Боресков А. В., Харламов А. А.

Основы работы с технологией CUDA

ДМК Пресс
издательство
Москва, 2010

УДК 32.973.26-018.2
ББК 004.4
Б82

Б82 **Боресков А. В., Харламов А. А.**
Основы работы с технологией CUDA. – М.: ДМК Пресс, 2010. – 232 с.: ил.
ISBN 978-5-94074-578-5

Данная книга посвящена программированию современных графических процессоров (GPU) на основе технологии CUDA от компании NVIDIA. В книге разбираются как сама технология CUDA, так и архитектура поддерживаемых GPU и вопросы оптимизации, включающие использование .PTX.

Рассматривается реализация целого класса алгоритмов и последовательностей на CUDA.

К книге прилагается CD, который содержит примеры решения на CUDA реальных задач с большим объемом вычислений из широкого класса областей, включая моделирование нейронных сетей, динамику движения элементарных частиц, геномные исследования и многое другое.

УДК 32.973.26-018.2
ББК 004.4

ISBN 978-5-94074-578-5

Содержание

Глава 1
Существующие многоядерные системы. Эволюция GPU. GPGPU

Одной из важнейших характеристик любого вычислительного устройства является его быстродействие. Для математических расчетов быстродействие обычно измеряется в количестве floating-point операций в секунду (*Flops*). При этом довольно часто рассматривается так называемое пиковое быстродействие, то есть максимально возможное число операций с вещественными (*floating-point*) величинами в секунду (когда вообще нет других операций, обращений к памяти и т. п.).

Реальные цифры по загрузке (пиковая vs реальная) по CPU и GPU

На самом деле реальное быстродействие всегда оказывается заметно ниже пикового, поскольку необходимо выполнять другие операции, осуществлять доступ к памяти и т. п.

Для персонального компьютера быстродействие обычно напрямую связано с тактовой частотой центрального процессора (CPU). Процессоры архитектуры x86 за время с момента своего появления в июне 1978 года увеличили свою тактовую частоту почти в 700 раз (с 4,77 МГц у Intel 8086 до 3,33 ГГц у Intel Core i7).

Таблица 1.1. Динамика роста тактовых частот

Год	Тактовая частота	Процессор
1978	4.77 MHz	Intel 8086
2004	3.46 GHz	Intel Pentium 4
2005	3.8 GHz	Intel Pentium 4
2006	2.333 GHz	Intel Core Duo T2700
2007	2.66 GHz	Intel Core 2 Duo E6700
2007	3 GHz	Intel Core 2 Duo E6800
2008	3.33 GHz	Intel Core 2 Duo E8600
2009	3.06 GHz	Intel Core i7 950

Однако если внимательно посмотреть на динамику роста частоты CPU, то становится заметно, что в последние годы рост частоты заметно замедлился, но зато появилась новая тенденция – создание многоядерных процессоров и систем и увеличение числа ядер в процессоре.

Это связано как с ограничениями технологии производства микросхем, так и с тем фактом, что энергопотребление (а значит, и выделение тепла) пропорцио-

нально четвертой степени частоты. Таким образом, увеличивая тактовую частоту всего в 2 раза, мы сразу увеличиваем тепловыделение в 16 раз. До сих пор с этим удавалось справляться за счет уменьшения размеров отдельных элементов микросхем (так, сейчас корпорация Intel переходит на использование технологического процесса в 32 нанометра).

Однако существуют серьезные ограничения на дальнейшую миниатюризацию, поэтому сейчас рост быстродействия идет в значительной степени на счет увеличения числа параллельно работающих ядер, то есть через параллелизм. Скорее всего, эта тенденция сохранится в ближайшее время, и появление 8- и 12-ядерных процессоров не заставит себя долго ждать.

Максимальное ускорение, которое можно получить от распараллеливания программы на N процессоров (ядер), дается законом Амдала (Amdahl Law):

$$S = \frac{1}{(1+P)+\frac{P}{N}}.$$

В этой формуле P – это часть времени выполнения программы, которая может быть распараллелена на N процессоров. Как легко видно, при увеличении числа процессоров N максимальный выигрыш стремится к $\frac{1}{1-P}$. Таким образом, если вы можем распараллелить 3/4 всей программы, то максимальный выигрыш составит 4 раза.

Именно поэтому крайне важно использование хорошо распараллеливаемых алгоритмов и методов.

1.1. Многоядерные системы

Рассмотрим в качестве иллюстрации несколько существующих многоядерных систем и начнем наше рассмотрение с процессоров Intel Core 2 Duo и Intel Core i7.

1.1.1. Intel Core 2 Duo и Intel Core i7

Процессор Intel Core 2 Duo содержит два ядра (P0 и P1), каждое из которых фактически является процессором Pentium M, со своим L1-кешем команд и L1-кешем данных (по 32 Кбайта каждое). Также имеется общий L2-кеш (размером 2 или 4 Мб), совместно используемый обоими ядрами (см. рис. 1.1).

Процессор Core i7 содержит уже четыре ядра (P0, P1, P2 и P3). Каждое из этих ядер обладает своими L1-кешем для данных и L1-кешем для команд (по

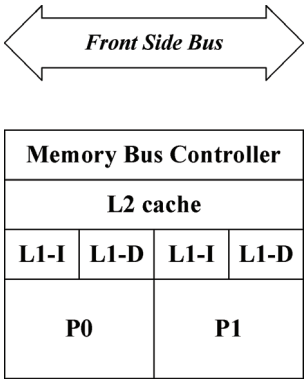

Рис. 1.1. Схема процессора Intel Core 2 Duo

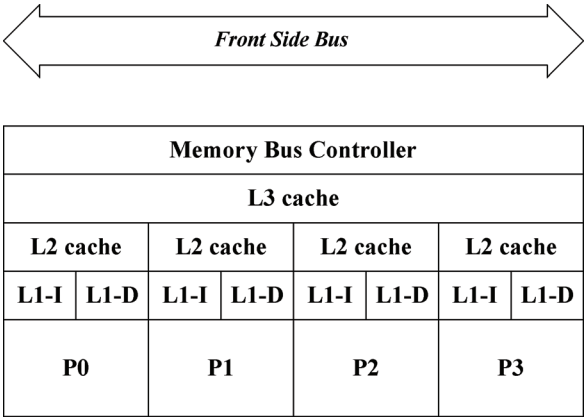

Рис. 1.2. Схема процессора Intel Core i7

32 Кбайта каждое) и L2-кешем (256 Кбайт). Кроме того, имеется общий для всех ядер L3-кеш размером 8 Мбайт.

1.1.2. Архитектура SMP

Кроме этих процессоров, есть также и другие многопроцессорные архитектуры. Одной из таких систем является система на основе симметричной мультипроцессорной архитектуры (Simmetric MultiProcessor Architecture, SMP).

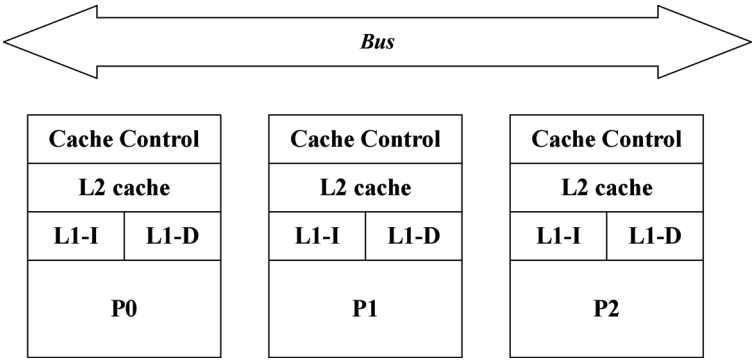

Рис. 1.3. Архитектура SMP-систем

В таких системах каждое ядро содержит свои кеши L1 и L2, и все ядра подсоединены к общей шине. Блок Cache Control отслеживает изменения памяти другими процессорами и обновляет соответствующим образом содержимое кешей – ведь если один и тот же участок памяти содержится в кеше сразу нескольких ядер, то любое изменение этого участка памяти одним из ядер должно быть немедленно передано в кеши других ядер, работающих с данным участком памяти.

Это типичная проблема архитектур с набором процессоров, подключенных к общей шине, – принципиальным моментом для всех многоядерных систем является то, что каждый процессор должен «видеть» целый и корректный образ памяти, что неизбежно ведет к необходимости для каждого процессора отслеживать обращения к памяти всех остальных процессоров для поддержания актуальности своих кешей. Подобная задача имеет квадратичную сложность от числа процессоров.

1.1.3. BlueGene/L

Еще одним примером многопроцессорной архитектуры является суперкомпьютер BlueGene/L. Он состоит из 65 536 двухъядерных узлов (*nodes*). Каждый узел содержит два 770 МГц процессора PowerPC. Каждый из них имеет свои кеши первого и второго уровней, специализированный процессор для *floating-point* вычислений *(Double Hammer FPU*). Оба процессора подключены к общему кешу третьего уровня (L3) размером 4 Мб и имеют собственный блок памяти размером 512 Мб (см. рис. 1.4).

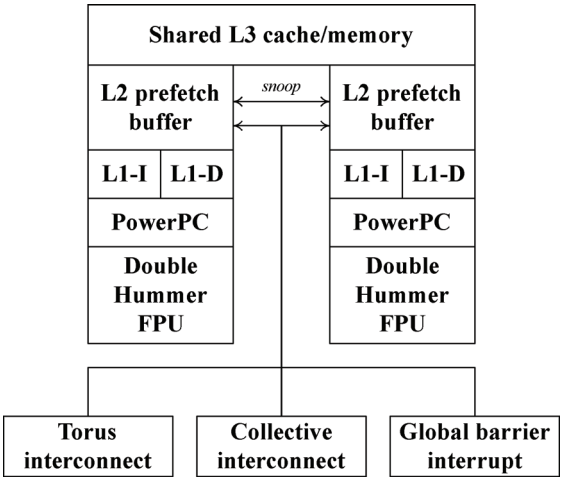

Рис. 1.4. Устройство одного узла в архитектуре BlueGene/L

Отдельные узлы могут соединяться между собой различными способами при помощи набора портов. Таким образом, каждый узел может непосредственно обратиться всего к небольшому числу других узлов, но за счет соединения всех узлов в сеть сообщение может быть передано любому другому узлу за небольшое количество шагов.

Для минимизации числа шагов, необходимых для передачи сообщения от одного узла другому, очень хорошо подходит топология тороидального куба. В ней все узлы образуют куб и у каждого узла есть ровно восемь соседей. При этом если узел лежит на одной из граней куба, то недостающих соседей он берет с противоположной грани (рис. 1.5).