем запустить таймер и начать выполнение ядра.

```
HANDLE_ERROR( cudaEventRecord( start, 0 ) );

dot<<<blocksPerGrid,threadsPerBlock>>>( size, dev_a, dev_b,
dev_partial_c );

HANDLE_ERROR( cudaThreadSynchronize() );
```

Несмотря на то что указатели dev_a, dev_b и dev_partial_c адресуют память CPU, ядру они представляются указателями на память GPU – благодаря обращениям к cudaHostGetDevicePointer(). Так как частичные результаты уже находятся в памяти CPU, то вызывать функцию cudaMemcpy() нет необходимости. Однако обратите внимание, что CPU и GPU синхронизируются путем обращения к функции cudaThreadSynchronize(). Содержимое нуль-копируемой памяти не определено во время выполнения ядра, которое может это содержимое изменять. После синхронизации есть гарантия, что ядро завершилось и нуль-копируемый буфер содержит результаты вычисления. Поэтому мы можем остановить таймер и довести вычисления до конца на CPU, как делали и прежде.

```
HANDLE_ERROR( cudaEventRecord( stop, 0 ) );
HANDLE_ERROR( cudaEventSynchronize( stop ) );
HANDLE_ERROR( cudaEventElapsedTime( &elapsedTime,
```

```
    start, stop ) );
```

```
// завершить вычисления на CPU
c = 0;
for (int i=0; i<blocksPerGrid; i++) {
    c += partial_c[i];
}
```

Осталось только прибраться за собой:

```
HANDLE_ERROR( cudaFreeHost( a ) );
HANDLE_ERROR( cudaFreeHost( b ) );
HANDLE_ERROR( cudaFreeHost( partial_c ) );

// уничтожить события
HANDLE_ERROR( cudaEventDestroy( start ) );
HANDLE_ERROR( cudaEventDestroy( stop ) );

printf( "Вычисленное значение: %f\n", c );

return elapsedTime;
}
```

Отметим, что независимо от флагов, переданных при вызове `cudaHostAlloc()`, освобождение памяти всегда производится одинаково – обращением к `cudaFreeHost()`.

Вот и все! Теперь посмотрим, как `main()` связывает все это воедино. Прежде всего нужно проверить, поддерживает ли устройство отображение памяти CPU. Делается это так же, как при проверке поддержки перекрытия в предыдущей главе, – путем обращения к `cudaGetDeviceProperties()`.

```
int main( void ) {
    cudaDeviceProp prop;
    int whichDevice;
    HANDLE_ERROR( cudaGetDevice( &whichDevice ) );
    HANDLE_ERROR( cudaGetDeviceProperties( &prop, whichDevice ) );
    if (prop.canMapHostMemory != 1) {
        printf( "Устройство не поддерживает отображение памяти.\n" );
        return 0;
    }
}
```

В предположении, что устройство поддерживает нуль-копирование, приводим исполняющую среду в состояние, в котором она сможет выделять нуль-копируемые буферы. Для этого нужно вызвать функцию `cudaSetDeviceFlags()` с флагом `cudaDeviceMapHost`, который требует разрешить устройству отображение памяти CPU:

```
HANDLE_ERROR( cudaSetDeviceFlags( cudaDeviceMapHost ) );
```

Вот, собственно, и весь код `main()`. Запускаем оба теста, печатаем затраченное время и выходим:

```

float elapsedTime = malloc_test( N );
printf( "Затрачено времени с cudaMalloc: %3.1f ms\n",
       elapsedTime );

elapsedTime = cuda_host_alloc_test( N );
printf( "Затрачено времени с cudaHostAlloc: %3.1f ms\n",
       elapsedTime );
}

```

Само ядро, по сравнению с главой 5, не изменилось, но для полноты приведем его код целиком:

```

#define imin(a,b) (a<b?a:b)

const int N = 33 * 1024 * 1024;
const int threadsPerBlock = 256;
const int blocksPerGrid =
    imin( 32, (N+threadsPerBlock-1) / threadsPerBlock );

__global__ void dot( int size, float *a, float *b, float *c ) {
    __shared__ float cache[threadsPerBlock];
    int tid = threadIdx.x + blockIdx.x * blockDim.x;
    int cacheIndex = threadIdx.x;
    float temp = 0;
    while (tid < size) {
        temp += a[tid] * b[tid];
        tid += blockDim.x * gridDim.x;
    }

    // сохранить значение в кэше
    cache[cacheIndex] = temp;

    // синхронизировать нити в этом блоке
    __syncthreads();

    // для успешной редукции threadsPerBlock должно быть степенью 2
    // из-за следующего кода
    int i = blockDim.x/2;
    while (i != 0) {
        if (cacheIndex < i)
            cache[cacheIndex] += cache[cacheIndex + i];
        __syncthreads();
        i /= 2;
    }
    if (cacheIndex == 0)
        c[blockIdx.x] = cache[0];
}

```

11.2.2. Производительность нуль-копирования

Какого выигрыша ожидать от использования нуль-копируемой памяти? Ответ различен для дискретных и интегрированных GPU. *Дискретным* называется графический процессор с собственным DRAM; обычно он находится не на одной плате с центральным процессором. Например, если вы установили в свой ПК графическую карту, то находящийся на ней GPU дискретный. *Интегрированными* называются графические процессоры, встроенные в системный набор микросхем; обычно они разделяют системную память с CPU. Во многие системы, построенные на базе медиакоммуникационных процессоров (MCP) NVIDIA nForce, уже интегрированы GPU с поддержкой CUDA. Помимо nForce, все нетбуки, ноутбуки и настольные компьютеры на базе новой платформы NVIDIA ION содержат интегрированные GPU с поддержкой CUDA. Для интегрированных GPU использование нуль-копируемой памяти *всегда* дает выигрыш, потому что эта память в любом случае физически разделяется с CPU. Объявление буфера нуль-копируемым просто предотвращает избыточное копирование данных. Но не забывайте, что бесплатных завтраков не бывает, и на нуль-копируемую память распространяются те же ограничения, что на любую закрепленную память: всякое закрепление отъедает кусочек доступной физической памяти, что в конечном итоге приведет к снижению быстродействия системы.

В тех случаях, когда входные и выходные буферы используются ровно один раз, повышение производительности будет наблюдаться и на дискретном GPU. Поскольку GPU проектируются так, чтобы эффективно скрывать задержки, обусловленные доступом к памяти, то этот механизм позволяет в какой-то мере компенсировать чтение и запись по шине PCI Express, что дает заметный выигрыш в производительности. Но так как доступ к нуль-копируемой памяти не кэшируется в GPU, то при многократном чтении из такой памяти мы будем уплачивать большой штраф, которого можно было бы избежать, с самого начала скопировав данные в память GPU.

Как узнать, является GPU интегрированным или дискретным? Можно, конечно, открыть компьютер и посмотреть, но для приложения такое решение вряд ли годится. Впрочем, тип GPU можно определить, опросив одно из полей структуры, которую возвращает функция `cudaGetDeviceProperties()`, а именно поле `integrated`, равное `true`, если GPU интегрирован, и `false` в противном случае.

Поскольку программа вычисления скалярного произведения удовлетворяет условию «читает и/или пишет ровно один раз», то не исключено, что мы сможем порадоваться ускорению от применения нуль-копируемой памяти. И действительно ускорение есть. На GeForce GTX 285 время выполнения уменьшилось на 45%, с 98,1 до 52,1 мс. На GeForce GTX 280 ускорение составило 34%: с 143,9 до 94,7 мс. Понятно, что характеристики производительности GPU разнятся из-за разного соотношения скорости вычислений и пропускной способности шины, а также из-за различий в эффективной пропускной способности PCI Express для разных наборов микросхем.

11.3. Использование нескольких GPU

В предыдущем разделе мы сказали, что графические процессоры бывают интегрированными и дискретными, причем первые являются частью системного набора микросхем, а последние обычно размещаются на плате расширения, которая вставляется в слот шины PCI Express. Все больше систем включают одновременно интегрированные и дискретные GPU, то есть оснащены несколькими процессорами с поддержкой CUDA. Компания NVIDIA продает также продукты, например карту GeForce GTX 295, которые изначально содержат более одного GPU. Так, GeForce GTX 295 физически занимает один слот расширения, но для приложений CUDA выглядит как два разных GPU. Да и сами пользователи могут вставлять дополнительные графические карты в слоты PCI Express, соединяя их мостами с помощью разработанной NVIDIA технологии *масштабируемого интерфейса соединений* (scalable link interface – SLI). Ввиду этих тенденций перестали быть редкостью приложения CUDA, работающие в системе с несколькими графическими процессорами. Поскольку основной целью приложений CUDA является распараллеливание работы, было бы замечательно задействовать все имеющиеся в системе CUDA-устройства для достижения максимальной производительности. Посмотрим, как это можно сделать.

Чтобы не изучать новый пример, давайте переработаем программу вычисления скалярного произведения, так чтобы она могла использовать несколько GPU. А чтобы упростить себе жизнь, соберем все необходимые для вычисления скалярного произведения данные в одну структуру. Почему это упрощает жизнь, вы поймете очень скоро.

```
struct DataStruct {
    int    deviceID;
    int    size;
    float *a;
    float *b;
    float returnValue;
};
```

Мы поместили в эту структуру идентификатор устройства, на котором будет вычисляться скалярное произведение, размер входных векторов и указатели на оба входных вектора, *a* и *b*. В последнем поле хранится величина скалярного произведения *a* и *b*.

Чтобы задействовать *N* GPU, неплохо бы знать, сколько *N*. Поэтому в начале программы мы обращаемся к функции *cudaGetDeviceCount()*, которая сообщает, сколько процессоров с поддержкой CUDA имеется в системе.

```
int main( void ) {
    int deviceCount;
    HANDLE_ERROR( cudaGetDeviceCount( &deviceCount ) );
```

```

if (deviceCount < 2) {
    printf( "Необходимо по меньшей мере два устройства с уровнем "
            "вычислительных возможностей не ниже 1.0, а найдено только %d\n",
deviceCount );
    return 0;
}

```

Поскольку пример предназначен для демонстрации работы с несколькими GPU, то при обнаружении всего одного CUDA-устройства мы просто выходим (хотя это никакая не ошибка). Понятно, что такой подход нельзя назвать рекомендуемым. Простоты ради для хранения входных векторов мы выделяем область в стандартной памяти CPU и заполняем ее данными так же, как раньше.

```

float *a = (float*)malloc( sizeof(float) * N );
HANDLE_NULL( a );
float *b = (float*)malloc( sizeof(float) * N );
HANDLE_NULL( b );

// заполнить память CPU данными
for (int i=0; i<N; i++) {
    a[i] = i;
    b[i] = i*2;
}

```

Вот теперь мы готовы заняться кодом для работы с несколькими GPU. Хитрость в том, что API исполняющей среды CUDA предполагает, что каждый GPU управляется отдельным потоком CPU¹. Так как раньше мы работали с одним GPU, то на этой тонкости не акцентировали внимание. Все сложности, относящиеся к многопоточному программированию, мы вынесли во вспомогательный заголовочный файл book.h. И поскольку от этих деталей мы избавлены, то осталось только поместить в структуру данные, необходимые для вычислений. Хотя в системе может быть любое количество GPU, большее 1, мы для простоты будем работать только с двумя:

```

DataStruct data[2];

data[0].deviceID = 0;
data[0].size = N/2;
data[0].a = a;
data[0].b = b;

data[1].deviceID = 1;
data[1].size = N/2;

```

¹ Если в контексте GPU мы переводили слово *thread* как «нить», то в контексте CPU будем применять более привычную терминологию – «поток» (вычислительный), а программу с несколькими потоками будем называть «многопоточной». – *Прим. перев.*

```
data[1].a = a + N/2;
data[1].b = b + N/2;
```

Далее указатель на переменную типа DataStruct передается вспомогательной функции start_thread() вместе с указателем на функцию, которая будет исполняться во вновь созданном потоке; в этом примере она называется routine(). Функция start_thread() создает поток и вызывает в нем указанную функцию, передавая ей в качестве единственного аргумента структуру DataStruct, полученную от вызывающей программы. Второй раз функция routine() вызывается из главного потока программы (стало быть, мы создали только один *дополнительный* поток).

```
CUTThread thread = start_thread( routine, &(data[0]) );
routine( &(data[1]) );
```

Перед тем как двигаться дальше, главный поток программы должен дождаться завершения второго потока, для чего вызывается функция end_thread().

```
end_thread( thread );
```

Поскольку в этой точке main() оба потока завершились, можно без опаски освободить более не нужную память и вывести результат.

```
free( a );
free( b );

printf( "Вычисленное значение: %f\n",
       data[0].returnValue + data[1].returnValue );

return 0;
}
```

Здесь мы складываем результаты, вычисленные обоими потоками. Это последний шаг редукции, необходимой для вычисления скалярного произведения. В каком-нибудь другом алгоритме для комбинирования нескольких результатов могли бы понадобиться дополнительные шаги. На самом деле иногда бывает, что два GPU выполняют совершенно разную обработку различных наборов данных. Но мы не стали усложнять.

Поскольку сама процедура вычисления скалярного произведения ничем не отличается от приведенной ранее, то мы ее опустили. Но вот код функции routine() представляет определенный интерес. Согласно объявлению, эта функция принимает и возвращает указатель типа void*, чтобы реализация start_thread() не зависела от конкретной функции потока. Конечно, нам хотелось бы получить приз за оригинальную идею, но в действительности это стандартное соглашение о функциях обратного вызова в языке C:

```
void* routine( void *pvoidData ) {
    DataStruct *data = (DataStruct*)pvoidData;
    HANDLE_ERROR( cudaSetDevice( data->deviceID ) );
```

В каждом потоке вызывается функция `cudaSetDevice()`, но передаются ей разные идентификаторы устройства. Поэтому каждый поток будет работать со своим GPU. У этих GPU может быть одинаковая производительность, как в случае карты GeForce GTX 295 с двумя GPU, или различная – если, например, в системе имеется интегрированный и дискретный GPU. Нашему приложению эти детали не важны, хотя вам они могут быть интересны. В частности, эта информация может оказаться полезной, если ядро нуждается в определенном минимальном уровне вычислительных возможностей или если вас обуревает желание равномерно распределить нагрузку на разные GPU, имеющиеся в системе. Если GPU различны, то придется разбить вычисления на части так, чтобы все GPU были заняты примерно одинаковое время. Но сейчас все эти мелкие детали нас не интересуют.

Если не считать обращения к `cudaSetDevice()`, то реализация `routine()` очень похожа на функцию `malloc_test()` из раздела 11.2.1 «Вычисление скалярного произведения с применением нуль-копируемой памяти». Мы выделяем в памяти GPU буферы для хранения копий входных векторов и частичных результатов, а затем с помощью `cudaMemcpy()` копируем входные векторы.

```
int size = data->size;
float *a, *b, c, *partial_c;
float *dev_a, *dev_b, *dev_partial_c;

// выделить память на CPU
a = data->a;
b = data->b;
partial_c = (float*)malloc( blocksPerGrid*sizeof(float) );

// выделить память на GPU
HANDLE_ERROR( cudaMalloc( (void**)&dev_a,
                           size*sizeof(float) ) );
HANDLE_ERROR( cudaMalloc( (void**)&dev_b,
                           size*sizeof(float) ) );
HANDLE_ERROR( cudaMalloc( (void**)&dev_partial_c,
                           blocksPerGrid*sizeof(float) ) );

// скопировать массивы 'a' и 'b' в память GPU
HANDLE_ERROR( cudaMemcpy( dev_a, a, size*sizeof(float),
                           cudaMemcpyHostToDevice ) );
HANDLE_ERROR( cudaMemcpy( dev_b, b, size*sizeof(float),
                           cudaMemcpyHostToDevice ) );
```

Далее запускаем вычислительное ядро, копируем результаты в память CPU и завершаем вычисления на CPU.

```
dot<<<blocksPerGrid,threadsPerBlock>>>( size, dev_a, dev_b,
                                             dev_partial_c );

// скопировать массив 'c' с GPU на CPU
HANDLE_ERROR( cudaMemcpy( partial_c, dev_partial_c,
                         blocksPerGrid*sizeof(float),
                         cudaMemcpyDeviceToHost ) );

// завершить вычисления на CPU
c = 0;
for ( int i=0; i<blocksPerGrid; i++ ) {
    c += partial_c[i];
}
```

Как обычно, освобождаем память GPU и возвращаем вычисленное скалярное произведение в поле `returnValue` структуры `DataStruct`.

```
HANDLE_ERROR( cudaFree( dev_a ) );
HANDLE_ERROR( cudaFree( dev_b ) );
HANDLE_ERROR( cudaFree( dev_partial_c ) );

// освободить память на CPU
free( partial_c );

data->returnValue = c;
return 0;
```

Как видите, если не считать деталей управления потоками, то работа с несколькими GPU не намного сложнее, чем с одним. А с применением нашего вспомогательного кода для создания потока и исполнения в нем функции задача еще упрощается. Если у вас есть собственные библиотеки для поддержки многопоточного программирования, можете пользоваться ими. Только не забывайте, что для каждого GPU нужен отдельный поток, и жизнь покажется медом.

11.4. Переносимая закрепленная память

И последняя важная деталь, относящаяся к работе с несколькими GPU, – использование закрепленной памяти. В главе 10 мы объяснили, что закрепленной называется память CPU, страницы которой заблокированы в физической памяти системы, то есть не могут быть ни выгружены на диск, ни перемещены. Но в действительности страница представляется закрепленной только одному потоку CPU. То есть она будет оставаться заблокированной в памяти, если ее закрепил *любой* поток, но *казалось* закрепленной она будет лишь тому потоку, который ее

выделил. Если указатель на эту область памяти разделяется несколькими потоками, то все остальные будут видеть ее как стандартную страничную память.

У этого поведения есть побочный эффект: если поток, который не выделял закрепленный буфер, попытается выполнить для него `cudaMemcpy()`, то копирование будет выполняться со скоростью, характерной для стандартной страничной памяти. В главе 10 мы видели, что она примерно в два раза меньше максимально достижимой скорости передачи. Хуже того, если вычислительный поток попытается поставить вызов `cudaMemcpyAsync()` в очередь потока CUDA, то произойдет ошибка, так как для выполнения этой операции необходим закрепленный буфер. А поскольку потоку, не выделявшему буфер, он кажется незакрепленным, то попытка с треском проваливается. Даже на будущее нет надежды!

Однако проблему можно решить. Закрепленную память можно выделить как *переносимую* (*portable*); это означает, что ей разрешено мигрировать из одного потока в другой, и каждый поток будет считать буфер закрепленным. Для этого мы воспользуемся нашим старым надежным другом, функцией `cudaHostAlloc()`, но с новым флагом: `cudaHostAllocPortable`. Этот флаг можно задавать вместе с другими, например `cudaHostAllocWriteCombined` и `cudaHostAllocMapped`. Стало быть, буфер в памяти CPU может быть переносимым, нуль-копируемым и с объединением записи – в любом сочетании.

Чтобы продемонстрировать использование переносимой закрепленной памяти, мы немного доработаем программу вычисления скалярного произведения. В основу будет положена версия с нуль-копируемой памятью, так что получится вариант, работающий на нескольких GPU с применением нуль-копирования. Как всюду в этой главе, сначала необходимо убедиться, что имеются по меньшей мере два GPU с поддержкой CUDA и оба могут работать с нуль-копируемыми буферами.

```
int main( void ) {
    int deviceCount;
    HANDLE_ERROR( cudaGetDeviceCount( &deviceCount ) );
    if (deviceCount < 2) {
        printf( "Необходимо по меньшей мере два устройства с уровнем "
                "вычислительных возможностей не ниже 1.0, а найдено только %d\n",
                deviceCount );
        return 0;
    }

    cudaDeviceProp prop;
    for (int i=0; i<2; i++) {
        HANDLE_ERROR( cudaGetDeviceProperties( &prop, i ) );
        if (prop.canMapHostMemory != 1) {
            printf( "Устройство не поддерживает отображение памяти.\n" );
            return 0;
        }
    }
}
```

В предыдущих примерах мы в этом месте приступали к выделению памяти CPU для хранения входных векторов. Но чтобы выделить переносимую закрепленную память, сначала необходимо выбрать CUDA-устройство, на котором мы собираемся работать. Поскольку мы намереваемся воспользоваться еще и нуль-копированием, то вслед за вызовом `cudaSetDevice()` обращаемся к функции `cudaSetDeviceFlags()`, как в разделе 11.2.1.

```
float *a, *b;
HANDLE_ERROR( cudaSetDevice( 0 ) );
HANDLE_ERROR( cudaSetDeviceFlags( cudaDeviceMapHost ) );
HANDLE_ERROR( cudaHostAlloc( (void**)&a, N*sizeof(float),
                           cudaHostAllocWriteCombined |
                           cudaHostAllocPortable |
                           cudaHostAllocMapped ) );
HANDLE_ERROR( cudaHostAlloc( (void**)&b, N*sizeof(float),
                           cudaHostAllocWriteCombined |
                           cudaHostAllocPortable |
                           cudaHostAllocMapped ) );
```

Ранее в этой главе мы уже вызывали `cudaSetDevice()`, но только после выделения памяти и создания потоков. Однако, чтобы выделить закрепленную память с помощью `cudaHostAlloc()`, требуется предварительно инициализировать устройство, обратившись к `cudaSetDevice()`. Также обратите внимание, что новый флаг, `cudaHostAllocPortable`, передается при выделении памяти для обоих буферов. Так как память выделялась после вызова `cudaSetDevice(0)`, то лишь CUDA-устройство с номером 0 считало бы эти буферы закрепленными – если бы мы вдобавок не указали, что они еще и переносимые.

Далее мы делаем то же, что и раньше: генерируем данные для входных векторов и подготавливаем структуры `DataStruct`, как в примере работы с несколькими GPU из раздела 11.3.

```
// заполнить память CPU данными
for (int i=0; i<N; i++) {
    a[i] = i;
    b[i] = i*2;
}

// подготовить все для запуска двух потоков
DataStruct data[2];
data[0].deviceID = 0;
data[0].offset = 0;
data[0].size = N/2;
data[0].a = a;
data[0].b = b;

data[1].deviceID = 1;
data[1].offset = N/2;
```

```
data[1].size = N/2;
data[1].a = a;
data[1].b = b;
```

Затем можно создать дополнительный поток и в обоих потоках вызвать функцию `routine()`, которая будет производить вычисления на том и другом устройстве.

```
CUTThread thread = start_thread( routine, &(data[1]) );
routine( &(data[0]) );
end_thread( thread );
```

Поскольку память CPU выделялась исполняющей средой CUDA, для ее освобождения нужно вызывать функцию `cudaFreeHost()`, а не `free()`. И это все, что есть в функции `main()`.

```
// освободить память CPU
HANDLE_ERROR( cudaFreeHost( a ) );
HANDLE_ERROR( cudaFreeHost( b ) );

printf( "Вычисленное значение: %f\n",
       data[0].returnValue + data[1].returnValue );

return 0;
}
```

Чтобы поддержать работу с переносимой закрепленной и нуль-копируемой памятью в версии программы для нескольких GPU, нужно внести два важных изменения в функцию `routine()`. Первое довольно тонкое, и уж интуитивно очевидным его точно не назовешь.

```
void* routine( void *pvoidData ) {
    DataStruct *data = (DataStruct*)pvoidData;
    if (data->deviceID != 0) {
        HANDLE_ERROR( cudaSetDevice( data->deviceID ) );
        HANDLE_ERROR( cudaSetDeviceFlags( cudaDeviceMapHost ) );
    }
}
```

Напомним, что в версии для нескольких GPU необходимо было вызывать `cudaSetDevice()` в функции `routine()`; это гарантировало, что каждый поток управляет своим GPU. С другой стороны, в этом примере мы уже вызывали `cudaSetDevice()` в главном потоке. Это было нужно для того, чтобы выделить закрепленную память в `main()`. Следовательно, осталось вызвать `cudaSetDevice()` и `cudaSetDeviceFlags()` только для тех устройств, для которых они еще не вызывались. Именно поэтому мы проверяем, `deviceID` не равно нулю. Код выглядел бы проще, если бы мы повторили вызовы для устройства с идентификатором 0, но это было бы ошибкой. После того как устройство один раз сконфигурировано

в некотором потоке, еще раз вызывать в нем функцию `cudaSetDevice()` нельзя, даже если передать ей тот же самый идентификатор устройства. Выделенное оператор `if` как раз и позволяет не сталкиваться с этой неприятной особенностью исполняющей среды CUDA. Теперь перейдем ко второму изменению в функции `routine()`.

Выделенная память CPU не только переносимая и закрепленная, но и нуль-копируемая, потому что мы хотим обращаться к буферам непосредственно из кода, работающего на GPU. Следовательно, вызывать `cudaMemcpy()`, как в первоначальной версии программы для нескольких GPU, уже не нужно, зато надо воспользоваться функцией `cudaHostGetDevicePointer()`, чтобы получить указатели на память CPU, отображенныe на адресное пространство устройства. Однако обратите внимание, что для хранения частичных результатов мы используем стандартную память GPU. Как обычно, эта память выделяется функцией `cudaMalloc()`.

```

int size = data->size;
float *a, *b, c, *partial_c;
float *dev_a, *dev_b, *dev_partial_c;

// выделить память на CPU
a = data->a;
b = data->b;
partial_c = (float*)malloc( blocksPerGrid*sizeof(float) );

HANDLE_ERROR( cudaHostGetDevicePointer( &dev_a, a, 0 ) );
HANDLE_ERROR( cudaHostGetDevicePointer( &dev_b, b, 0 ) );
HANDLE_ERROR( cudaMalloc( (void**)&dev_partial_c,
                         blocksPerGrid*sizeof(float) ) );

// сместиться в те части массивов 'a' и 'b', откуда данный
// GPU берет данные для обработки
dev_a += data->offset;
dev_b += data->offset;

```

Вот теперь все готово, можно запустить ядро, а затем скопировать результаты в память CPU.

```

dot<<<blocksPerGrid,threadsPerBlock>>>( size, dev_a, dev_b,
                                             dev_partial_c );

// скопировать массив 'c' из памяти GPU в память CPU
HANDLE_ERROR( cudaMemcpy( partial_c, dev_partial_c,
                         blocksPerGrid*sizeof(float),
                         cudaMemcpyDeviceToHost ) );

```

Как и во всех версиях скалярного произведения, функция завершается сложением частичных результатов на CPU, освобождением временной памяти и возвратом в `main()`.

```
// завершить вычисления на CPU
c = 0;
for (int i=0; i<blocksPerGrid; i++) {
    c += partial_c[i];
}

HANDLE_ERROR( cudaFree( dev_partial_c ) );

// освободить память CPU
free( partial_c );

data->returnValue = c;
return 0;
}
```

11.5. Резюме

Мы рассмотрели новые способы выделения памяти CPU, причем все они реализуются одним обращением к функции `cudaHostAlloc()`. С помощью различных комбинаций флагов можно сделать выделенную память нуль-копируемой, переносимой или с объединенной записью – в любом сочетании. Нуль-копируемая память применяется для того, чтобы избежать явного копирования в память GPU и обратно; этот прием может повысить быстродействие широкого класса приложений. С помощью библиотеки поддержки многопоточного программирования мы научились работать с несколькими GPU в одном приложении, что позволило распределить вычисление скалярного произведения между несколькими устройствами. Наконец, мы видели, что несколько GPU могут разделять одну область закрепленной памяти, если она выделена как *переносимая*. В последнем примере использовались переносимая закрепленная память, несколько GPU и нуль-копируемые буферы; в результате получилась «разогнанная» версия программы вычисления скалярного произведения, с которой мы начали экспериментировать в главе 5. Системы с несколькими графическими процессорами становятся все более популярными, и овладение этими приемами позволит вам в полной мере задействовать вычислительную мощь имеющейся платформы.



Глава 12. Последние штрихи

Поздравляем! Надеемся, что вам понравился язык CUDA C и вы получили удовольствие от экспериментов с GPU-вычислениями. Путешествие было длинным, поэтому давайте присядем и вспомним, с чего мы начинали и куда пришли. Начав с краткого обзора программирования на C и C++, мы научились использовать синтаксическую конструкцию с угловыми скобками для запуска многочисленных экземпляров ядра на произвольном числе мультипроцессоров. Далее мы перешли к блокам и нитям, способным обрабатывать входные данные сколь угодно большого объема. При этом мы применяли механизмы межнитевой коммуникации для работы со специальной находящейся на кристалле разделяемой памятью и пользовались примитивами синхронизации, которые гарантируют корректное поведение в окружении, поддерживающем тысячи параллельных нитей (причем такое количество считается нормой).

Овладев основными идеями параллельного программирования на CUDA C в архитектуре NVIDIA CUDA, мы перешли к изучению более сложных концепций и API. В контексте GPU-вычислений весьма полезно специализированное графическое оборудование GPU, поэтому мы рассмотрели применение текстурной памяти для ускорения некоторых приложений, характеризующихся определенным способом доступа к памяти. Поскольку многие пользователи добавляют GPU-вычисления в графические приложения, мы исследовали интероперабельность ядер на CUDA C со стандартными графическими библиотеками OpenGL и DirectX. Атомарные операции с глобальной и разделяемой памятью обеспечивают безопасный многонитевой доступ к совместно используемым адресам памяти. Продолжая движение от простого к сложному, мы перешли к потокам, которые позволяют максимально полно загрузить систему за счет одновременного выполнения операций копирования памяти и исполнения ядер. И наконец, мы узнали, как выделять и использовать нуль-копируемую память для ускорения приложений, работающих на интегрированных GPU. Мы завершили путь рассказом о том, как инициализировать несколько устройств и выделять переносимую закрепленную память, что позволяет писать на CUDA C программы, которые в полной мере используют возможности набирающих популярность систем с несколькими GPU.

12.1. О чём эта глава

Прочитав эту главу, вы:

- узнаете о некоторых инструментах, полезных при разработке на CUDA C;
- узнаете о дополнительных ресурсах (руководствах и программах), которые позволяют подняться на новый уровень владения CUDA C.

12.2. Инструментальные средства CUDA

В этой книге мы опирались на несколько компонентов системы программирования CUDA C. Для преобразования написанных нами приложений в код, используемый на NVIDIA GPU, мы активно пользовались компилятором CUDA C. За конфигурирование устройств, запуск ядер и взаимодействие с GPU отвечала исполняющая среда CUDA. А среда CUDA, в свою очередь, обращается к драйверу CUDA для работы непосредственно с аппаратными подсистемами. Помимо этих компонентов, с которыми мы уже хорошо знакомы, компания NVIDIA предоставляет немало другого ПО, упрощающего разработку на CUDA C. Настоящий раздел не следует рассматривать как руководство пользователя по этим продуктам; мы лишь хотим проинформировать вас об их существовании и потенциальной полезности.

12.2.1. CUDA Toolkit

Почти наверняка пакет программных средств CUDA Toolkit уже установлен на машину, где вы ведете разработку. Мы так уверены в этом, потому что одним из основных компонентов этого пакета является компилятор CUDA C. Если CUDA Toolkit на вашей машине нет, значит, можно с большой долей уверенности утверждать, что вы ни разу не пытались откомпилировать код CUDA C. Вот мы тебя и вывели на чистую воду, обманщик! Вообще-то ничего страшного в этом нет (хотя мы не вполне понимаем, зачем вы тогда прочли всю эту книгу). С другой стороны, если вы проработали все примеры, приведенные в этой книге, то располагаете и библиотеками, о которых пойдет речь ниже.

12.2.2. Библиотека CUFFT

В состав CUDA Toolkit входят две¹ очень важные библиотеки служебных функций, которые могут понадобиться при включении GPU-вычислений в собственные приложения. Во-первых, NVIDIA предоставляет оптимизированную библиотеку для вычисления быстрого преобразования Фурье, *CUFFT*. В версии 3.0 CUFFT поддерживает целый ряд полезных возможностей, в том числе:

- одно-, дву- и трехмерные преобразования вещественных и комплексных входных данных;
- пакетный запуск для параллельного выполнения большого числа одномерных преобразований;
- двумерные и трехмерные преобразования с размером массива от 2 до 16 384 по любому измерению;
- одномерные преобразования набора, содержащего до 8 миллионов элементов;

¹ В последнюю версию CUDA входят еще несколько библиотек.

- ❑ преобразования с размещением результата в той же или в другой области памяти для вещественных и комплексных данных.

NVIDIA поставляет библиотеку CUFFT бесплатно по лицензии, разрешающей ее использование в любом приложении – в личных, коммерческих или академических целях.

12.2.3. Библиотека CUBLAS

Помимо библиотеки CUFFT, NVIDIA предлагает библиотеку функций для линейной алгебры, в которой реализован широко известный пакет основных подпрограмм для линейной алгебры (Basic Linear Algebra Subprograms – BLAS). Эта библиотека, *CUBLAS*, также бесплатна и поддерживает большое подмножество полного пакета BLAS. Каждая функция представлена в нескольких вариантах: для входных данных одинарной и двойной точности, вещественных и комплексных. Поскольку пакет BLAS изначально был написан на языке FORTRAN, NVIDIA приложила все усилия к тому, чтобы обеспечить максимальную совместимость с ожиданиями пользователей. Соответственно, массивы в CUBLAS хранятся по столбцам, а не строкам, как принято в С и С++. На практике это обычно не очень существенно, но позволяет пользователям стандартного пакета BLAS с минимальными усилиями перевести свои программы на библиотеку CUBLAS, ускоренную за счет GPU. NVIDIA распространяет также интерфейс к CUBLAS из программ на FORTRAN, чтобы продемонстрировать, как можно скомпоновать существующие FORTRAN-приложения с библиотеками CUDA.

12.2.4. Комплект NVIDIA GPU Computing SDK

Отдельно от драйверов NVIDIA и пакета CUDA Toolkit поставляется *GPU Computing SDK*, который содержит многие десятки примеров приложений с GPU-вычислениями. Мы уже упоминали этот SDK выше, потому что он прекрасно дополняет материал первых 11 глав. На случай, если вы пока не удосужились в него заглянуть, скажем, что NVIDIA предлагает примеры для пользователей с различным уровнем владения CUDA C, разбив их на несколько категорий, а именно:

- ❑ Основы CUDA
- ❑ Дополнительные средства CUDA
- ❑ Интеграция CUDA с системой
- ❑ Алгоритмы распараллеливания по данным
- ❑ Интероперабельность с графикой
- ❑ Текстуры
- ❑ Стратегии оптимизации производительности
- ❑ Линейная алгебра
- ❑ Обработка изображений и видео
- ❑ Финансовый инжиниринг
- ❑ Сжатие данных
- ❑ Моделирование физических процессов

Все примеры работают на любой платформе, где работает CUDA C, и могут послужить отличной отправной точкой для ваших собственных приложений. Читателей, хорошо знакомых с той или иной предметной областью, заранее предупреждаем: в NVIDIA GPU Computing SDK не стоит искать передовые реализации своих любимых алгоритмов. Это не код промышленного качества, а учебные материалы на тему функционирования программ, написанных на CUDA C. В этом отношении они несильно отличаются от примеров в этой книге.

12.2.5. Библиотека NVIDIA Performance Primitives

Помимо библиотек CUFFT и CUBLAS, NVIDIA поддерживает также библиотеку функций на CUDA для ускоренной обработки данных, которая называется NVIDIA Performance Primitives (NPP). В настоящее время NPP ориентирована главным образом на обработку изображений и видео и широко применяется специалистами, работающими в этих областях. В планах NVIDIA стоит расширение NPP на широкий спектр вычислительных задач в различных областях. Если вас интересует высокопроизводительная обработка изображений и видео, то обязательно ознакомьтесь с библиотекой NPP, которую можно бесплатно скачать со страницы по адресу www.nvidia.com/object/npp.html (или найти с помощью вашего любимого поисковика).

12.2.6. Отладка программ на языке CUDA C

Нам доводилось слышать из разных источников, что изредка программы при первом запуске работают не так, как ожидалось. Одни неправильно считают, другие не завершаются, а третьи даже вводят компьютер в ступор, из которого его можно вывести, только нажав на кнопку выключения питания. Хотя авторы данной книги сами *никогда* не писали такой код, они все же понимают, что некоторым программистам необходимы средства для отладки ядер на CUDA C. К счастью, NVIDIA предлагает инструменты, делающие этот процесс менее болезненным.

CUDA-GDB

Программа *CUDA-GDB* – один из самых полезных инструментов для программистов на CUDA C, работающих в системах на базе Linux. NVIDIA дополнила отладчик GNU (*gdb*) с открытыми исходными текстами, реализовав возможность прозрачно отлаживать исполняемый устройством код в режиме реального времени, не порывая с привычным интерфейсом *gdb*. До появления CUDA-GDB не существовало способа отлаживать исполняемый устройством код, помимо эмуляции его работы на CPU. Но это было крайне медленно и зачастую давало лишь очень грубое приближение к реальному выполнению ядра графическим процессором. NVIDIA CUDA-GDB позволяет отлаживать код непосредственно на GPU, предоставляя в распоряжение программиста все средства, знакомые им по работе с отладчиками, ориентированными на CPU. Перечислим некоторые возможности CUDA-GDB.

- Просмотр состояния CUDA, в частности информации об установленных GPU и их вычислительных возможностях.
- Установка контрольных точек в исходном коде на CUDA C.
- Просмотр памяти GPU – как глобальной, так и разделяемой.
- Перечисление блоков и нитей, исполняемых на GPU.
- Пошаговое выполнение варпа нитей.
- Подключение к работающему приложению, в том числе зависшему или находящемуся в состоянии взаимоблокировки.

Вместе с отладчиком NVIDIA поставляет средство CUDA Memory Checker, к которому можно обращаться из CUDA-GDB или из автономно запускаемой программы `cuda-memcheck`. Поскольку в состав архитектуры CUDA входит устройство управления памятью, встроенное непосредственно в аппаратуру, то любое недопустимое обращение к памяти обнаруживается и предотвращается аппаратно. Попытка нарушения защиты памяти означает, что программа работает неправильно, поэтому крайне желательно видеть такие ошибки. CUDA Memory Checker обнаруживает все попытки нарушения защиты глобальной памяти и доступа к невыровненному адресу глобальной памяти со стороны ядра и выдает о них гораздо более подробные и информативные сообщения, чем было возможно раньше.

NVIDIA Parallel Nsight

Хотя CUDA-GDB – зрелый и весьма мощный инструмент для отладки написанных на CUDA C ядер в реальном времени, NVIDIA понимает, что не все в мире программисты работают в Linux. Пользователю Windows тоже нужно отлаживаться, если, конечно, он не подстраховался, отложив деньжат на открытие собственного магазинчика «Товары для животных». В конце 2009 года NVIDIA выпустила NVIDIA Parallel Nsight (предварительное кодовое название Nexus), первый интегрированный отладчик GPU/CPU для Microsoft Visual Studio. Как и CUDA-GDB, Parallel Nsight поддерживает отладку CUDA-приложений с тысячами нитей. В любом месте исходного кода на CUDA C можно ставить контрольные точки, в том числе срабатывающие при записи по указанному адресу памяти. Есть возможность просматривать память GPU прямо в окне Memdump и обнаруживать выход за границы выделенной памяти. Во время работы над этой книгой продукт был представлен публике в виде бета-версии, но скоро должна выйти окончательная версия.

12.2.7. CUDA Visual Profiler

Мы не раз расхваливали архитектуру CUDA как замечательный фундамент для высокопроизводительных вычислений. К сожалению, реальность такова, что после выкорчевывания всех ошибок оказывается, что приложение, которое, по искренней задумке автора, должно было осуществлять «высокопроизводительные вычисления», осуществляет просто «вычисления». Как же часто мы в недоумении восклицаем: «Ну почему, черт побери, она работает так медленно?!» В такой си-

туации очень полезно запустить ядро под пристальным наблюдением профилировщика. NVIDIA предлагает и такой инструмент, его можно отдельно скачать в разделе сайта CUDA Zone. На рис. 12.1 показано, как Visual Profiler используется для сравнения двух реализаций транспонирования матрицы. Даже не заглядывая в код, вы сразу видите, что и по доступу к памяти, и по числу выполненных команд ядро `transpose()` заметно превосходит ядро `transpose_naive()`. (Впрочем, было бы нечестно ожидать слишком много от функции, в имени которой встречается слово *naive*.)

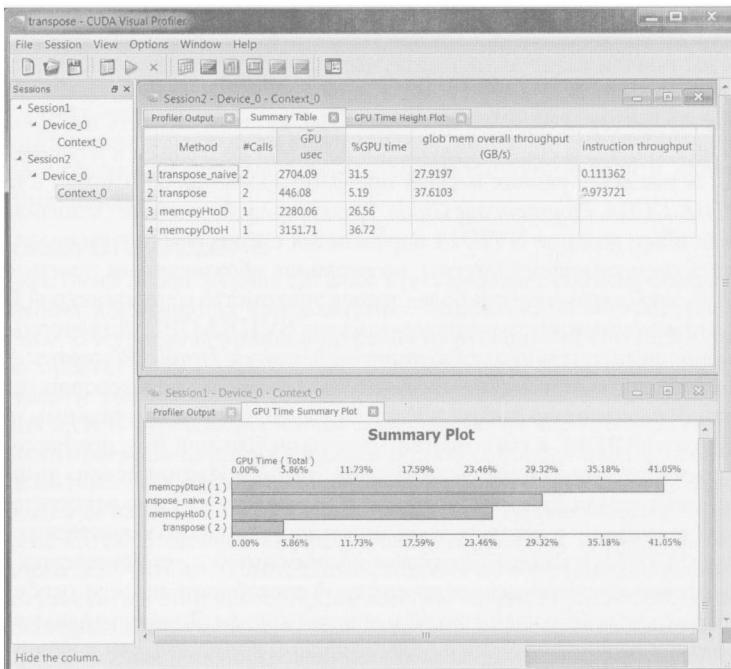


Рис. 12.1. Применение CUDA Visual Profiler для профилирования транспонирования матрицы

Профилировщик CUDA Visual Profiler исполняет приложение, анализируя специальные счетчики производительности, встроенные в GPU. По завершении программы профилировщик обрабатывает собранные данные и представляет их в виде отчетов. Он знает, сколько времени было затрачено на выполнение каждого ядра, сколько было запущено блоков, объединялись ли операции доступа к памяти со стороны ядра, сколько расходящихся ветвей было выполнено в варпах и т. д. Если вы столкнетесь с какой-нибудь неочевидной проблемой, касающейся производительности, рекомендуем попробовать CUDA Visual Profiler.

12.3. Текстовые ресурсы

Если вас еще не тошнит от этой книги, то, быть может, вы захотите почитать еще что-нибудь. Мы знаем, что кое-кто предпочел бы продолжить обучение, экспериментируя с кодом, ну а для остальных упомянем о дополнительных источниках информации, которые помогут стать более квалифицированным программистом на CUDA C.

12.3.1. Programming Massively Parallel Processors: A Hands-On Approach

Если вы прочли главу 1, то знаете, что эта книга не задумывалась как учебник по параллельным архитектурам. Конечно, мы упоминали такие термины, как *мультипроцессор* и *варп*, но основная цель книги – научить вас программированию на CUDA C с соответствующим API. Мы подходили к изучению языка CUDA C, оставаясь в рамках модели программирования, описанной в руководстве *NVIDIA CUDA Programming Guide*, и по большей части не останавливались на том, как оборудование NVIDIA справляется с предложенными задачами. Но чтобы стать по-настоящему крутым, всесторонне образованным программистом на CUDA C, необходимо свести более тесное знакомство с архитектурой CUDA и некоторыми нюансами внутренних механизмов NVIDIA GPU. Для этого мы рекомендуем проштудировать книгу *Programming Massively Parallel Processors: A Hands-on Approach* (Программирование массивно-параллельных процессоров: практический подход), написанную Дэвидом Кирком, ранее работавшим главным научным сотрудником в NVIDIA, в соавторстве с доктором Вен-мэй Хву, профессором факультета электротехники и вычислительной техники Иллинойского университета, получателем гранта Джерри Сандерса III. Здесь вы не только встретитесь с уже знакомыми терминами и концепциями, но и многое узнаете о технических деталях архитектуры CUDA, в том числе о планировании нитей и устойчивости к задержкам, эффективном использовании пропускной способности памяти, особенностях обработки чисел с плавающей точкой и многом другом. Тема параллельного программирования рассматривается в более широком контексте, чем в данной книге, поэтому вы станете лучше понимать, как строятся параллельные решения больших и сложных задач.

12.3.2. CUDA U

Как это ни печально, но кое-кто из нас окончил университет до того, как появился удивительный мир GPU-вычислений. Ну а для тех, у кого эта счастливая пора еще впереди, скажем, что сейчас курсы по CUDA читаются в 300 университетах по всему миру. Но не торопитесь садиться на жесткую диету, чтобы втиснуть свои телеса в тесную студенческую форму, есть другой выход! На сайте CUDA Zone имеется раздел *CUDA U* – это, по сути дела, онлайновый университет, в котором преподают технологию CUDA. А можете сразу перейти по ссылке www.nvidia.com.

com/object/cuda_education. Хотя на онлайновых лекциях в CUDA U вы узнаете немало полезного о программировании GPU, но вот онлайновые вечеринки после занятий до сих пор еще не организованы.

Материалы университетских курсов

Среди множества источников знаний о CUDA один из самых примечательных – полный курс программирования на CUDA C, читаемый в Иллинойском университете. Компания NVIDIA совместно с Иллинойским университетом бесплатно опубликовала его в сети в формате M4V для iPod, iPhone и совместимых видеоплееров. Знаем, о чем вы думаете: «Ну наконец-то я смогу изучать CUDA, стоя в очереди в автоинспекцию!» И, наверное, недоумеваете, почему мы упомянули о существовании видеоверсии этой книги только в самом конце. Извините, что так долго тянули, но ведь фильм даже отдаленно не сравнится с книгой, правда? Помимо самих материалов курсов Иллинойского университета и Калифорнийского университета в Дэвисе, вы найдете также материалы из учебных подкастов по CUDA и ссылки на учебные курсы и консультационные услуги, оказываемые сторонними фирмами.

Журнал DR. DOBB'S

Вот уже более 30 лет журнал *Dr. Dobb's* рассказывает обо всех сколько-нибудь существенных достижениях компьютерных технологий, и NVIDIA CUDA – не исключение. В журнале опубликована серия из нескольких статей, прорубивших широкую просеку в зарослях CUDA. Эта серия под общим названием *CUDA, Supercomputing for the Masses* (CUDA, суперкомпьютинг в массы) начинается с введения в GPU-вычисления и быстро переходит от первого ядра к другим частям модели программирования в CUDA. В статьях из журнала *Dr. Dobb's* рассмотрены такие темы, как обработка ошибок, производительность глобальной памяти, разделяемая память, CUDA Visual Profiler, текстурная память, CUDA-GDB, библиотека CUDPP базовых параллельных операций, а также многие другие. Здесь можно получить дополнительные сведения по тому материалу, который мы пытались осветить в этой книге. Кроме того, вы найдете практическую информацию о некоторых инструментах, которые мы смогли упомянуть лишь мимоходом, например о средствах профилирования и отладки. На эту серию статей есть ссылка со страницы CUDA Zone, но при желании можете выйти на нее, задав в поисковике запрос *Dr Dobbs CUDA*.

12.3.3. Форумы NVIDIA

Даже вдоволь порывшись в документации NVIDIA, вы можете не найти ответа на свой вопрос. Быть может, вам интересно, не сталкивался ли еще кто-то со странным поведением, которое вы наблюдаете. Или вы собираетесь устроить посвященную CUDA вечеринку и ищете единомышленников. В любом случае настоятельно рекомендуем посетить форумы на сайте NVIDIA по адресу <http://forums.nvidia.com>. Здесь вы можете задать вопросы другим пользователям CUDA. Скажем больше, прочитав эту книгу, вы и сами сможете помочь другим, если за-

хотите! На эти форумы регулярно заходят сотрудники NVIDIA, так что на самые зажигательные вопросы вы быстро получите авторитетный ответ прямо из источника технологии. Мы также приветствуем предложения по новым функциям и отзывы о том, что мы в NVIDIA делаем хорошо, плохо и вообще никуда не годно.

12.4. Программные ресурсы

Хотя комплект NVIDIA GPU Computing SDK – истинная сокровищница примеров, он ограничивается лишь педагогическими целями. Если вас интересуют библиотеки или исходный код на базе CUDA промышленного качества, то придется продолжить поиски. К счастью, существует обширное сообщество разработчиков для CUDA, создавшее продукты высочайшего качества. Два из них описаны ниже, но призываю вас поискать в сети то, что вам нужно. Да, кстати, может, когда-нибудь вы и сами внесете вклад в копилку сообщества пользователей CUDA C!

12.4.1. Библиотека *CUDA Data Parallel Primitives Library*

Компания NVIDIA совместно с исследователями из Калифорнийского университета в Дэвисе выпустила библиотеку CUDA Data Parallel Primitives Library (CUDPP), содержащую базовые примитивы параллельной обработки данных. В их число входят алгоритмы параллельного префиксного суммирования (*scan*), параллельной сортировки и параллельной редукции. Подобные примитивы составляют основу для широкого спектра алгоритмов параллельной обработки данных, включая сортировку, сжатие потока, построение структур данных и др. Если вы собираетесь написать даже умеренно сложный алгоритм, велики шансы, что в библиотеке CUDPP уже есть в точности то, что вам нужно, или очень близкое к тому. Скачайте ее со страницы по адресу <http://code.google.com/p/cudpp>.

12.4.2. *CULatools*

В разделе 12.2.3 «Библиотека CUBLAS» мы отметили, что NVIDIA предлагает реализацию пакета BLAS в составе CUDA Toolkit. Читателям, интересующимся более полным решением для линейной алгебры, мы рекомендуем ознакомиться с разработанной компанией EM Photonics реализацией промышленного стандарта Linear Algebra Package (LAPACK) на основе CUDA. Эта реализация называется *CULatools* и содержит более сложные функции, написанные на базе библиотеки NVIDIA CUBLAS. Бесплатный пакет Basic содержит алгоритмы LU-декомпозиции, QR-факторизации, решения систем линейных уравнений, сингулярного разложения, а также решатель по методу наименьших квадратов с ограничениями и без. Скачать пакет Basic можно по адресу www.culatools.com/versions/basic. Кроме того, EM Photonics предлагает премиальную и коммерческую лицензии, открывающие доступ к гораздо большему подмножеству LAPACK

и разрешающие распространять собственные коммерческие приложения на основе библиотеки CULAtools.

12.4.3. Интерфейсы к другим языкам

Все программы в этой книге написаны на языках С и С++, но понятно, что существуют сотни проектов на других языках. К счастью, сторонние разработчики написали обертки, позволяющие обращаться к технологии CUDA из языков, которые NVIDIA официально не поддерживает. Сама компания NVIDIA предоставляет привязки к языку FORTRAN для своей библиотеки CUBLAS, но по адресу www.jcuda.org можно найти привязки к Java для нескольких библиотек на основе CUDA. Проект PyCUDA по адресу <http://mathematician.de/software/pycuda> содержит обертки, позволяющие использовать написанные на CUDA С ядра в программах на языке Python. Наконец, в проекте CUDA.NET по адресу www.hoopoe-cloud.com/Solutions/CUDA.NET вы найдете привязки к среде Microsoft .NET.

Хотя все эти проекты официально не поддерживаются NVIDIA, существуют их версии для нескольких версий CUDA, все они бесплатны, и у многих есть вполне успешные пользователи. В общем, мы хотим сказать, что если вашим любимым языком (или любимым языком вашего шефа) является не С или С++, не стоит сразу же отказываться от GPU-вычислений, сначала поищите, нет ли в сети нужных вам языковых привязок.

12.5. Резюме

Вот мы и добрались до конца. На этих 11 главах, посвященных CUDA С, жизнь не заканчивается; существует еще множество ресурсов, которые можно прочитать, скачать, откомпилировать и запустить. Сейчас самое подходящее время для изучения GPU-вычислений, поскольку мы на пороге эры гетерогенных вычислительных платформ. Надеемся, что вы получили удовольствие от изучения одной из самых распространенных сред для параллельного программирования. И еще мы надеемся, что вы прониклись увлекательными возможностями, открывающимися для разработки новых удивительных средств работы с компьютером и обработки все возрастающих объемов информации. Именно ваши идеи и созданные вами необыкновенные технологии выведут GPU-вычисления на новый уровень.



Приложение А. Еще об атомарных операциях

В главе 9 мы рассмотрели несколько способов использования атомарных операций, позволяющих сотням нитей одновременно и безопасно обновлять разделяемые данные. В этом приложении мы поговорим еще об одном применении атомарных операций для реализации структур данных, обеспечивающих блокировку. На первый взгляд, эта тема не выглядит сложнее всего, что мы изучали до сих пор. И в общем-то так оно и есть. Вы изучили в этой книге немало сложных тем, и блокировка ничуть не сложнее. Так почему же мы упратяли этот материал в приложение? Мы не хотим открывать наши тайны сразу, поэтому, если вы заинтригованы, читайте дальше, и все постепенно выяснится.

А.1. И снова скалярное произведение

В главе 5 мы рассматривали задачу о вычислении скалярного произведения векторов с применением CUDA C. Эта задача относится к семейству алгоритмов, известных под общим названием *редукция*. Напомним, что алгоритм устроен следующим образом.

1. Каждая нить в каждом блоке перемножает два соответственных элемента входных векторов и сохраняет результат в разделяемой памяти.
2. В блоке оказывается несколько произведений, но затем запускаются нити, каждая из которых складывает два произведения и записывает результат назад в разделяемую память. На каждом шаге количество слагаемых уменьшается вдвое (отсюда и название *редукция*, то есть уменьшение!).
3. Когда в каждом блоке остается одно значение – частичная сумма, блок записывает его в глобальную память и завершает работу.
4. Если ядро запустило N параллельных блоков, то CPU суммирует N получившихся значений, в результате чего получается окончательное скалярное произведение.

Это всего лишь высокоуровневый обзор алгоритма вычисления скалярного произведения, поэтому если вы успели подзабыть детали или пропустили пару стаканчиков Шардонне, то имеет смысл вернуться к главе 5. Если же вы чувствуете себя достаточно уверенно, то обратите внимание на шаг 4. Хотя в память CPU не копируется большой объем данных и от CPU не требуется

производить сложные вычисления, все равно завершение вычислений на CPU представляется нелепостью.

Но дело не только в нелепом шаге алгоритма или в неэлегантности решения в целом. Рассмотрим случай, когда вычисление скалярного произведения – лишь один этап длинной цепочки операций. Если вы хотите, чтобы *все* операции производились на GPU, поскольку CPU занят другими задачами, то считайте, что вам не повезло. В том виде, в каком код написан сейчас, вам придется остановить вычисления на GPU, скопировать промежуточные результаты в память CPU, закончить вычисление на CPU, а затем вернуть результат на GPU и запустить следующее ядро. Поскольку это приложение посвящено атомарным операциям и мы так долго объясняли, в чем недостатки исходного алгоритма, вы, наверное, уже догадались, к чему мы клоним. Мы намереваемся с помощью атомарных операций исправить алгоритм скалярного произведения, так чтобы вычисления производились только на GPU, освободив CPU для других задач. В идеале, вместо того чтобы завершать ядро на шаге 3 и возвращаться на CPU на шаге 4, мы хотели бы, чтобы каждый блок прибавлял свой конечный результат к итогу, хранящемуся в глобальной памяти. Если бы сумма вычислялась атомарно, то можно было бы не беспокоиться о возможных коллизиях и неопределенных результатах. Поскольку мы уже пользовались операцией `atomicAdd()` в задаче о вычислении гистограммы, кажется естественным применить ее и сейчас.

К сожалению, если уровень вычислительных возможностей GPU ниже 2.0, то операция `atomicAdd()` применима только к целым числам. Конечно, если исходные векторы содержат только целочисленные компоненты, то все нормально, но гораздо чаще компонентами являются числа с плавающей точкой. Увы, большинство процессоров NVIDIA не поддерживают атомарных операций над числами с плавающей точкой¹! Тому есть вполне разумное объяснение, поэтому не торопитесь выбрасывать свой GPU в мусорный ящик!

Атомарность операции над значением в памяти гарантирует лишь, что пока последовательность чтение–модификация–запись в одной нити не завершится полностью, никакая другая нить не сможет ни прочитать, ни изменить модифицируемое значение. Не дается никаких гарантий относительно порядка выполнения операций нитями. Если сложение производят три нити, то иногда будет вычисляться сумма $(A+B)+C$, а иногда $A+(B+C)$. Для целых чисел это приемлемо, потому что целочисленное сложение ассоциативно, то есть $(A+B)+C = A+(B+C)$. Но сложение чисел с плавающей точкой *не* ассоциативно из-за округления промежуточных результатов, поэтому $(A+B)+C$ часто не равно $A+(B+C)$. А в результате атомарные арифметические операции над числами с плавающей точкой – венец сомнительная, так как при наличии очень большого количества нитей, как в случае GPU, получаются недетерминированные результаты. Существует много приложений, в которых получение разных результатов при двух прогонах совершенно

¹ На момент публикации книги на русском языке это уже неверно, книга была написана до выхода Fermi.

недопустимо. Поэтому поддержка атомарной арифметики для чисел с плавающей точкой не была приоритетной задачей для первых версий оборудования.

Однако если вы готовы смириться с толикой недетерминизма, то редукцию все же можно выполнить целиком на GPU. Но сначала нужно придумать, как обойти отсутствие атомарной арифметики с плавающей точкой. Решение состоит в том, чтобы все-таки воспользоваться атомарными операциями, но не для выполнения арифметических действий.

A.1.1. Атомарные блокировки

Функция `atomicAdd()`, которой мы пользовались при построении гистограмм на GPU, гарантирует, что последовательность чтение–модификация–запись не будет прерываться другими нитями. Если мыслить низкоуровневыми категориями, то можно представить, что оборудование блокирует ячейку памяти, к которой применяется операция, а пока ячейка блокирована, никакая другая нить не может ни прочитать, ни изменить ее. Если бы существовал способ имитировать такую блокировку в ядре, то мы могли бы производить произвольные операции над соответствующей ячейкой памяти или структурой данных. Сам механизм блокировки работал бы в точности как типичный *мьютекс* (*mutex*) на CPU. Если вы незнакомы с понятием взаимного исключения (это и есть мьютекс), не расстраивайтесь. Оно не сложнее того, что вы уже изучили.

Основная идея заключается в том, чтобы выделить небольшую область памяти, которая будет использоваться как *мьютекс*. Он будет играть роль семафора, управляющего доступом к некоторому ресурсу. Сам ресурс может быть структурой данных, буфером или просто адресом в памяти, который мы хотим модифицировать атомарно. Прочитав из мьютекса значение 0, нить интерпретирует его как «зеленый свет», означающий, что никакая другая нить сейчас не использует защищаемую им память. Поэтому нить может заблокировать эту память и делать с ней все, что угодно, не опасаясь вмешательства со стороны других нитей. Чтобы заблокировать участок памяти, нить записывает в мьютекс значение 1, которое играет роль «красного света» для других, потенциально конкурирующих нитей. Теперь конкурирующие нити должны ждать, пока владелец мьютекса не запишет в него 0, после чего смогут обратиться к ранее заблокированной памяти.

Простой код, реализующий блокировку, мог бы выглядеть так:

```
void lock( void ) {
    if( *mutex == 0 ) {
        *mutex = 1; // записать в мьютекс 1
    }
}
```

К сожалению, этот код не годится. Однако возникающая проблема нам уже хорошо знакома: что произойдет, если другая нить запишет 1 в мьютекс после того, как наша нить прочитала из него значение 0? Получается, что обе нити проверили мьютекс и увидели, что он равен нулю. Затем обе записывают в него 1, сообщая

другим нитям, что структура заблокирована и недоступна для модификации. После этого обе думают, что владеют памятью или структурой данных и приступают к небезопасной модификации. Катастрофа неминуема!

Интересующая нас операция довольно проста: сравнить значение мьютекса с 0 и записать в него 1 тогда и только тогда, когда мьютекс равен 0. Чтобы все было корректно, эта операция должна выполняться атомарно, то есть никакая другая нить не должна вмешиваться, пока мы не закончим проверять и изменять значение мьютекса. В языке CUDA С эту операцию можно выполнить с помощью функции `atomicCAS()` – атомарно сравнить и обменять. Она принимает указатель на ячейку памяти, значение, которое нужно сравнить со значением в этой ячейке, и значение, которое нужно сохранить в ячейке, если сравнение было успешным. С помощью этой функции реализовать блокировку на GPU можно следующим образом:

```
__device__ void lock( void ) {
    while( atomicCAS( mutex, 0, 1 ) != 0 );
}
```

Функция `atomicCAS()` возвращает значение, которое нашла по адресу `mutex`. Поэтому цикл `while` продолжается до тех пор, пока `atomicCAS()` не обнаружит 0 по адресу `mutex`. Если там находится 0, то сравнение завершается успешно, и нить записывает в мьютекс 1. Таким образом, нить «крутится» в цикле `while`, пока не заблокирует структуру данных. Мы воспользуемся этим механизмом блокировки, чтобы реализовать на GPU хеш-таблицу. Но сначала инкапсулируем код в структуру, чтобы его было проще использовать при вычислении скалярного произведения:

```
struct Lock {
    int *mutex;
    Lock( void ) {
        int state = 0;
        HANDLE_ERROR( cudaMalloc( (void**)& mutex,
                                  sizeof(int) ) );
        HANDLE_ERROR( cudaMemcpy( mutex, &state, sizeof(int),
                               cudaMemcpyHostToDevice ) );
    }
    ~Lock( void ) {
        cudaFree( mutex );
    }
    __device__ void lock( void ) {
        while( atomicCAS( mutex, 0, 1 ) != 0 );
    }
    __device__ void unlock( void ) {
        atomicExch( mutex, 0 );
    }
};
```

Обратите внимание, что мы восстанавливаем значение мьютекса с помощью функции `atomicExch(mutex, 0)`. Функция `atomicExch()` читает значение по адресу `mutex`, обменивает его со вторым аргументом (в данном случае 0) и возвращает значение, которое прочитала. Зачем надо использовать атомарную функцию? Почему бы просто не записать в мьютекс значение 0?

```
*mutex = 0;
```

Если вы ищете какую-то тонкую, глубоко скрытую причину, по которой этот метод не работает, вынуждены вас разочаровать: он тоже годится. Тогда почему мы его не используем? Потому что для выполнения атомарных транзакций и обычных операций с глобальной памятью в GPU предусмотрены разные пути. Поэтому применение к одному и тому же адресу в глобальной памяти как атомарных, так и обычных операций может привести к кажущейся рассинхронизации `unlock()` с последующими попытками захватить мьютекс с помощью `lock()`. С функциональной точки зрения, поведение будет по-прежнему правильно, но, чтобы сохранить интуитивно очевидное представление приложения о мьютексе, лучше во всех операциях доступа к нему использовать один и тот же путь. Поскольку мы вынуждены применять атомарную операцию для захвата мьютекса, то и для его освобождения применяем такую же операцию.

A.1.2. Возвращаясь к скалярному произведению: атомарная блокировка

Единственная часть показанной ранее программы вычисления скалярного произведения, которую следовало бы изменить, – последний шаг редукции, выполняемый на CPU. В предыдущем разделе мы показали, как реализовать мьютекс на GPU. Структура `Lock`, содержащая реализацию мьютекса, находится в файле `lock.h`, который включен в начало новой версии кода:

```
#include "../common/book.h"
#include "lock.h"

#define imin(a,b) (a<b?a:b)

const int N = 33 * 1024 * 1024;
const int threadsPerBlock = 256;
const int blocksPerGrid =
    imin( 32, (N+threadsPerBlock-1) / threadsPerBlock );
```

За двумя исключениями, этот код совпадает с кодом ядра из главы 5. Оба исключения касаются сигнатуры ядра:

```
_global_ void dot( Lock lock, float *a, float *b, float *c )
```

В обновленной версии мы передаем ядру, помимо входных векторов и выходного буфера, еще и структуру Lock. Она контролирует доступ к выходному буферу на последнем шаге суммирования. Второе изменение по сигнатуре не заметно, но тем не менее связано с ней. Раньше аргумент float *c представлял буфер на N чисел типа float, в котором каждый из N блоков сохранял вычисленную им частичную сумму. Затем этот буфер копировался в память CPU для вычисления окончательной суммы. Теперь же с указывает не на временный буфер, а на одно число с плавающей точкой, в котором сохраняется полное скалярное произведение векторов a и b. Впрочем, и после этих изменений код ядра начинается так же, как в главе 5.

```
__global__ void dot( Lock lock, float *a,
                     float *b, float *c ) {
    __shared__ float cache[threadsPerBlock];
    int tid = threadIdx.x + blockIdx.x * blockDim.x;
    int cacheIndex = threadIdx.x;

    float temp = 0;
    while (tid < N) {
        temp += a[tid] * b[tid];
        tid += blockDim.x * gridDim.x;
    }

    // сохранить значение в кэш
    cache[cacheIndex] = temp;

    // синхронизировать нити в этом блоке
    __syncthreads();

    // для успешной редукции threadsPerBlock должно быть степенью 2
    // из-за следующего кода
    int i = blockDim.x/2;
    while (i != 0) {
        if (cacheIndex < i)
            cache[cacheIndex] += cache[cacheIndex + i];
        __syncthreads();
        i /= 2;
    }
}
```

В этой точке все 256 нитей в каждом блоке просуммировали свои 256 попарных произведений и записали сумму в элемент cache[0]. Теперь каждый блок должен прибавить это значение к величине, хранящейся в c. Чтобы сделать это безопасно, мы воспользуемся мьютексом, контролирующим доступ к этому адресу. Перед тем как обновить значение *c, любая нить должна захватить мьютекс, а прибавив частичную сумму, вычисленную своим блоком, освободить мьютекс, дав возможность сделать то же самое другим нитям. После того как частичная сумма прибавлена к окончательному результату, блоку больше делать нечего, поэтому ядро может завершиться.

Осталось только объявить переменную типа Lock, запустить ядро и скопировать результат в память CPU.

```
Lock lock;
dot<<<blocksPerGrid,threadsPerBlock>>>( lock, dev_a,
                                              dev_b, dev_c );

// скопировать с из памяти GPU в память CPU
HANDLE_ERROR( cudaMemcpy( &c, dev_c,
                           sizeof(float),
                           cudaMemcpyDeviceToHost ) );
```

В главе 5 в этом месте находился цикл for, в котором складывались частичные суммы. Поскольку теперь это делается на GPU с помощью атомарных блокировок, то можно сразу переходить к печати ответа и очистке:

```
#define sum_squares(x) (x*(x+1)*(2*x+1)/6)
printf( "Значение, вычисленное GPU %.6g = %.6g?\n", c,
        2 * sum_squares( (float)(N - 1) ) );

// освободить память GPU
cudaFree( dev_a );
cudaFree( dev_b );
cudaFree( dev_c );

// освободить память CPU
free( a );
free( b );
}
```

Так как мы не можем предсказать точный порядок, в котором блоки прибавляют свои частичные суммы к окончательному результату, то весьма вероятно (почти наверняка), что итоговая сумма вычисляется не в том порядке, что на CPU. Из-за того, что сложение чисел с плавающей точкой не ассоциативно, очень может статься, что результаты, вычисленные на CPU и на GPU, будут немного отличаться. Тут ничего не поделаешь, если не пытаться добавить далеко не тривиальный код, гарантирующий детерминированный порядок захвата мьютекса блоками, – такой же, как при сложении частичных сумм на CPU. Если очень хочется, попробуйте. А мы пока рассмотрим, как атомарные блокировки можно применить для реализации структуры данных, безопасной относительно нитей.

A.2. Реализация хеш-таблицы

Хеш-таблица – одна из самых востребованных структур данных в информатике. Она играет важную роль во многих приложениях. Для читателей, не знакомых с хеш-таблицами, скажем несколько вступительных слов. Вообще-то структуры

данных заслуживают более углубленного изучения, чем мы можем здесь предложить, но, чтобы не задерживаться надолго, ограничимся кратким обзором. Если вы и так все знаете о хеш-таблицах, можете сразу перейти к разделу А.2.2 «Реализация хеш-таблицы на CPU».

A.2.1. Обзор хеш-таблиц

Хеш-таблица – это структура, предназначенная для хранения пар *ключ/значение*. Примером хеш-таблицы может служить толковый словарь. Каждое слово в нем – *ключ*, а определение слова – значение, то есть слово вместе со своим определением составляет пару *ключ/значение*. Но чтобы такая структура данных была полезна, необходимо минимизировать время поиска значения по ключу. В общем случае поиск должен занимать постоянное время, не зависящее от того, сколько пар *ключ/значение* хранится в таблице.

На абстрактном уровне можно сказать, что хеш-таблица помещает значения в тот или иной «кластер», зависящий от ключа. Способ сопоставления ключей с кластерами часто называют *функцией хеширования*. Хорошая функция хеширования распределяет все множество возможных ключей равномерно между кластерами, поскольку именно так можно удовлетворить требованию постоянства времени поиска при любом размере хеш-таблицы.

Рассмотрим, к примеру, хеш-таблицу, соответствующую словарю. Существует очевидная функция хеширования, распределяющая ключи по 26 кластерам – по одному на каждую букву алфавита. Она просто помещает ключ в кластер, соответствующий его первой букве. На рис. А.1 показано, как эта функция распределяет несколько слов.

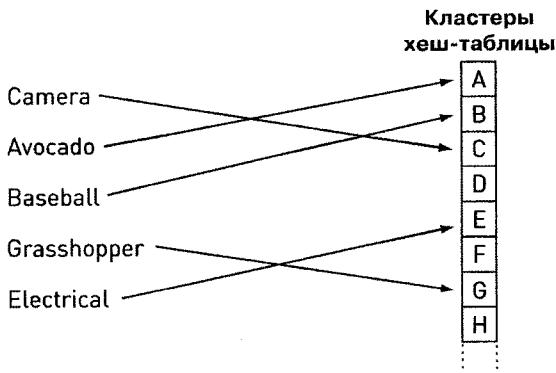


Рис. А.1. Распределение слов по кластерам

Но если принять во внимание информацию о распределении слов в английском языке, то окажется, что эта функция хеширования оставляет желать лучшего, поскольку распределяет слова по 26 кластерам вовсе не равномерно. В одних кла-

стерах будет совсем немного пар, а в других гораздо больше. Следовательно, поиск слова, которое начинается с часто встречающейся буквы, например S, займет куда больше времени, чем поиск слова, начинающегося, скажем, с буквы X. Поскольку мы ищем функцию хеширования, которая дает постоянное время поиска, то такой эффект крайне нежелателен. Изучению функций хеширования посвящено огромное количество работ, но даже краткий обзор применяемых методов выходит за рамки этой книги.

Последняя часть нашей структуры данных – сами кластеры. Если бы функция хеширования была совершенной, то каждому ключу соответствовал бы уникальный кластер. Тогда можно было бы просто поместить пары ключ/значение в массив, каждый элемент которого как раз и являлся бы *кластером*. Однако на практике, как бы хороша ни была функция хеширования, почти всегда возникают *коллизии*. Коллизией называется ситуация, когда нескольким разным ключам сопоставляется один кластер, как, например, словам *avocado* и *aardvark* в хеш-таблице, описывающей словарь. Проще всего хранить все значения, отображаемые на один кластер, в ассоциированном с этим кластером списке. Обнаружив коллизию, например добавление слова *aardvark* в словарь, который уже содержит слово *avocado*, мы помещаем новое значение (*aardvark*) в конец списка, ассоциированного с кластером «A», как показано на рис. А.2.

После добавления слова *avocado* на рис. А.2 в списке для первого кластера будет находиться одна пара ключ/значение. Затем, при добавлении слова *aardvark*, обнаружится коллизия с *avocado*, так как оба слова начинаются на букву A. На рис. А.3 показано, что коллизия разрешается путем добавления нового слова в конец списка, ассоциированного с первым кластером.

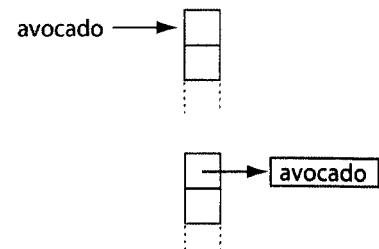


Рис. А.2. Добавление слова *avocado* в хеш-таблицу

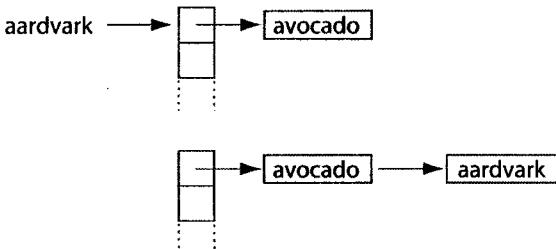


Рис. А.3. Разрешение коллизии при добавлении слова *aardvark*

Теперь, разобравшись в том, что такое *функция хеширования* и *разрешение коллизий*, мы готовы приступить к реализации хеш-таблицы.

A.2.2. Реализация хеш-таблицы на CPU

В предыдущем разделе мы объяснили, что хеш-таблица состоит из двух частей: функции хеширования и структуры данных для представления кластеров. Кластеры мы реализуем, в точности как было описано: выделим массив длины N , в каждом элементе которого будет находиться список пар ключ/значение. Прежде чем переходить к функции хеширования, приведем код получившихся структур данных:

```
#include "../common/book.h"

struct Entry {
    unsigned int key;
    void* value;
    Entry* next;
};

struct Table {
    size_t count;
    Entry** entries;
    Entry* pool;
    Entry* firstFree;
};
```

Как мы и сказали во введении, структура `Entry` содержит ключ и значение. В нашем приложении ключами будут беззнаковые целые, а значениями – произвольные данные, представленные типом `void*`. Смысл нашего приложения в том, чтобы реализовать структуру данных для хеш-таблицы, поэтому в поле `value` мы ничего хранить не собираемся. А включили его лишь для полноты, на случай, если вам захочется использовать этот код в своих программах. Последнее поле в структуре `Entry`, описывающей одну запись хеш-таблицы, – указатель на следующую запись `Entry`. После ряда коллизий в одном кластере окажется несколько записей, и мы решили организовать их в виде связанного списка. Поэтому каждая запись содержит указатель на следующую запись в том же кластере, образуя список записей, хешируемых в один и тот же кластер. В последней записи указатель `next` будет равен `NULL`.

Сама структура `Table` представляет собой массив `entries` кластеров длиной `count`, в котором каждый элемент является указателем на `Entry`. Чтобы не выделять память при каждом добавлении записи `Entry` в таблицу (что негативно скажется на производительности), мы заведем большой пул предварительно выделенных записей, представленный массивом, на который указывает поле `pool`. В поле `firstFree` хранится указатель на следующий доступный элемент `Entry`, поэтому, когда нам понадобится добавить запись в таблицу, мы возьмем элемент, на который указывает `firstFree`, после чего сдвинем указатель. Заодно это упростит код очистки, поскольку для освобождения всех записей достаточно одного вызова `free()`. А если бы мы выделяли память для каждой записи по мере необходимости,

то пришлось бы обойти всю таблицу и вызывать `free()` для каждой хранящейся в ней записи.

Определившись со структурами данных, рассмотрим служебный код:

```
void initialize_table( Table &table, int entries,
                      int elements ) {
    table.count = entries;
    table.entries = (Entry**)calloc( entries, sizeof(Entry*) );
    table.pool = (Entry*)malloc( elements * sizeof( Entry ) );
    table.firstFree = table.pool;
}
```

Инициализация таблицы сводится в основном к выделению и обнулению памяти, предназначенной для массива кластеров. Кроме того, выделяется память для пула записей, а в поле `firstFree` записывается указатель на первую запись в пуле.

Перед выходом из приложения мы должны освободить выделенную память, что и делает функция очистки:

```
void free_table( Table &table ) {
    free( table.entries );
    free( table.pool );
}
```

В введении мы сказали несколько слов о функции хеширования. В частности, было отмечено, что именно качество этой функции отличает хорошую реализацию хеш-таблицы от плохой. В нашем примере ключами служат беззнаковые целые числа, которые требуется отобразить на индексы массива кластеров. Для этого проще всего взять кластер с индексом, равным ключу, то есть поместить запись `e` в элемент `table.entries[e.key]`. Правда, нет никакой гарантии, что все ключи будут меньше длины массива кластеров. Однако эта проблема легко решается:

```
size_t hash( unsigned int key, size_t count ) {
    return key % count;
}
```

Если функция хеширования настолько важна, то правильно ли мы делаем, выбрав такое нехитрое решение? В идеале мы хотели бы, чтобы ключи распределялись между кластерами равномерно, а здесь просто берется ключ по модулю длины массива. В реальных программах функции хеширования посложнее, но мы пишем всего лишь демонстрационный пример, для которого ключи будут генерироваться случайным образом. Если предположить, что распределение ключей более-менее равномерное, то такая функция хеширования равномерно распределит их по кластерам хеш-таблицы. Когда будете реализовывать свою хеш-таблицу, возможно, придется выбрать более сложную функцию хеширования.

Разобравшись со структурой хеш-таблицы и функцией хеширования, мы можем обратиться к процедуре добавления в таблицу пары ключ/значение. Она состоит из трех шагов:

1. Вычислить функцию хеширования от ключа, определив тем самым, в какой кластер поместить новую запись.
2. Выбрать структуру `Entry` из предварительно выделенного пула и инициализировать в ней поля `key` и `value`.
3. Вставить новую запись в начало списка, ассоциированного с кластером.

Перевод словесного описания алгоритма в код не вызывает сложностей.

```
void add_to_table( Table &table, unsigned int key, void* value )
{
    // Шаг 1
    size_t hashValue = hash( key, table.count );

    // Шаг 2
    Entry *location = table.firstFree++;
    location->key = key;
    location->value = value;

    // Шаг 3
    location->next = table.entries[hashValue];
    table.entries[hashValue] = location;
}
```

Если вы никогда не встречались со связанными списками (или успели подзабыть), то шаг 3 может показаться непонятным. Указатель на первый узел списка хранится в элементе массива `table.entries[hashValue]`. Поэтому для вставки нового узла в начало списка нужно выполнить два действия. Во-первых, в поле `next` новой записи нужно записать указатель на прежний первый узел списка. А во-вторых, новую запись нужно вставить в кластер таким образом, чтобы именно она стала первым узлом списка.

Неплохо бы знать, работает ли написанный код, поэтому мы реализовали функцию, проверяющую корректность данных в хеш-таблице. Сначала мы проходим по таблице и исследуем каждый узел. Для ключа в этом узле вычисляется функция хеширования и проверяется, что узел находится в том кластере, где должен быть. Проверив узлы, мы смотрим, равно ли количество узлов, *оказавшихся* в таблице, количеству узлов, *добавленных* нами в таблицу. Если это не так, значит, либо мы по ошибке поместили узел в несколько кластеров, либо вставили его неправильно.

```
#define SIZE (100*1024*1024)
#define ELEMENTS (SIZE / sizeof(unsigned int))

void verify_table( const Table &table ) {
    int count = 0;
```

```
for (size_t i=0; i<table.count; i++) {
    Entry *current = table.entries[i];
    while (current != NULL) {
        ++count;
        if (hash( current->value, table.count ) != i)
            printf( "%d хешируется в %ld, но обнаружено "
                    "в %ld\n", current->value,
                    hash( current->value, table.count ), i );
        current = current->next;
    }
}
if (count != ELEMENTS)
    printf( "Количество элементов в хеш-таблице: %d. Должно быть %ld\n",
            count, ELEMENTS );
else
    printf( "Количество элементов в хеш-таблице: %d. Так и должно быть.\n", count );
}
```

С инфраструктурным кодом всё, теперь можно взглянуть на `main()`. Как и в большинстве примеров в этой книге, трудная работа делается во вспомогательных функциях, поэтому в коде `main()` разобраться сравнительно просто:

```
#define HASH_ENTRIES 1024

int main( void ) {
    unsigned int *buffer =
        (unsigned int*)big_random_block( SIZE );
    clock_t start, stop;
    start = clock();

    Table table;
    initialize_table( table, HASH_ENTRIES, ELEMENTS );

    for (int i=0; i<ELEMENTS; i++) {
        add_to_table( table, buffer[i], (void*)NULL );
    }

    stop = clock();
    float elapsedTime = (float)(stop - start) /
        (float)CLOCKS_PER_SEC * 1000.0f;
    printf( "Время построения хеш-таблицы: %.3f ms\n", elapsedTime );

    verify_table( table );

    free_table( table );
    free( buffer );
    return 0;
}
```

Вначале выделяется большой блок со случайными числами. Сгенерированные случайным образом беззнаковые целые станут ключами, вставляемыми в хеш-таблицу. После того как все ключи сгенерированы, мы запрашиваем системное время, чтобы можно было измерить производительность нашей реализации. Далее мы инициализируем хеш-таблицу и в цикле for вставляем в нее все сгенерированные ключи. Затем снова запрашиваем системное время и вычисляем общее время инициализации и добавления ключей. А перед тем как выйти, проверяем корректность хеш-таблицы и освобождаем выделенную память.

Вероятно, вы заметили, что во всех парах ключ/значение в качестве значения фигурирует NULL. В нормальном приложении вместе с ключом, вероятно, сохранились бы какие-то полезные данные, но нас интересует только сама реализация хеш-таблицы, а значения безразличны.

A.2.3. Многонитевая реализация хеш-таблицы

В реализации хеш-таблицы на CPU сделаны некоторые предположения, которые перестают быть верными при переходе на GPU. Во-первых, мы предполагали, что в каждый момент времени в таблицу может добавляться только один узел. Если же узлы одновременно добавляются несколькими нитями, то может возникнуть та же проблема, что в примере многонитевого сложения, рассмотренном в главе 9.

Давайте вернемся к случаю добавления слов *avocado* и *aardvark* и представим, что их пытаются добавить нити A и B. Нить A вычисляет функцию хеширования от *avocado*, а нить B – функцию хеширования от *aardvark*. Обе приходят к выводу, что ключи попадают в один и тот же кластер. Чтобы добавить новую запись в список, нити A и B изменяют значение в поле *next* записи, так чтобы оно указывало на первый узел текущего списка (см. рис. A.4).

Затем обе нити пытаются заменить запись в массиве кластера своей записью. Но сохраняется только изменение, сделанное нитью, которая завершилась последней, поскольку она затирает изменение, совершенное предыдущей нитью. Предположим, что нить A заменяет запись *altitude* записью *avocado*. А как только

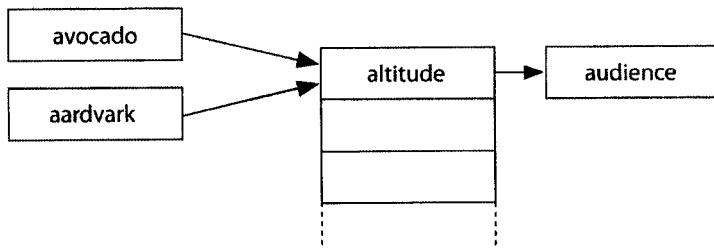


Рис. A.4. Несколько потоков пытаются добавить узел в один и тот же кластер

она это сделала, нить В заменяет то, что она считает записью *altitude*, своей записью *aardvark*. Увы, на самом деле она заменила запись *avocado*, что привело к ситуации, изображенной на рис. А.5.

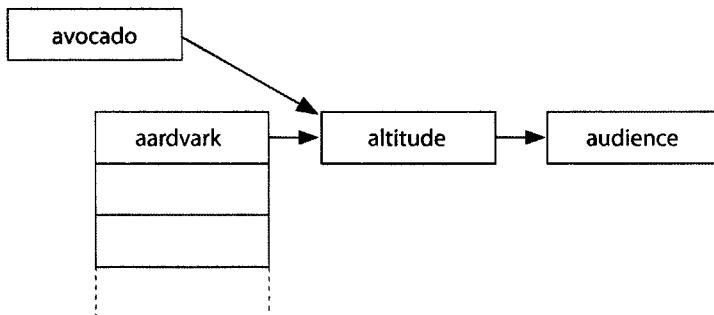


Рис. А.5. Так выглядит хеш-таблица после неудачной одновременной модификации двумя нитями

Записи, добавленной нитью A, не повезло – она выпала из таблицы. К счастью, наша функция проверки корректности обнаружит эту ситуацию и сообщит о проблеме – ведь она насчитает в таблице меньше узлов, чем ожидается. Но вопрос-то остается: как построить хеш-таблицу на GPU?! Принципиальным здесь является тот факт, что в любой момент времени модифицировать кластер разрешено только одной нитью. Здесь наблюдается аналогия с вычислением скалярного произведения, когда лишь одна нить может безопасно прибавлять частичную сумму к итоговому результату. Если бы с каждым кластером была ассоциирована атомарная блокировка, то можно было бы гарантировать, что изменения в кластер вносит только одна нить.

A.2.4. Реализация хеш-таблицы на GPU

Зная, как гарантировать безопасный доступ к хеш-таблице со стороны нескольких нитей, мы можем переработать для GPU код, написанный в разделе A.2.2 для CPU. Необходимо включить файл `lock.h`, в котором содержится определение структуры `Lock` (см. раздел A.1.1), и добавить в объявление функции хеширования квалификатор `__device__`¹. В остальном структуры данных и функции хеширования ничем не отличаются от реализации для CPU.

¹ Функция объявлена как `__device__ __host__ size_t hash(unsigned int value, size_t count)`. В этой книге мы не раз видели функции с квалификатором `__device__`, но нигде не объяснили, что такое квалификатор `__host__`. Если в объявлении функции присутствуют оба квалификатора `__device__` и `__host__`, то компилятор NVIDIA генерирует два варианта функции: для GPU и для CPU. Функция для GPU исполняется GPU, и вызывать ее можно только из кода, работающего на GPU. Аналогично функция для CPU исполняется CPU и может быть вызвана только из кода, работающего на CPU. Это удобный способ написать единственный вариант кода, который будет исполняться как GPU, так и CPU.

```
#include "../common/book.h"
#include "lock.h"

struct Entry {
    unsigned int key;
    void* value;
    Entry* next;
};

struct Table {
    size_t count;
    Entry** entries;
    Entry* pool;
    Entry* firstFree;
};

__device__ __host__ size_t hash( unsigned int value,
                               size_t count ) {
    return value % count;
}
```

Инициализация и освобождение хеш-таблицы устроены так же, как в версии для CPU, только теперь мы используем функции, предоставляемые исполняющей средой CUDA. Память для массива кластеров и пула записей выделяется функцией `cudaMalloc()`, а для обнуления массива кластеров вызывается функция `cudaMemset()`. Чтобы освободить память перед выходом из программы, мы используем функцию `cudaFree()`.

```
void initialize_table( Table &table, int entries,
                      int elements ) {
    table.count = entries;
    HANDLE_ERROR( cudaMalloc( (void**)&table.entries,
                           entries * sizeof(Entry*) ) );
    HANDLE_ERROR( cudaMemset( table.entries, 0,
                           entries * sizeof(Entry*) ) );
    HANDLE_ERROR( cudaMalloc( (void**)&table.pool,
                           elements * sizeof(Entry) ) );
}

void free_table( Table &table ) {
    cudaFree( table.pool );
    cudaFree( table.entries );
}
```

В варианте для CPU мы написали функцию для проверки корректности хеш-таблицы. Аналогичная функция нужна и для GPU, и тут у нас есть два варианта. Можно написать специальную версию `verify_table()` для GPU или воспользоваться той же функцией, что для CPU, но добавить функцию, которая скопирует

хеш-таблицу из памяти GPU в память CPU. Хотя задачу решают оба варианта, мы предпочли второй по двум причинам. Во-первых, повторно используется уже написанная функция; это экономит время и гарантирует, что в будущем любые изменения придется делать только в одном месте. Во-вторых, реализация функции копирования вскрывает одну интересную проблему, решение которой может пригодиться вам в будущем.

Как и было обещано, функция `verify_table()` почти ничем не отличается от версии для CPU, но для удобства мы повторим ее код:

```
#define SIZE (100*1024*1024)
#define ELEMENTS (SIZE / sizeof(unsigned int))
#define HASH_ENTRIES 1024

void verify_table( const Table &dev_table ) {
    Table table;
    copy_table_to_host( dev_table, table );

    int count = 0;
    for (size_t i=0; i<table.count; i++) {
        Entry *current = table.entries[i];
        while (current != NULL) {
            ++count;
            if (hash( current->value, table.count ) != i)
                printf( "%d хешируется в %ld, но обнаружено "
                        "в %ld\n", current->value,
                        hash( current->value, table.count ), i );
            current = current->next;
        }
    }

    if (count != ELEMENTS)
        printf( "Количество элементов в хеш-таблице: %d. Должно быть %ld\n",
                count, ELEMENTS );
    else
        printf( "Количество элементов в хеш-таблице: %d. Так и должно быть.\n", count );
    free( table.pool );
    free( table.entries );
}
```

Поскольку мы решили повторно использовать реализацию `verify_table()` для CPU, то необходима функция, которая будет копировать таблицу из памяти GPU в память CPU. Она состоит из трех шагов, причем первые два более-менее очевидны, а последний похитрее. Сначала мы выделяем память CPU для данных хеш-таблицы, затем с помощью `cudaMemcpuy()` копируем в эту память данные из памяти GPU. То и другое делалось уже много раз, так что ничего неожиданного в следующем коде нет:

```
void copy_table_to_host( const Table &table, Table &hostTable) {
    hostTable.count = table.count;
    hostTable.entries = (Entry**)calloc( table.count,
                                         sizeof(Entry*));
    hostTable.pool = (Entry*)malloc( ELEMENTS *
                                    sizeof( Entry ) );
    HANDLE_ERROR( cudaMemcpy( hostTable.entries, table.entries,
                            table.count * sizeof(Entry*),
                            cudaMemcpyDeviceToHost ) );
    HANDLE_ERROR( cudaMemcpy( hostTable.pool, table.pool,
                           ELEMENTS * sizeof( Entry ),
                           cudaMemcpyDeviceToHost ) );
}
```

А хитрость связана с тем, что часть копируемых данных – указатели. Просто копировать указатели в память CPU нельзя, потому что это адреса в памяти GPU, то есть на CPU они будут указывать совсем не туда. Однако же относительные смещения элементов, адресуемых этими указателями, верны и для CPU, и для GPU. Поскольку любая структура `Entry`, адресуемая указателем в памяти GPU, – это какой-то элемент массива `table.pool[]`, поэтому, чтобы хеш-таблица в памяти CPU была правильной, указатели должны быть направлены на соответственные элементы массива `hostTable.pool[]`.

Если известен указатель `X` на память GPU, то для получения правильного указателя на память CPU мы должны прибавить к адресу `hostTable.pool` смещение `X` от начала массива `table.pool`. Иными словами, новый указатель вычисляется по формуле

$$(X - \text{table.pool}) + \text{hostTable.pool}$$

Эту коррекцию нужно применить ко всем указателям на структуры `Entry`, скопированным из памяти GPU, то есть к указателям в массиве `hostTable.entries` и к указателям, хранящимся в поле `next` всех записей `Entry` в пуле:

```
for (int i=0; i<table.count; i++) {
    if (hostTable.entries[i] != NULL)
        hostTable.entries[i] =
            (Entry*)((size_t)hostTable.entries[i] -
                     (size_t)table.pool + (size_t)hostTable.pool);
}

for (int i=0; i<ELEMENTS; i++) {
    if (hostTable.pool[i].next != NULL)
        hostTable.pool[i].next =
            (Entry*)((size_t)hostTable.pool[i].next -
                     (size_t)table.pool + (size_t)hostTable.pool);
}
```

Итак, со структурами данных, функцией хеширования, инициализаций и очисткой и проверкой мы разобрались. Остался самый важный кусок – подключение атомарных операций. Ядро `add_to_table()` принимает в качестве аргументов массивы ключей и значений, добавляемых в хеш-таблицу, саму хеш-таблицу и массив структур `Lock`, используемых для блокировки каждого кластера таблицы. Поскольку входными данными являются два массива, которые будут индексироваться нитями, необходима уже навязшая в зубах линеаризация индексов:

```
__global__ void add_to_table( unsigned int *keys, void **values,
                             Table table, Lock *lock ) {
    int tid = threadIdx.x + blockIdx.x * blockDim.x;
    int stride = blockDim.x * gridDim.x;
```

Нити обходят входные массивы точно так же, как в примере вычисления скалярного произведения. Для каждого ключа в массиве `keys[]` нить вычисляет функцию хеширования, чтобы определить, в какой кластер поместить пару ключ/значение. Определив кластер, нить блокирует его, добавляет свою пару и разблокирует кластер.

```
while (tid < ELEMENTS) {
    unsigned int key = keys[tid];
    size_t hashValue = hash( key, table.count );
    for (int i=0; i<32; i++) {
        if ((tid % 32) == i) {
            Entry *location = &(table.pool[tid]);
            location->key = key;
            location->value = values[tid];
            lock[hashValue].lock();
            location->next = table.entries[hashValue];
            table.entries[hashValue] = location;
            lock[hashValue].unlock();
        }
    }
    tid += stride;
}
```

Но в этом коде есть одна примечательная особенность. На первом взгляд, цикл `for` и следующее за ним предложение `if` кажутся решительно лишними. В главе 6 мы ввели понятие *варпа*. Напомним, что варпом называется набор из 32 нитей, выполняемых синхронно. Хотя детали аппаратной реализации выходят за рамки данной книги, отметим, что в любой момент времени захватить блокировку может только одна нить из варпа, и если позволить всем 32 входящим в него нитям одновременно конкурировать за блокировку, то мы не обернемся неприятностей. Мы пришли к выводу, что эту проблему лучше решить на уровне программы – после-

Оставшаяся часть main() аналогична версии для CPU: добавляем все ключи в хеш-таблицу, останавливаем таймер, проверяем корректность таблицы и прибираемся за собой:

```
add_to_table<<<60,256>>>( dev_keys, dev_values,
    table, dev_lock );

HANDLE_ERROR( cudaEventRecord( stop, 0 ) );
HANDLE_ERROR( cudaEventSynchronize( stop ) );
float elapsedTime;
HANDLE_ERROR( cudaEventElapsedTime( &elapsedTime,
    start, stop ) );
printf( "Время построения хеш-таблицы: %3.1f ms\n", elapsedTime );

verify_table( table );

HANDLE_ERROR( cudaEventDestroy( start ) );
HANDLE_ERROR( cudaEventDestroy( stop ) );
free_table( table );
cudaFree( dev_lock );
cudaFree( dev_keys );
cudaFree( dev_values );
free( buffer );
return 0;
}
```

A.2.5. Производительность хеш-таблицы

На машине с процессором Intel Core 2 Duo программа, представленная в разделе A.2.2, построила хеш-таблицу, содержащую 100 Мб данных, за 360 мс. При компиляции был задан флаг -O3, обеспечивающий максимальную оптимизацию кода для CPU. Многонитевой версии программы для GPU из раздела A.2.4 для решения той же задачи потребовалось 375 мс. При разнице в 5% возникает законный вопрос: почему массивно-параллельный графический процессор уступил однопоточной версии той же программы для CPU? Будем честны – потому что GPU не оптимизированы для многонитевого доступа к сложным структурам данных, таким как хеш-таблицы. Поэтому строить на GPU хеш-таблицу или иную структуру данных в расчете получить выигрыш в быстродействии не имеет смысла. Это лучше делать на CPU.

С другой стороны, иногда встречаются ситуации, когда длинная последовательность вычислений включает один-два этапа, где GPU не дает выигрыша в производительности по сравнению с CPU. Тогда есть три (очевидных) варианта действий:

- выполнять все этапы последовательности на GPU;
- выполнять все этапы последовательности на CPU;
- выполнять часть этапов на GPU, а часть – на CPU.

Вроде бы последний вариант объединяет все лучшее из обоих миров, но на практике это означает, что в каждой точке приложения, где нужно переключаться с GPU на CPU или наоборот, необходимо производить синхронизацию между GPU и CPU. А синхронизация вкупе с последующей передачей данных между GPU и CPU может съесть весь выигрыш, который вы надеялись получить от смешанного подхода.

В такой ситуации, быть может, имеет смысл производить на GPU все вычисления, пусть даже для некоторых шагов алгоритма GPU неидеален. И вот тогда заполнение хеш-таблицы на GPU может избавить от необходимости синхронизировать GPU и CPU и свести к минимуму передачу данных, освободив CPU для других вычислений. Не исключено, что при этом общая производительность реализации на GPU окажется выше, чем при смешанном подходе, пусть даже на некоторых шагах GPU не быстрее CPU (а то и заметно медленнее).

A.3. Резюме

Мы показали, как с помощью атомарной операции сравнить—и—обменять можно реализовать мьютекс на GPU. Воспользовавшись блокировкой, построенной на базе мьютекса, мы смогли улучшить прежнюю версию программы вычисления скалярного произведения, перенеся ее целиком на GPU. Далее мы развили эту идею и реализовали многонитевую хеш-таблицу, в которой для предотвращения небезопасных одновременных модификаций со стороны нескольких нитей применяется массив блокировок. На самом деле разработанный нами мьютекс годится для различных параллельных структур данных, и мы надеемся, что вы найдете ему применение в собственных приложениях. Разумеется, производительность приложений, в которых GPU используется для реализации структур данных с мьютексами, следует тщательно измерять. Наша версия хеш-таблицы на GPU по скорости уступила однопоточной версии, работающей на CPU, так что задействовать GPU для такого рода задач имеет смысл лишь в определенных ситуациях. Не существует общего правила, отвечающего на вопрос, какой подход окажется лучше: только GPU, только CPU или смешанный, но, зная, как применять атомарные операции, вы сможете принять правильное решение в каждом конкретном случае.



Предметный указатель

Символы

`_constant`, модификатор, 95
`_device`, квалификатор, 56, 219
`_global`, квалификатор, 33, 47
`syncthreads()`
вычисление скалярного произведения, 75, 81
нежелательные последствия, 82
растровое изображение в разделяемой памяти, 83

А

`add_to_table()`, ядро, хеш-таблица на GPU, 223
`add()`, функция, 45
`anim_and_exit()`, метод, анимация волн на GPU, 68
`animExit()`, функция, 130
`anim_gpu()`, функция, текстурная память, 110, 114
`atomicAdd()`
атомарные блокировки, 206
вычисление гистограмм в глобальной памяти, 152
отсутствие поддержки для чисел с плавающей точкой, 205
`atomicCAS()`, функция, 207
`atomicExch()`, функция, 208

Б

BLAS (Basic Linear Algebra Subprograms), 196
`blend_kernel()`, текстурная память, 113, 116
`blockDim`, переменная
анимация волн, 69
вычисление гистограммы, 151, 153
вычисление скалярного произведения, 73, 79, 82, 182, 209
изображение в разделяемой памяти, 84
интероперабельность с графикой, 126
моделирование теплообмена, 107
несколько потоков CUDA, 166
реализация хеш-таблицы на GPU, 223
сложение векторов, 62, 64
текстурная память, 116
трассировка лучей на GPU, 93

С

`cacheIndex`, неправильная оптимизация скалярного произведения, 82
`CPUAnimBitmap`, класс, анимация волн, 67
CPU (центральный процессор)
вычисление гистограммы, 145
как хозяин, 33
реализация хеш-таблиц, 214
сложение векторов, 44
тактовая частота, 17
увеличения числа ядер, 17

управление потоками, 68
CUBLAS, библиотека, 196
`cuComplex`, структура, 52, 56
`cudaBindTexture2D()`, текстурная память, 118
`cudaBindTexture()`, текстурная память, 112
CUDA C
CUDA Development Toolkit, 28
графические процессоры с поддержкой, 26
драйвер устройства NVIDIA, 28
обзор, 32
отладка, 197
первая программа, 32
передача параметров, 34
получение информации об устройстве, 36
приступая к работе, 26
стандартный компилятор, 30
`cudaChannelFormatDesc()`, привязка двумерных текстур, 118
`cudaChooseDevice()`, 41, 123
GPUAnimBitmap, 130
`cudaD3D9SetDirect3DDevice()`,
интероперабельность с DirectX, 138
`cudaDeviceMapHost`, скалярное произведение с применением нуль-копируемой памяти, 181
`cudaDeviceProp`, структура, 37, 41
несколько потоков CUDA, 167
работа с `cudaChooseDevice()`, 123
`cudaEventCreate()`, измерение производительности, 98
`cudaEventDestroy()`, 101
`cudaEventElapsedTime()`, 101
`cudaEventRecord()`, 98
`cudaEventSynchronize()`, 99
`cudaFree()`, 36
`cudaFreeHost()`, 160
CUDA-GDB, отладчик, 197
`cudaGetDeviceCount()`, 37, 41
интегрированный или дискретный GPU, 183
`cudaGetDeviceProperties()`, 41
`cudaGLSetGLDevice()`, 123, 130
`cudaGraphicsGLRegisterBuffer()`, 124, 131
`cudaGraphicsMapFlagsNone`, флаг, 124
`cudaGraphicsMapFlagsReadOnly`, флаг, 124
`cudaGraphicsMapFlagsWriteDiscard`, флаг, 125
`cudaGraphicsUnmapResources()`, 125
`cudaHostAlloc()`
закрепленная память, 157
и `malloc()`, 156
нуль-копируемая память, 179
потоки CUDA, 165
`cudaHostAllocDefault`, флаг, 160, 176
`cudaHostAllocMapped`, флаг, 176, 179, 189
`cudaHostAllocPortable`, флаг, 189
`cudaHostAllocWriteCombined`, флаг, 180, 189
`cudaHostGetDevicePointer()`, 180, 192
`cudaMalloc()`, выделение памяти устройства, 35

cuda-memcheck, 198
cudaMemcpyAsync(), 164, 171
cudaMemcpyDeviceToHost, 47
cudaMemcpyHostToDevice, 47
cudaMemcpyToSymbol(), константная память, 96
cudaMemcpy(), копирование данных между хостом и устройством, 36
CUDA Memory Checker, 198
cudaMemset(), 147
CUDA.NET, проект, 203
cudaSetDevice(), 41, 187, 190
cudaSetDeviceFlags(), 181, 190
cudaStreamCreate(), 163, 168
cudaStreamDestroy(), 165
cudaStreamSynchronize(), 165, 170
cudaThreadSynchronize(), 180
CUDA Toolkit, 28, 195
cudaUnbindTexture(), 119
CUDA Zone, 142
CUDPP (CUDA Data Parallel Primitives Library), 202
CUFFT библиотека, 195
CULAtools, 202

D

dim3, тип, 54
DirectX
 GeForce 8800 GTX, 20
 интероперабельность с графикой, 138
DRAM, 183

E

end_thread(), 186
EXIT_FAILURE, код возврата, 35

F

fAnim(), функция, 130
FORTRAN, приложения
 библиотека CUBLAS, 196
 интерфейс из CUDA C, 203
free(), стандартная функция С
 сравнение с **cudaFree()**, 36

G

GeForce 256, 19
GeForce 8800 GTX, 20
generate_frame(), анимация волн на GPU, 68, 133
generate_frame(), анимация рыб на GPU, 69
glBindBuffer()
 интероперабельность с графикой, 127
 создание пиксельного буфера, 124
glBufferData(), 124
glDrawPixels(), 127, 134
glGenBuffers(), 124
GL_PIXEL_UNPACK_BUFFER_ARB,
 интероперабельность с OpenGL, 131
GLUT (GL Utility Toolkit)
 инициализация, 130

инициализация драйвера OpenGL, 123
 настройка интероперабельности с графикой, 125
glutIdleFunc(), 130
glutInit(), 130
glutMainLoop(), 125
GPUAnimBitmap, структура
 анимация волн на GPU, 132
 моделирование теплообмена, 135
 создание, 129
gpu_anim.h, 132
gridDim, переменная, 58

H

HANDLE_ERROR(), макрос, 35
hit, метод, трассировка лучей на GPU, 91
HOOMD (Highly Optimized Object Oriented Many-particle Dynamics), 24

I

idle_func(), функция-член структуры
GPUAnimBitmap, 133
IEEE требования, 21

J

julia(), функция, 51, 55

K

key_func(), 125

L

LAPACK (Linear Algebra Package), 202
Linux, стандартный компилятор, 30
Lock структура, 208

M

Macintosh OS X, стандартный компилятор, 31
malloc()cudaHostAlloc(), 156
malloc()cudaMalloc(), 35
maxThreadsPerBlock, свойство устройства, 62
memcpy(), стандартная функция С, 36
memset(), стандартная функция С, 148

N

nForce, медиа-коммуникационные процессоры (MCP), 183

O

OpenGL
 3D-графика, 18
 взаимодействие с, 123
 генерация данных изображения, 122
 ранние этапы GPU-вычислений, 19
 создание структуры **GPUAnimBitmap**, 129

P

Parallel Nsight, отладчик, 198
PCI Express шина, 184

PyCUDA, проект, 203
Python, языковые привязки, 203

S

SC. Cm. Servis Console
Silicon Graphics, библиотека OpenGL, 18
start_thread(), 186

T

tex1Dfetch(), текстурная память, 113
tex2D(), текстурная память, 116
threadIdx, переменная, 60
threadsPerBlock
выделение разделяемой памяти, 73
вычисление скалярного произведения, 75

V

VCB. *Cm.* Consolidated Backup
Visual Profiler, 199
Visual Studio C, компилятор, 30
vMA. *Cm.* vSphere Management Assistant
VMkernel. *Cm.* Гипервизор

A

АЛУ (арифметически-логическое устройство)
архитектура CUDA, 21
константная память, 88
Анимация
волны на GPU, 67
моделирование теплообмена, 108
Анимация волн, 67
интероперабельность с графикой, 128
Аппаратные ограничения
количество нитей в блоке, 62
количество одновременно запущенных блоков, 50
сложение векторов произвольной длины, 64
Архитектура CUDA
вычислительная гидродинамика, 23
науки об окружающей среде, 24
обработка медицинских изображений, 22
образовательные ресурсы, 200
определение, 21
Асинхронный вызов
cudaMemcpyAsync(), 164
события, 98
Атомарные блокировки
обзор, 206
хеш-таблица на GPU, 224
Атомарные операции, 140
блокировки, 206
вычисление гистограммы на GPU, 151, 153
вычислительные возможности NVIDIA GPU, 140
дополнительные сведения, 204
скалярное произведение, 204

B

Блоки
аппаратные ограничения, 50

определенение, 58

расщепление на нити, 59

сложение векторов на GPU, 48

фрактал Джулия, 53

Блокировки атомарные, 206

Буферы, объявление в разделяемой памяти, 73

Быстрое преобразование Фурье, библиотека NVIDIA, 195

B

Барьеры, чтение константной памяти, 96

Вычисление скалярного произведения

некорректная оптимизация, 81

нуль-копируемая память, 177

разделяемая память, 73

целиком на GPU с применением атомарных операций, 205, 208

Вычислительные возможности

cudaChooseDevice(), 123

NVIDIA GPU, 141

задание минимального уровня при компиляции, 142

обзор, 123

определение, 140

G

Гистограмма, вычисление

на CPU, 145

на GPU, 147

обзор, 145

применение атомарных операций

с глобальной и разделяемой памятью, 153

применение атомарных операций

с глобальной памятью, 151

Графические процессоры (GPU)

вычисление гистограммы, 147

вычисление гистограммы с применением

атомарных операций с глобальной

и разделяемой памятью, 153

вычисление гистограммы с применением

атомарных операций с глобальной

памятью, 151

дискретные и интегрированные, 183

измерение производительности с помощью событий, 97

название устройства, 33

освобождение памяти. См. cudaFree()

планирование задач, 171

ранние этапы, 19

реализация хеш-таблицы, 219

с поддержкой CUDA, 26

трассировка лучей, 90

фрактал Джулия, 53

Графические процессоры, несколько

использование, 184

нуль-копируемая память CPU, 176

обзор, 176

переносимая закрепленная память, 188

производительность шуля-копирования, 183

Д

Дискретные GPU, 183
Драйвер устройства, 28

З

Закрепленная память
обзор, 156
ограничения, 157
переносимая, 188
потоки CUDA, 163

И

Инструментальные средства CUDA
CUBLAS библиотека, 196
CUDA Toolkit, 195
CUFFT библиотека, 195
GPU Computing SDK, 196
NVIDIA Performance Primitives, 197
Visual Profiler, 198
обзор, 195
отладка CUDA C, 197
Интегрированные GPU, 183
Интероперабельность с графикой, 121
DirectX, 138
анимация волн, 128
генерация данных изображения в ядре, 122
передача данных изображения OpenGL, 124

К

Камера, трассировка лучей, 89
Кластеры, хеш-таблица
концепция, 212
реализация на GPU, 220
Ключи хеш-таблицы, 212
Коллизии, хеш-таблица, 213
Компилятор
С стандартный, 30
минимальный уровень вычислительных возможностей, 142
Константная память
производительность, 97
трассировка лучей на GPU, 90, 94
ускорение приложений, 88
Кэширование на кристалле. См. Константная память, Текстурная память

М

Масштабируемый интерфейс соединений (SLI), 184
Медиа-коммуникационные процессоры (MCP), 183
Моделирование теплообмена
анимация, 108
двумерная текстурная память, 116
интероперабельность с графикой, 134
обновление температур, 106
применение текстурной памяти, 112
простая модель, 105

Мосты, соединение GPU, 184

Мультиядерная революция, 18
Мютекс, 206

Н

Нити
анимация волн на GPU, 67
аппаратное ограничение числа, 62
и константная память, 96
и разделяемая память. См. Разделяемая память
несколько GPU, 185
обзор, 59
расхождение, 82
синхронизация, 180
сложение векторов на GPU, 60
сложение векторов произвольной длины на GPU, 64
сложение длинных векторов на GPU, 62
трассировка лучей на GPU, 93
 чтение-модификация-запись, 143
ядро вычисления гистограммы, 151
Нуль-копируемая память
выделение, 176
определение, 176
производительность, 183

О

Объект пиксельного буфера, OpenGL, 124
Отладка CUDA C, 197

П

Память
закрепленная, 156
константная. См. Константная память
освобождение. См. cudaFree(), free()
получение информации об устройстве, 36
разделяемая. См. Разделяемая память
текстурная. См. Текстурная память
устройства, выделение. См. cudaMalloc()
Перекрытие операций, 162, 166
Пиксельные шейдеры, 19
Планирование задач на GPU, 171
Полуварпы, чтение из константной памяти, 97
Потоки CUDA
несколько, 166, 173
обзор, 161
один, 161
планирование задач на GPU, 171
Производительность
вычисление гистограммы, 149
закрепленная память, 157
измерение с помощью событий, 97
константная память, 96
нуль-копирование, 183
хеш-таблицы, 225
эволюция CPU, 17
Пространственная локальность
анимация модели теплообмена, 112

Предметный указатель

- текстурные кэши, 105
- Прямой доступа к закрепленной памяти (DMA), 157

- P**
- Разделяемая память
 - атомарные операции, 142, 153
 - вычисление скалярного произведения, 73, 81
 - и синхронизация, 72
 - растровое изображение, 83
- Распараллеливание по задачам, 156
- Растеризация, 89
- Редукция
 - обзор, 204
 - разделяемая память и синхронизация, 76
 - скалярное произведение как пример, 78

- C**
- Свойства
 - maxThreadsPerBlock, 62
 - использование, 40
 - распечатка, 39
 - структура cudaDeviceProp.
 - См. cudaDeviceProp, структура
- Сетка
 - как совокупность параллельных блоков, 49
 - определение, 58
 - трехмерная, 54
- Синхронизация
 - и разделяемая память. См. Разделяемая память
 - нитей, 180
 - потоков, 165
 - событий. См. cudaEventSynchronize()
- Сложение векторов
 - длинных, 62
 - на CPU, 44
 - на GPU, 46
 - произвольной длины, 64
 - с использованием нитей, 60
- События
 - время между двумя зарегистрированными событиями. См. cudaEventElapsedTime()
 - вычисление гистограммы на GPU, 147
 - измерение производительности трассировщика лучей, 99
 - обзор, 97
 - регистрация. См. cudaEventRecord()
 - создание. См. cudaEventCreate()
 - уничтожение. См. cudaEventDestroy()
- Среда разработки, 26
- Стандартный компилятор C
 - вызов ядра, 33
 - задание минимального уровня вычислительных возможностей, 142
 - среда разработки, 30

T

- Текстурная память
 - анимация моделирования, 108
 - двумерная, 116
 - моделирование теплообмена, 105
 - обзор, 104
 - определение, 104
 - применение, 112
- Трассировка лучей, 89
 - измерение производительности, 99
 - с применением константной памяти, 94

У

- Унифицированный шейдерный конвейер, 21
- Ускорители двумерной графики, 18
- Устройство
 - определение термина в этой книге, 33
 - получение информации о, 36
 - свойства, 40

Ф

- Форумы NVIDIA, 201
- Фрактал Джулиа
 - вычисление на CPU, 51
 - вычисление на GPU, 53
 - обзор, 50
- Функция хеширования, 212

Х

- Хеш-таблицы
 - концепция, 212
 - многонитевые, 218
 - производительность, 225
 - реализация на CPU, 214
 - реализация на GPU, 219
- Хозяин, использование термина в книге, 33

Ч

- Чередование операций, 144
- Числа с плавающей точкой
 - в архитектуре CUDA, 21
 - поддержка атомарных операций, 205
- Чтение-модификация-запись
 - атомарная, 143
 - использование атомарных блокировок, 206

Ш

- Шейдерные языки, 20

Я

- Ядро
 - вызов, 33
 - определение, 33
 - передача параметров, 34
 - переменная blockIdx.x, 48
- Языковые привязки, 203

Книги издательства «ДМК Пресс» можно заказать в торгово-издательском холдинге «АЛЬЯНС-КНИГА» наложенным платежом, выслав открытку или письмо по почтовому адресу: **123242, Москва, а/я 20** или по электронному адресу: **orders@aliants-kniga.ru**.

При оформлении заказа следует указать адрес (полностью), по которому должны быть высланы книги; фамилию, имя и отчество получателя. Желательно также указать свой телефон и электронный адрес.

Эти книги вы можете заказать и в Интернет-магазине: www.dmk.ru.

Оптовые закупки: тел. (499) 725-50-27, 725-54-09; электронный адрес books@aliants-kniga.ru.

Дж. Сандерс, Э. Кэндрот

Технология CUDA в примерах
Введение в программирование графических процессоров

Главный редактор *Мовчан Д. А.*
dm@dmk-press.ru

Научный редактор *Боресков А. В.*
Перевод с английского *Слинкин А. А.*
Корректор *Синяева Г. И.*
Верстка *Чаннова А. А.*
Дизайн обложки *Мовчан А. Г.*

Подписано в печать 10.11.2013. Формат 60×90 1/16 .

Гарнитура «Петербург». Печать офсетная.

Усл. печ. л. 21,75. Тираж 100 экз.

Web-сайт издательства: www.dmk-press.ru

ТЕХНОЛОГИЯ **CUDA** В ПРИМЕРАХ

CUDA – вычислительная архитектура, разработанная компанией NVIDIA и предназначенная для разработки параллельных программ. В сочетании с развитой программной платформой архитектура CUDA позволяет программисту задействовать невероятную мощь графических процессоров для создания высокопроизводительных приложений, включая научные, инженерные и финансовые приложения. Книга написана двумя старшими членами команды группы по разработке программной платформы CUDA. Новая технология представлена в ней с точки зрения программиста. Авторы рассматривают все аспекты разработки на CUDA, иллюстрируя изложение работающими примерами. После краткого введения в саму платформу и архитектуру CUDA, а также беглого обзора языка CUDA C начинается подробное обсуждение различных функциональных возможностей CUDA и связанных с ними компромиссов. Вы узнаете, когда следует использовать то или иное средство и как писать программы, демонстрирующие поистине выдающуюся производительность.

ОСНОВНЫЕ ТЕМЫ:

- параллельное программирование;
- взаимодействие нитей;
- константная память и события;
- текстурная память;
- интероперабельность с графикой;
- атомарные операции;
- потоки CUDA;
- использование CUDA C при наличии нескольких ГП;
- дополнительные ресурсы, относящиеся к CUDA.

*“Эта книга –
обязательное чтение
для всех работающих
с вычислительными
системами,
содержащими
ускорители.”*

— из предисловия
Джека Донгарра,
Университет
штата Теннеси,
национальная
лаборатория Оух Ридж

Все программное обеспечение, которое необходимо для работы с книгой, можно бесплатно скачать с сайта компании NVIDIA.

ДЖЕЙСОН САНДЕРС (JASON SANDERS) – старший инженер-программист в группе CUDA Platform в компании NVIDIA. Принимал участие в разработке первых версий системы CUDA и в работе над спецификацией OpenCL 1.0, промышленном стандарте гетерогенных вычислений. Ранее работал в ATI Technologies, Apple и Novell.

ЭДВАРД КЭНДРОТ (EDWARD KANDROT) – старший инженер-программист в группе CUDA Algorithms в компании NVIDIA. Больше 20 лет занимается оптимизацией кода и повышением производительности различных программ, в том числе в компаниях Adobe, Microsoft, Google и Autodesk.

ISBN 978-5-94074-889-2



9 785940 748892 >



Addison-Wesley
Pearson Education

Internet-магазин
www.dmk-press.ru
Книга-почтой:
orders@aliens-kniga.ru
Оптовая продажа:
“Альянс-книга”
(499)725-5409
books@aliens-kniga.ru