GPU – доступный билет в мир супервычислений

В рамках ограниченных бюджетов российские предприятия остро нуждаются в инновационных решениях, которые позволяют эффективно внедрять передовые технологии производства. Ограниченный бюджет и передовые технологии. Казалось бы, несовместимые понятия. Обычно, для того чтобы выйти на новый технологический уровень, компаниям сначала нужно сделать значительные вложения в обновление инфраструктуры и существующего парка оборудования. Однако в сфере информационных технологий периодически появляются разработки, которые становятся исключением из этого правила. Одна из таких технологий построена на использовании графических процессоров (Graphics Processing Unit, GPU) для высокопроизводительных вычислений. Используя GPU в качестве вычислителя, можно получать результаты в десятки и даже сотни раз быстрее, а значит, появляется возможность браться за решение задач, которые прежде считались неразрешимыми.

Именно поэтому крупнейшие суперкомпьютерные центры мира начали активное внедрение решений на GPU. В конце прошлого года Национальная Лаборатория Окриджа, которая располагает самым быстрым в мире суперкомпьютером (ему принадлежит первое место в рейтинге TOP500), заявила о том, что ее новый суперкомпьютер, который должен стать в 10 раз мощнее, чем существующий, будет построен именно на графических процессорах.

Гибридные системы (именно в них GPU, наряду с центральным процессором (Central Processing Unit, CPU), используется как полноправный вычислитель) обладают колоссальной вычислительной мощью и при этом в несколько раз дешевле традиционных решений, использующих только возможности центрального процессора. Кроме того, они обеспечивают значительную экономию потребляемой электроэнергии. Например, крупнейшему европейскому банку BNP Paribas после перехода на гибридные технологии удалось сократить энергопотребление в 190 раз.

Достижение подобного эффекта позволяет говорить о новой вехе в развитии вычислительных систем.

CPU и GPU – два мира, две судьбы

До последнего времени ключевым компонентом систем для высокопроизводительных вычислений, включая кластеры, был центральный процессор. Однако в последние 2-3 года у него появился серьезный конкурент – графический процессор.

Исключительная вычислительная мощь GPU объясняется особенностями его архитектуры. В отличие от CPU,

CPU vs GPU – во многих задачах графический процессор обеспечивает лучшую производительность системы, меньшее энергопотребление и экономию площадей

который состоит из нескольких ядер (на большинстве современных систем от двух до четырех), графический процессор изначально создавался как многоядерная структура, в которой количество вычислительных блоков измеряется сотнями. Разница в архитектуре обусловливает и разницу в принципах работы. Если архитектура CPU предполагает последовательную обработку информации (этот принцип был заложен несколько десятков лет назад Джоном фон Нейманом), то GPU исторически предназначался для обработки компьютерной графики, поэтому рассчитан на массивно параллельные вычисления.

Каждая из этих двух архитектур имеет свои достоинства. CPU лучше работает с последовательными задачами. При большом объеме обрабатываемой информации очевидное преимущество имеет GPU. Условие только одно – в задаче должен присутствовать параллелизм.

Поскольку требования к производительности вычислителей в науке и промышленности уже давно опережают возможности традиционных вычислительных систем (даже несмотря на недавний переход на многоядерные центральные процессоры), несколько лет назад возникла идея использования GPU для вычислений общего назначения. Так появилась концепция GPGPU (General-Purpose computing on GPUs), или, как ее еще называют, GPU Computing.

До последнего времени основная проблема заключалась в том, как "научить" графический процессор выполнять несвойственные ему задачи (все-таки изначально основной работой GPU была обработка графики). Ее удалось решить в 2006 году, когда американская компания NVIDIA, мировой лидер в области визуальных технологий, представила платформу CUDA (Compute Unified Device Architecture), позволяющую запускать произвольный код на GPU.

CUDA предоставляет разработчику возможность по своему усмотрению организовывать доступ к набору инструкций графического ускорителя и управлять его памятью, организовывать на нем сложные параллельные

вычисления. Данная архитектура применяется на практике уже несколько лет, и компания продолжает активно работать над ее дальнейшим развитием. Сегодня программная модель CUDA поддерживает практически все наиболее популярные языки и интерфейсы программирования: C, Fortran, Open CL, Direct Compute и др.

NVIDIA Tesla – профессиональные супервычислители

Опробовав архитектуру CUDA на массовых решениях, NVIDIA разработала специальную серию вычислителей Tesla. Процессоры NVIDIA Tesla предназначены исключительно для высокопроизводительных вычислений и не занимаются обработкой графики.

Сегодня суперкомпьютерной индустрии NVIDIA предлагает два типа решений:

- для супервычислений в датацентрах и масштабных вычислительных установках (Tesla S);
- для персональных вычислений за рабочим столом (Tesla C).

В конце прошлого года компания представила новую архитектуру GPU под кодовым названием Fermi, а также новую линейку решений Tesla, созданных на базе этой архитектуры. Разработка новой архитектуры велась в первую очередь с прицелом на их использование для решения сверхсложных вычислительных задач.

Для выполнения ответственных вычислительных задач в Fermi добавлена функция коррекции ошибок (ECC) на всех уровнях системной памяти. Кроме того,

Tesla Personal Supercomputer – персональный суперкомпьютер NVIDIA Tesla – оборудован 960 вычислительными ядрами

значительно возросло быстродействие нового процессора NVIDIA в приложениях, требующих операций с двойной точностью.

С появлением Fermi архитектура графических процессоров впервые стала поддерживать требования стандарта по обработке чисел с плавающей запятой IEEE 754-2008, обеспечивая выполнение совмещенной операции умножения-сложения (fused multiply-add) не только для чисел с двойной точностью (что было реализовано в предыдущем графическом чипе NVIDIA), но и

для чисел одинарной точности. Также впервые появилась возможность исполнять код, написанный на языке C++. Эта возможность призвана существенно облегчить программирование графических чипов.

Графические процессоры Tesla 20-й серии (на базе новой архитектуры Fermi) стали самыми передовыми в мире среди немногих аналогичных решений.

Производительность новых процессоров Tesla серии C при выполнении операций над числами с двойной точностью достигает 630 Гфлопс, а энергопотребление не превышает 225 Вт. Персональная система на базе процессора NVIDIA Tesla в 250 раз мощнее обычных рабочих станций, потребляя при этом в 20 раз меньше электроэнергии, чем современные системы с четырехядерными CPU с той же производительностью.

Такие системы выпускают как глобальные производители компьютеров (Asus, Cray, Dell, Lenovo и др.), так и российские сборщики: Arbyte, Depo Computers, STSS.

На российский рынок вычислители новой серии должны поступить уже во втором квартале текущего года. Примерно в те же сроки ожидается и появление на рынке новых четырехпроцессорных 1U-серверов Tesla S, предназначенных для создания кластерных систем и размещения в центрах обработки данных. Серверы разработаны NVIDIA совместно с компаний SuperMicro. Производительность 1U-серверов при выполнении операций над числами с двойной точностью достигает 2,5 Тфлопс, а энергопотребление не превышает 1200 Вт.

Наука, финансы, промдизайн, нефтегаз…

Персональные системы и кластерные решения на базе графических процессоров уже хорошо себя зарекомендовали в исследовательских институтах по всему миру. Среди них Институт имени Макса Планка, Иллинойский университет, Кембриджский университет, российские МГУ, МФТИ, Объединенный институт ядерных исследований и др. Параллельные вычисления на базе GPU помогли исследователям добиться результатов, на которые прежде они не могли рассчитывать, или, по крайней мере, позволили существенно сократить время их получения. Сегодня вычислители Tesla используются в биоинформатике и биомедицине, вычислительной химии, различных областях физики, электродинамике, гидродинамике, медицине, молекулярной динамике, при моделировании погодных и атмосферных явлений, в астрофизике и других областях современной науки.

Решения на базе NVIDIA Tesla S позволяют модернизировать вычислительные центры без изменения существующей инфраструктуры и потребляемой электроэнергии

Колоссальная производительность, доступная цена, высокая энергоэффективность все чаще привлекают внимание к таким решениям со стороны коммерческих организаций и промышленных предприятий. Уже сегодня параллельные вычисления на GPU успешно применяются в финансах, инженерных расчетах, облачных 3D-вычислениях, при анализе и кодировании информации, в геологоразведке, промышленном дизайне и т.д.

В финансовой сфере Tesla применяют BNP Paribas, Bloomberg, Hanweck Associates и другие известные компании. Так, благодаря возможностям GPU Hanweck Associates смогла провести оценку всего рынка опционов США в режиме реального времени. BNP Paribas построила датацентр на базе Tesla. По внутренним оценкам компании, аналогичное по производительности решение на базе традиционных технологий обошлось бы ей в 10 раз дороже.

В нефтегазовой отрасли технологии NVIDIA также позволили более чем на порядок улучшить показатели эффективности датацентров (соотношение мощности с расходуемой электроэнергией, занимаемым пространством и стоимостью). Компаниям Hess и Petrobrass удалось увеличить производительность кластеров более чем в 20 раз, при этом во столько же раз сократив энергопотребление.

В России гибридные технологии только пробивают себе дорогу. Наиболее технологичные компании уже обратили на них свое внимание. "Лаборатория Касперского" для повышения уровня защиты своих клиентов использует высокопроизводительные вычислительные системы на базе графических процессоров NVIDIA Tesla предыдущего поколения. Эти системы позволили компании повысить скорость идентификации неизвестных файлов и ускорить реагирование на новые угрозы. Графический процессор Tesla S1070 обеспечил в 360 раз более высокую скорость работы алгоритма определения схожести по сравнению с решением на популярном в России Intel Core 2 Duo 2,6 ГГц.

С информационной безопасностью связан еще один пример успешного применения гибридных технологий NVIDIA в России. Компания "Элкомсофт", которая специализируется на разработке решений, призванных улучшить продуктивность и безопасность пользователей в среде Microsoft Windows, использовала вычислители NVIDIA Tesla для проведения аудита паролей. В результате переноса задачи на GPU скорость обработки информации увеличилась многократно без дополнительных инвестиций в расширение инфраструктуры.

Подобных примеров инновационного подхода к решению сложных задач должно быть намного больше в России. У российских ученых, промышленников, разработчиков приложений имеется огромный потенциал для обеспечения прорывов во многих областях, что крайне необходимо сегодня. Параллельные вычисления на базе архитектуры CUDA вполне могут быть использованы для реализации этого потенциала.

Дмитрий Желвицкий