НА ПУТИ К ОСВОЕНИЮ ГЕТЕРОГЕННЫХ СУПЕРВЫЧИСЛЕНИЙ В ГАЗОВОЙ ДИНАМИКЕ

Горобец Андрей Владимирович, ИПМ им. М. В. Келдыша РАН, к. ф.-м. н., с.н.с.
Суков Сергей Александрович, ИПМ им. М. В. Келдыша РАН, к. ф.-м. н., с.н.с.

Ключевые слова: аэроакустика, газовая динамика, MPI, OpenMP, OpenCL, GPU

Как известно, задачи вычислительной газовой динамики представляют собой один из наиболее вычислительно-емких классов задач механики сплошной среды. Одновременно требуются численные схемы повышенной точности, высокое пространственное разрешение для качественного моделирования турбулентности, длительный период интегрирования по времени для качественного осреднения статистики течения и так далее. Все эти требования приводят к заоблачной вычислительной стоимости, которая может быть под силу только крупным вычислительным системам.

В связи с этим, эффективное использование современных вычислительных систем является одной из наиболее актуальных проблем вычислительной газовой динамики. Особенно с учетом того, что архитектура суперкомпьютеров быстро меняется, и необходима адаптация алгоритмов к новым системам. Еще совсем недавно кластеры с многоядерными узлами пришли на смену системам на одноядерных процессорах. Затем процессоры с 6-ю, 8-ю (Intel Xeon X6550, X7550) и даже с 16-ю (AMD Opteron 6262 HE, 6282 SE) ядрами пришли на смену 2-х и 4-х ядерным процессорам. Это мотивировало переход на более сложную двухуровневую модель, сочетающую распределенную и общую память. Теперь эволюция архитектур, с одной стороны, продолжает идти по пути от “multi-core” к “many-core” в сторону существенно-многоядерных узлов, а суммарное число ядер в крупнейших системах уже перевалило за миллион (Sequoia – IBM BlueGene/Q, USA). С другой стороны, эволюция пошла в сторону применения массивно-параллельных ускорителей в дополнение к CPU. Среди десяти самых мощных суперкомпьютеров мира уже можно видеть несколько систем такой гибридной (гетерогенной) архитектуры.

В гибридной вычислительной системе, в отличие от "классической", в дополнение к CPU установлены вычислительные ускорители, которые можно рассматривать как математические сопроцессоры. В настоящее время наибольшее распространение в качестве таких ускорителей получили графические процессоры GPU (graphics processing units). Данная работа посвящена мучительному переходу к гетерогенным вычислениям и сфокусирована на разработке программных ядер для базовых операций газодинамических и аэроакустических алгоритмов на неструктурированных сетках. Конечная цель состоит в адаптации к гибридным суперкомпьютерам программного обеспечения, изначально ориентированного на “классические” суперкомпьютеры с десятками тысяч процессоров.

Для расчетов на гибридных системах предлагается использовать наиболее современную многоуровневую параллельную модель. На первом уровне используется MPI (Message Passing Interface – интерфейс передачи сообщений) для объединения узлов суперкомпьютера в рамках модели с распределенной памятью. На втором уровне для распараллеливания по CPU ядрам внутри узлов используется интерфейс прикладного программирования OpenMP (Open Multi-Processing) для модели с общей памятью. Применение более сложной параллельной модели

MPI+OpenMP дает возможность задействовать существенно большее число CPU ядер за счет многократного сокращения количества параллельных процессов, обменивающихся данными через коммуникационную среду суперкомпьютера. На третьем уровне задействуются массивно-параллельные ускорители, в которых реализуется принципиально иной тип параллелизма на уровне потоковых процессоров – SIMD (Single Instruction Multiple Data – одиночный поток команд, множественные потоки данных) параллелизм.

Методика адаптации программной реализации алгоритма подразумевает следующую последовательность действий. Численный алгоритм раскладывается на составляющие его базовые операции – реконструкция конвективной части потоков через грани контрольных объемов, расчет диссипативной части потоков, расчет узловых градиентов, полиномиальных коэффициентов, граничных условий, шаг интегрирования по времени и так далее. Каждая базовая операция отчуждается в виде модельных тестов с наборами входных данных, полученных из реальных задач. Затем для каждой операции, составляющих газодинамический алгоритм, выполняются поиск оптимальной реализации на архитектурах различных типов ускорителей NVIDIA и AMD. Полученные программные ядра (kernel код) внедряются в основной расчетный код. При этом каждая операция получает две реализации – исходную (для CPU) и новую, работающую на ускорителях. Каждая операция снабжается функцией проверки корректности вычислений путем сравнения выходных данных с исходной CPU версией, для того чтобы гарантировать отсутствие ошибок в адаптированной программе и иметь возможность в процессе счета проверять с некоторым временным шагом корректность работы самих ускорителей.

В качестве средства разработки был выбран открытый стандарт OpenCL (Open Computing Language) [2], который поддерживается большинством основных производителей вычислительного оборудования, включая Intel, AMD, NVIDIA, Sony, Apple и другие. Основная цель доклада состоит в том, чтобы изложить опыт в адаптации операций численных алгоритмов к различным GPU архитектурам. Будет представлено сравнение GPU от AMD и NVIDIA, сравнение возможности средств разработки OpenCL и CUDA. Будут описаны основные оптимизационные подходы, которые позволили получить чистую фактическую производительность более 15% от пика на GPU, что близко к теоретическому пределу для такого типа алгоритмов, существенно ограниченных пропускной способностью памяти и к тому же имеющих нерегулярный шаблон доступа к памяти.

Библиография:

1. И.А. Абалакин, П.А. Бахвалов, А.В. Горобец, А.П. Дубень, Т.К. Козубская, Комплекс программ NOISEtte для супервычислений в области аэродинамики и аэроакустики, XIII международный семинар «Супервычисления и математическое моделирование», с 3 по 7 октября 2011 года, г. Саров.
2. OpenCL - The open standard for parallel programming of heterogeneous systems, http://www.khronos.org/opencl/, 2012.