Аналитическая модель оценки эффективности выполнения параллельного кода на GPU

М.Д. Нгуен

Московский Государственный Университет им. М.В. Ломоносова
Факультет Вычислительной Математики и Кибернетики

В данной работе описывается аналитическая модель A3, предназначенная для оценки эффективности выполнения параллельного кода на GPU. Теоретическая часть модели была разработана в рамках исследования по распараллеливанию метода SCF, входящий в состав пакета Siesta [1]. Поскольку модель A3 будет внедрена в систему автоматизации распараллеливания САПФОР [2] в качестве одного из инструментов для предсказания эффективности при распараллеливании последовательных Fortran программ, алгоритмы реализации модели были изменены с учетом требований САПФОР. Модель может быть использована для оценки возможности распараллеливания с помощью высокоуровневого языка Fortran DVMH и также для выбора оптимальных схем распараллеливания. Предлагаемая модель является востребованным инструментом для распараллеливания больших программ, время выполнения которых занимает от нескольких часов и более.

A3 статически анализирует регионы кода исходной программы, которые могут выполняться параллельно на GPU, и оценивает эффективность их выполнения без программирования и запуска на реальном GPU. По сравнению с существующей моделью оценки, входящей в состав фреймворка Grophecy [**Error! Reference source not found.**], A3 более универсальна, т.к. она не требует участия программиста в процессе работы. Такое преимущество возможно благодаря мощному анализатору системы САПФОР. Более того, A3 анализирует не только CWP (Compute Warp Parallelism) и MWP (Memory Warp Parallelism), но и ILP (Instruction Level Parallelism) и TLP (Thread Level Parallelism) – совокупность всех этих четырех факторов позволяет значительно быстрее сократить пространство поиска возможных схем распараллеливания. Другим преимуществом является статическое моделирование объема используемых регистров, что позволяет оценить степень занятости GPU до получения параллельного кода и его компиляции.

В модели A3 каждый параллельный регион рассматривается как отдельная функция-ядро (далее ядро). Для каждого ядра собираются характеристики, влияющие на возможность его распараллеливания и на эффективность выполнения. Эффективность выполнения ядра – это время его выполнения на GPU. Оптимальная схема распараллеливания ядра – это схема с минимальным временем выполнения. Оптимальная схема распараллеливания всей программы – это множество схем распараллеливания ядер, при которых время выполнения программы минимально.

Применение модели A3 позволяло распараллелить метод SCF за 11 дней. Ускорение на тестах, связанных с молекулой h2o (32h2o, h2o, h2o_2, h2o_4, h2o_basis, и др.) составило от 6x и выше. В данный момент ведется работа над внедрением модели A3 в систему САПФОР.

Литература

1. Siesta's Official Website. URL: http://icmab.cat/leem/siesta/Documentation/Publications/index.html (дата обращения: 02.12.2012).
2. Бахтин В.А., Бородич И.Г., Катаев Н.А., Клинов М.С., Ковалева Н.В., Крюков В.А., Поддерюгина Н.В. Диалог с программистом в системе автоматизации распараллеливания САПФОР. // Труды Международной научной конференции "Научный сервис в сети Интернет: экзафлопсное будущее", Новороссийск, сентябрь 2011 – М.: Изд-во МГУ, 2011, С. 67-70.
3. J. Meng, V. A. Morozov, K. Kumaran, V. Vishwanath, T. D. Uram. GROPHECY: GPU performance projection from CPU code skeletons // Proceedings of 2011 International Conference for High Performance Computing, Networking, Storage and Analysis, November 2011.