ОЦЕНКА ВЫЧИСЛИТЕЛЬНОЙ ПРОИЗВОДИТЕЛЬНОСТИ ЛОКАЛЬНО-РЕКУРСИВНЫХ НЕЛОКАЛЬНО-АСИНХРОННЫХ АЛГОРИТМОВ НА ГЕТЕРОГЕННЫХ СИСТЕМАХ

И.А. Горячев

Интерес к использованию графических процессоров общего назначения (GPGPU) для высокопроизводительных вычислений в последние годы резко возрос в связи с достижением параметров производительности устройств на графических процессорах на порядок больших, чем у традиционных процессоров как по пиковой производительности, так и по пропускной способности глобальной и локальной памяти.

Наряду с обычными для любой новой технологии трудностями внедрения, такими как недостаточная разработка инструментов программирования и недостаточная подготовка программистов к овладению этими инструментами, также стоит отметить сложности, связанные с ограниченным ресурсным запасом памяти вычислительного узла и способом хранения данных больших объемов, что напрямую влияет на эффективность вычислений. Для лучшего понимания способов решения обозначенных проблем и задач создания алгоритмов для гетерогенных систем в первую очередь необходимо провести анализ теоретической производительности и построение программной модели существующих аппаратных устройств на графических процессорах. После этого можно приступать к построению соответствующих данной модели алгоритмов.

Для специализированных устройств вычислений общего назначения (GPGPU) можно выделить следующее иерархическое строение логических уровней памяти: глобальная память (3072 MB), L2 cache (768 KB), local/share memory (896 MB), регистры (1.75 MB). Данные приведены для графических ускорителей Tesla C2050 на базе кластера Келдыш-100. С учетом этих параметров проведен анализ архитектуры подсистемы памяти и построена пирамидальная модель вычислительной системы, включающей GPGPU [[pic1]].

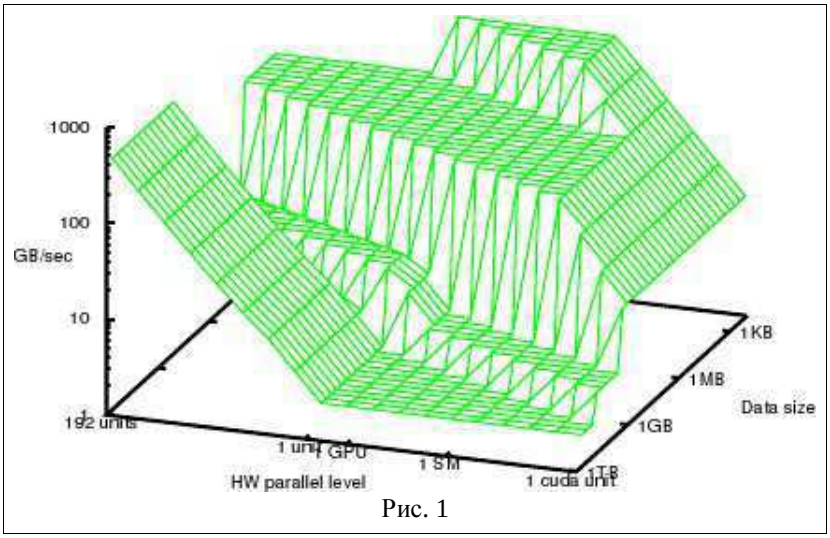

Рис. 1

Помимо архитектуры уровней системы памяти подобная модель учитывает такие параметры, как пропускная способность каждого из обозначенных уровней памяти и аппаратно выделенные HW-уровни параллельности с их темпом вычислений. Каждый HW-уровень можно классифицировать по вычислительным ядрам (448 cuda cores per card), мультипроцессорам (14 SM в видеокарте), видеокартам (3 видеокарты на узел) и вычислительным узлам кластера (192 узлов). Пропускная способность передачи данных оценивалась исходя из предложенных производителем характеристик и представлена на [[pic1]] в логарифмическом масштабе. По предложенной диаграмме можно определить ключевые параметры, напрямую влияющие на эффективность алгоритма и его характеристики: коэффициент локальности — отношение числа операции к числу данных — и коэффициент асинхронности, отвечающий эффективности выполнения параллельного алгоритма в зависимости от числа процессоров на идеальном вычислителе, следуя закону Амдала.

Для оценки производительности и оценки скорости доступа к памяти в зависимости от размера обрабатываемых данных произведены вычисления с применением локально-рекурсивных нелокально-асинхронных (LRnLA) алгоритмов [1], использующих трехуровневое разбиение по программным потокам (threads), блокам (blocks) и сеткам (grid).

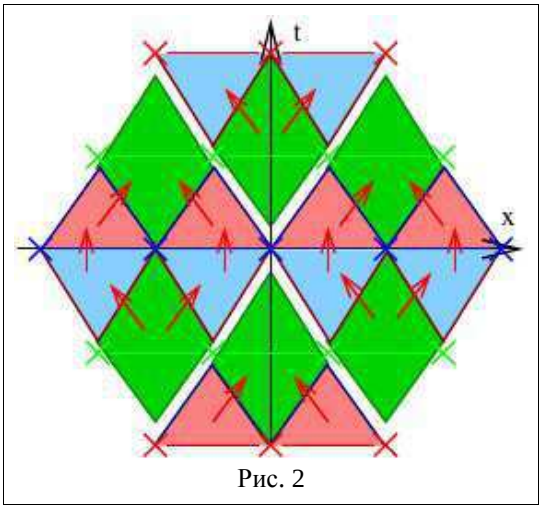

Рис. 2

LRnLA-алгоритмы базируются на принципе «разделяй-и-властвуй» — обрабатываемые данные рекурсивным образом разбиваются на подобные подструктуры [[pic2]], что позволяет локализовать их на более близких к вычислительным ядрам уровнях иерархии памяти. Таким образом, можно говорить о разбиении основной задачи на подзадачи и решать их параллельно, независимо друг от друга на разных HW-уровнях параллелизма.

Помимо этого повышение эффективности вычислений достигается путем максимального обращения к данным, локализованных на близких к вычислительным ядрам уровнях иерархии памяти, чтобы сократить время доступа к ним. Учитывая обычные требования численных методов, LRnLA алгоритмы удовлетворяют обозначенному требованию, что обеспечивается путем проведения дискретизации на подобласти зависимости и влияния данных ([[pic2]]), и на основе их локальности по пространству и зависимости по времени осуществляется декомпозиция вычислений.

Для реализации представленного алгоритма на гетерогенных системах выбрана классическая задача перемножения матриц размером 2^{Rank} x 2^{Rank} с вычислительной сложностью ~2^{3*Rank}, а также произведено сравнение ([[pic3]]) со стандартной библиотекой CUBLAS [2], предоставляемой корпорацией NVIDIA.

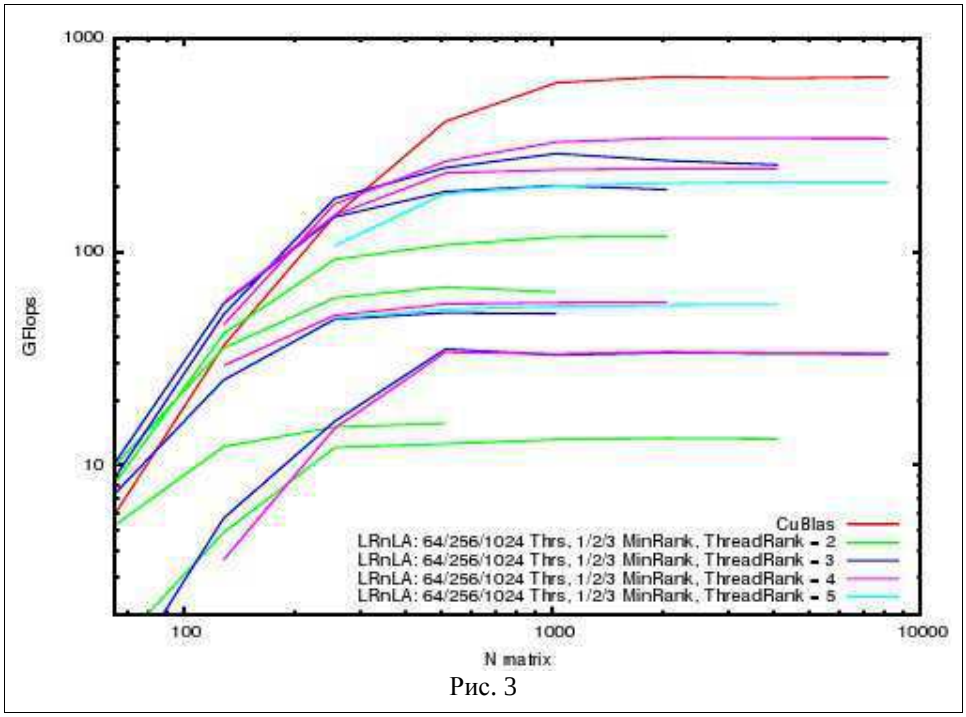

Рис. 3

Вместо традиционного способа хранения массивов и доступа к памяти использовался обход Мортона [3], разработанный в виде структуры данных CubeLR, интерфейс которого реализован с помощью шаблонов C++. CubeLR позволяет хранить элементы в массиве таким образом, чтобы участвующие в одной операции данные были расположены как можно ближе друг к другу как в операционной памяти, так и непосредственно в структуре массива. Это позволяет уменьшить количество запросов для загрузки данных, сокращает среднюю латентность и увеличивает локальность.

Алгоритм перемножения матриц имеет трехуровневое разбиение:

- На первом уровне вычислительные потоки (thread) объединяются в отдельные группы размером $2^{ThreadRank}$, что согласовано с аппаратным объединением потоков (warp). Тем самым гарантируется бесконфликтное обращение потоков в локальную память выровненного адресного пространства.
- Уровнем выше происходит разбиение программного блока на подуровни размером $2^{CurMinRank}$, каждый из которых обрабатывается одной группой потоков, следуя локально-рекурсивному обходу CubeLR. Указанное разбиение также согласовано с аппаратным выполнением каждого отдельного блока на отдельном мультипроцессоре. Это позволяет активно задействовать регистровую память.
- Третий уровень соответствует разбиению матрицы на программные блоки (block). Количество блоков при этом получается равным $2^{BlockRank}$. Обход матрицы по блокам также реализуется локально-рекурсивно. При выполнении каждого блока обрабатываемые данные естественным образом локализуются в кэшах разных уровней.

При этом параметры каждого уровня разбиения подчиняются соотношению Rank = ThreadRank + CurMinRank + BlockRank. Однако подобный алгоритм не ограничивается лишь тремя уровнями и может быть расширен до многоярусности количеством равным Rank.

Таким образом, осуществляется перемножение соответствующих элементов матриц по известной формуле линейной алгебры, учитывая указанный выше обход, и запуск на аппаратных мультипроцессорах (stream multiprosessors) на разных устройствах GPGPU.

При этом без проведения каких-либо оптимизаций достигнута производительность на уровне 25% от пиковой или порядка 50% от результата выполнения CUBLAS ([[pic3]]) для матриц в виде двумерных массивов. Показано, что LRnLA алгоритмы применимы для GPGPU, при этом для оптимального использования регистровой памяти требуется выбрать промежуточное количество равное не более, чем 256 тредов на блок. В соответствии с этими параметрами будет выбран подходящий алгоритм семейства LRnLA для достижения предельной производительности для гетерогенных вычислений.

LRnLA алгоритмы являются одним из методов, направленных на моделирование физических процессоров на параллельных высокопроизводительных системах с развитой иерархией памяти. Разрабатываемые модели и алгоритмы ориентированы на решение 3D3V задач прямого моделирования многомерных нестационарных явлений и процессов в физических средах. В целом, задача моделирования ставится как задача Коши для систем дифференциальных уравнений в частных производных гиперболического типа сеточными методами в многомерной области с числом узлов сетки более 10^9. Примером таких задач является возникновение неустойчивостей и турбулентности [4] в замагниченной плазме (FDTD и метод PIC). Решение подобных задач планируется перенести на гетерогенные системы с использованием графических процессоров.

ЛИТЕРАТУРА:

1. В.Д. Левченко «Асинхронные параллельные алгоритмы как способ достижения эффективности вычислений», Информационные технологии и вычислительные системы, 1, 2005.
2. NVIDIA CUDA Programming Guide v1.1
3. Morton, G. M. (1966), A computer Oriented Geodetic Data Base; and a New Technique in File Sequencing, Technical Report, Ottawa, Canada: IBM Ltd.
4. И.А. Горячев, В.Д. Левченко, А.Ю. Перепелкина, «Трехмерная полностью кинетическая численная модель замагниченной плазмы канала холловского двигателя»//Тезисы докладов с XXXIX международной Звенигородской конференции по физике плазмы и управляемому термоядерному синтезу, 2011.