

СУПЕРУСКОРЕНИЕ ГИДРОДИНАМИЧЕСКИХ РАСЧЕТОВ С ПОМОЩЬЮ ПРИМЕНЕНИЯ ГРАФИЧЕСКИХ ПРОЦЕССОРОВ NVIDIA И ТЕХНОЛОГИЙ ПРОГРАММИРОВАНИЯ CUDA

Максимов Д.Ю. , Кудряшов И.Ю., Марченко Н.А.
Национальный центр развития инновационных технологий

SUPERACCELERATION OF HYDRODYNAMIC CALCULATIONS BY MEANS OF THE GRAPHIC PROCESSORS NVIDIA AND PROGRAMMING TECHNOLOGIES CUDA

Maksimov D.Yu., Kudryashov I.Yu., Marchenko N.A.
The National Center for Development of Innovative Technologies

Высокоточная обработка и интерпретация сейсмических данных в последнее время приводит к созданию гигантских гидродинамических моделей, размеры которых выходят далеко за рамки возможностей обычных персональных компьютеров.

Одним из путей ускорения расчетов таких моделей является перенос части программы на платформу видеокарты, имеющейся на рабочей станции, например GTX 295 и Tesla C1060 компании NVIDIA. Рассматриваются преимущества интеграции кода симулятора с видеоускорителем на разных уровнях, начиная с наиболее ресурсоемкого этапа решения задачи – расчета системы линейных алгебраических уравнений. Даны оценки ожидаемого прироста производительности.

High-accurate seismic data processing and interpretation has recently led to the creation of giant hydrodynamic models. Their dimensions exceed PC's capabilities.

One of the ways of calculation acceleration of these models is a transfer of program part to the video card platform on the workstation, for example, GTX 295 and Tesla C1060 (NVIDIA). Advantages of simulator code integration with video acceleration at different levels are considered, starting from the most resource-demanding task solution stage – calculation of linear algebraic equations system. Estimations of productivity increase expected are given.

Гидродинамическое моделирование представляет собой проведение многочисленных расчетов по адаптации и еще большего числа прогнозных расчетов. На реализацию больших и сложных задач уходит очень много времени, и уже сейчас нужно искать решения в использовании программно-аппаратных комплексов моделирования.

Куда уходит время при гидродинамическом расчете? Подготовка данных, вывод и организация итераций условно занимают четверть всего времени. Большая же часть времени уходит на расчет матрицы для решения линейных уравнений, которые приводят к специфическим матрицам, и это происходит фактически во всех симуляторах. Кратко обозначим особенности матриц, которые возникают для задач подземной гидродинамики. В зависимости от числа компонент они могут содержать 14 диагоналей и 21 диагональ. Наличие активных и неактивных ячеек ухудшает регулярность матрицы. Также специфику задачи определяет наличие скважин, так как это означает наличие дополнительных связей между уравнениями системы. Зачастую бывают нестыкованные сетки, что также нарушает регулярность матриц, подаваемых на линейный солвер.

Кратко иллюстрация одной из матриц показана на рисунке 1. Здесь семь диагоналей, в которых каждый эле-

мент представляет собой подматрицу размером 3x3, если рассматривать трехфазную задачу. Отдельные точки – это вклад от скважин. Самые стандартные методы решения – это итерационные схемы плюс заранее выбранный преобуславливатель. В качестве самого простого примера приведен ILUTP+GMRES, которые позволяют решить некоторые матрицы за какое-то время. Подробнее рассмотрим распределения времени вычислительных ресурсов при решении такой матрицы. Помимо подготовки элементов матрицы, это прекодиционер и итерационная схема. В обоих случаях на прекодиционер и саму схему уходит весь ресурс шины памяти, а непосредственно на вычислительные ресурсы процессора отводится незначительная часть.

Как такой стандартный подход можно ускорить? Здесь можно предложить два принципиальных пути. Во-первых, использовать имеющуюся дополнительную информацию о задаче, использовать знания о геометрии, знания о предыдущих итерациях по Ньютону или по времени. Это так называемые алгоритмические методы. Мы частично используем знания о прекодиционере, частично – знания с предыдущего шага. Если брать отдельно решения матрицы, ее можно увеличить в 2,4 раза. Суммарный прирост

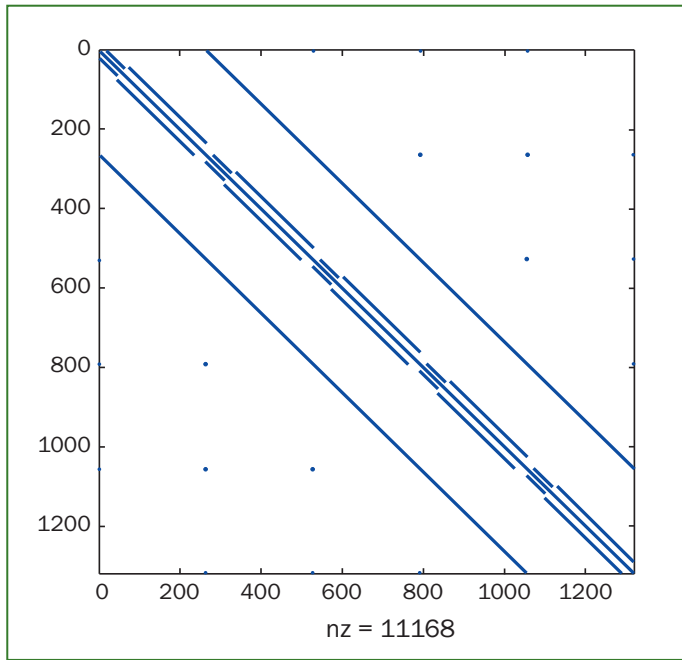


Рисунок 1.
Структура матрицы. В данном представлении каждому элементу соответствует блок 3x3

решения всей задачи в этом случае составит 60% с теми или иными вариациями.

Второй путь состоит в использовании других компьютерных ресурсов с большей производительностью, в первую очередь, с большим быстродействием памяти. Это новая технология для решения таких задач, и пока в западных симуляторах такая технология на промышленном уровне не применяется. В качестве примера выступают ви-

деокарты компании NVIDIA: GTX 295 и Tesla C1060 с довольно приемлемой стоимостью. Выбор пал на них, поскольку они поддерживают операции с двойной точностью. Конечно, существует множество других видеокарт, но далеко не все поддерживают двойную точность, что необходимо для решения таких задач.

Приведенные выше видеокарты позволяют серьезно ускорить доступ к памяти, а операции с числами с двойной точностью ускоряются в пять раз. В таблице 1 приведены некоторые технические характеристики видеокарты. В правом столбце приведены характеристики видеокарты новой архитектуры, которая еще не поступила в продажу, но уже была проанонсирована компанией NVIDIA. Она содержит большое количество вычислительных ядер и имеет высокую скорость доступа к памяти, так что операции с числами двойной точности будут еще больше ускорены. Обращаем внимание, что алгоритмы, которые переносятся на видеокарту, должны обладать очень хорошим параллелизмом. Также NVIDIA предлагает среду программирования для разработки приложений, которые будут исполняться на видеокарте. Сами программы пишутся на языке C, с необходимыми расширениями, которые называются библиотекой CUDA. Одно из применений, которое уже реально сделано, – это расчет задач сейсмологии компанией Paradigm. Обработка сейсмических данных здесь была перенесена на ресурсы вычислительных графических ускорителей.

Что же можно ожидать при применении графических ускорителей для расчетов? Сейчас практически в любом компьютере свободно можно подключить видеокарты, например 2 видеокарты GTX 295 с двойной точностью. Этап факторизации ускорится примерно в 5-10 раз, также, как и итерационная схема. Суммарно на весь симулятор мы получаем ускорение примерно в 2 раза и более. Это то, что

CPU	G80	GT200	Fermi
Transistors	681 million	1.4 billion	3.0 billion
CUDA Cores	128	240	512
Double Precision Floating Point Capability	None	30 FMA ops/clock	256 FMA ops/clock
Single Precision Floating Point Capability	128 MAD ops/clock	240 MAD ops/clock	512 FMA ops/clock
Special Function Units (SFUs)/SM	2	2	4
Warp schedulers (per SM)	1	1	2
Shared Memory (per SM)	16 KB	16 KB	Configurable 48 KB or 16 KB
L1 Cache (per SM)	None	None	Configurable 16 KB or 48 KB
L2 Cache	None	None	768 KB
ECC Memory Support	No	No	Yes
Concurrent Kernels	No	No	Up to 16
Load/Store Address Width	32-bit	32-bit	64-bit

Таблица 1.
Технические характеристики графических процессоров NVIDIA



касается непосредственно решения матрицы, представляя собой некоторый замкнутый этап. Если перенести на видеокарту и процесс построения матрицы, то ускорение будет еще больше в 5-10 раз.

Конечно, можно использовать и «обычную» видеокарту с видеовыходом, либо специальную промышленную видеокарту Tesla, которая ориентирована на крупномасштабные цели и даже не имеет выхода на монитор, но отличается тем, что она гораздо более тщательно протестирована на задачах повышенной сложности. Также сейчас существует проект в МСЦ, направленный на интеграцию видеокарт с обычными узлами кластера. Таким образом, создается комбинированный кластер, что также ускорит расчеты.

В итоге появилась технология, которая позволяет серьезно ускорить гидродинамические расчеты, – это программно-аппаратный комплекс на основе симулятора MKT и графических ускорителей NVIDIA, означающий возможность расчета больших моделей за приемлемый срок.

Затраты времени при гидродинамическом расчёте:

25% – Расчёт PVT-, ОПФ-свойств, вывод данных, организация нелинейных итераций.

75% – Линейный solver – решение СЛАУ (систем линейных алгебраических уравнений).

Компьютерные ресурсы с большим быстродействием памяти

Затраты:

- Два GTX 295: +500 евро
- Tesla C1060: +1300 евро

Ускорение:

- Доступ к общей памяти: в 100 раз.
- Операции с числами двойной точности: в 5 раз.

Оценка прироста для двух GTX 295

Можно ожидать:

- 5–10-кратный прирост производительности на факторизации;
- 10–20-кратный прирост производительности на итерационной схеме.

Суммарный прирост производительности – 1.5–2.2 раз.

Дальнейшее ускорение – 5-10 раз и выше при переносе генерации элементов матрицы на графические ускорители.

Заключение

Разработана технология, позволяющая создать супербыструю программу для решения задач большой сложности в рамках обычной рабочей станции.