РАСПРЕДЕЛЕННАЯ ОБРАБОТКА ДАННЫХ[1]

Вл.В. Воеводин

НИВЦ МГУ

Удивительно быстро развивается вычислительная техника. Еще недавно производительность в один гигафлопс прочно ассоциировалась с суперкомпьютерами, а сегодня это обычная характеристика обычной персоналки. Для того чтобы иметь представление о современных компьютерах с максимальной производительностью, два раза в год формируется список пятисот самых мощных компьютеров мира. В ноябре 2003 года опубликована 22-я редакция списка, определившая лидеров нашего времени. На первом месте находится японский компьютер Earth Simulator производства фирмы NEC с производительностью на тесте Linpack 35,86 Тфлопс. Вторую позицию занимает компьютер ASCI Q производства Hewlett Packard с производительностью на тесте Linpack 13,88 Тфлопс. На третьем месте в новом списке оказался кластер X, состоящий из 2200 процессоров Apple G5 2,0 GHz, соединенных коммуникационной сетью Infiniband. Данный кластер стал третьим компьютером, преодолевшим рубеж 10 Тфлопс, показав на тесте Linpack производительность 10,28 Тфлопс. На данный момент в списке находится 131 система с производительностью на тесте Linpack более 1 Тфлопс (в предыдущей редакции списка было лишь 59 систем). Суммарная производительность 500 компьютеров из списка возросла за полгода с 375 до 528 Тфлопс. Для того чтобы попасть в список, на данный момент требуется производительность более 403,4 Гфлопс (по сравнению с 245,1 Гфлопс в июне 2003 года). Фантастика…

Интересно, что лидер нынешнего списка компьютер Earth Simulator держится на своей позиции уже второй год. В нем все уникально: архитектура, характеристики, возможности. Его создание, безусловно, стало возможным только в результате выполнения проекта действительно общенационального масштаба. Производство компьютера было закончено в феврале 2002 года, после чего он был установлен в Японском центре морских наук и технологий (Japan Marine Science and Technology Center). Сам компьютер и все его технологическое окружение (системы электропитания, кондиционирования, освещения, сейсмозащиты и т.п.) занимают здание размером 50×65×17 метров.

[1] Работа выполняется при поддержке РФФИ, грант № 02-07-90442.

Earth Simulator содержит 640 процессорных узлов, соединенных между собой через высокоскоростной переключатель. В состав узла входит 8 векторных арифметических процессоров, работающих над общей для каждого узла оперативной памятью, коммуникационный процессор и процессор для операций ввода/вывода. Оперативная память каждого узла разделена на 2048 банков и имеет объем 16 Гбайт. Пиковая производительность одного арифметического процессора равна 8 Гфлопс, поэтому пиковая производительность всего компьютера, объединяющего 5120 процессоров, равна 40 Тфлопс. Попутно заметим, что производительность на тесте Linpack составляет 89,6% от теоретического максимума – это также рекордный показатель среди участников списка Top500.

И, наконец, еще одной отличительной особенностью компьютера является то, что его архитектура объединяет все три базовых принципа построения высокопроизводительных систем. Если рассматривать его в целом, то Earth Simulator является массивно-параллельным компьютером с распределенной памятью. Вместе с тем каждый процессорный узел построен на принципах SMP-архитектуры, причем основа каждого процессора – это векторно-конвейерная обработка. Все лучшие качества основных классов современных высокопроизводительных систем были использованы для создания безусловного лидера – суперкомпьютера Earth Simulator.

Но лидер всегда один и уникален, а что же сегодня чаще всего используется на практике? В настоящее время самый динамичный сегмент принадлежит кластерным системам. Общее число кластеров в списке Top500 продолжает быстро увеличиваться и составляет уже 208 систем (149 в июне 2003 года), причем в первую десятку наиболее мощных компьютеров сейчас входит 7 кластеров.

Если попытаться определить коротко, то вычислительный кластер – это совокупность компьютеров, объединенных в рамках некоторой сети для решения одной задачи. В качестве вычислительных узлов обычно используются доступные на рынке однопроцессорные компьютеры, двух- или четырехпроцессорные SMP-серверы. Каждый узел работает под управлением своей копии операционной системы, в качестве которой чаще всего используются стандартные ОС: Linux, NT, Solaris и т.п. Состав и мощность узлов могут меняться даже в рамках одного кластера, давая возможность создавать неоднородные системы. Выбор конкретной коммуникационной среды определяется многими факторами: особенностями класса решаемых задач, доступным финансированием, необходимостью последующего расширения кластера и некоторыми другими. Возможно включение в конфигурацию специализированных компьютеров, например файл-сервера, и, как правило, предоставлена возможность удаленного доступа на кластер через Интернет.

Простор для творчества при проектировании кластеров огромен. Рассматривая крайние точки, кластером можно считать как пару ПК, связанных локальной Ethernet-сетью, так и упоминавшийся выше кластер X, занимающий третье место в списке Top500.

По “кластерному” пути в 1998 году пошли и мы в Научно-исследовательском вычислительном центре МГУ. Именно тогда было принято решение сделать основную ставку на кластерные системы, и время показало, что это решение было абсолютно правильным. Первым мы построили кластер на основе 24 процессоров Intel Pentium-III/500 MHz с использованием сети SCI в качестве вычислительной коммуникационной среды. Через полгода решили провести модернизацию и добавить еще шесть узлов (12 процессоров). И вот здесь, казалось бы, на столь тривиальном действии, нас поджидала неожиданность – узлов на базе использованных процессоров и нужных нам материнских плат в продаже уже не было. Нашли вариант на базе Intel Pentium-III/550 MHz, проблему модернизации кластера введением легкой неоднородности решили, но одновременно появилось четкое представление о серьезной задаче на ближайшее будущее: процессоры быстро совершенствуются, значит, неоднородность кластеров НИВЦ МГУ будет столь же быстро расти. Можно ли будет в таких условиях использовать весь потенциал вычислительных ресурсов суперкомпьютерного центра или придется для каждой задачи выделять лишь некоторое число процессоров какого-либо одного кластера?

Задача оказалась исключительно актуальной. Ниже показана современная структура вычислительного комплекса НИВЦ МГУ, причем в ближайшее время ряды кластеров пополнятся новыми установками:

• Кластер SCI, 36 CPUs, 18 узлов: 2 × Intel Pentium III/500 MHz, 1GB, HDD 3,2 GB, Fast Ethernet + SCI, 18 Gflops,

• Кластер SKY, 40 CPUs, 20 узлов: 2 × Intel Pentium III/850 MHz, 1GB, 2 × HDD 15 GB, Fast Ethernet, 34 Gflops,

• Кластер AQUA, 40 CPUs, 24 узла: 2 × Intel Pentium III/1 GHz, 1GB, HDD 20 GB, Fast Ethernet, 48 Gflops,

• Кластер LEO, 32 CPUs, 16 узлов: 2 × Intel Xeon 2.6 GHz, 2GB, 2 × HDD 40 GB, Fast Ethernet + SCI (D335), 166 Gflops.

Как обеспечить преемственность вычислительной техники для тех пользователей, которым необходимы максимально возможные ресурсы? Как сделать так, чтобы появление новых систем не приводило бы к забвению старых, а просто давало пользователю приращение мощности за счет добавления новых процессоров к уже имеющемуся парку компьютеров?

Заметим, что похожая картина в настоящее время складывается и в сети Интернет: скоростные характеристики сети растут, число процессоров в сети огромно – чем не параллельный компьютер? Причем самый мощный в мире…. Но вот как его использовать для решения одной задачи? В

этом и состоит проблема, которую мы решаем с помощью разработанной в НИВЦ МГУ системы X-Com.

Основные требования к проектируемой системе были сформулированы следующим образом:

• ориентация на вычислительные задачи;

• работа через Интернет, возможность использования всех доступных в сети вычислительных ресурсов различной мощности;

• минимум дополнительных действий и системного вмешательства на используемых ресурсах;

• масштабируемость системы, устойчивость к неоднородности и изменению конфигурации вычислительной среды;

• простота адаптации прикладных программ.

К настоящему времени система прошла апробацию в ходе нескольких масштабных экспериментов с использованием широкого спектра ресурсов: от простых домашних компьютеров до мощных параллельных вычислительных систем, от использования сети до работы на выделенном гигабитном полигоне Московского университета. В каждом случае на подготовку и разворачивание распределенного эксперимента требовалось не более 1 часа.

Совместно с группой специалистов из Центра «Биоинженерия» РАН, в распределенном режиме решалась задача определения скрытой периодичности в генетических последовательностях. Основную проблему в данном случае представляли высокая вычислительная сложность алгоритмов и огромный объем входных данных. По самым скромным подсчетам для обработки материала, имеющего реальное научное значение, на одном процессоре потребовалось бы обработать несколько гигабайт входной информации и многие месяцы, а то и годы, непрерывных вычислений. На вход программы подавались описания структуры генетических последовательностей, закодированные стандартным набором символов a-t-g-c, а на выходе появлялась информация о найденных во входных последовательностях повторах тех или иных фрагментов.

Все входные данные были взяты из специализированного банка данных – центрального хранилища актуальной информации о генетической структуре различных организмов (http://ncbi.nih.gov). Во всех расчетах были задействованы географически удаленные вычислительные ресурсы, объединенные с помощью системы X-Com.

Для решения исходной задачи в среде X-Com все входные последовательности разбивались на порции, содержащие от нескольких десятков до нескольких сотен тысяч символов. Часть порций нарезалась с “перехлестом”. В этом случае конец предыдущей порции совпадал с началом последующей, что было необходимо для нахождения повторов в самой зоне разрезания. Количеством символов в порции можно было регулировать объем

вычислений, необходимый для ее обработки в разумное время. В проведенных экспериментах время обработки одной порции менялось в пределах нескольких десятков минут. Для сокращения сетевого трафика выполнялась online-компрессия всех передаваемых данных. Общий объем написанного для интеграции с системой X-Com дополнительного кода составил менее 200 строк.

Проводя серию экспериментов, одновременно с решением основной прикладной задачи нами преследовались две дополнительные цели: проверить потенциальные возможности X-Com и определить особенности ее работы при пиковых нагрузках. Остановимся подробнее на описании двух наиболее интересных экспериментов.

Цель первого эксперимента – географический размах: 14 удаленных вычислительных систем, принадлежащих 10 организациям, расположенным в 8 городах и 6 различных часовых поясах. В разные моменты времени к эксперименту подключалось 407 процессоров, причем максимальное число одновременно работающих процессоров составило 385, а в среднем по эксперименту эта величина оказалась равной 296. Общее время эксперимента составило около 64 часов. Один средний компьютер справился бы с подобной задачей лишь за два года непрерывной работы. Взаимодействие между компьютерами шло по обычным каналам связи, и суммарно было передано 9 Гбайт данных. Максимальная загрузка исходящего канала данных от центрального сервера составила около 7 Мбит/с. Никаких специальных действий, направленных на обеспечение гарантированной полосы пропускания или увеличение стабильности связи, не предпринималось.

Таблица 1

Вычислительные ресурсы “географического” эксперимента

Организация и кластер	Город	Платформа	Число	
			процессоров	выполненных заданий
ЛИТ ОИЯИ. DN1	Дубна	Linux / Intel	16	3824
ЛИТ ОИЯИ. DN2	Дубна	Linux / Intel	16	3712
ИММ УрО РАН	Екатеринбург	Linux / Intel	16	2108
НИИЯФ МГУ	Москва	Linux / Intel	44	4270
НИВЦ МГУ. AQUA	Москва	Linux / Intel	40	10464
НИВЦ МГУ. SCI	Москва	Linux / Intel	32	3366
НИВЦ МГУ. SKY	Москва	Linux / Intel	40	7404
Офисные компьютеры	Нью-Йорк	Windows / Intel	5	834
ИПС РАН	Переславль-Залесский	Linux / Intel	32	11904
Физфак МГУ	Москва	Linux / Intel	10	488
ИМПБ РАН	Пущино	Linux / Intel	40	3656
ГТУ	Тверь	Linux / Intel	16	45
УГАТУ	Уфа	Linux / Intel	64	490
ВМиК МГУ	Москва	Linux / Intel	36	3067

В табл. 1 показаны вычислительные системы, работавшие в этом эксперименте. Число выполненных каждой системой заданий определялось как их мощностью, так и длительностью участия в эксперименте. Так, кластеры Дубны и Переславля-Залесского работали все время, а системы из Твери и Уфы подключались лишь на несколько часов.

Несмотря на значительную географическую удаленность использованных вычислительных систем, эффективность работы системы в целом превысила 95% (отношение суммарного времени полезных вычислений на узлах к общему времени использования данных узлов в эксперименте).

Задача второго эксперимента – работа в условиях максимальной неоднородности. Организаторы эксперимента: НИВЦ МГУ и Межведомственный суперкомпьютерный центр (МСЦ). В расчете участвовали четыре типа вычислительных ресурсов: отдельные компьютеры, учебный класс, традиционные вычислительные кластеры и суперкомпьютер МВС-1000М. Часть компьютеров работала под управлением Windows, а часть – под Linux (как правило, Red Hat). На одной части компьютеров использовались процессоры Intel, а на другой – Alpha. Одновременно в эксперименте применялись три различных режима использования ресурсов (табл. 2):

• МОНО – монопольное использование, при котором вычислительная система полностью отдавалась под проведение данного эксперимента;

• ОЧЕР – работа через стандартные системы управления очередями заданий. Данный режим не требует менять политику администрирования и использования вычислительной системы при ее подключении и после ее выхода из эксперимента;

• ЗАН – режим “по занятости” позволяет использовать свободные ресурсы вычислительной системы. С помощью специальных алгоритмов система X-Com определяет, что узел в данный момент не используется, и запускает на нем прикладную задачу. Как только любая другая программа на узле проявляет активность, X-Com приостанавливает свою работу.

Эксперимент продолжался 14,5 часов. Среднее число одновременно работающих процессоров составило 125 (максимум зафиксирован на значении 150). Всего к эксперименту подключалось 437 различных процессоров. Казалось бы слишком много, учитывая среднее число 125. Но объясняется это очень просто. Суперкомпьютер МВС-1000М объединяет 768 процессоров, и система управления заданиями при распределении программы на реальные процессоры не обязана учитывать, куда были распределены предыдущие варианты этой же программы.

Всего за время этого эксперимента было передано 26,8 Гбайт данных. Время обработки одной порции данных менялось от 1 секунды до 10 минут. Несмотря на значительную коммуникационную нагрузку и специально созданную неоднородность, эффективность работы всей системы составила 92,9%. В табл. 2 приводятся характеристики вычислительных си-

стем, режимов их использования и агрегированные данные хода вычислений.

Таблица 2

Вычислительные ресурсы “неоднородного” эксперимента

Кластер	Платформа	Режим	Выдано заданий	Обработано заданий	Оборванные сессии	
					число	%
МСЦ. KAPPA	Linux / Intel	МОНО	2485	2476	9	0,36
МСЦ. МВС-1000М	Linux / Alpha	ОЧЕР	3203	2796	407	12,70
НИВЦ. AQUA	Linux / Intel	МОНО	9689	9645	44	0,45
НИВЦ. SCI	Linux / Intel	ЗАН	501	467	34	6,78
НИВЦ. SKY	Linux / Intel	ЗАН	5184	5184	49	0,93
НИВЦ. Уч.классы	Windows / Intel	МОНО	1642	1599	43	2,61

Естественно, что наиболее эффективно используются системы в монопольном режиме. Небольшие накладные расходы, возникшие при использовании режима “по занятости”, во многом объясняются малым изменением вычислительной нагрузки на соответствующих кластерах во время данного эксперимента. Самый большой процент оборванных сессий оказался при использовании стандартных систем управления очередями. Объясняется это тем, что при постановке в очередь система заказывает для программы некоторое время. После того как программу распределяют на реальные узлы, она начинает штатную обработку порций данных и продолжает это до того момента, пока не исчерпает свой лимит по времени, после чего всегда происходит разрыв сессии между узлом и сервером.

Проведенные эксперименты убедительно показали, насколько эффективно можно решать большие задачи с использованием существующих распределенных вычислительных ресурсов. Конечно же, X-Com не является, да и не может быть универсальным решением, одинаково хорошо соответствующим любой задаче и любой вычислительной среде. Но то, что она является одним из эффективных и реально применимых на практике подходов, это точно. Что использовать в каждом случае, определяется прежде всего самим пользователем и стоящими перед ним задачами. Сейчас важно осознать, что в подобных неоднородных распределенных вычислительных средах заложен колоссальный потенциал, и очень важно не опоздать с его освоением.

Литература

1. Воеводин В.В., Воеводин Вл.В. Параллельные вычисления. – СПб.: БХВ-Петербург, 2002. – 609 с.
2. http://www.parallel.ru/ – Информационно-аналитический центр по параллельным вычислениям в сети Интернет.