NVIDIA

CUDA© АЛЬМАНАХ

МАРТ 2014

СОДЕРЖАНИЕ

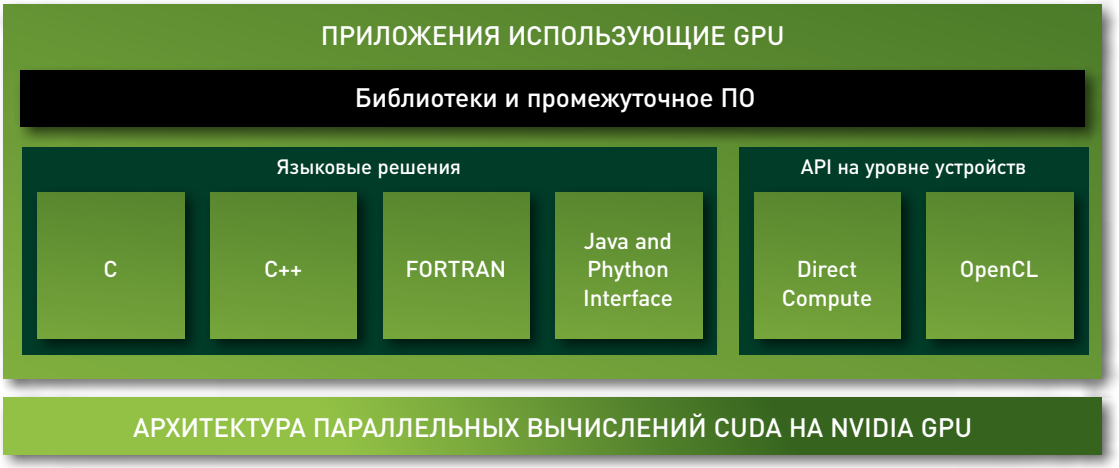
ЧТО ТАКОЕ CUDA АЛЬМАНАХ?

CUDA АЛЬМАНАХ — это периодическое издание от NVIDIA, содержащее научные работы, в которых используется архитектура паралельных вычислений CUDA.

CUDA используется в различных областях, включая обработку видео и изображений, вычислительную биологию и химию, моделирование динамики жидкостей, восстановление изображений, полученных путем компьютерной томографии, сейсмический анализ, трассировку лучей и многое другое.

Приложения, базирующиеся на архитектуре CUDA, можно разрабатывать на различных языках и API, включая C, C++, Fortran, OpenCL и directCompute. Архитектура CUDA подразумевает сотни ядер, способных исполнять тысячи параллельных потоков, а модель программирования CUDA позволяет программистам сосредоточиться на распараллеливании своих алгоритмов.

Архитектура CUDA текущего поколения под названием Kepler — это самая передовая архитектура вычислений на GPU. Построенные на свыше семи миллиардов транзисторах, GPU Kepler делают универсальными вычисления на GPU и CPU для широкого спектра вычислительных приложений. Поддержка C++ упрощает разработку ПО для параллельных вычислений и повышает производительность широчайшего спектра приложений.



Архитектура параллельных вычислений CUDA с комбинацией ПО и аппаратной части.

Всего за несколько лет вокруг архитектуры CUDA возникла целая экосистема программного обеспечения — от различных языковых решений до широкого спектра библиотек, компиляторов и связующего ПО, которые помогают пользователям оптимизировать приложения для GPU. Разнообразие оптимизированных программных средств ускоряет научные открытия и расчет моделей во многих областях, включая математику, бионауки и производство.

Подробнее

НОВОСТИ NVIDIA CUDA

ДЕНЬ ТЕХНОЛОГИЙ NVIDIA В МГУ ИМ. ЛОМОНОСОВА – 10 АПРЕЛЯ 2014

10 апреля Московский Государственный Университет имени М.В.Ломоносова, NVIDIA, факультет ВМК, кафедра Суперкомпьютеров и квантовой информатики проводят День Технологий NVIDIA.

День Технологий NVIDIA — это уникальная возможность узнать о решениях компании NVIDIA в области компьютерных игр и визуализации, инновационных мобильных технологий и параллельных вычислений. У вас будет возможность пообщаться со специалистами NVIDIA, узнать о возможностях работы в компании, поучаствовать в демонстрации технологий, а также выиграть полезные призы.

> **Место проведения школы**:
> 2-й Гуманитарный корпус МГУ.

ПРОГРАММА МЕРОПРИЯТИЯ:

16:00–16:30	Приветственный кофе
16:30–16:40	Вступительное слово
16:40–17:10	Инновационные решения и технологии NVIDIA
17:10–17:40	Возможности работы в NVIDIA
17:40–18:10	Технологии компьютерных игр
18:10–18:40	Кофе-брейк с демонстрацией технологий NVIDIA
18:40–19:00	Решения NVIDIA для высокопроизводительных вычислений
19:00–19:30	Вопросы, ответы и розыгрыш призов от NVIDIA

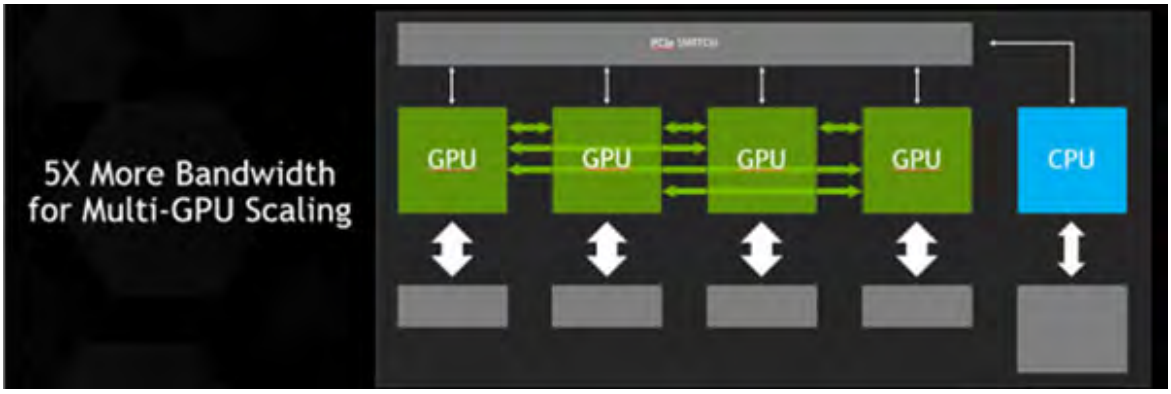
РЕГИСТРАЦИЯ ОБЯЗАТЕЛЬНА

* Регистрация открыта до 7-го апреля
Участие бесплатное

NVIDIA ПРЕДСТАВЛЯЕТ ВЫСОКОСКОРОСТНОЙ ИНТЕРФЕЙС ДЛЯ GPU, ОТКРЫВАЯ ДВЕРЬ В МИР ЭКЗАФЛОПСНЫХ ВЫЧИСЛЕНИЙ

NVIDIA анонсировала новый высокоскоростной интерфейс NVIDIA® NVLink™который появится в будущих графических процессорах, что ускорит передачу данных между GPU и CPU в 5-12 раз. Таким образом, новый интерфейс откроет путь к новому поколению экзафлопсных суперкомпьютеров, которые станут в 50-100 быстрее самых мощных систем сегодняшнего дня.

NVIDIA добавит технологию NVLink в свою новую графическую архитектуру Pascal, которая придет на смену архитектуре NVIDIA Maxwell . Ее выпуск на рынок запланирован на 2016 год. Новый интерфейс был разработан совместно с компанией IBM, которая включит его в будущие версии своих процессоров POWER.

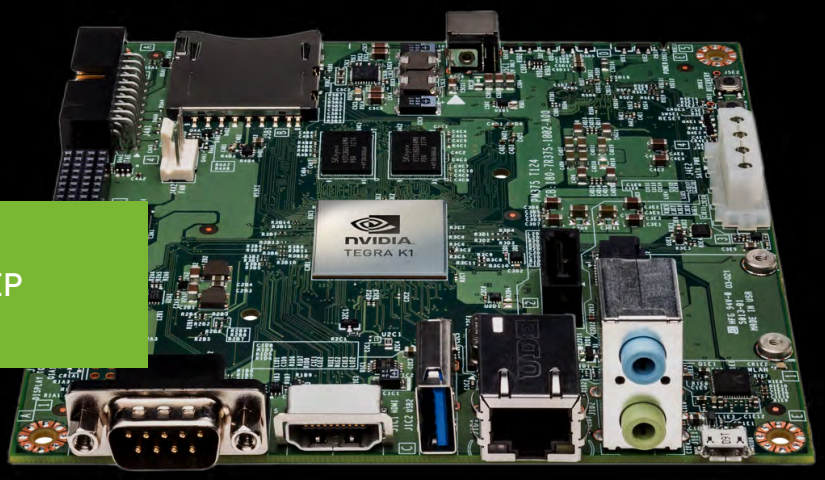


“Технология NVLink полностью раскрывает потенциал GPU, значительно улучшая передачу данных между CPU и GPU и сокращая время ожидания окончания обработки данных для графического процессора”, — говорит Брайан Келлехер (Brian Kelleher), старший вице-президент по разработке GPU в NVIDIA.

“NVLink обеспечивает быстрый обмен данными между CPU и GPU, таким образом улучшая пропускную способность вычислительной системы и устраняя ключевую проблему сегодняшних ускоренных вычислений, — говорит Брэдли МакКреди (Bradley McCredie), вице-президент IBM. — С NVLink разработчикам будет легче модифицировать высокопроизводительные приложения и программы анализа данных под ускоренные системы на базе CPU и GPU. Мы считаем, что эта технология вносит весомый вклад в нашу экосистему OpenPOWER”.

Благодаря технологии NVLink, крепко связывающей CPU IBM POWER и GPU NVIDIA Tesla®, ЦОД POWER смогут максимально использовать потенциал GPU для различных приложений, таких как высокопроизводительные вычисления, анализ данных и машинное обучение.

NVIDIA ПРЕДСТАВЛЯЕТ ПЕРВЫЙ МОБИЛЬНЫЙ СУПЕРКОМПЬЮТЕР ДЛЯ ВСТРАИВАЕМЫХ СИСТЕМ



NVIDIA открывает дверь в мир приложений нового поколения, использующих машинное зрение, обработку изображений и данных в реальном времени, с представлением платформы для разработчиков на базе первого в мире мобильного суперкомпьютера для встраиваемых систем.

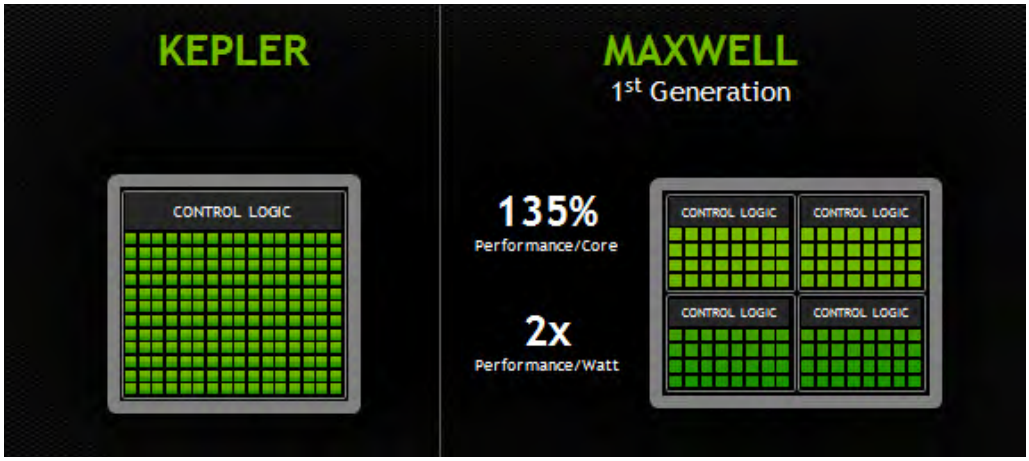
Платформа NVIDIA® Jetson™ TK1 предоставляет разработчикам инструменты для создания систем и приложений, которые обеспечат роботам плавное управление, позволят врачам проводить мобильное УЗИ, беспилотникам избегать движущихся объектов, а автомобилям определять пешеходов.

С несравненной производительностью в 326 гигафлопс, что почти втрое выше, чем у аналогичного класса встраиваемых платформ, набор инструментов разработчика Jetson TK1 включает полноценный набор для C/C++ на базе архитектуры NVIDIA CUDA®, самой популярной платформы и модели программирования параллельных вычислений. Представленная платформа намного упрощает разработку приложений по сравнению с процессорами FPGA, ASIC и DSP, широко используемыми в сегодняшних встраиваемых системах.

5 ВЕЩЕЙ, КОТОРЫЕ НЕОБХОДИМО ЗНАТЬ РАЗРАБОТЧИКУ О НОВОЙ ГРАФИЧЕСКОЙ АРХИТЕКТУРЕ MAXWELL

Недавний анонс графических процессоров NVIDIA на архитектуре первого поколения Maxwell — это крайне волнующий момент в индустрии GPU-вычислений. Первые продукты на архитектуре Maxwell, такие, как GeForce GTX 750 Ti, основаны на чипе GM107 и предназначены для использования в малопотребляющих устройствах — ноутбуках, компактных компьютерах и не только. Ключевым моментом Maxwell для разработчиков HPC и других GPU-приложений является большой скачок в энергоэффективности: почте вдвое по сравнению с архитектурой Kepler, что делает Maxwell отличной базой для будущих продуктов в линейке NVIDIA Tesla.

В этом посте мы расскажем о пяти главных вещах про Maxwell, которые следует знать разработчику GPU-приложений. Среди них — преимущества архитектуры, специфика нового потокового процессора Maxwell, руководства по настройке и ссылки на дополнительные ресурсы.



1 СЕРДЦЕ MAXWELL: БОЛЕЕ ЭФФЕКТИВНЫЕ МУЛЬТИПРОЦЕССОРЫ

Потоковый процессор (SM) в Maxwell — его называют SMM — создан с нуля и обладает значительно большей энергоэффективностью по сравнению с предшественниками. Стоит отметить, что Kepler SMX был достаточно эффективен для своего поколения. В результате его создания инженеры NVIDIA увидели новые возможности в повышении эффективности архитектуры GPU, которые и были реализованы в SM Maxwell. Улучшения коснулись механизмов распределения управляющей логики и нагрузки, гранулярности алгоритмов энергосбережения, планирования инструкций и количества исполняемых инструкций за такт, а также многих других аспектов, позволивших SM Maxwell намного опередить Kepler SMX по эффективности. Новая архитектура SM Maxwell позволила увеличить количество SM до пяти в GM107, в отличие от двух в GK107, при увеличении площади матрицы всего на 25%.

Улучшенное планирование инструкций

Количество ядер CUDA в одном SM сократилось, однако с учетом возросшей эффективности исполнения в Maxwell (прирост производительность в расчете на SM составляет в пределах 10% от производительности Kepler) и более эффективных размеров SM, общее число ядер CUDA на GPU будет намного выше, чем у Fermi и Kepler. В Maxwell SM осталось то же самое количество планировщиков инструкций и уменьшены задержки на арифметических операциях по сравнению с Kepler.

Как и в SMX, в каждом SMM есть четыре warp-планировщика, но в отличие от SMX, все ключевые функциональные блоки SMM привязаны к определенному планировщику, а не делятся между ними. Количество ядер на один раздел теперь равно степени двойки, что упрощает планирование — каждый планировщик использует свой собственный набор ядер количеством равным размеру warp`a. Warp-планировщик может по-прежнему за один такт выполнять две инструкции (например, выполняя математическую операцию на CUDA-ядрах одновременно с выполнением операции обращения к памяти в блоке load/store), однако теперь можно полностью загрузить CUDA-ядра даже если планировщик отправляет на выполнение по одной инструкции.

Увеличенная загрузка потоковых процессоров

SMM по многим аспектам похож на SMX архитектуры Kepler, при этом ключевые изменения нового типа процессоров направлены на повышение эффективности без необходимости значительного увеличения параллелизма в расчете на SM в приложении. Размер регистрового файла (64K 32-битных регистров), максимально количество warp`ов на SM (64 warp`a) и максимальное количество регистров (255 регистров) остались прежними. Максимальное количество блоков на потоковый мультипроцессор SMM удвоилось до 32, что должно привести к автоматическому увеличению загрузки для ядер, которые использую малый размер блока — 64 или меньше — в предположении, что регистры и разделяемая память не ограничивают загрузку мультипроцессора. В таблице 1 представлены в сравнении ключевые характеристики Maxwell GM107 и предшественника Kepler GK107.

Уменьшены задержки при выполнении арифметических инструкций

Еще одним значительным преимуществом SMM является уменьшение задержек выполнения арифметических инструкций. Так как загрузка мультипроцессора (которая преобразуется в параллелизм на уровне warp`ов) у SMM такая же или лучше, чем у SMX, сокращенные задержки улучшают использование CUDA-ядер и повышают скорость работы ядра.

2 УВЕЛИЧЕННАЯ ВЫДЕЛЕННАЯ ОБЩАЯ ПАМЯТЬ

В архитектуре Maxwell предусмотрено 64 кбайт разделяемой памяти, в то время как в Fermi или Kepler эта память делится между L1-кэшом и разделяемой памятью. В Maxwell один блок по-прежнему может использовать до 48 кбайт разделяемой памяти, причем увеличение общего объема разделяемой памяти может привести к увеличению загрузки мультипроцессора. Это стало возможным благодаря объединению функциональности L1-кэша и текстурного-кэша в отдельном блоке.

3 БЫСТРЫЕ АТОМАРНЫЕ ОПЕРАЦИИ В РАЗДЕЛЯЕМОЙ ПАМЯТИ

В архитектуре Maxwell появились встроенные атомарные операции над 32-битными целыми числами в разделяемой памяти, а также CAS-операции над 32-битными и 64-битными значениями в разделяемой памяти — с помощью них можно реализовать другие атомарные функции. В случае Kepler и Fermi приходилось использовать сложный принцип "Lock/Update/Unlock", что приводило к дополнительным издержкам.

4 ДИНАМИЧЕСКИЙ ПАРАЛЛЕЛИЗМ

Динамический параллелизм, появившийся в Kepler GK110, позволяет GPU самому создавать задачи для себя. Поддержка этой функции была впервые добавлена в CUDA 5.0, позволяя нитям на GK110 запускать дополнительные ядра на том же GPU.

Теперь динамический параллелизм поддерживается во всей продуктовой линейке, включая даже такие экономичные чипы, как GM107. Разработчикам это на руку, так как теперь для приложений не требуется создавать специальные алгоритмы для high-end GPU, отличающиеся от тех, которые используются на графических процессорах более низкого уровня.

5 ПОДРОБНЕЕ О ПРОГРАММИРОВАНИИ MAXWELL

Подробнее об архитектуре и оптимизации кода под Maxwell смотрите в Руководстве по настройке Maxwell и Руководстве по совместимости Maxwell, которые уже доступны для зарегистрированных разработчиков CUDA. Авторизуйтесь или бесплатно вступайте в сообщество уже сегодня.

ПРЕДЛОЖЕНИЯ ОТ NVIDIA

CUDA И OPENACC – БЕСПЛАТНЫЙ ОНЛАЙН КУРС

NVIDIA предлагает вам пройти бесплатные online курсы по программированию массивно-параллельных процессоров. Пройдя предлагаемый курс, вы получите широкий спектр практических навыков, которые позволят Вам к концу занятий овладеть основами программирования современных графических процессоров (GPU) NVIDIA, а также ознакомитесь с директивным программированием GPU ускорителей (стандарт OpenACC) и особенностями использования нескольких GPU видеокарт для решения Ваших задач. Тем, кто продемонстрирует высокие показатели при прохождении заданий курса, предлагается также бесплатная сертификация по программированию GPU, поддерживающих технологию CUDA в учебном центре Applied Parallel Computing.

КРАТКОЕ СОДЕРЖАНИЕ КУРСА:

Лекция 1. Введение в CUDA.
Лекция 2. Модель исполнения CUDA.
Лекция 3. Иерархия памяти. Глобальная, локальная и регистровая память.
Лекция 4. Иерархия памяти. Разделяемая память.
Лекция 5. Прикладные CUDA библиотеки.
Лекция 6. Библиотека Thrust.
Лекция 7. Оптимизация CUDA программ.
Лекция 8. Стандарт директивного программирования OpenACC.

ОНЛАЙН КУРС

Applied Parallel Computing

ЗАРЕГИСТРИРОВАТЬСЯ

УСКОРЯЙТЕ ВАШИ НАУЧНЫЕ ПРИЛОЖЕНИЯ С OPENACC

БЕСПЛАТНАЯ ЛИЦЕНЗИЯ ОТ PGI НА 30 ДНЕЙ

Получив доступ к бесплатной 30-дневной версии компилятора PGI, вы сможете воспользоваться вычислительными мощностями GPU и стандартом программирования OpenACC.

OpenACC — это:

- **Легкость:** простота добавления директив в исходный код своей программы.
- **Открытость:** используйте единый исходный код как для CPU так и для GPU.
- **Мощность:** получите быстрый доступ к вычислительной мощности GPU.

ВЕРСИЯ 2014 УЖЕ ДОСТУПНА!

ЗАРЕГИСТРИРОВАТЬСЯ

ПРОВЕДИТЕ ТЕСТ-ДРАЙВ УСКОРИТЕЛЯ TESLA K20/K40 GPU

Воспользуйтесь нашим предложением провести простой и бесплатный тест-драйв ускорителей Tesla K20/K40 GPU.

Самые быстрые в мире ускорители Tesla K20/K40 GPU созданы на основе архитектуры Kepler и обеспечивают высокую производительность и энергоэффективность ваших приложений.

ЗАРЕГИСТРИРОВАТЬСЯ

НАУЧНЫЕ РАБОТЫ С ИСПОЛЬЗОВАНИЕМ ВЫЧИСЛЕНИЙ НА CUDA

ЧИСЛЕННОЕ МОДЕЛИРОВАНИЕ РАСПРОСТРАНЕНИЯ СЕЙСМИЧЕСКИХ ВОЛН В СЛОЖНО ПОСТРОЕННЫХ СРЕДАХ НА ГИБРИДНОМ КЛАСТЕРЕ

А.Ф. Сапетина

Институт вычислительной математики и математической геофизики СО РАН, Новосибирск

ПОДРОБНЕЕ

Одним из путей изучения строения геологических объектов является активный вибросейсмический мониторинг. Проведение натурных геофизических экспериментов позволяет получить некоторые представления о скоростных параметрах упругой среды, а также о геометрии изучаемого объекта. Но в процессе обработки результатов полевого эксперимента могут возникнуть интересные эффекты, требующие дальнейшего исследования. Создание математических моделей изучаемых объектов и дальнейшее моделирование сейсмических полей помогают в изучении реальных объектов. В связи с тем, что реальная область исследования зачастую имеет довольно сложный рельеф, не всегда удается поставить площадную систему наблюдения для решения обратной задачи геофизики. Поэтому приходится решать набор прямых задач с целью определения параметров изучаемой среды, соответствующих экспериментальным наблюдениям на поверхности изучаемых грязевых вулканов. Программный комплекс, решающий поставленную задачу на основе выбранного математического аппарата, должен иметь возможность моделирования упругой неоднородной среды со специфической геометрией. В связи со сложностью и масштабом моделируемой области, решение задачи численного моделирования распространения упругих волн от сосредоточенного источника может требовать значительных вычислительных ресурсов. Поэтому необходима разработка параллельных программ для уменьшения времени расчета и возможностью моделирования «больших» 3D-моделей упругих сред.

Д.А. Караваевым были разработаны и обоснованы методы численного моделирования сейсмических полей для 3D сложно построенных сред, характерных для грязевых вулканов. Им же разработан инструментарий для решения прикладных задач численного моделирования сейсмических полей, включающий построитель 3D-моделей неоднородных упругих сред и параллельную программу для численного моделирования распространения упругих волн, реализованную на кластерах с MPP-архитектурой, с использованием технологий MPI и OpenMP.

В свете приведенных фактов стала актуальной адаптация существующих и создание новых алгоритмов и программного обеспечения для решения задачи численного моделирования распространения упругих волн на гибридном кластере.

ОРГАНИЗАЦИЯ ПОИСКА ДАННЫХ В СУПЕРКОМПЬЮТЕРНЫХ СИСТЕМАХ НА БАЗЕ TOSQL-ПОДХОДА И ТЕХНОЛОГИИ NVIDIA CUDA

Н. П. Васильев, М. М. Ровнягин

Национальный Суперкомпьютерный Форум (НСКФ-2013)

ПОДРОБНЕЕ

В настоящее время для решения множества промышленных задач используются информационно-аналитические системы на основе классических SQL-решений. Однако, вследствие возрастания общего объема хранимой информации требуется все большее число вычислительных ресурсов для ее обработки Так, например, в одном из филиалов крупнейшего оператора мобильной связи в мире — китайской компании ChinaMobile - в 2006 году для предсказания поведения пользователей с помощью клиентской базы этого филиала использовалось коммерческое решение, включающее сервер с 8 процессорными ядрами, 32 Гб оперативной памяти и дисковым массивом хранения данных. Этот сервер был способен анализировать лишь 10% от получаемой ежедневно информации, и при этом имел стоимость порядка 1 млн. долларов США (с учетом стоимости ПО и расходов на сопровождение).

Размер файлов, байт	Простая замена		
	GPU ms	CPU ms	Прирост раз
80	0,822	0	*нет*
3056428	6,878	180	*26,172*
6631243	12,654	403	*31,847*
121097812	196,013	7140	*36,426*
337166183	576,618	20020	*34,720*

Прирост производительности по сравнению с последовательной реализацией

Для увеличения общей пропускной способности различных информационно-аналитических систем, а также ускорения процесса разработки распределенных приложений начиная с 2008 года крупнейшие мировые компании, такие как Yahoo, Last.fm, Facebook, TheNewYorkTimes используют технологию Hadoop. Она базируется на основе концепции MapReduce в соответствие с которой данные хранятся на узлах вычислительного кластера распределенным образом, перераспределяясь во время вычислений, чтобы снизить число межузловых передач данных. В целом, можно говорить о том, что работает концепция перемещения обрабатывающего кода к данным, на не наоборот, как это делалось ранее. Для хранения информации и обеспечения доступа к файлам применяется распределенная файловая система HDFS. Основным языком разработки параллельных приложений является Java. Применение технологии Hadoop позволило сократить расходы на наладку и поддержание работоспособности вычислительного кластера филиала ChinaMobile в 6 раз. Вместе с этим производительность Hadoop-решения оказалась почти на два порядка выше.

АЛГОРИТМ ПАРАЛЛЕЛЬНОЙ АГРЕГАЦИИ ДАННЫХ ДЛЯ ВИЗУАЛИЗАЦИИ ДАННЫХ О ВЕРБАЛЬНОМ И НЕВЕРБАЛЬНОМ ПОВЕДЕНИИ ЧЕЛОВЕКА

Б.А. Князев

Инженерный журнал: наука и инновации, 2013, вып. 11

ПОДРОБНЕЕ

Вербальное и невербальное поведения могут рассматриваться как процессы, изменяющиеся во времени. Для решения таких задач, как безопасность, медицинская и психологическая диагностика, робототехника и др., необходима объективная оценка параметров данных процессов. Оценка может осуществляться с помощью интерпретации этих параметров в виде временных и частотных графиков. При этом частота движений частей тела и элементов лица не превышает 10...12 Гц (⩽ 12 Гц для пальцев рук, ⩽ 10 Гц для жестов рук и движений тела в целом и ⩽ 4 Гц для изменения мимики лица); около 90 % энергетической составляющей речевого сигнала находится в диапазоне 100...5000 Гц. Таким образом, из теоремы Котельникова следует, что для исключения значительных потерь исходного сигнала частота дискретизации исследуемых в данной работе невербальных и вербальных сигналов должна быть ⩾ 25 кадров/с и ⩾ 10 КГц соответственно.

Длительность исследований, записанных на видео-и/или аудионосители, может достигать нескольких часов. Следовательно, количество отсчетов данных для визуализации N' равно объему данных N:

$N = 3600\,F\,L$ точек,

где L — длительность исследования, ч; F — частота отсчетов, с.

Для эффективного (с частотой ⩾ 10 операций в секунду) обзора данных необходимо снижение их размерности с помощью либо методов аппроксимации кривой, либо агрегирования данных. При этом под агрегированием в общем случае понимается объединение нескольких элементов в единое целое. При размере данных, достигающих 109 отсчетов, для выполнения быстрого масштабирования и фильтрации данных также необходимы высокоэффективные агрегирующие функции или специальные масштабируемые типы данных.

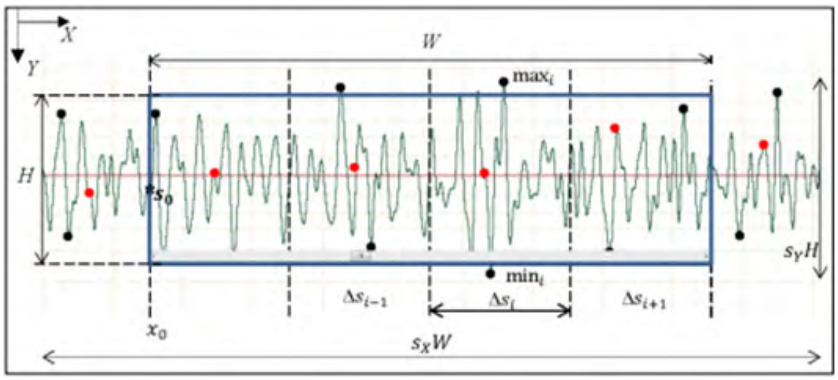


График исходных данных $S{:}W{\times}H$ — видимая область отображения графика (черными точками показаны отсчеты, которые следует выбрать из соответствующих блоков данных; красными – неточно выбранные отсчеты)

Для задачи агрегирования также могут быть использованы методы параллельных вычислений, поскольку:

- предназначены для вычисления независимых блоков данных;
- графические процессоры (GPU) имеют необходимую для данной задачи пропускную способность (до 63,4 Гб/с для процессора G80);
- при корректном использовании ресурсов GPU значительно (до 100 раз) превосходят производительность центрального процессора (CPU).

БИБЛИОТЕКА PRAND: ГЕНЕРАЦИЯ ПАРАЛЛЕЛЬНЫХ ПОТОКОВ СЛУЧАЙНЫХ ЧИСЕЛ ДЛЯ РАСЧЕТОВ МОНТЕ-КАРЛО С ИСПОЛЬЗОВАНИЕМ GPU

Л.Ю. Бараш, Л.Н. Щур

Computer Physics Communications
Volume 185, Issue 4, April 2014, Pages 1343–1353

ПОДРОБНЕЕ

Разработаны библиотека RNGSSELIB и библиотека PRAND по параллельной генерации псевдослучайных чисел для расчетов Монте-Карло. Библиотека RNGSSELIB содержит только реализации для CPU и не требует при использовании наличия GPU или компилятора CUDA, в то время как библиотека PRAND содержит все разработанные реализации и предназначена для использования с CUDA версии 5.0 или более поздней. В библиотеку включены тщательно отобранные современные и надежные генераторы случайных чисел, обладающие лучшими показателями по совокупности важнейших свойств, таких как, в частности, длина периода, размерность, для которой выполняется равнораспределение вероятности, скорость работы генератора, результаты статистического тестирования, возможность инициализации параллельных потоков. Именно, в библиотеку включены генераторы, основанные на параллельной эволюции автоморфизмов тора (GM19, GM31, GM61, GM29.1, GM55.4, GQ58.1, GQ58.3, GQ58.4), генератор MRG32K3A, генератор LFSR113, генератор MT19937. Использование массивного параллелизма современных GPU и SIMD-параллелизма современных CPU значительно увеличивает производительность генераторов.

Для каждого из генераторов реализованы:

- инициализация параллельных потоков методом расщепления блока; имеется возможность инициализировать до 10^{19} независимых потоков.
- эффективные версии для CPU с использованием SIMD-параллелизма и SSE-команд;
- однонитевые версии, которые можно использовать в вычислениях Монте-Карло на графических процессорах, распределенных по нитям и вычислительным узлам произвольным образом;
- параллельные версии с использованием множества нитей графического процессора для ускорения вычислений.
- Поддержка Си и Фортрана; имеются примеры использования на Фортране и на Си.

Библиотека протестирована, в частности, на суперкомпьютере «Ломоносов» НИВЦ МГУ им. М.В. Ломоносова для многомерного численного интегрирования методом Монте-Карло с одновременным использованием более 750 графических процессоров.

ПАРАЛЛЕЛЬНЫЙ ПРОГРАММНЫЙ КОМПЛЕКС ДЛЯ РЕШЕНИЯ ЗАДАЧ ГАЗОВОЙ ДИНАМИКИ В ОБЛАСТЯХ СО СЛОЖНОЙ ГЕОМЕТРИЕЙ НА СОВРЕМЕННЫХ ГИБРИДНЫХ ВЫЧИСЛИТЕЛЬНЫХ СИСТЕМАХ

П.В. Павлухин, И.С. Меньшов

Суперкомпьютерные технологии в науке, образовании и промышленности
Архив заседний семинара 18 марта 2014 г.

ПОДРОБНЕЕ

В докладе представлена параллельная реализация метода LU-SGS на Multi-GPU системах. Этот метод используется в программном комплексе для решения трехмерных задач газовой динамики в рамках модели уравнений Навье-Стокса. Как правило, на GPU реализуются методы на основе явных схем с простыми вычислительными ядрами, условие устойчивости для которых накладывает достаточно сильные ограничения на выбор временного шага интегрирования с ростом сеточного разрешения расчетной области. Неявные схемы позволяют ослабить это ограничение для некоторых классов задач, но построение и реализация параллельных алгоритмов, особенно под массивно-параллельные архитектуры GPU, для них намного сложнее.

Одна из важных особенностей рассматриваемого параллельного алгоритма — точное соблюдение работы последовательного прототипа для неявной схемы с масштабируемостью до нескольких сотен узлов. Представленный алгоритм позволяет совмещать во времени вычислительную работу с обменом данными между графическими ускорителями, обеспечивая при этом корректность и повторяемость получаемого решения. Его реализация с помощью CUDA и MPI позволила получить близкий к линейному рост производительности при использовании более 700 графических ускорителей. Расчетная область представляется несвязной декартовой сеткой, а для описания сложной геометрии на ней используется метод погруженной границы с введением специальной правой части в исходную систему уравнений.

ВАКАНСИИ CUDA

Вакансия: **Инженер по обработке медицинских изображений**

Компания: **Samsung Research Center**

Город: **Москва**

ОПИСАНИЕ

Обязанности старшего инженера по разработке алгоритмов обработки медицинских изображений включают, но не ограничиваются следующим списком:

- Решение прикладных задач обработки МРТ изображений;
- Цифровая обработка сигналов (спектральный анализ, итд);
- Compressive sensing, разреженное кодирование, обучение словаря;
- Алгоритмы восстановления изображений.

Должностные обязанности:

- Проведение прикладных исследований и разработки технологий в области обработки медицинских изображений;
- Работа в тесном сотрудничестве с другими членами исследовательской группы;
- Написание патентов, охватывающих разработанные технологии;
- Реализация разработанных алгоритмов;
- Презентация и демонстрация результатов исследований на внутренних (внешних) мероприятиях.

Квалификация:

- Отличные аналитические способности;
- Кандидатская степень (крайне желательно) в области технических или физико-математических наук;
- 3+ года опыта в области исследований и разработок в рамках широкой области обработки [медицинских] изображений;
- Хороший уровень математической подготовки;
- Опыт разработки на C++;
- Опыт в области разработки с использованием научных языков программирования (Matlab/Octave, Python w Numpy, Scipy);
- Хорошие навыки командной работы и коммуникации, энтузиазм, творческие способности, продуктивность и обучаемость.

Один или несколько из следующих пунктов являются сильным плюсом:

- Знание принципов работы МРТ;
- Навыки работы со специализированными библиотеками и программным обеспечением (ITK, VTK, Slicer, и т.п.);
- Опыт разработки с использованием GPGPU(CUDA, OpenCL);
- Опыт разработки библиотек для Linux.

Условия:

- Офис А класса в 15 мин пешком от м. Савеловская и м. Марьина Роща;
- Полное соблюдение ТК РФ;
- Гибкий график работы;
- Для аспирантов 1-го курса есть возможность 4-х дневной рабочей недели;
- Бесплатные вкусные обеды (карта в столовую на 5000 NET/месяц);
- ДМС; 100% оплата 7 дней больничного в году (в т.ч. без оформления больничного листа);
- Фитнес room (+ настольный теннис, настольный футбол), бассейн;
- Курсы английского (online and on-site);
- Возможны командировки в Корею;
- Компания оплачивает релокацию (оплачивается также дорога для прохождения собеседования);
- Премии за патенты, передовые идеи и проекты;
- Возможность карьерного роста (компания активно расширяется);
- Хорошая возможность углубления технических знаний (есть возможность посещение конференций, в т.ч. с докладами по проведенным исследованиям, обширная библиотека журналов и книг, есть возможность часть времени тратить на самообразование и открытые проекты);
- Возможность участвовать в обсуждении идей для будущих продуктов Samsung.

Подробнее

Вакансия: **Программист C++ (математика)**

Компания: **Ракурс, Фирма, ЗАО**

Город: **Москва, м. ВДНХ**

ОПИСАНИЕ

Обязанности:

- Разработка алгоритмов решения сложных геометрических и алгебраических задач в области фотограмметрии и обработки изображений.

Требования:

- Образование — высшее техническое по математической специальности, желательна ученая степень.
- Отличное знание линейной алгебры и геометрии, математического анализа, численных методов решения задач оптимизации.
- Опыт программирования на C++.
- Опыт разработки и реализации численных алгоритмов решения сложных математических задач.
- Приветствуются знания из Computer Vision, знание библиотек IPP, MKL, CUDA/OpenCL, OpenCV.
- Способность работать в команде.
- Технический английский.

Условия:

- Оформление в соответствии с ТК РФ
- Занятость: 40-часовая рабочая неделя, гибкий график
- В перспективе возможна удаленная работа
- Место работы: новое офисное здание в 2-х минутах ходьбы от м. ВДНХ
- Бесплатные обеды
- Ежегодный оплачиваемый отпуск
- Возможность участия в профильных семинарах и конференциях на территории России и за рубежом
- Дружный коллектив.

Подробнее

КОНТАКТЫ И ПОЛЕЗНЫЕ ССЫЛКИ

Если вы хотите, чтобы ваша статья появилась в следующем выпуске CUDA Альманах пишите нам на: landreeva@nvidia.com

По вопросам обучения CUDA обращайтесь в наш тренинговый центр: www.parallel-compute.ru

По вопросам приобретения NVIDIA GPU и по прочим техническим вопросам пишите нам на: adzhoraev@nvidia.com

Протестируйте PGI OpenACC compiler бесплатно в течение месяца: www.nvidia.ru/openacc

Узнайте подробнее про CUDA: www.nvidia.ru/cuda

Полный каталог приложений, ускоряемых на CUDA: http://www.nvidia.ru/gpuapps

Copyright © 2014 NVIDIA логотип NVIDIA, являются торговыми знаками или зарегистрированными торговыми знаками компании NVIDIA в США и других странах. Все другие названия компаний и/или продуктов могут являться товарными знаками и/или зарегистрированными товарными знаками соответствующих владельцев. Все права защищены.

NVIDIA Corporation, 2701 San Tomas Expressway, Santa Clara, CA 9505