Открытые системы

Открытые Системы № 7, 2011

Graph500: адекватный рейтинг

На конференции SuperComputing 2010 был предложен список Graph500, претендующий на более адекватное, чем в Top500, отражение рейтинга суперкомпьютеров, способных обрабатывать большие массивы данных. Что побудило создать очередной тест и в чем его суть?

Леонид Эйсымонт, Александр Фролов, Александр Семенов

Изначально суперкомпьютеры, как известно, создавались для решения вычислительных задач: моделирования физических процессов, инженерных расчетов, баллистики, криптоанализа и т. п., для которых характерна хорошая пространственно-временная локализация данных в памяти, что позволяет эффективно использовать быструю кэш-память процессоров. Для оценки производительности суперкомпьютеров при выполнении такого класса задач (Cache Friendly, CF) хорошо подходит тест Linpack, положенный в основу списка Top500, в 36-й версии которого впервые оказался китайский суперкомпьютер Tianhe-1A с пиковой производительностью 4,702 PFLOPS.

Критика теста Linpack, применяемого для составления рейтингового списка, который приобретает все больше политическое значение, звучала достаточно давно и предлагались разные методики более объективной оценки суперкомпьютеров, например оценочный тестовый набор HPC Challenge. Однако такой популярности, как Top500, ни одна оценка до сих пор не получила в силу сложности тестовых программ и трудности восприятия широкой общественностью. В то же время все большую важность для экономик развитых стран и систем обеспечения национальной безопасности стали приобретать задачи, отличные от CF-класса, — новые типы вычислительных задач и приложения обработки больших пулов информации: анализ социальных сетей, выявление заданных и характерных ситуаций (в том числе и террористических угроз) посредством анализа накопленных в неструктурированных базах графового типа оперативно получаемых данных и т. п. Такие задачи сначала называли *DIS (Data Intensive System)*, а сегодня рассматривают их более широко, добавляя еще и затраты на обработку входных/выходных потоков данных — *DIC (Data Intensive Computing)*.

Вычисления с акцентом на данные

Необходимость обработки петабайтных массивов данных вызвала к жизни новые подходы, которые получили название Data-Intensive Computing.

Как финансируется DIC

Сегодня в России активно верстаются государственные программы финансирования НИОКР по стратегическим направлениям, в число которых входят и решения в обла-

сти Data Intensive Computing. Однако помимо финансовой поддержки таких работ, необходимо также решать множество организационных и управленческих проблем, и здесь еще многому надо поучиться.

Появление списка Graph500 отражает созревшее в обществе признание важности DIC-задач, и, хотя по форме предложения Graph500 явно противопоставляется Top500, первый не исключает, а дополняет второй для получения объективной оценки возможностей суперкомпьютера, причем в такой же простой и доходчивой форме.

Специфика задач DIC-класса

Для задач DIC-класса характерны: работа с наборами данных, объем которых значительно превышает память вычислительного узла современного суперкомпьютера (до нескольких петабайтов); высокая интенсивность выполнения операций над данными по отношению к вычислительным операциям; высокая непредсказуемость нерегулярно разбросанных по памяти адресов данных; возможность сильного распараллеливания с использованием взаимодействующих друг с другом процессов. Поэтому средств и приемов оптимального выполнения CF-задач для DIC-задач недостаточно, хотя попытки их применения встречаются. Например, по нашему мнению, за редким исключением не имеет смысла использовать различные ускорители на программируемых логических матрицах (FPGA). Как следствие, базовые характеристики суперкомпьютеров, подходящих для решения DIC-задач, должны быть иными, например очень важна способность быстрой передачи вычислений к данным на удаленных узлах посредством аппаратно поддержанного механизма удаленного вызова процедур. Метрики быстродействия также иные, например более адекватны единицы GUPS [1], определяемые на тесте RandomAccess. Однако эти важные для профессиональной оценки характеристики относятся к архитектурному уровню и явно не связаны с приложениями, поэтому не годятся для составления воспринимаемого общественностью рейтингового списка.

Многие DIC-задачи сводятся к типовым алгоритмам на графах: поиск кратчайшего пути между заданными вершинами; поиск вхождения графа заданного вида в другой граф; поиск вширь и вглубь на графах — задачи BFS (breadth-first search) и DFS (depth-first search); нахождение вершин, через которые проходит наибольшее количество кратчайших путей, — задача BWC (Betweenness Centrality), и др. Некоторые прикладные задачи естественным образом приводятся к задачам на графах, а другие сводятся к графовым. Например, задача BWC может непосредственно использоваться в приложениях реального времени в электроэнергетике для определения наиболее ответственных участков сети и анализа непредвиденных ситуаций, а также в графах, моделирующих в финансовой области зависимости стоимости акций друг от друга. Другой пример — использование графов, моделирующих взаимодействие белков, где требуется находить небольшие графы, изоморфные заданным, чтобы определять свойства и моделировать молекулярные процессы в клетках.

В инициативную группу Graph500 входят Cray, SGI, Intel, IBM, AMD, nVidia, Oracle, LexisNexis и др. Кроме того, в группу входят три национальные лаборатории Министерства энергетики США и лаборатория по созданию оружия в Великобритании, три правительственные организации — Национальный научный фонд США, Пентагон и Агентство перспективных оборонных исследований (DARPA), а также 11 университетов и суперкомпьютерных центров.

Руководителем проекта Graph500 является Ричард Мерфи. Он занимается компьютерными архитектурами в Национальной лаборатории Sandia министерства энергетики США, возглавляя одну из групп разработчиков, включенных в исполнители новой программы UHPC (инициированная DARPA программа «вездесущих» высокопроизводительных вычислений Ubiquitous High Performance Computing). Кстати, задачи на графах, причем в более сложном варианте, когда они еще и динамически изменяются, – одно из основных приложений, которое должно успешно решаться на создаваемых в рамках этой программы суперкомпьютерах с революционной архитектурой. Таким образом, за

Graph500 стоят серьезные ведомства, и к этому рейтингу стоит относиться соответственно.

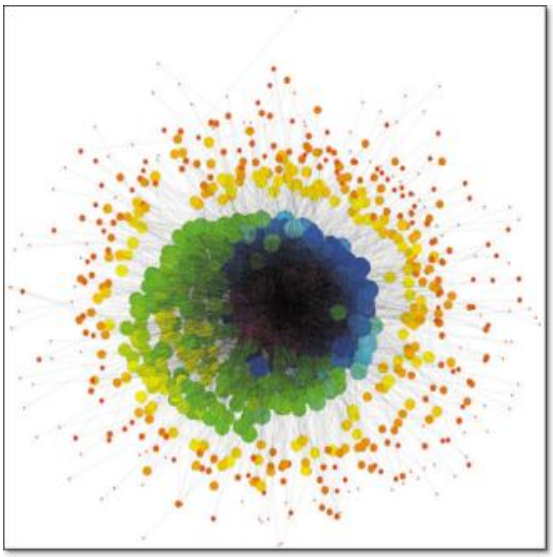
DARPA UHPC — дорога к экзафлопсам

В августе 2010 года появились сообщения о начале работ по программе DARPA UHPC, предусматривающей создание принципиально новых высокопроизводительных компьютеров экзафлопсного уровня. Ожидается, что эта программа определит направления работ в области суперкомпьютеров на предстоящее десятилетие.

Методика Graph500

Тест, используемый для формирования списка Graph500, включает два ядра. Сначала генератор создает граф в виде списка дуг, затем первое ядро порождает его внутреннее представление, которое используется впоследствии вторым ядром, выполняющим алгоритм поиска вширь в графе, – решается задача BFS.

Тест Graph500 имеет два параметра: SCALE и edgefactor; первый задает общее количество (N) вершин графа, которое равно 2^{SCALE}, а второй определяет количество дуг (M) в графе, равное edgefactor*N. По умолчанию edgefactor = 16.



Пример графа, генерируемого тестом Graph500 (Университет шт. Индиана)

Графы, используемые в приложениях и моделирующие реальный мир (например, социальные сети), имеют большое количество вершин и малое число соседей либо небольшое количество вершин и много соседей. Число вершин, имеющих k соседей, обозначаемое P(k), в таких графах убывает в соответствии с законом $P(k) = ck^{-a}$ (как правило, $2 < a < 3$). В таких графах содержится небольшое количество вершин, имеющих настолько большое количество связей, что именно это обеспечивает высокую среднюю связность графа. Такие вершины называются «хабами» (hub). Хабы с разным количеством связей образуют некоторую иерархическую структуру, в которой связность вершин-хабов с переходом на более высокий уровень увеличивается. На рисунке приведен пример сгенерированного в Graph500 графа, в котором цветом и размером вершин обозначено количество соседей у каждой вершины; чем больше вершина и темнее оттенок цвета (используется цветовая модель HSL), тем больше у нее соседей. Знание таких особенностей графов может быть важно при оптимизации теста Graph500.

При генерации графов в Graph500 используется генератор Кронекера, очень похожий на генератор графов типа Recursive MATrix (R-MAT), который в процессе работы использует матрицу смежности создаваемого графа. При добавлении каждой дуги матрица смежности NxN рекурсивно дробится до тех пор, пока не будет получена матрица из одного элемента — это и есть выбранная дуга. Такой процесс повторяется M раз. Матрица на каждом шаге такого рекурсивного процесса дробится на четыре равные части: A, B, C и D. Для каждой из этих частей изначально задана вероятность, с которой происходит выбор именно ее при добавлении новой дуги. По умолчанию вероятности выбора частей матрицы равны: P(A) = 0,57; P(B) = 0,19; P(C) = 0,19; P(D) = 1-(A+B+C) = 0,05.

В результате такого процесса создается список дуг графа. При этом генератор может породить небольшое количество петель и кратных дуг, они могут не учитываться впоследствии, но обязательно попадают в итоговый список. После генерации дуг вершины графа нумеруются случайным образом и также случайным образом перемешиваются дуги в списке дуг каждой вершины. Генератор графа может быть параллельным.

Внутреннее представление графа, создаваемое из списка дуг в первом ядре, может быть любым, но его модификация в следующем ядре, выполняющем поиск, запрещена. Внутреннее представление включает размещение вершин графа в памяти вычислительных узлов суперкомпьютера, если реализация распределенная. В базовой реализации Graph500 номер вычислительного узла-хозяина вершины определяется как остаток от деления номера вершины на количество вычислительных узлов. Таким образом, вершины распределены по вычислительным узлам циклически. Для каждой вершины, принадлежащей данному узлу, на этом узле хранится список номеров вершин, смежных с данной.

Алгоритм BFS, выполняемый вторым ядром, для заданного графа и заданной вершины r находит все вершины, расположенные от нее на расстоянии в одну дугу, две дуги и т. д. Особенностью алгоритма является то, что он не должен анализировать вершины, отстоящие от вершины r на расстоянии s+1 до тех пор, пока не проанализирует все вершины, отстоящие от вершины r на расстоянии s дуг. Повторно вершины графа не рассматриваются. В результате работы алгоритма BFS получается дерево, корнем которого является исходная вершина r, а узлы этого дерева — множества вершин графа, находящихся на таком расстоянии от r, которое соответствует уровню этого узла в дереве.

Производительность алгоритма BFS измеряется количеством *пройденных дуг графа в секунду* (Traversed Edges Per Second, TEPS). Используются обозначения ME/c и GE/c – миллионы и миллиарды пройденных дуг в секунду соответственно.

Рейтинг Graph500 составляется с учетом сначала размера графа, а затем развиваемой в процессе поиска вширь производительности. Существует шесть диапазонов размеров графов, соответствующих значениям параметра SCALE: toy, mini, small, medium, large и huge. Объем данных для toy соответствует примерно 10^{10} байт (17 Гбайт, SCALE=26), mini — 10^{11} байт (140 Гбайт, SCALE=29) и т. д., объем данных для huge составляет 1,1 Пбайт (SCALE=42). При SCALE=30 количество вершин графа около миллиарда, а самый большой граф при SCALE=42 – четыре триллиона узлов.

По словам Мерфи, планируется создать версию Graph500 для включения в тесты SPEC и планируется включение в тест дополнительных ядер.

Реализация

Дистрибутив Graph500 включает две последовательные реализации на языке Си, версию на языке высокого уровня GNU Octave, параллельную версию с использованием OpenMP, две параллельные версии для Cray XMT (одна базовая, другая оптимизированная), а также две MPI-версии (одна базовая, с использованием функций двустороннего взаимодействия, другая написана с использованием функций одностороннего взаимодействия нового стандарта MPI 2.0).

Базовая версия для Cray XMT похожа на OpenMP-реализацию с использованием формата CRS (Compress Row Storage) для хранения строк разреженной матрицы смежности, соответствующих вершинам графа. Кроме того, для синхронизации тредов используются теговые биты ячеек памяти и атомарные операции. Cray XMT обладает уникальной возможностью работы с памятью разных вычислительных узлов через единое виртуальное адресное пространство – общую память, это и позволяет использовать OpenMP.

В оптимизированной версии для Cray XMT очередной список достижимых вершин следующего уровня (уровень s+1), который в конечном итоге размещается в общей памяти, формируется в два этапа. Сначала формируются фрагменты этого списка в локальной памяти вычислительных узлов. Затем эти фрагменты включаются в список вершин следующего уровня, находящийся в общей памяти. Суть оптимизации в том, что при добавлении очередного фрагмента счетчик достижимых вершин увеличивается на количество

вершин во фрагменте, а не на единицу, если бы каждую достижимую вершину добавляли в отдельности. Это позволяет избежать узкого места, связанного с ожиданием выполнения атомарной операции увеличения счетчика.

В MPI-версиях теста также используется формат хранения разреженной матрицы CRS. В базовой версии на MPI 1.0 узлы обмениваются списками доступных вершин следующего уровня посредством асинхронных посылок сообщений. Перед такими посылками в каждом узле номера вершин сначала накапливаются в сообщениях, а после этого производится посылка. В версии на MPI 2.0 производятся поэлементные односторонние записи номеров вершин в список следующего уровня посредством асинхронной функции одностороннего взаимодействия MPI_Accumulate. При этом в конце обработки каждого уровня посредством функции MPI_Win_fence производится барьерная синхронизация для завершения всех выданных ранее асинхронных записей.

Характерно, что в реализациях теста Graph500 сразу используется множество средств и стилей программирования, что отражает одну из сторон современной проблематики решения DIC-задач.

Для получения результата можно использовать одну из реализаций, но при этом поощряется создание собственных версий. Использование стандартной версии помечается в списке как Reference, использование своей – Optimized.

Для проверки созданной версии в тесте существует валидация, которая проверяет корректность построенного дерева при поиске вширь. Для этого при выполнении поиска вширь заполняется специальный массив, в котором хранятся вершины-родители для каждой вершины в построенном дереве.

Первые результаты

Пока в списке Graph500 всего девять суперкомпьютеров (см. таблицу), но даже при таком небольшом количестве удивляет разнообразие представленных архитектур. На первом месте - IBM BlueGene/P, на котором при использовании 8192 узлов получено 6,6 GE/c для графа medium (17 Тбайт). На втором месте с результатом 5,22 GE/c на графе small (1 Тбайт) довольно старый суперкомпьютер Cray XT4 с узлами на базе коммерческого процессора AMD Opteron и коммуникационной сетью Cray Seastar2. Оба этих суперкомпьютера имеют заказные сети с топологией 3D-тор.

В списке Graph500 представлено три системы Cray XMT – суперкомпьютера на базе 128-тредовых микропроцессоров Threadstorm разработки Cray, соединенных сетью Cray Seastar2. Главная особенность Cray XMT – глобально адресуемая виртуальная память, эффективность работы с которой обеспечивается за счет большой мультитредовости микропроцессоров Threadstorm и принятых специальных архитектурных решений, позволяющих одновременно выполнять более тысячи обращений к памяти от одного микропроцессора. Сегодня Cray XMT позиционируется как наиболее удачный по архитектуре суперкомпьютер для решения аналитических, в том числе графовых, задач.

В списке также присутствуют суперкомпьютеры, построенные с использованием коммерческой сети Infiniband и коммерческих микропроцессоров. Лучший результат для таких кластерных суперкомпьютеров находится на пятом месте.

Мерфи высказал предположение, что постепенно облачные архитектуры займут значительную часть списка, но господствовать в его верхней части будут суперкомпьютеры с экзотическими архитектурами. Собственно говоря, это не новость — именно это было целью федеральной программы США прошлого десятилетия DARPA HPCS и является задачей программы следующего десятилетия DARPA UHPC.

Черты лидера Graph500

В ближайшем будущем следует ожидать “героических” реализаций теста Graph500 на суперкомпьютерах, которые, вообще говоря, не предназначены для решения графовых задач. Однако идея этого рейтингового списка в другом – пробудить понимание важности и активизировать исследования в области новых архитектур для DIC-задач. Сейчас уже

готовы некоторые суперкомпьютеры с улучшенной организацией как вычислительных узлов, так и коммуникационных сетей. Наиболее интересны суперкомпьютеры с мощными многоядерно-мультитредовыми узлами и многосвязными коммуникационными сетями: IBM Blue Waters (32-тредовые микропроцессоры Power 7, 32-процессорные вычислительные узлы, иерархическая полносвязная коммуникационная сеть на 48-портовых маршрутизаторах HUB); Tianhe-1A (64-тредовые микропроцессоры FT-1000 собственного производства, четырехсокетные платы с этими микропроцессорами, сеть Arch). Кластерные суперкомпьютеры также должны улучшить результаты за счет появляющихся четырех- и восьмисокетных вычислительных узлов на базе новых многоядерных микропроцессоров и сети Infiniband нового поколения. В работе [2] для четырехсокетного узла с восьмиядерными Nehalem EX для R-MAT-графа с 200 млн вершин и 1 млрд дуг получено 550 ME/c, что сравнимо с результатами для Reference-реализаций в текущем списке Graph500.

По мнению Мерфи, имеются три ключевые архитектурные особенности систем, помогающие эффективно решать DIC-задачи:

- поддержка возможности одновременного выполнения большого количества обращений к памяти, что позволяет работать с памятью со скоростью, определяемой не задержками выполнения операций с ней, а темпом их выдачи/приема;
- аппаратная поддержка большого количества выполняющихся потоков, что позволяет получить мощный поток обращений к памяти;
- поддержка мелкозернистых средств взаимодействия и синхронизации параллельных процессов.

Всеми перечисленными свойствами обладает Cray XMT, и ожидается появление следующего суперкомпьютера Cray XMT-2 с более современным мультитредовым процессором и новой коммуникационной сетью Cray Gemini. Машину Cray XMT/XMT-2 планируется применять совместно с большими базами данных в различных системах поддержки принятия решений. Похожие архитектурные решения были приняты и в российском проекте [3].

Дальнейшее развитие в области архитектур ведется, например, в рамках работ по программе UHPC, одно из направлений которой возглавляет Мерфи — ученик и последователь известного гуру в области компьютерных архитектур Питера Когги. Эта школа на протяжении ряда лет развивает направление интеллектуализации памяти посредством добавления в модули процессоров обработки данных (Processor In Memory, PIM). Проект X-caliber, в котором участвует Мерфи, нацелен на разработку 3D-структур с кристаллами памяти и процессоров, и ожидается, что на разных уровнях этой структуры могут оказаться подложки с многоядерными процессорами разного типа, скорее всего, со множеством асинхронных тредов и множеством синхронных тредов. Пиковая пропускная способность 2048 узлов суперкомпьютера X-caliber будет на два порядка выше (260 Пбайт/с), чем, например, у системы с таким же количеством узлов (Cray Seastar2 и Cray Gemini): явно сделана заявка на очень сильную архитектуру, в первую очередь предназначенную для решения DIC-задач.

***

Graph500 замышляется как масштабный проект, отражающий, с одной стороны, научные и коммерческие задачи с интенсивным нерегулярным доступом, а с другой — влияющий на развитие архитектур суперкомпьютеров революционной архитектуры. Удастся ли Graph500 стать таким же успешным, как Top500, и оправдают ли ожидания перспективные архитектуры наподобие Cray XMT-2 и X-caliber, покажет будущее.

Литература

1. Дмитрий Волков, Александр Фролов. Оценка быстродействия нерегулярного доступа к памяти // Открытые системы. – 2008. – №1. – С. 15-19.

2. V.Agarwal, F.Petrini, D.Pasetto, D. Bader. Scalable Graph Exploration on Multicore Processors // The 22nd IEEE and ACM Supercomputing Conference (SC10), New Orleans, LA, November 13-19, 2010.

3. Аеатолий Слуцкин, Леонид Эйсымонт. Российский суперкомпьютер с глобально адресуемой памятью // Открытые системы. – 2007. – №9. – С. 42-51.

Александр Семенов *(semenov@nicevt.ru),* ***Александр Фролов*** *(frolov@nicevt.ru) – сотрудники ОАО НИЦЭВТ,* ***Леонид Эйсымонт*** *(verger-lk@yandex.ru) – сотрудник «НИИ «Квант» (Москва).*