Адаптация геостатистического симулятора к современным гибридным вычислительным системам на основе графических процессоров NVIDIA

М.В. Андреев, Р.К. Газизов, А.Л. Штангеев, А.В. Юлдашев, А.А. Яковлев
Уфимский государственный авиационный технический университет

В настоящем докладе представлены новые результаты по адаптации геостатистического симулятора, реализующего построение обусловленных геостохастических геологических моделей по скважинным данным с учетом анизотропности и нестационарности пластовых свойств, для эффективного выполнения на гибридных вычислительных системах на основе графических процессоров (GPU) компании NVIDIA.

Нами была разработана параллельная версия геостатистического симулятора [1], способная выполняться на гибридных вычислительных системах с общей памятью и графическими процессорами NVIDIA с поддержкой CUDA. Распараллеливание в модели общей памяти реализовано средствами интерфейса OpenMP, а с помощью технологии PGI Accelerator [2] произведена адаптация для выполнения на GPU наиболее трудоемких участков кода программы, выделенных в виде функций, решающих следующие вычислительные задачи: построение матрицы ковариации, выполнение быстрого преобразования Фурье, а также обусловленного стохастического моделирования.

Было проведено тестирование производительности полученной версии симулятора при расчете реальной модели на двухпроцессорной рабочей станции FSC CELSIUS V840 на базе процессоров AMD Opteron 2214 с различными графическими ускорителями NVIDIA: GeForce 295, 460, 470, 480 GTX и Tesla C1060. Отслеживалось ускорение суммарного времени выполнения адаптированных участков кода при расчете на CPU+GPU относительно суммарного времени их выполнения при расчете на 4 ядрах центральных процессоров (CPU). В результате максимальное ускорение было достигнуто при использовании GeForce 480 GTX и составило более 8 раз.

Далее возникла задача адаптации симулятора к более современным гибридным вычислительным системам на основе графических процессоров NVIDIA. Благодаря предоставленному нам доступу к вычислительным ресурсам Межведомственного суперкомпьютерного центра РАН, были проведены работы по тестированию производительности и оптимизации симулятора на высокопроизводительном вычислительном сервере Kraftway Science KT25. Данный сервер имеет в своем составе два современных шестиядерных процессора Intel Xeon X5660, а также четыре профессиональных графических ускорителя NVIDIA Tesla M2050 последнего поколения, что позволяет говорить о его пиковой производительности более двух терафлопс на вычислениях двойной точности. Необходимо отметить несомненное достоинство установленных в сервере ускорителей Tesla серии 20: большой объем видеопамяти с поддержкой технологии коррекции ошибок ECC.

Первые тесты, проведенные на сервере Kraftway, выявили ряд особенностей. Во-первых, при тестировании на моделях, время расчета которых достаточно мало, было отмечено ощутимое (около 4-10 секунд) время инициализации при первом обращении к GPU в программе. В связи с этим инициализация в программе была вынесена отдельно от адаптированных функций и далее при замерах времени выполнения не учитывалась. Во-вторых, ускорение при выполнении на CPU+GPU относительно CPU оказалось существенно более низким, чем на ранее рассмотренной рабочей станции CELSIUS V840. Одной из причин снижения ускорения, безусловно, явилось использование более производительных процессоров Xeon X5660. Кроме того, сравнительный анализ характеристик карт GeForce 480 GTX и Tesla M2050 с сайта производителя, показал, что пер-

вая имеет в своем составе большее количество ядер и ее аппаратные блоки работают на более высоких частотах, что приводит к более высокой пиковой производительности GeForce 480 GTX в сравнении с Tesla M2050.

Профилирование процесса выполнения симулятора на различных графических процессорах, проведенное средствами NVIDIA Visual Profiler из CUDA Toolkit, показало, что при выполнении некоторых адаптированных функций, ресурсы GPU используются не более чем на десять процентов. Это и послужило толчком для поиска путей оптимизации программного кода в целях более эффективного использования графических процессоров.

Оказалось, что в ряде рассмотренных функций имел место неэффективный порядок доступа к данным в памяти GPU. Например, в функции обусловленного стохастического моделирования было замечено, что при выборке элементов трехмерного массива данных во вложенном цикле, нити, выполняющиеся на графическом процессоре в рамках одного блока, осуществляют выборку разнесенных в памяти элементов массива. Проблема была устранена за счет перегруппировки данных в массиве и изменения порядка обращения к элементам массива, что позволило сократить время выполнения данного участка кода более чем в три раза.

Ниже приведены результаты тестирования симулятора после проведенных оптимизаций при расчете реальной модели. Отметим, что сборка симулятора для CPU осуществлялась компилятором Intel C++ Compiler, а для CPU+GPU – компиляторами PGI C/C++.

В таблице 1 представлены суммарные времена выполнения адаптированных функций в зависимости от числа потоков, порождаемых в OpenMP-программе при выполнении на CPU, а также CPU+GPU. Причем при запуске на CPU число потоков варьировалось от 1 до 12 в соответствии с числом процессорных ядер в системе, а при запуске на CPU+GPU от 1 до 4 для того, чтобы каждый поток, выполняющийся на одном из процессорных ядер, имел возможность монопольно использовать один из четырех графических ускорителей.

Таблица 1. Время выполнения адаптированных функций на CPU и CPU+GPU.

	Время выполнения, с				
Платформа/Количество потоков OpenMP	1	2	4	8	12
CPU	1 285,19	644,11	324,82	170,75	112,49
CPU+GPU	48,30	24,54	12,39	-	-

Для наглядности на рисунке 1 представлены зависимости ускорения расчета на CPU, а также CPU+GPU в зависимости от числа OpenMP-потоков по сравнению со временем расчета одним потоков на одном ядре центрального процессора. Несложно увидеть, что время расчета адаптированных функций снижается линейно с увеличением количества потоков. В случае одного потока, подключение к расчету графического ускорителя позволяет сократить время в 26 раз. Кроме того сравнение минимальных времен выполнения адаптированных функций на CPU и CPU+GPU показывает, что за счет использования графических процессоров минимальное время выполнения (расчет на всех 12 процессорных ядрах) может быть сокращено более чем в 9 раз (расчет на 4 процессорных ядрах при поддержке 4 графических ускорителей).

Таким образом, применение современных гибридных вычислительных систем на основе графических процессоров позволяет значительно ускорить процесс построения обусловленных геостохастических геологических моделей.

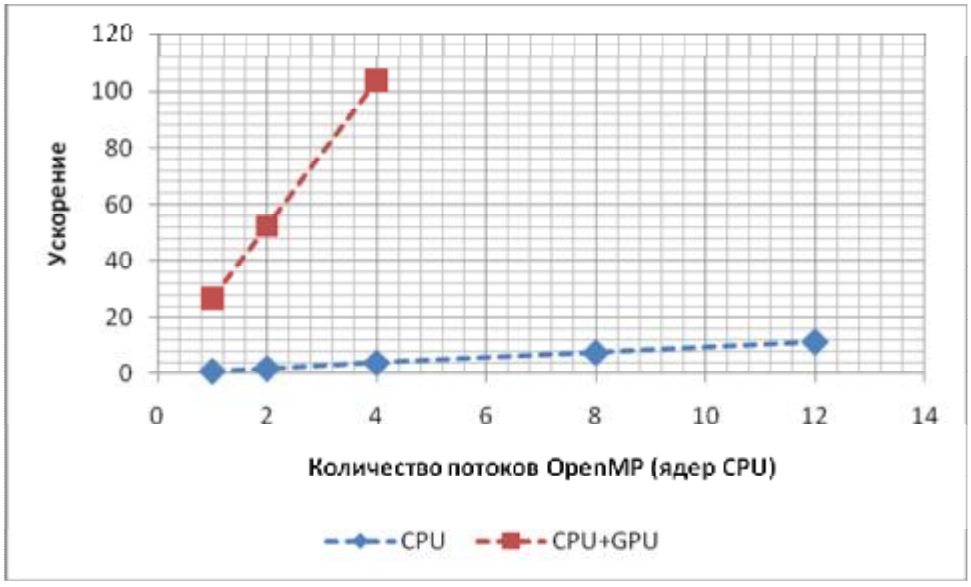

Рис. 1. Зависимости ускорения от количества потоков OpenMP при использовании CPU и CPU+GPU относительно времени расчета одним потоком на CPU

В дальнейшем планируется рассмотреть возможность перехода на вычисления с одинарной точностью в целях дальнейшего ускорения времени расчетов. Кроме того планируется оценить эффективность реализации алгоритмов решения задач геостатистического моделирования с помощью высокоуровневой технологии PGI Accelerator в сравнении с наиболее распространенной технологией программирования графических процессоров – CUDA.

Литература

1. Андреев М.В., Галеев Э.Р., Штангеев А.Л., Юлдашев А.В., Яковлев А.А. Опыт портирования геостатистического симулятора на архитектуру GPU средствами технологии PGI Accelerator // Высокопроизводительные параллельные вычисления на кластерных системах (HPC-2010): Материалы X международной конференции (Пермь, 1 – 3 ноября 2010 г.), в двух томах – Пермь: Издательство ПГТУ, 2010. – Том 1. С. 37-41.
2. PGI Fortran & C Accelerator Programming Model white paper. URL: http://www.pgroup.com/lit/whitepapers/pgi_accel_prog_model_1.3.pdf (дата обращения: 11.02.2011).