УЧЕБНО-НАУЧНЫЙ ЦЕНТР МГУ ПО ВЫСОКОПРОИЗВОДИТЕЛЬНЫМ ВЫЧИСЛЕНИЯМ*

Вл.В.Воеводин

НИВЦ МГУ

Введение

9 ноября 2001 года была опубликована 18-я редакция списка 500 мощнейших компьютеров мира – Top500. Лидером списка по-прежнему остается корпорация IBM, которой принадлежит 32% установленных систем и 37% от общей производительности. Интересной новостью стало появление Hewlett-Packard на втором месте по количеству систем (30%). При этом поскольку все эти системы относительно небольшие, то их суммарная производительность составила всего 15% от общего показателя. Далее по количеству компьютеров в списке идут SGI, Cray и Sun.

Самым мощным суперкомпьютером мира является по-прежнему система ASCI White в Ливерморской лаборатории (США) с производительностью в 7,2 Тфлопс на тесте LINPACK (58% от пиковой производительности). Этот компьютер содержит 8192 процессора Power3 (512×16), 4 Тбайта оперативной памяти, дисковый массив в 160 Тбайт, пиковая производительность его 12 Тфлопс (flops – floating operations per second, операции с плавающей точкой в секунду, Т(тера) = 10^{12}).

На втором месте теперь стоит система Compaq AlphaServer SC, объединяющая 3024 процессора. Она установлена в Питтсбургском суперкомпьютерном центре и имеет производительность 4 Тфлопс. Замыкает список Top500 116-процессорная вычислительная система Cray T3E с производительностью на LINPACK в 94 Гфлопс.

Список включает уже 16 систем с производительностью более 1 Тфлопс, 8 из которых установлено IBM. Стабильно увеличивается число систем, представляющих собой кластеры из небольших SMP-блоков – сейчас в списке уже 43 такие системы. Однако большинство в списке по-прежнему за массивно-параллельными системами (50%), за которыми идут кластеры, состоящие из больших SMP-систем (29%).

По каким же направлениям идет развитие высокопроизводительной вычислительной техники в настоящее время? Основных направлений четыре.

* Данная работа выполняется при поддержке РФФИ, грант №02-07-90442.

1. Векторно-конвейерные компьютеры. Конвейерные функциональные устройства и набор векторных команд – это две особенности таких машин. В отличие от традиционного подхода, векторные команды оперируют целыми массивами независимых данных, что позволяет эффективно загружать доступные конвейеры. В таких системах команда вида A=B+C может означать сложение двух массивов, а не только двух чисел. Характерным представителем данного направления является семейство векторно-конвейерных компьютеров Cray, куда входят, например, Cray EL, Cray J90, Cray T90. Также сюда относятся компьютеры фирм NEC, Fujitsu и др.

2. Массивно-параллельные компьютеры с распределенной памятью. Идея построения компьютеров этого класса тривиальна: возьмем серийные микропроцессоры, снабдим каждый своей локальной памятью, соединим посредством некоторой коммуникационной среды – вот и все. Достоинств у такой архитектуры масса: если нужна высокая производительность, то можно добавить еще процессоры, если ограничены финансы или заранее известна требуемая вычислительная мощность, то легко подобрать оптимальную конфигурацию, и т.п. Однако есть и решающий "минус", сводящий многие "плюсы" на нет. Дело в том, что межпроцессорное взаимодействие в компьютерах этого класса идет намного медленнее, чем происходит локальная обработка данных самими процессорами. Именно поэтому написать эффективную программу для таких компьютеров очень сложно, а для некоторых алгоритмов иногда просто невозможно. К данному классу можно отнести компьютеры Intel Paragon, IBM SP1, Parsytec, МВС 100/1000. В какой-то степени к этому классу можно причислить IBM SP2 и Cray T3D/T3E, хотя в этих компьютерах влияние указанного фактора ослаблено. К этому же классу можно отнести вычислительные кластеры и сети компьютеров, которые все чаще рассматривают как дешевую альтернативу крайне дорогим суперкомпьютерам.

3. Параллельные компьютеры с общей памятью (SMP-компьютеры). Вся оперативная память таких компьютеров разделяется несколькими одинаковыми процессорами. Это снимает проблемы предыдущего класса, но добавляет новые – число процессоров, имеющих доступ к общей памяти, по чисто техническим причинам нельзя сделать большим. В данное направление входят многие современные многопроцессорные SMP-компьютеры или отдельные узлы компьютеров класса HP V-Class или Sun StarFire.

4. Последнее направление, строго говоря, не является самостоятельным, а скорее представляет собой комбинации предыдущих трех. Из нескольких процессоров (традиционных или векторно-конвейерных) и общей для них памяти сформируем вычислительный узел. Если полученной вычислительной мощности недостаточно, то объединим несколько узлов высокоскоростными

каналами. По такому принципу построены Cray SV1, HP Superdome, Sun StarFire, NEC SX-5, последние модели IBM SP2 и др.

“Кластерное” направление в настоящее время является наиболее перспективным для конструирования компьютеров с рекордными показателями производительности. В последнем выпуске Top500 за кластерами 43 позиции. На сегодня самый «быстрый» Linux-кластер “IBM+NCSA” имеет производительность 2Тфлопс. Кластеры бывают «классическими», бездисковыми, многосетевыми, неоднородными и т.п. Они могут объединять самые разные процессоры: Pentium, Itanium, Alpha и др. Для их соединения используются сети Fast Ethernet, SCI, Myrinet, cLAN, ServerNet, Gigabit Ethernet и т.п. Как правило, кластеры работают под управлением операционной системы Linux, иногда под Windows NT.

1. Кластерные системы в России

В настоящее время в России насчитывается более 40 кластерных систем, среди которых наиболее мощными являются:

- вычислительная система Межведомственного суперкомпьютерного центра: 768CPUs, 2×Alpha 21264 (EV67/667MHz), 2Gb RAM, 15Gb HDD, Myrinet+Fast Ethernet, RedHat Linux;
- вычислительные системы НИВЦ МГУ им.М.В.Ломоносова:

а) 36CPUs, 2×PentiumIII/500MHz, 1Gb RAM, 9Gb HDD, SCI+Fast Ethernet, RedHat Linux;

б) 40CPUs, 2×PentiumIII/850MHz, 1Gb RAM, 2×15Gb HDD, Fast Ethernet, RedHat Linux;

- вычислительная система МФТИ:

24CPUs, 2×PentiumIII/600MHz, 256Mb RAM, 10Gb HDD, SCI+Fast Ethernet, ASP Linux;

- вычислительная система СПбГТУ:

16CPUs, 2×PentiumIII/550MHz, 256Mb RAM, Myrinet+Fast Ethernet, Linux, HDD U2W SCSI 72Gb RAID.

2. Учебно-научный центр МГУ

В 2000-2001 годах при поддержке ФЦП "Интеграция", РФФИ и ректората МГУ в Московском университете создан уникальный и на настоящий день самый мощный среди вузов России вычислительный центр (рис.1). Комбинация вычислительных мощностей с накопленным опытом в области использования высокопроизводительной техники и теоретическими исследованиями в области параллельных вычислений позволяет обеспечить полноцен-

ную среду для поддержки фундаментальных научных исследований и учебного процесса, а также экспериментальный полигон для разработки и апробирования кластерных технологий.

Коллектив лаборатории параллельных информационных технологий НИВЦ МГУ уже много лет работает в области параллельных вычислений и супер-ЭВМ. Мы достаточно хорошо представляем данную предметную область, знакомы со многими коллективами и отдельными учеными, работающими в данной области, участвуем в совместных проектах. На этой основе в мае 1998 года мы создали сервер по параллельным вычислениям в сети Интернет, затем выросший в Информационно-аналитический центр PARALLEL.RU. В 1999 году работа центра была поддержана Российским фондом фундаментальных исследований. Тематика центра охватывает следующие разделы предметной области: обзор новостей мира высокопроизводительных вычислений, описание архитектур компьютеров, технологии параллельного программирования, задачи, решаемые на суперкомпьютерах, обзор источников информации в данной области, списки конференций по данной тематике, списки крупнейших суперкомпьютерных центров, информацию о российских ресурсах в данной области, очерки по истории высокопроизводительных вычислений и многое другое. Задача, которая ставилась нами в самом начале, – структуризация предметной области и построение справочника на базе HTML, практически выполнена.

Задача-максимум для PARALLEL.RU состоит в создании полноценной энциклопедии предметной области на основе веб-технологий, что требует привлечения к проекту гораздо большего числа авторов, специалистов в отдельных областях. Другая задача центра – это создание среды для постоянного общения и взаимодействия российских специалистов по параллельным вычислениям и пользователей высокопроизводительных вычислительных ресурсов. Собирается информация о российских персоналиях и организациях, вовлеченных в данную деятельность, поддерживается "карта российских проектов". Организована рассылка по электронной почте новостей центра и мира высокопроизводительных вычислений, охватывающая все основные коллективы России и ближнего зарубежья, занимающиеся данной тематикой (на октябрь 2001 года более 1600 подписчиков). Функционирует дискуссионный клуб, в котором обсуждаются все интересующие вопросы данной тематики, оказывается информационная поддержка научных конференций, семинаров и других мероприятий вокруг тематики параллельных вычислений. На базе центра реализуются элементы дистанционного обучения. Кроме предоставления учебных материалов и заданий, разработаны также варианты интерактивного взаимодействия с обучаемыми, такие как онлайн-тестирование, сдача зачета через Интернет и др. В 1999 году был проведен пробный кон-

курс по параллельному программированию для студентов ведущих российских вузов.

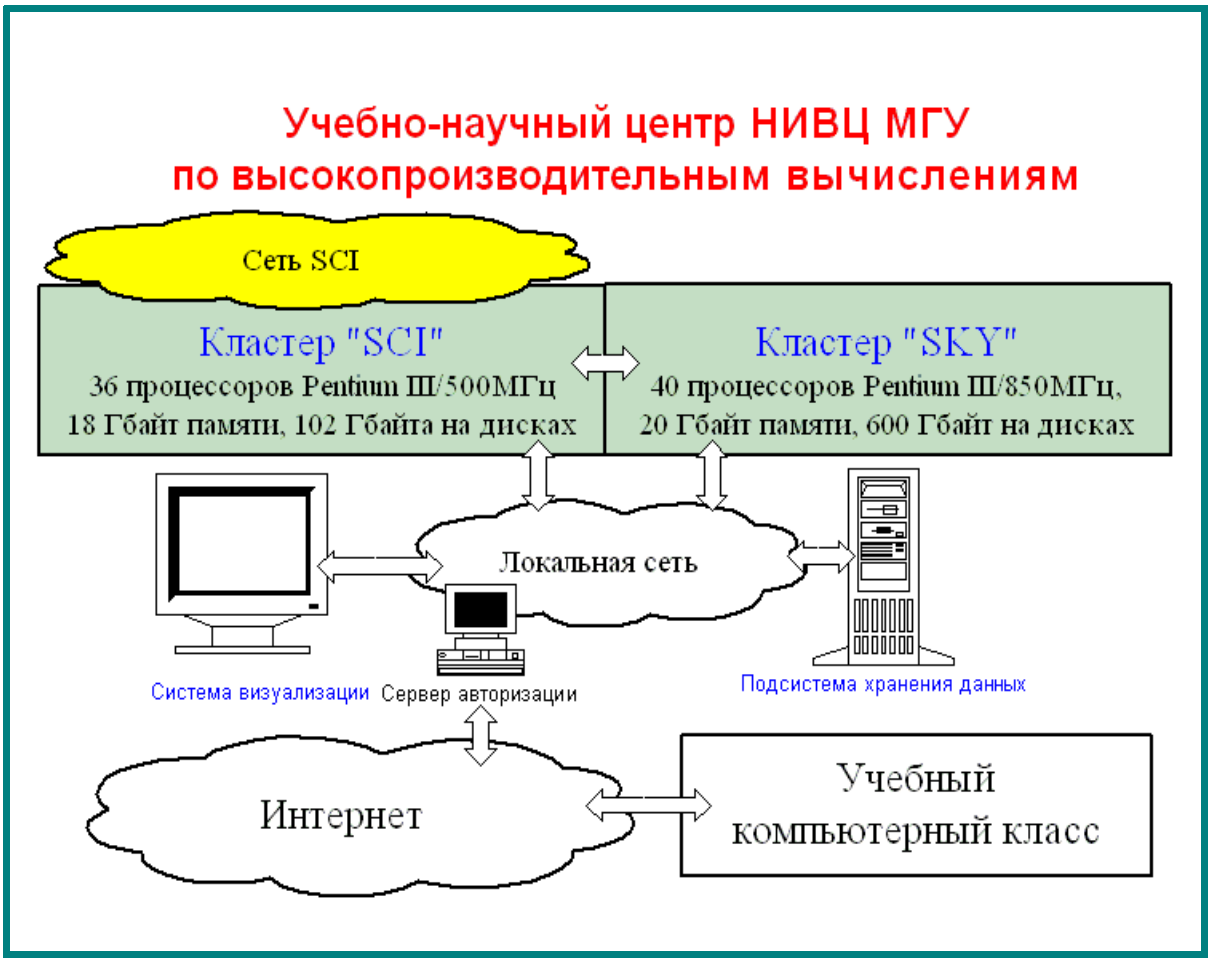

Рис.1

3. Высокопроизводительные вычислительные ресурсы МГУ

В феврале 2000 года в НИВЦ МГУ был установлен первый вычислительный кластер (рис.2), оборудование для которого было закуплено частично за счет ФЦП "Интеграция", частично за счет собственных средств НИВЦ.



Этот кластер состоял из 12 двухпроцессорных ПК-серверов российской сборки (компании "Инел"). Использование двухпроцессорных узлов считается более эффективным по технологическим и экономическим соображениям по сравнению с одно- и четырехпроцессорными узлами. При проектировании кластера с учетом технических требований и имеющихся средств рассматривалось несколько вариантов построения коммуникационной сети, в том числе Fast Ethernet, Gigabit Ethernet, Myrinet и SCI. В результате оценки имеющихся данных по цене, надежности программного обеспечения и производительности этих технологий была выбрана технология SCI (Scalable Coherent Interface), несмотря на то, что стоимость коммуникационного оборудования в расчете на один узел (1700 долларов США) превосходила стоимость самого вычислительного узла (1300 долларов).

Рис.2

Для построения коммуникационной сети были закуплены программно-аппаратные комплекты Wulfkit производства норвежских компаний “Dolphin

Interconnect Solutions” (сетевые адаптеры) и “Scali Computer” (программное обеспечение). Технология SCI отличается от других тем, что не использует специальных коммутаторов, а узлы кластера объединяются в топологию "двухмерного тора", состоящего из однонаправленных колец SCI. При такой организации сети каждый узел имеет по 2 входящих и выходящих канала SCI с очень высокой скоростью передачи и присутствует одновременно в двух кольцах.

При дальнейшем расширении аппаратных ресурсов в начале 2001 года мы сочли необходимым учесть потребности наших пользователей и провели среди них опрос – какие компоненты нужно нарастить в первую очередь? По результатам этого опроса сначала был увеличен объем оперативной памяти на всех узлах до 1 Гбайта, а затем увеличено количество узлов. Таким образом, кластер “SCI” в настоящее время включает 18 двухпроцессорных узлов с процессорами Pentium III 500 МГц (и 550 на новых узлах) и на каждом узле установлено по 1 Гбайту оперативной памяти. Узлы объединены в сеть SCI с топологией двухмерного тора 3×6, а управление кластером производится с отдельной головной машины по сети Fast Ethernet. Кроме того, установлен отдельный файл-сервер с RAID-массивом, на котором хранятся домашние директории пользователей, доступные со всех машин по протоколу NFS. Временные файлы программы пользователей могут создавать на локальных дисках каждого узла. Для поддержки непрерывной работы кластера на всех узлах установлены специальные платы watch-dog, разработанные в ИПС РАН (г. Переславль-Залесский), которые позволяют автоматически перезагружать зависшие узлы.

Весной 2001 года был установлен новый 40-процессорный кластер "SKY" с более мощными процессорами (Pentium III/850 МГц) на базе традиционной сети Fast Ethernet. На этой конфигурации успешно работают многие параллельные приложения, в частности расчет больших квантовохимических систем с помощью PC-GAMESS. Для поддержки больших расчетов на каждом узле кластера установлено по 2 жестких диска, которыми независимо могут пользоваться два процесса параллельной задачи.

4. Характеристики производительности

По результатам измерения характеристик сети скорость обменов данными по сети SCI на уровне MPI-приложений составила более 80 Мбайт/с, а латентность (время задержки сообщений) – примерно 5,6 мкс. Эти результаты практически не зависят от взаимного расположения узлов. При больших размерах сообщения скорость обменов по сети SCI даже превышает скорость обменов внутри одного SMP-узла. При обменах по сети Fast Ethernet пропуск-

ная способность не превышает 10-11 Мбайт/с, а латентность составляет более 130 мкс. Это ограничивает круг параллельных приложений, которые могут эффективно работать на таком кластере.

Традиционным тестом производительности параллельных компьютеров, используемым при составлении списка Top500, является тест решения системы линейных уравнений LINPACK HPC. Лучшие результаты по LINPACK составляют 10 Гфлопс (или 278 Мфлопс на процессор) для кластера “SCI” и 13 Гфлопс (или 325 Мфлопс на процессор) для кластера “SKY” (на задачах размера 44000×44000 и 47000×47000 соответственно). Таким образом, за счет более мощных процессоров кластер “SKY” на тесте LINPACK показывает лучшую производительность. Эти данные говорят не о том, что сеть SCI плохо справляется с нагрузкой, а о том, что тест LINPACK при больших объемах вычислений не предъявляет высоких требований к производительности коммуникационной сети. При фиксированном размере задачи (16000×16000) кластер на базе SCI показывает лучшую масштабируемость при большом числе процессоров.

За счет высоких коммуникационных характеристик сети SCI достигается очень высокая масштабируемость кластера на многих прикладных тестах. В частности, на некоторых тестах пакета NPB 2.3 результаты кластера оказались лучше, чем результаты суперкомпьютера Cray T3E/900 с аналогичным числом процессоров.

5. Технологии программирования

На кластере поддерживаются все стандартные технологии программирования для систем с распределенной памятью. Установлены реализации стандарта MPI (Message Passing Interface) для работы поверх сети SCI и Fast Ethernet. Поддерживается программирование на языках C/C++ и Фортран 77/90. Большинство параллельных программ пользователи создают на языке Фортран с использованием технологии MPI. Кроме этого, установлены и реально используются российские программные системы DVM и НОРМА, разработанные в ИПМ РАН. Проводились эксперименты с системой динамического автоматического распараллеливания программ "T-система", разработанной в ИПС РАН, а также с ParJava, параллельным расширением языка Java, разрабатываемой в ИСП РАН. Установлены однопроцессорные и параллельные библиотеки процедур линейной алгебры и преобразований Фурье: BLAS, MKL, SCALAPACK, FFTW.

Осуществляется перенос "Библиотеки численного анализа", разработанной в НИВЦ МГУ для машин серии БЭСМ-6, на кластер.

6. Особенности работы центра

Среди особенностей функционирования центра по высокопроизводительным вычислениям в МГУ стоит отметить следующие:

- необходимо поддерживать множество пользователей с различными задачами и различными требованиями к вычислительным ресурсам;
- большинство пользователей работают в удаленном режиме;
- требуется качественная поддержка учебного процесса;
- задачи пользователей написаны на разных языках с применением различных технологий параллельного программирования;
- нужно поддерживать запуск как параллельных, так и последовательных программ;
- в центре установлены множественные вычислительные ресурсы с различными параллельными средами.

7. Система управления заданиями

С учетом описанных выше требований нами была разработана и продолжает развиваться система управления заданиями для вычислительного кластера (см. parallel.ru/cluster/). В качестве вариантов ПО управления кластером рассматривались многие распространенные пакеты, такие как OpenPBS, но по разным причинам они не стали использоваться. Среди возможностей системы управления заданиями, разрабатываемой в НИВЦ МГУ, стоит отметить такие, как:

- поддержка множественных подкластеров с независимыми очередями и с возможностью запуска задач на всем кластере;
- независимые настройки различных очередей;
- гибкая настройка системы на различные параллельные среды;
- поддержка приоритетов задач;
- возможности временной блокировки очередей, задач и узлов кластера;
- настройка различных ограничений и прав пользователей;
- средства оптимального планирования загрузки с учетом предполагаемого времени работы задач;
- внешние модули для поддержки различных стратегий распределения процессоров;
- средства сбора статистики по использованию кластера;
- веб-интерфейс для просмотра состояния кластеров и очередей;
- переносимость кода; система реализована на языке Perl с отдельными элементами на языке Си;
- развитая документация для пользователя и администратора.

8. Поддержка пользователей

Большое внимание в работе центра уделяется взаимодействию с пользователями и комплексной поддержке их работы. В качестве элементов такой деятельности стоит выделить:

- подготовку справочной и методической литературы, документации для пользователей; публикацию в рамках Интернет-центра *Parallel.Ru;*
- оперативную поддержку пользователей по электронной почте и по телефону;
- список рассылки оперативной информации для пользователей;
- обязательный учет требований и пожеланий пользователей при расширении аппаратного и программного обеспечения, при настройке параметров системы управления заданиями.