ДОКЛАДЫ АКАДЕМИИ НАУК, 2017, том 473, № 3, с. 277–281

ИНФОРМАТИКА

УДК 004.75

ПРОГРАММНАЯ ПЛАТФОРМА МАССОВОГО СУПЕРКОМПЬЮТИНГА

© 2017 г. Е. В. Биряльцев[1, 2], М. Р. Галимов[2], А. М. Елизаров[1, 3, *]

Представлено академиком РАН А.Б. Жижченко 14.11.2016 г.
Поступило 05.12.2016 г.

Описан опыт создания комплексной программно-аппаратной платформы для организации высокопроизводительных вычислений при решении современных геофизических задач на основе методов полноволновой инверсии. Обсуждены проблемы создания массовых высокопроизводительных программных комплексов для широкого использования в промышленности.

DOI: 10.7868/S0869565217090043

В настоящий момент времени развитие суперкомпьютерных вычислений достигло очередного качественного рубежа – производительность высокопроизводительных кластеров превысила планку в десятки петафлопс, и научное сообщество обсуждает вопросы будущих эксафлопных вычислений [1]. Повышение конкурентоспособности предприятий и экономики в целом во многом определяется активным массовым использованием суперкомпьютерных технологий [2]. В такой ситуации широкое внедрение суперкомпьютерных систем в промышленности становится еще более актуальным, тем более что такое внедрение наблюдается сегодня лишь в крупных корпорациях и стратегических отраслях.

Как известно, в 2000-х годах произошло резкое снижение стоимости и сложности высокопроизводительных вычислительных комплексов. Этому в значительной мере способствовали появление (в 2008 г.) и дальнейшее развитие технологии графических ускорителей (GPGPU). На конец 2015 г. в мировом списке Топ 500 [3] наиболее мощных суперкомпьютеров 90 из них имеют GPGPU, в том числе занимающий первую строку рейтинга суперкомпьютер Tiahe-2, причем число таких комплексов ежегодно увеличивается. За 7 лет, прошедших с момента ее появления, технология GPGPU достигла такой степени зрелости, которая сделала возможным ее применение не только в уникальных установках для решения научных и стратегических задач, но и для массового использования в прикладных областях. Одним из таких примеров служит высокопроизводительный вычислительный комплекс, созданный при участии авторов и описанный ниже.

1. Высокопроизводительные вычисления в сейсморазведке. Сегодня нефтегазовая промышленность характеризуется активным поиском и внедрением новых технологий, что объясняется, в частности, постоянным и значительным увеличением стоимости разведки и добычи углеводородного сырья. В этой ситуации происходит поиск новых методов геологоразведки, эффективных и одновременно малозатратных, которые позволили бы увеличить качество геологического прогноза и экономическую обоснованность принимаемых технических решений.

Технологии полноволновой инверсии [4] являются перспективными методами моделирования в геологии и позволяют восстанавливать свойства среды в любых геологических условиях, в отличие от технологий, основанных на базовом способе современной сейсморазведки – методе ОГТ (общей глубинной точки), эффективном при плоскопараллельной структуре исследуемой геологической среды. Частным случаем полноволновой инверсии можно считать разработанную в России технологию пассивной низкочастотной сейсморазведки нефти и газа [5], при которой анализируются изменения характеристик естественных микросейсм над нефтегазовыми залежами. Эта технология основана на расчетах распространения сейсмических волн в сложных геологических средах с помощью сеточных моделей с размерностями (в настоящее время) до 10^9. Для промышленного применения этой технологии необходимо выполнять расчеты для численных

[1] *Казанский (Приволжский) федеральный университет*
[2] *ООО «Градиент технолоджи», Казань*
[3] *Казанское отделение Межведомственного суперкомпьютерного центра Российской Академии наук – Филиал Федерального научного центра «Научно-исследовательский институт системных исследований Российской Академии наук»*
**E-mail: amelizarov@gmail.com*

моделей такой размерности со скоростями, сопоставимыми с темпом выполнения производственных работ (десятки часов), и экономически оптимальным способом. Для решения этой задачи специалисты казанской геофизической компании ЗАО "Градиент" совместно с Казанским филиалом Федерального научного центра "НИИ системных исследований РАН" и научно-производственной компанией ООО "Градиент технолоджи" (г. Казань), начиная с 2005 г., провели исследования по поиску наиболее эффективных программно-аппаратных средств для выполнения массовых математических вычислений при моделировании в геологии. При этом рассматривались различные аппаратные и программные платформы, было разработано специализированное программное обеспечение (ПО) для расчетов, в том числе с использованием кластеров GPU.

2. Программно-аппаратный комплекс высокопроизводительных вычислений. Результатом проведенных работ стало создание программно-аппаратного комплекса на основе графических ускорителей компании AMD. В последней (28 сентября 2015 г.) 23-й редакции Топ 50 российских суперкомпьютеров [6] этот вычислительный комплекс, предназначенный для решения геофизических задач, занял 38 место с производительностью 36,61 терафлопс по тесту Linpack на основе всего 6 узлов с графическими ускорителями. Кроме того, данный комплекс характеризуется в этом рейтинге максимальной производительностью на вычислительный узел, а наличие лишь 6 вычислительных узлов с 24 графическими картами позволяет считать его компактным и экономичным решением. С одной стороны, такие вычислительные устройства вполне доступны по цене и сложности технического обслуживания не только для крупных, но и для средних компаний. С другой стороны, они имеют достаточную вычислительную мощность для решения прикладных задач с использованием численных моделей размером до миллиарда ячеек.

Опыт разработки и внедрения этого комплекса показал, что в настоящее время созданная аппаратная база готова для промышленного внедрения. Непосредственный процесс сборки вычислительных мощностей, развертывания и настройки базового ПО не представлял особой сложности. Кроме того, в настоящий момент на российском ИТ-рынке присутствуют различные производители (Т-Платформы, Meijin, IntellectDigital, ARBYTE SC, STSS), которые готовы на заказ произвести сборку кластера определенной вычислительной конфигурации, стоимость которого вполне доступна для промышленного сектора. Также существует возможность аренды необходимых вычислительных мощностей, в том числе с GPU, в облачных системах (например, Amazon). Таким образом, задачу создания технической базы массового суперкомпьютинга для промышленного применения, поставленную в [2], можно считать в достаточной степени решенной.

Как видится, на ближайшее десятилетие архитектура массового суперкомпьютера будет представлять собой набор относительно небольшого (от единиц до первых сотен) количества гетерогенных CPU/GPU узлов с доступом GPU к памяти через PCIe или NVLink, соединенных высокоскоростной сетью Infiniband или сетью OmniPath, разрабатываемой сегодня компанией Intel. С прогнозируемым в ближайшем будущем увеличением доступности технологии SSD стандарта SAS 12000 Mbod каждый узел также может включать, согласно архитектуре DAS, высокоскоростное локальное хранилище из 8–16 накопителей емкостью в несколько десятков терабайт. Таким образом, можно с достаточной уверенностью полагать, что массовый суперкомпьютер будет представлять собой, согласно определению Джима Грея [7], набор однородных, легко заменяемых и наращиваемых "вычислительных кирпичей" совокупной мощностью 0,1–10 Пфлопс и совокупным объемом постоянного хранилища 0,1–10 Пбайт.

Вместе с тем массовый суперкомпьютинг, преодолев барьер высокой стоимости и сложности аппаратной базы, сталкивается с проблемой сложности разработки ПО. Этот барьер находится в стадии осмысления мировыми ИТ-лидерами, в частности, компания "Интел" имеет партнерскую программу Advanced Computing Program для поддержки разработки программных решений для массового суперкомпьютинга.

Проблема сложности разработки ПО для суперкомпьютерных вычислительных кластеров имеет два существенных аспекта: гетерогенность архитектуры вычислительной платформы и распределенность вычислений.

Для минимизации влияния гетерогенности архитектуры кластеров с GPU сейчас разрабатываются методы унификации программирования классических многоядерных центральных процессоров и систем с массовым параллелизмом GPU. В этом направлении наиболее перспективной выглядит среда OpenCL, которую поддерживают как производители центральных процессоров, так и производители GPGPU. Для межузлового взаимодействия достаточно распространенным применяемым инструментом служит протокол MPI. Связка протоколов MPI+OpenCL признается перспективной для низкоуровневого программного обеспечения суперкомпьютинга вплоть до экзафлопного масштаба (см. [2]). Таким

образом, проблема гетерогенности вычислительной платформы решается достаточно активно.

Распределенность вычислений, изначально заложенная в любом проекте суперкомпьютинга, – другой существенный аспект сложности разработки ПО. Традиционно суперкомпьютеры применялись для решения научных и стратегических задач на пределе имеющихся технических возможностей, основное внимание при этом обращалось на эффективность собственно численного моделирования. Однако с точки зрения известной модели программного обеспечения "Model-View-Controller" собственно расчетная часть системы численного моделирования является лишь одной из трех составляющих программного комплекса – развитию средств визуализации и управления данными численных моделей уделялось недостаточно внимания, в результате эти программные компоненты оказались недостаточно развиты. При внедрении же суперкомпьютерных технологий в стратегические отрасли промышленности применялась организационная модель разработки ПО привлекаемыми крупными компаниями или научно-исследовательскими институтами, обладающими высококвалифицированными кадрами. Это позволяло преодолевать сложности разработки и собственно системы моделирования, и подсистем взаимодействия с пользователем и управления данными.

3. Предлагаемый подход. Для массового применения численного моделирования в разнообразных областях централизованная разработка прикладного ПО невозможна. Применение суперкомпьютерных технологий в легкой, пищевой, строительной и других нестратегических отраслях базируется на многочисленных вариациях многомасштабных мультифизических моделей, всю гамму которых невозможно разработать централизованно. Поэтому развитие в этих отраслях базируется на творчестве большого числа малых инновационных компаний, не имеющих возможности нанимать дорогостоящие коллективы высококвалифицированных специалистов. Большое значение имеют также стоимость и сроки разработки и модернизации ПО. Для эффективного использования суперкомпьютерных технологий в названных отраслях представляется актуальной разработка программной платформы, резко снижающей барьер сложности создания прикладных систем с использованием суперкомпьютинга до уровня сложности программирования на персональных компьютерах. Такая платформа должна поддерживать сквозную работу с модельными данными (модели и результаты моделирования), а не только сам процесс моделирования.

Подходы и программные реализации для организации хранения и визуализации больших объемов данных, безусловно, известны. Однако эти подходы базируются на собственных моделях данных. Визуализация сложных трехмерных сцен активно используется в компьютерных играх, конструкторских задачах, современном кинематографе. Как правило, программные средства для решения этих задач базируются на полигональной модели. Хранение больших данных наиболее развито в настоящее время для больших бесструктурных текстовых/бинарных файлов или их множеств в системах поддержки социальных сетей, электронной почты и т.п.

Учитывая большой планируемый объем обрабатываемых данных, мы предлагаем строить программную платформу массового суперкомпьютинга на базе единой модели данных для минимизации занимаемого ими объема, трансформаций и снижения сложности освоения. Такая структура должна быть субоптимальной, в данном случае допустимой без существенного снижения производительности и других критических свойств ПО, для подсистем математического моделирования, визуализации и управления хранением. Рассмотрим модели данных, допустимость и эффективность которых при моделировании, визуализации и хранении в суперкомпьютерных вычислениях в георазведке подтверждена практикой использования.

Численное моделирование физических процессов базируется в настоящее время преимущественно на сеточных методах. Моделируемый объект представляется набором пространственно фиксированных узлов или ячеек (сеткой), на котором рассчитываются потоки физических параметров (величин). Этот подход хорошо разработан методологически и широко применяется для моделирования аэродинамических, прочностных, гидродинамических процессов, в том числе в подземной гидромеханике, тепловых и физико-химических расчетах с использованием методов конечных элементов и его вариаций. Практика решения на суперкомпьютерах научных и стратегических задач на пределе технических возможностей требовала оптимизации расчетных сеток по объему. Это привело к тенденции строить сложные нерегулярные сеточные конструкции с уникальными алгоритмами конструирования сеток, необходимостью анализа на обусловленность и построением предобусловливателей. Массовый суперкомпьютинг требует упрощения конструирования расчетных сеток, пусть и за счет некоторой потери оптимальности в объемах данных и времени расчетов. Известен подход к построению универсальных сеток с локальным измельчением на основе октодерева (см., например, [8]).

Для промышленного использования численных моделей в многопользовательском режиме актуальна динамическая балансировка нагрузки. Если для решения уникальных однократных задач возможно заранее спланировать оптимальное распределение подобластей по узлам кластера, то для асинхронного использования вычислительного кластера несколькими пользователями, которые могут запускать и снимать свои задачи, требующие различных ресурсов в произвольные моменты времени, актуально динамическое изменение числа узлов, доступных для решения задачи. Такие алгоритмы достаточно очевидны, в частности, возможна предварительная сегментация на максимально возможное число расчетных подобластей при динамическом управлении количеством смежных подобластей в рамках одного узла. Универсальная модель сеток на основе октодерева применима для параллельных расчетов [8].

Результаты моделирования в сеточных методах не соответствуют полигональной модели, наиболее популярной для визуализации в машинной графике – как правило, исследователя интересует распределение моделируемых полей внутри модельного объема. Естественным графическим представлением таких объектов являются воксельные модели, связывающие с вокселем, элементарным объемом пространства модели, набор параметров цвета и прозрачности. В нашем случае объект в виде сеточной модели уже существует, и для преобразования в визуальные характеристики достаточно связать моделируемые параметры со шкалами цвета и прозрачности. Воксельная графика применяется при представлении томографических исследований, в геофизике, а также для создания реалистичных компьютерных игр. Воксельное представление объектов значительно более объемное по сравнению с полигональным сопоставимой детальности, в связи с чем для кодирования воксельной графики применяют сжатие данных на основе октодеревьев [9].

Значительной проблемой для визуализации результатов численного моделирования является также большой объем получающейся информации. Кроме той очевидной проблемы, что модель большого объема, рассчитываемая на многих узлах, может не поместиться в графическую карту или их SLI-связку, узким местом является передача информации из моделирующих узлов в узел визуализации, особенно при визуализации динамических процессов. Визуализация данных численного моделирования должна быть распределенной. Рендеринг и захват должны производиться непосредственно в узлах моделирования, без пересылки данных на отдельные рабочие места или рендер-фермы.

Еще одной проблемой является организация хранения и доступа к численным моделям, в том числе результатам динамического численного моделирования. Численная модель может иметь размер в сотни гигабайт, а результаты динамического моделирования – десятки терабайт. При решении обратных задач и оптимизации технических решений и модель, и результаты моделирования могут существовать в десятках вариантов, как правило мало отличающихся друг от друга. Результаты моделирования записываются в постоянное хранилище для дальнейшего анализа десятками моделирующих узлов, в общем случае асинхронно. Считывание моделей и результатов моделирования из хранилища для визуализации и постобработки должно производиться в темпе онлайн-работы, т.е. максимум за десятки секунд.

Кластерные файловые системы HDFS, Lustre и аналогичные системы рассчитаны в первую очередь на надежное хранение неограниченного количества небольших файлов. Поскольку внешняя архивация больших данных практически невозможна, надежность обеспечивается многократным дублированием внутри системы. Кластерные файловые системы имеют также механизмы хранения файлов большого объема, превышающего размеры дисков, большие файлы можно сегментировать на блоки, хранящиеся на различных узлах кластера. Однако в существующих системах эта фрагментация никак не связана со структурой данных модели, что вызывает сложности при записи и считывании. Недостаточная адекватность такого подхода применительно к данным, имеющим внутреннюю структуру, в том числе данным численных моделей, хорошо известна [10].

Кластерные базы данных, такие как HBase, рассчитаны на хранение разреженных плоских файлов и могут их сегментировать по ключу, однако не поддерживают иерархические схемы, возникающие при моделировании при локальном измельчении сетки. Естественной моделью сегментации данных сеточной модели является пространственная сегментация на основе пространственного *R*-индекса, представимого, в частности, для двумерных моделей квадро-, а для пространственных – октодеревом. Пространственные базы данных хорошо известны, однако модель хранения данных в них не связана с распределением кубов данных по узлам кластера. Для платформы массового суперкомпьютинга на основе сеточных моделей оптимальной моделью хранения данных является модель с пространственной сегментацией, автоматическим распределением сегментов верхних уровней по узлам с одновременным обеспечением надежности по модели RAID с управляемым уровнем избыточности.

Таким образом, рассматриваемая платформа массового суперкомпьютинга с использованием численного моделирования должна иметь архитектуру на основе сквозной модели данных, удовлетворяющей задачам моделирования, визуализации и манипулирования данными, в том числе хранения, доступа и балансировки вычислений. Субоптимальной общей моделью данных являются 2k-деревья, в том числе квадродеревья – для статических двумерных моделей, октодеревья – для статических трехмерных моделей и динамических двумерных, а также 16-деревья – для динамических трехмерных моделей. Можно рассматривать деревья большей кратности для управления данными в пространстве версий, уровней доступа и прочих потенциально возможных размерностей без изменения основных алгоритмов обработки.

З а к л ю ч е н и е . Одним из основных потребителей новейших информационно-коммуникационных технологий традиционно является нефтегазодобывающая отрасль. С начала 2010-х годов в нефтегазовой сейсморазведке развивается технология полноволновой инверсии, работоспособная в геологических средах любой сложности и постепенно заменяющая классическую сейсморазведку (метод ОГТ). Полноволновая инверсия основана на численных методах решения обратных задач и ориентирована на высокопроизводительную обработку. В отличие от традиционных систем обработки методом ОГТ, технологическая база полноволновой инверсии находится пока на начальном этапе своего развития. Мировые лидеры информационного рынка совместно с ведущими геофизическими компаниями только ищут формы организации массовых прикладных программных систем, основанных на численном моделировании [11]. В России также ведутся разработки интегрированных прикладных систем [8, 12], основанных на суперкомпьютерном численном моделировании и одновременно включающих остальные компоненты полного цикла обработки информации, такие как визуализация и управление данными и процессом численных расчетов.

Важнейшим фактором для успешного применения массового суперкомпьютинга является максимально доступная для освоения программная платформа, сформированная на основе модели данных, единой для всего цикла обработки информации, субоптимальной для собственно расчетов, управления данными и визуализации. Универсализация структур данных позволит сократить трансформации многотерабайтных массивов данных в системе и, самое главное, минимизировать сложность освоения программной платформы в прикладных разработках и увеличить скорость разработки и модификации прикладных систем.

Работа выполнена при финансовой поддержке РФФИ (проекты 15–07–08380, 15–47–02343).

СПИСОК ЛИТЕРАТУРЫ

1. *Da Costa G.* Exascale Machines Require New Programming Paradigms and Runtimes // Supercomputing Frontiers and Innovations. 2015. V. 2. P. 6–27.
2. *Бетелин В. Б., Велихов Е. П., Кушниренко А. Г.* Массовые суперкомпьютерные технологии – основа конкурентоспособности национальной экономики в XXI веке // Информ. технологии и вычисл. системы. 2007. № 2. С. 3–10. http://www.jitcs.ru/index.php?option=com_content&view=article&id=178
3. http://www.top500.org/
4. *Haffinger P. R.* Seismic Broadband Full Waveform Inversion by Shot/Receiver refocusing. Delft: Delft Univ. Technol. 2013. http://repository.tudelft.nl/view/ir/uuid%3Ad2d8d264–5037–4573–8418-a079afa8d1e7
5. *Шабалин Н. Я., Рыжов В.А, Биряльцев Е. В.* Пассивная низкочастотная сейсморазведка – мифы и реальность // Приборы и системы разведочной геофизики. 2013. № 2. С. 46–53.
6. http://top50.supercomputers.ru/?page=rating
7. The Fourth Paradigm: Data-Intensive Scientific Discovery. http://research.microsoft.com/en-us/UM/redmond/about/collaboration/fourthparadigm/4th_PARADIGM_BOOK_complete_HR.pdf
8. *Василевский Ю. В., Коньшин И. Н., Копытов Г. В., Терехов К. М.* INMOST – программная платформа и графическая среда для разработки параллельных численных моделей на сетках общего вида. М.: Изд-во МГУ, 2012. 144 с.
9. *Tero Karras, Samuli Laine, Gregory J. Ward.* Efficient Sparse Voxel Octrees – Analysis, Extensions, and Implementation. http://mediatech.aalto.fi/samuli/publications/laine2010tr1_paper.pdf
10. *Stonebraker M., Kepner J.* Possible Hadoop Trajectories. http://cacm.acm.org/blogs/blog-cacm/149074-possible-hadoop-trajectories/fulltext
11. Global Technology Leaderin Visualizationand Visual Compute. http://hue.no/sites/default/files/Hue_Brochure_US_web.pdf
12. *Биряльцев Е. В., Богданов П. Б., Галимов М. Р., Демидов Д. Е., Елизаров А. М.* Программно-техническая платформа высокопроизводительных вычислений для нефтегазовой промышленности // Программные системы: теория и приложения. 2016. № 1 (28). С. 15–27. http://psta.psiras.ru/2016/01(028)/ч5/ч5–7_.html