

УДК 004.3'144:004.383.5CUDA
ББК 32.973.26-04
Р60

Рутш, Грегори.

Р60 CUDA Fortran для инженеров и научных работников. Рекомендации по эффективному программированию на языке CUDA Fortran / Г. Рутш, М. Фатика ; пер. с англ. А. А. Слинкина. — 2-е изд., эл. — 1 файл pdf : 365 с. — Москва : ДМК Пресс, 2023. — Систем. требования: Adobe Reader XI либо Adobe Digital Editions 4.5 ; экран 10". — Текст : электронный.

ISBN 978-5-89818-540-4

Fortran — один из важнейших языков программирования для высокопроизводительных вычислений, для которого было разработано множество популярных пакетов программ для решения вычислительных задач. Корпорация NVIDIA совместно с The Portland Group (PGI) разработали набор расширений к языку Fortran, которые позволяют использовать технологию CUDA на графических картах NVIDIA для ускорения вычислений.

Книга демонстрирует всю мощь и гибкость этого расширенного языка для создания высокопроизводительных вычислений. Не требуя никаких предварительных познаний в области параллельного программирования авторы скрупулезно шаг за шагом раскрывают основы создания высокопроизводительных параллельных приложений, попутно поясняя важные архитектурные детали современного графического процессора — ускорителя вычислений.

Издание предназначено для инженеров, научных работников, программистов, в также будет полезно студентам вузов соответствующих специальностей.

УДК 004.3'144:004.383.5CUDA
ББК 32.973.26-04

Электронное издание на основе печатного издания: CUDA Fortran для инженеров и научных работников. Рекомендации по эффективному программированию на языке CUDA Fortran / Г. Рутш, М. Фатика ; пер. с англ. А. А. Слинкина. — Москва : ДМК Пресс, 2014. — 364 с. — ISBN 978-5-97060-065-8. — Текст : непосредственный.

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.

В соответствии со ст. 1299 и 1301 ГК РФ при устранении ограничений, установленных техническими средствами защиты авторских прав, правообладатель вправе требовать от нарушителя возмещения убытков или выплаты компенсации.

ISBN 978-5-89818-540-4

© 2014 Gregory Ruetsch/NVIDIA Corporation and
Massimiliano Fatica/NVIDIA Corporation. Published
by Elsevier Inc. All rights reserved.
© Оформление, перевод на русский язык, издание,
ДМК Пресс, 2014



ОГЛАВЛЕНИЕ

| | |
|----------------------------|-----------|
| Благодарности | 10 |
| Предисловие | 11 |

ЧАСТЬ I

| | |
|---|-----------|
| Программирование на CUDA Fortran | 13 |
|---|-----------|

| | |
|--------------------------------|-----------|
| Глава 1. Введение | 14 |
|--------------------------------|-----------|

| | |
|---|----|
| 1.1. Краткая история вычислений на GPU | 14 |
| 1.2. Параллельные вычисления | 16 |
| 1.3. Основные идеи | 17 |
| 1.3.1. Первая программа на CUDA Fortran | 18 |
| 1.3.2. Обобщение на большие массивы | 22 |
| 1.3.3. Многомерные массивы | 25 |
| 1.4. Определение возможностей и ограничений оборудования с поддержкой CUDA | 27 |
| 1.5. Обработка ошибок | 38 |
| 1.6. Компиляция программы на CUDA Fortran | 39 |
| 1.6.1. Раздельная компиляция | 43 |

| | |
|---|-----------|
| Глава 2. Измерение производительности и метрики производительности | 48 |
|---|-----------|

| | |
|--|----|
| 2.1. Измерение времени выполнения ядра | 48 |
| 2.1.1. Синхронизация хоста и устройства и таймеры CPU | 49 |
| 2.1.2. Хронометраж с помощью событий CUDA | 50 |
| 2.1.3. Командный профилировщик | 51 |
| 2.1.4. Профилировщик nvprof | 53 |
| 2.2. Ядра, производительность которых, ограничена вычислениями, пропускной способностью памяти и задержкой | 54 |
| 2.3. Пропускная способность памяти | 58 |
| 2.3.1. Теоретически максимальная пропускная способность | 58 |
| 2.3.2. Эффективная пропускная способность | 60 |

Глава 3. Оптимизация..... 63

| | |
|--|-----|
| 3.1. Передача данных между хостом и устройством..... | 63 |
| 3.1.1. Зафиксированная область памяти | 64 |
| 3.1.2. Объединение мелких операций передачи в один пакет | 69 |
| 3.1.3. Асинхронная передача данных (дополнительная тема)..... | 72 |
| 3.2. Память устройства | 83 |
| 3.2.1. Объявление данных в коде, выполняемом на устройстве | 85 |
| 3.2.2. Объединенный доступ к глобальной памяти | 85 |
| 3.2.3. Текстовая память..... | 99 |
| 3.2.4. Локальная память..... | 105 |
| 3.2.5. Константная память | 108 |
| 3.3. Внутрикристалльная память..... | 112 |
| 3.3.1. L1-кэш..... | 112 |
| 3.3.2. Регистры | 113 |
| 3.3.3. Разделяемая память | 115 |
| 3.4. Пример оптимизации работы с памятью: | |
| транспонирование матрицы | 122 |
| 3.4.1. Недогрузка разделов (дополнительная тема)..... | 128 |
| 3.5. Конфигурация выполнения | 133 |
| 3.5.1. Параллелизм на уровне потоков..... | 133 |
| 3.5.2. Параллелизм на уровне команд..... | 137 |
| 3.6. Оптимизация команд | 140 |
| 3.6.1. Встроенные функции устройства | 141 |
| 3.6.2. Флаги компилятора..... | 142 |
| 3.6.3. Расходящиеся варпы | 143 |
| 3.7. Директивы генерации ядра из цикла | 144 |
| 3.7.1. Редукция в CUF-ядрах | 147 |
| 3.7.2. Потоки CUDA в CUF-ядрах | 147 |
| 3.7.3. Параллелизм на уровне команд в CUF-ядрах..... | 148 |

Глава 4. Программирование компьютера с несколькими GPU 150

| | |
|--|-----|
| 4.1. Средства CUDA для работы с несколькими GPU | 150 |
| 4.1.1. Связь между равноправными устройствами..... | 152 |
| 4.1.2. Прямая передача данных между равноправными устройствами..... | 157 |
| 4.1.3. Транспонирование матрицы с применением равноправного доступа..... | 168 |
| 4.2. Программирование нескольких GPU с применением библиотеки MPI | 177 |
| 4.2.1. Сопоставление устройств рангам MPI | 179 |
| 4.2.2. Транспонирование матрицы с применением MPI | 186 |
| 4.2.3. Транспонирование матрицы с применением MPI, поддерживающей GPU..... | 188 |

ЧАСТЬ II**Примеры задач 191****Глава 5. Метод Монте-Карло 192**

5.1. Библиотека CURAND 193

5.2. Вычисление π с помощью CUF-ядер 198

5.2.1. Стандарт IEEE-754 (дополнительная тема) 202

5.3. Вычисление π с помощью ядер редукции 2055.3.1. Редукция с атомарными блокировками
(дополнительная тема) 211

5.4. Точность суммирования 213

5.5. Опционное ценообразование 220

Глава 6. Метод конечных разностей 2296.1. Девятиточечный шаблон конечно-разностной схемы
для вычисления первой производной 2296.1.1. Повторное использование данных и разделяемая
память 2316.1.2. Ядро производной по x 2326.1.3. Производные по y и z 237

6.1.4. Неравномерные сетки 242

6.2. Двумерное уравнение Лапласа 246

**Глава 7. Приложения быстрого преобразования
Фурье 254**

7.1. Библиотека CUFFT 254

7.2. Спектральное дифференцирование 263

7.3. Свертка 267

7.4. Решение уравнения Пуассона 276

ЧАСТЬ III**Приложение 283****Приложение А. Технические характеристики
Tesla 284****Приложение В. Управление системой
и окружением 287**

В.1. Переменные окружения 287

В.1.1. Общие переменные окружения 287

В.1.2. Командный профилировщик 288

В.1.3. JIT-компиляция 288

| | |
|--|------------|
| В.2. Интерфейс управления системой nvidia-smi | 289 |
| В.2.1. Включение и выключение режима ECC..... | 290 |
| В.2.2. Режим вычислений | 292 |
| В.2.3. Инерционный режим..... | 293 |
| Приложение С. Вызов CUDA C из CUDA Fortran..... | 295 |
| С.1. Вызовы библиотеки, написанной на CUDA C..... | 295 |
| С.2. Вызов написанной пользователем функции на CUDA C... | 298 |
| Приложение D. Исходный код | 300 |
| D.1. Текстурная память | 300 |
| D.2. Транспонирование матрицы | 304 |
| D.3. Параллелизм на уровне потоков и команд | 311 |
| D.4. Программирование с использованием нескольких GPU | 315 |
| D.4.1. Транспонирование с применением равноправного доступа к памяти..... | 316 |
| D.4.2. Транспонирование с применением библиотеки MPI для передачи данных между хостами | 322 |
| D.4.3. Транспонирование с применением библиотеки MPI для передачи данных между устройствами | 327 |
| D.5. Программирование метода конечных разностей | 332 |
| D.6. Решение уравнения Пуассона спектральным методом ... | 352 |
| Литература..... | 357 |
| Предметный указатель | 359 |