Биомедицинская химия, 2004, том 50, приложение №1, с.101-116

МЕТОДЫ, ПРОГРАММЫ, БАЗЫ ДАННЫХ

УДК 681.324+539.192

ВЫЧИСЛИТЕЛЬНЫЕ РЕСУРСЫ КЛАСТЕРОВ И ИХ ПРИМЕНЕНИЕ В ХИМИЧЕСКИХ РАСЧЕТАХ

М.Б.Кузьминский, А.С.Мендкович

Институт органической химии РАН им. Н.Д. Зелинского (ИОХ РАН)
119991, ГСП-1 Москва, Ленинский проспект, 47
тел. (8-095) 135-63-88, (8-095) 135-63-77; факс (8-095) 135-53-88; эл. почта: kus@free.net

Дан обзор основных технических характеристик кластеров Beowulf (как их узлов, так и межсоединения). Приведены данные о производительности, в т.ч. распараллеливании в этих кластерах задач квантовой химии и молекулярной динамики. Проведено тестирование серверов на базе 64-разрядных x86-совместимых микропроцессоров нового поколения - AMD Opteron и показана перспективность их применения в кластерах.

Ключевые слова: Кластеры, распараллеливание, квантовая химия, молекулярная динамика

ВВЕДЕНИЕ. Современные методы вычислительной химии (главным образом, квантовой химии, молекулярной механики и молекулярной динамики) являются в настоящее время одним из основных потребителей ресурсов суперЭВМ. Актуальность этих методов связана с их широким применением не только в различных областях химии, но и в ряде смежных областей - например, физике (физика молекул, физика конденсированного состояния и поверхностных явлений и др.), биохимии, конструировании лекарств. Проведение таких вычислений также позволяет уменьшить финансовые расходы на проведение более дорогостоящих экспериментов.

Очевидным недостатком применения суперкомпьютеров, как и вообще высокопроизводительных вычислительных систем традиционной архитектуры, является их высокая стоимость. В последние годы активно развивается альтернативный подход, связанный с применением кластеров [1-3].

Мы будем называть кластерами систему компьютеров, соединенных между собой каналами связи (межсоединением), например, через локальную сеть. Отличия кластеров от компьютеров массивно-параллельной (МРР) архитектуры с распределенной между узлами (как физически, так и логически) оперативной памятью (ОП) носят скорее маркетинговый характер: как сами компьютеры, входящие в кластер, так и аппаратные средства межсоединения могут быть приобретены у разных производителей/поставщиков как отдельная продукция. А в МРР-компьютерах как их узлы, так и межсоединение специально разработаны фирмой-производителем для данного типа компьютеров и по отдельности, как правило, не поставляются.

Кластеры могут использоваться для разных целей, например, для обеспечения отказоустойчивости [2]. Нас интересуют кластеры, которые создаются с целью увеличения производительности путем интеграции вычислительных ресурсов компьютеров, образующих кластер (узлов кластера). Это увеличение производительности достигается либо за счет распределения вычислительной нагрузки между узлами кластера путем решения независимых задач на разных узлах, либо путем распараллеливания выполнения одной (каждой) задачи одновременно на разных узлах. В дальнейшем мы рассматриваем только последний случай, что и является альтернативой высокопроизводительным многопроцессорным компьютерам традиционной архитектуры (вплоть до суперкомпьютеров).

В кластеры могут объединяться как дорогие высокопроизводительные многопроцессорные системы, так и дешевые одно- и двухпроцессорные системы вплоть до ПК. Последние кластеры, где стоимость узлов обычно не превышает нескольких тысяч долларов США, называют еще Beowulf-кластерами. Подобным кластерам и их применению для задач вычислительной химии и посвящена данная статья.

1. Технические основания для использования кластеров в высокопроизводительных вычислительных системах.

Собственно, причины обращения к кластерам довольно очевидны: если имеется две даже достаточно мощные высокопроизводительные системы, их можно объединить в кластер, и это будет дешевле приобретения одной общей, более мощной, чем исходные, компьютерной системы. Однако мы будем рассматривать конкретно Beowulf-кластеры, в основном на базе x86-совместимых ПК-серверов, которые, как мы покажем ниже, обладают наилучшим отношением стоимость/производительность и наиболее распространены.

Причина сегодняшней популярности кластеров Beowulf состоит в том, что они имеют высокую производительность, обеспечивают большую пропускную способность (ПС) и емкость ОП и внешних устройств (в первую очередь НЖМД), а также хорошие возможности масштабирования по всем этим параметрам. Мы покажем ниже, что по всем этим характеристикам кластеры Beowulf сопоставимы с наиболее мощными современными суперкомпьютерами, имеющими MPP-архитектуру. Однако стоимость Beowulf-кластеров гораздо ниже.

Узким местом в архитектуре кластеров Beowulf часто являются характеристики "производительности" межсоединения, что непосредственно влияет на уровень распараллеливания задач (ускорение расчета по сравнению с однопроцессорным случаем), но, как мы покажем ниже, и по этим параметрам современные высокопроизводительные Beowulf-кластеры приближаются к MPP-системам.

1.1 Анализ характеристик процессоров

Начнем сопоставление с характеристик производительности процессоров суперкомпьютеров и микропроцессоров (МП), применяемых в кластерах Beowulf (табл. 1,2). Мы будем анализировать только производительность с плавающей запятой, поскольку именно она важна для рассматриваемых нами химических приложений - квантовой химии, молекулярной механики и молекулярной динамики.

Таблица 1. Производительность микропроцессоров на тестах SPECfp 2000 (www.specbench.org)

№	Микропроцессор	Примечание*1	Частота, (МГц)	Кэш L2, Мбайт	SPECfp2000*2
1	IBM Power 4	eServer pSeries 690	1700	1,4 (L2)+128 (L3)	1699/1598
2	Compaq Alpha 21364	GS1280	1150	1,75	1482/1124
3	Intel Itanium 2	hp rx4610	1500	6(L3)	н/д /2119
4	Intel Pentium 4	Dual DDR400	3000	0,512	1229/1213
5	AMD Opteron	Dual DDR 333	2000	1	1293/1219
6	Sun UltraSparcIII	Blade 2000	1200	8	1106/945
7	HP PA-8700+	hp c3750	875	1,5 (L1)	674/600
8	SGI R14K	Origin 3200	600	8	529/499

Примечания: 1. Компьютер/оперативная память. 2. Пиковое/базовое значение

Прежде всего, если отвлечься от многопроцессорных векторно-параллельных (PVP) суперкомпьютеров, использующих специализированные векторные процессоры [3,4], то можно отметить, что в суперкомпьютерах MPP-архитектуры используются в точности такие же МП, которые могут применяться и в узлах Beowulf-кластеров. Например, это могут быть самые быстрые на сегодня МП - Intel Itanium (Madison)/1.5 ГГц и IBM Power4+/1.7 ГГц (напомним, что мы говорим о производительности с плавающей запятой).

Как видно из данных таблицы 2, они обгоняют векторные процессоры на тестах Linpack с короткими векторами (N=100), а на длинных векторах (N=1000) уступают им примерно в 2 раза. Тесты Linpack (решение системы линейных уравнений с n неизвестными) хорошо локализуются в кэше и слабо зависят от ПС ОП. Эти МП стоят гораздо дешевле, чем векторные процессоры, в т.ч. и из-за массовости производства.

Таблица 2. Производительность процессоров на тестах Linpack (MFLOPS)

Система/процессор	N=100	N=1000	Пиковая производительность	
Fujitsu VPP5000 (3.3 нс)	1156	8784	9600	Векторные процессоры
NEC SX-6 (2 нс)	1161	7575	8000	
IBM eServer pSeries 690 (Power 4 1.7 ГГц)	1462	3817	6800	
Intel Pentium 4 (2.53 ГГц)	1190	2355	5060	
HP rx5670 (Intel Itanium 2, 1 ГГц)	1102	3534	4000	
Compaq ES45/Alpha 21264 1 ГГц	824	1542	2000	
AMD Opteron/1600 МГц	818	2103	3200	
HP Superdome (PA8700/750 МГц)	669	2099	3000	

Примечания: (по данным: - J.Dongarra, http://netlib2.cs.utk.edu; IBMeServer pSeries and IBM RS6000 Performance Report, May 6, 2003;данные для Opteron получены авторами).

Еще более дешевые x86-совместимые МП (Intel Pentium 4/Xeon, AMD Opteron) в свою очередь, немного отстают от них по производительности (табл. 1,2). Отметим, что результаты тестов SPECfp2000, представляющих собой смесь программ из реальных практических приложений (в которую входит, кстати, одна программа из области молекулярной динамики), в отличие от Linpack более сильно зависят также от ПС ОП. x86-совместимые МП, наиболее широко применяемые в кластерах, опережают по производительности большинство RISC-МП (табл. 1, 2), традиционно используемых в многопроцессорных вычислительных серверах. Применение x86-совместимых МП в кластерах Beowulf часто позволяет получать наилучшее отношение стоимость/производительность.

Итак, с точки зрения производительности процессоров кластеры Beowulf не сильно уступают современным суперкомпьютерам, особенно MPP-архитектуры, наиболее распространенной среди самых мощных суперкомпьютеров (см., например, список top500 крупнейших суперкомпьютерных инсталляций, www.top500.org).

В заключение нашего обсуждения следует упомянуть еще о специализированных процессорах MD-GRAPE, ориентированных на некоторые задачи молекулярной динамики (www.research.ibm.com/grape). Часто основное время расчета обусловлено вычислениями вкладов парных взаимодействий, которые обратно пропорциональны расстояниям и требуют соответственно расчетов величин типа обратного квадратного корня. На подобные вычисления (ван-дер-ваальсовские и кулоновские взаимодействия, суммирование по Эвальду) и нацелены процессоры MD-GRAPE, которые в этом случае достигают очень высокой производительности, однако и стоят они весьма дорого.

1.2 Анализ характеристик ОП

Обратимся теперь к технологии ОП. Если раньше в суперкомпьютерах использовалась быстродействующая, но дорогая ОП, например, типа SSRAM, то в настоящее время в многопроцессорных серверах и суперкомпьютерах как MPP, так и PVP-архитектуры применяется та же ОП, что и в ПК, базирующаяся на DDR SDRAM или RDRAM [5-7]. Для повышения ПС ОП в больших многопроцессорных вычислительных системах используются, например, схемы с многоканальной ОП [4-7]. Однако ПК-серверы имеют даже свои определенные преимущества: вследствие гораздо более низкого срока разработки последние технологические новинки (например, ОП DDR-типа с более высокой тактовой частотой) появляются сейчас в ПК даже раньше, чем в суперкомпьютерных системах, что позволяет поднять ПС ОП в расчете на 1 МП.

ПС ОП за последние 10 лет росла не столь быстро, как производительность процессоров (последняя, в соответствии с законом Мура, растет экспоненциально по времени). К настоящему моменту производительность многих приложений, в т.ч. на ряде типовых задач неэмпирической квантовой химии, ограничивается ПС ОП [8], и поэтому данная характеристика важна не менее, чем производительность процессора.

В последние 2-3 года производители стали уделять больше внимания ПС ОП, и пиковая ее величина в небольших одно- или двухпроцессорных серверах в настоящее время обычно соответствует пиковой ПС соответствующей шины МП. Если отвлечься от возможных конфликтов МП по доступу в ОП (см. ниже), то в качестве первого приближения можно сказать, что по ПС ОП (в частности, в расчете на 1 МП) кластеры Beowulf не отличаются от больших многопроцессорных серверов и MPP-систем, использующих те же МП.

На практике при выборе ПК-серверов для узлов кластеров Beowulf следует обращать внимание как на теоретическую (пиковую) ПС ОП, обеспечиваемую самим МП и собственно подсистемой ОП (северным мостом в наборе микросхем), так и на особенности реализации северного моста (контроллера ОП), сильно влияющие на реально достижимую ПС ОП [9]. В некоторых случаях (особенно для МП Pentium 4/Xeon) может быть целесообразным применение более низкочастотного МП, но с большей ПС ОП.

Для сопоставления различных МП и векторных процессоров между собой по характеристикам ПС ОП используются тесты STREAM [10]. По ПС ОП векторные процессоры по-прежнему сильно опережают обычные высокопроизводительные МП универсального назначения, и это является основным преимуществом PVP-систем. Прямых данных по сопоставлению ПС ОП наиболее мощных современных МП (Power4+, Itanium/Madison) и x86-совместимых МП на сайте этих тестов (www.streambench.org) в настоящее время нет.

Емкость ОП в расчете на 1 МП в узлах кластеров Beowulf также может быть достаточно велика; с 32-разрядными x86-совместимыми процессорами в Linux задаче может доступно до 2 или 3 Гбайт ОП в зависимости от настройки ядра. Однако по сравнению с многопроцессорными вычислительными серверами с общим полем памяти SMP- или ccNUMA-архитектуры [1-3, 5-7] в кластере каждый МП имеет доступ только к ОП своего узла, которая имеет гораздо меньшую допустимую емкость. Это является недостатком кластеров Beowulf: приложения, требующие очень большой емкости ОП, должны быть переписаны таким образом, чтобы каждый из параллельно выполняющихся (на разных узлах кластера) процессов данной задачи работал бы только с ОП своего узла, т.е. всю требуемую ОП следует "распределить" между узлами кластеров. Впрочем, этот недостаток характерен не только для кластеров, но и для MPP-систем с (физически и логически) распределенной ОП.

Среди задач вычислительной химии, требующих объемов ОП больше 2-3 Гбайт, можно указать на эффективные с точки зрения производительности "in-core" методы неэмпирической квантовой химии, реализованные, например, в известных комплексах программ серии Gaussian. Так, для in-core методов ССП и MP2 требуется ОП емкостью O(N**4), где N - размерность базиса. Для in-core метода ССП при N=200 требуется 1.6 Гбайт ОП, а для N=300 - уже 8.1 Гбайт. Однако у этих методов имеются и альтернативные численные схемы, требующие ОП меньшей емкости, например, O(N**3) для прямого метода MP2 (включая расчет градиентов); большинство типичных методов расчета по этим программам требует объемов ОП O(N**2) [11].

1.3 Анализ подсистем ввода-вывода

С точки зрения специфики интересующих нас задач вычислительной химии подсистема ввода-вывода (I/O) важна для организации эффективной работы с НЖМД. Сетевые возможности (межсоединение) узлов кластеров рассматриваются ниже отдельно. Что касается НЖМД, то в качестве примера приложений вычислительной химии, где их применение может оказаться узким местом (как с точки зрения емкости, так и с точки зрения производительности), сошлемся на некоторые задачи неэмпирической квантовой химии. Так, в convetional-методах, предполагающих однократный расчет и последующее хранение двухэлектронных интегралов на НЖМД, ПС дисков может лимитировать производительность [11].

Подсистемы I/O современных вычислительных систем - от ПК до суперкомпьютеров - используют одни и те же шины I/O, базирующиеся на стандарте PCI (в последний год начал применяться PCI-X, а в ближайшем будущем ожидается PCI Express). В слоты PCI вставляются одни и те же платы шин внешних устройств (Fibre Channel, UltraSCSI и др.), к которым можно подсоединять одни и те же НЖМД, и т.д.

Отличие небольших одно-двухпроцессорных узлов кластеров Beowulf от высокопроизводительных многопроцессорных вычислительных систем PVP, SMP или ccNUMA-архитектур (узлы во многих MPP-системах с распределенной ОП, как мы отмечали выше, по сути близки к кластерам Beowulf) обусловлены числом поддерживаемых шин PCI, числом доступных слотов этих шин и ПС магистралей, соединяющих подсистему I/O с процессорной частью. Конечно, по этим параметрам узлы Beowulf-кластеров сильно уступают. Более того, могут строиться бездисковые кластеры, в которых все узлы, кроме фронтального, вообще не содержат НЖМД (аналогичные "чисто вычислительные" узлы бывают и в MPP-системах).

Однако, подобно MPP-системам, возможности I/O в кластерах формально линейно масштабируются с числом узлов. Здесь встает, правда, задача, аналогичная использованию распределенной между узлами ОП: для использования этих возможностей приложение должно быть написано так, чтобы его процессы, одновременно выполняемые в узлах кластера, могли задействовать каждый ресурс I/O для работы с НЖМД своего узла. Для этого можно использовать и "параллельные" (распределенные) файловые системы (см., например, www.parl.clemson.edu/pvfs/).

В заключение нашего анализа I/O отметим и некоторые плюсы, возникающие при использовании ПК-серверов в узлах Beowulf-кластеров, аналогичные отмеченным выше при рассмотрении ОП. Подобно новым технологиям ОП, новые технологии шин I/O быстрее внедряются в ПК, чем в большие многопроцессорные системы. Так, в ПК-серверах раньше стали использоваться высокоскоростные шины PCI-X (среди распространенных MPP-систем они, по-видимому, доступны сегодня только в SGI Altix [5]). Вероятно, PCI Express также появится раньше в ПК-серверах.

1.4 Анализ межсоединения узлов кластера

Основным слабым местом Beowulf-кластеров по сравнению с традиционными высокопроизводительными вычислительными системами для рада задач вычислительной химии может оказаться межсоединение узлов, характеристики которого непосредственно влияют на уровень распараллеливания задач и, следовательно, эффективность использования кластера.

Но прежде чем проанализировать современные технологии соединения узлов, необходимо указать на общие преимущества кластеров по сравнению с большими многопроцессорными серверами SMP-архитектуры. Последние имеют общим своим недостатком ограниченную масштабируемость по всем основным характеристикам - числу МП (ограничено числом слотов процессорных плат), емкости и ПС ОП (ограничены соответственно числом слотов плат ОП и возможностями системной шины или коммутатора), ПС I/O (ограничена, например, числом шин и слотов PCI) и др. Кластеры, как и MPP-системы, с ростом числа узлов линейно масштабируются по всем этим параметрам.

Однако задействовать возможности такого масштабирования в кластерах и MPP-системах с распределенной ОП достаточно сложно (см.выше). MPP-системы ccNUMA-архитектуры имеют физически распределенную между узлами ОП, но логически она является общей для всех узлов (это свойство поддерживается аппаратно). Кроме очевидных преимуществ, в т.ч. потенциального упрощения распараллеливания, в последние несколько лет сделавших эту архитектуру очень популярной, она имеет и некоторые недостатки [2], в частности, существует проблема уменьшения производительности при работе с удаленной ОП.

В системах с общим полем памяти (SMP, ccNUMA) можно эффективно организовать иные, чем в кластерах, парадигмы распараллеливания, например, используя стандарт OpenMP [12]. Но даже при применении типичных для кластеров моделей распараллеливания с обменом сообщениями, использование поля общей памяти на нижнем уровне может быть эффективнее.

Возвращаясь собственно к межсоединению (каналам связи узлов кластеров), следует отметить, что, в условиях экспоненциального роста производительности МП, характеристики производительности этих каналов также должны быстро улучшаться: более быстрые МП справляются со своей порцией вычислительной нагрузки за более короткие интревалы времени и чаще должны обмениваться данными с МП других узлов.

Основные характеристики аппаратуры межсоединения узлов - это ПС и задержки (грубо говоря, время передачи сообщений нулевой длины).

В случае SMP-систем неким аналогом межсоединения при реальной пересылке сообщений является высокопроизводительная системная шина (или коммутатор). На

рисунке 1 представлены уровни распараллеливания на тестах Linpack (n=1000) в зависимости от числа процессоров N для современных многопроцессорных систем разной архитектуры. Видно, что для систем с общим полем памяти SMP- или ccNUMA-архитектуры (HP Superdome, SGI Origin 2000, IBM pSeries 690), имеющих более высокопроизводительное межсоединение, ускорение оказалось выше, чем для MPP-систем IBM SP2 и Cray T3E.

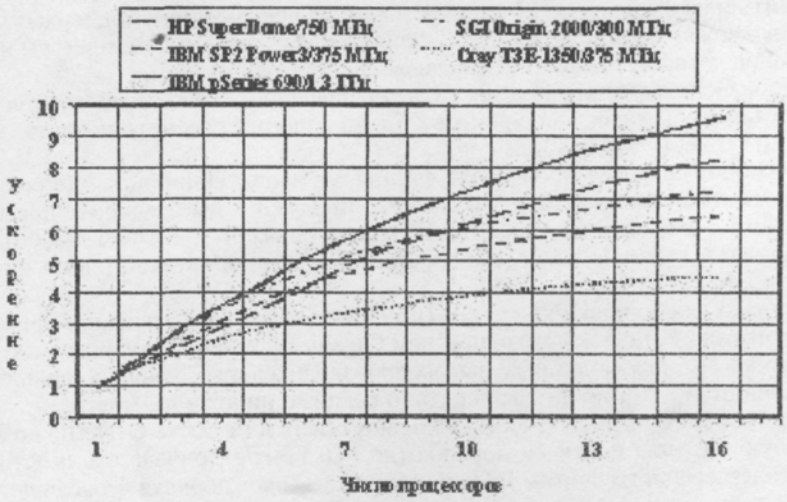

Рисунок 1.
Данные тестов Linpack (N=1000)

В таблице 3 приведены пиковая (аппаратная) ПС для ряда "типовых" каналов связи, используемых в кластерах, а также аналогичные значения ПС для связи узлов в MPP-системах. Из этой таблицы видно, что, хотя MPP-системы, использующие собственные, специально для каждой системы разработанные (и потому более дорогие) межсоединения имеют ПС, как правило, выше, чем в кластерах, кластерные каналы последнего поколения (Infiniband 8x, 10 Gigabit Ethernet) вплотную приблизились к MPP.

Наиболее широко используемыми в кластерах Beowulf каналами сегодня являются, вероятно, Fast Ethernet. В последнее время в связи с резким уменьшением цен на Gigabit Ethernet (меньше 100 долларов в расчете на порт коммутатора, до 50 долларов за сетевую плату) эта технология получает преимущества с точки зрения отношения ПС/стоимость даже по сравнению с использованием "сдвоенных" каналов Fast Ethernet для связи между узлами, и даже при использовании более медленных 32-разрядных шин

Таблица 3. Пропускная способность некоторых типовых соединений узлов кластера

Каналы	Пиковая пропускная способность(*), Мбайт/с	
Fast Ethernet	12.5	Применяются в кластерах
Gigabit Ethernet	125	
GigaNet CLAN	150	
HiPPI	200	
Myrinet 2000	250	
QsNET	340	
SCI	667	
Cray T3E	480	MPP
IBM SP2	500	
Cray/SGI Numalink 3	1600	

Примечания: (*) Все каналы дуплексные, поэтому полная пропускная способность при передаче сразу в двух направлениях - в 2 раза больше. Исключением являются однонаправленные каналы SCI.

PCI [13,14]. ПС при работе с Gigabit Ethernet существенно возрастает при использовании 64-разрядных шин PCI, доступных в серверных материнских платах. При этом удается достигнуть ПС (на уровне TCP) свыше 110 Мбайт/с [14].

Наши исследования [13,14] и другие данные показывают, что наилучшими характеристиками обладают сетевые платы Fast Ethernet производства Intel (с Linux-драйверами e100 и e1000 соответственно). Что касается соответствующих коммутаторов, то при их выборе следует учитывать ПС внутренней "системной шины" (чтобы обеспечить максимальную ПС одновременно на всех портах), поддержку "коммутации на лету", поддержку больших пакетов ("jumbo frames", для Gigabit Ethernet) и др. параметры, обсуждение которых выходит за пределы данной публикации.

Производство плат GigaNet CLan прекращено, HiPPI используется еще реже, и в качестве более дорогих конкурентов Gigabit Ethernet рассматриваются обычно Myrinet, SCI [15] и QsNet (www.quadrics.com) .

Это оборудование производят фирмы Myricom, Dolphinics Interconnect Solutions и Quadrics соответственно. Каждый из этих продуктов имеет свои плюсы и минусы, но производство каждого из них только одной фирмой и низкие объемы производства обуславливают высокие цены (стоимость только одной сетевой платы может превышать стоимость ПК).

Однако для больших кластеров (при числе узлов 64-96 и выше) применение Myrinet может стать экономически выгоднее, чем Gigabit Ethernet. Наконец, отметим, что реально достигаемая ПС для подобных высокопроизводительных каналов зависит не только от типа PCI-шин [15,16], но и от конкретной реализации южного моста.

Применение Infiniband 8x и 10 Gigabit Ethernet (а также QsNetII) позволяет поднять аппаратную ПС еще почти на порядок (до ~10 Гбит/с), однако стоимость, например, 10 Gigabit Ethernet- платы фирмы Intel сегодня составляет порядка 8 тысяч долларов США (и выше), и превышает стоимость небольшого ПК-кластера.

Прежде чем обсудить вопрос о задержках межсоединения, необходимо отметить, что разные межсоединения дают, вообще говоря, разные топологии соединения узлов (рис.2). Наиболее часто используется соединение узлов через коммутаторы (Ethernet, Myrinet, Infiniband и др.), чему в простейшем случае отвечает топология звезды.

Поскольку коммутаторы с очень большим числом портов дороги или вообще отсутствуют на рынке, для более крупных кластеров может понадобиться применение сразу нескольких соединенных между собой коммутаторов (рис. 2б).Такое соединение коммутаторов (trunking) в качестве узкого места может иметь канал связи между коммутаторами, поэтому желательно, чтобы его ПС была гораздо выше ПС портов. В некоторых вариантах соединения коммутаторов используется стекирование, при котором коммутаторы как бы объединяются в один с общей "внутренней шиной", но ее суммарная ПС при этом не меняется, а в расчете на 1 порт уменьшается.

Применение топологии звезды во многих случаях естественным образом отображается в математическую модель распараллеливания при обмене сообщениями.

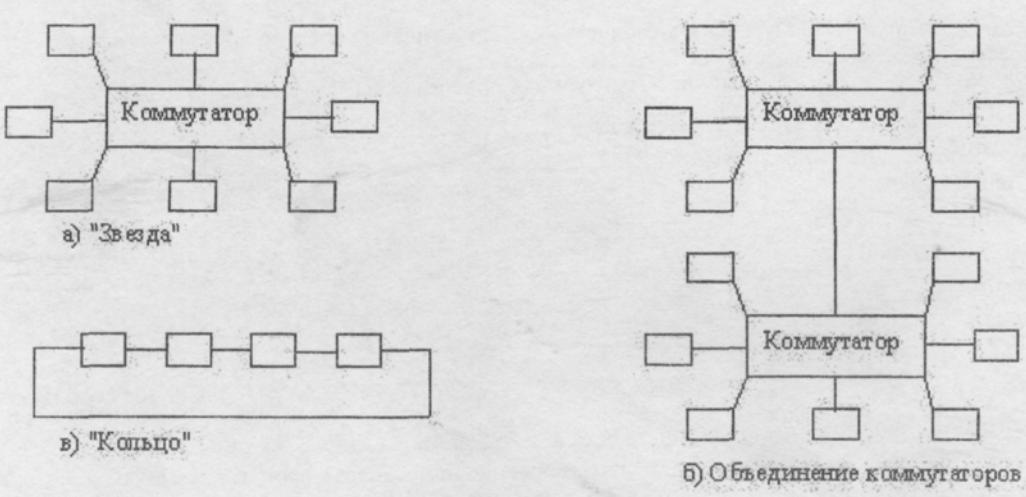

Рисунок 2 .
Типичные топологии соединения узлов

Недостатком применения коммутаторов являются дополнительные финансовые расходы на их приобретение и возможное увеличение задержки при прохождении портов коммутатора пакетами данных.

В "кольцеобразных" соединениях, примером которых может быть SCI, коммутаторов нет. В SCI имеются однонаправленные кольца, по которым передаются пакеты данных, однако прохождение пакета через каждый узел вносит небольшую дополнительную задержку. В кольце узлы разделяют ПС каналов связи, а применение SCI-коммутатора может уменьшить это разделение [15] и увеличивает реально допустимое число узлов.

В зависимости от используемых конфигураций SCI-плат (при отсутствии коммутатора) возможно применение топологий 1-, 2- и 3-мерного тора. Программные средства распараллеливания, используемые с Fast Ethernet/Gigabit Ethernet, обычно применяют протоколы TCP/IP. Этот стек протоколов, а также обращение к функциям ОС при обмене сообщениями между процессами, параллельно выполняющимися в узлах кластеров (и, в частности, копирование данных в "адресное пространство" ядра), существенно увеличивают задержки при работе с Ethernet (см., однако, ниже резюме по поводу задержек) . Эти задержки в Fast Ethernet и Gigabit Ethernet близки по порядку величины и варьируются в широком диапазоне от 20 до 90, а иногда даже до 200 мкс (см., например, [16]). Величина задержки может сильно зависеть, в частности, и от конкретной сетевой платы, и от реализации южного моста (особенно это характерно для Gigabit Ethernet; известно, например, что AMD 760MPX имеет задержки в 2 раза выше, чем Intel E7500).

Современные альтернативы каналам Ethernet - SCI, Myrinet, QsNet, Infiniband - имеют гораздо более низкие аппаратные задержки. При обмене сообщениями они работают непосредственно с адресным пространством пользователя, не используют средства TCP/IP и обходят лишние обращения к ОС. Кроме того, все эти технологии обеспечивают работу в режиме RDMA и соответственно аппаратную поддержку т.н. односторонним коммуникациям (см., например, [17]), что также способствует уменьшению результирующих эффективных величин задержек. Данные о задержках при использовании этих межсоединений см., например, [15-16, 18], а также на сайтах www.quadrics.com, nowlab.cis.ohio-state.edu/projects/mpi-iba .

Наинизшими задержками могут, естественно, обладать MPP-системы (например, 1-2 мкс в Cray T3E). Отличные показатели, на уровне 4-5 мкс, обеспечивает QsNet, чуть выше задержки в SCI (4-8 мкс), задержки в Myrinet2000 составляют 7-10 мкс, в Infiniband на уровне MPI достигаются близкие значения (8-9 мкс).

Для ОС Linux имеется ряд программных усовершенствований, позволяющих уменьшить задержки при работе с MPI в средах Fast Ethernet/Gigabit Ethernet (GAMMA, M-VIA, EMP и др.) при отказе от использования TCP/IP. Это позволяет уменьшить задержки при работе с Fast Ethernet до 14-30 мкс, с Gigabit Ethernet - до 9 мкс. Однако эти проекты имеют ограниченную применимость: они обеспечивают поддержку небольшого числа типов сетевых плат; некоторые из этих проектов завершили свое развитие; GAMMA имеет проблемы с поддержкой SMP-режима и т.д.

На типовых задачах квантовой химии, зависящих в большей степени от ПС, чем от задержек межсоединения, в наших исследованиях оказалось выгоднее применять спаривание каналов Fast Ethernet, чем уменьшать задержку путем использования M-VIA. Однако для типичных приложений молекулярной динамики (MM5, Gromacs, Amber, Charmm) характерен обмен короткими сообщениями, когда важны величины задержек. При этом использование более дорогих технологий (SCI, Myrinet и т.д.) может оказаться предпочтительным.

В качестве резюме по поводу задержек можно сказать, что альтернативы Ethernet либо дороги, либо имеют ограниченную применимость.

Выше мы говорили о том, что используемые в Beowulf-кластерах МП, в частности, x86-совместимые, сопоставимы по производительности с самыми быстродействующими процессорами современных суперкомпьютеров (табл. 1,2). Данные таблицы 4 (результаты тестов Linpack parallel) подтверждают сделанные выше выводы о конкурентоспособности по производительности Beowulf-кластеров и высокопроизводительных вычислительных систем традиционных архитектур, вплоть до самых мощных суперкомпьютеров. В качестве другой иллюстрации можно указать на то, что кластер Beowulf с 1536 двухпроцессорными узлами на базе Intel Xeon/2.4 ГГц и межсоединением от Quadrics занимает третье место в списке top500 крупнейших суперкомпьютеров мира - с производительностью 7.6 TFLOPS (второе место также занимает кластер, но на базе мощных многопроцессорных серверов с МП HP Alpha).

Таблица 4. Данные тестов производительности Linpack parallel (J.Dongarra, netlib2.cs.utk.edu)

ЭВМ,процессоры	Частота, МГц	Число процессоров	Производительность, GFLOPS	
Earth Simulator (NEC SX)	500	5120	35860	MPP
ASCI White-Pacific, IBM Power3	375	8000	7226	
Compaq AlphaServer SC ES45*1/EV68	1000	3016	4463	
Intel ASCI Red, Pentium II/Xeon	333	9632	2380	
NEC SX-5	3,2 нс	128	1192	век
Fujitsu VPP5000	3,3 нс	100	886	
Atipa Tech.,Pentium 4 Myrinet	1,8 ГГц	1024	2207	кластеры
Legend Deep Comp, Pentium 4/Myrinet	2,0 ГГц	512	1046	
Presto III, Athlon MP 1900+/Myrinet	1,6 ГГц	480	716	
NCSA Titan, Itanium/Myrinet	800	320	678	

Примечание: 1. Кластер SMP-систем

Что касается стоимости, то достаточно привести пару примеров. Так, 16-процессорный (32 процессорных ядра Power4, [19]) суперкомпьютер IBM pSeries 690, приобретенный недавно факультетом ВМК в МГУ, имеет стоимость 1,6 миллиона евро, или по 100 тыс. евро в расчете на 1 МП (ComputerWorld/Россия, 2003, N27), или 50 тыс. евро на процессорное ядро. Четырехпроцессорные SMP-системы на базе IBM Power4 (эквивалентно 8 процессорным ядрам) могут стоить в 5 раз дешевле - около 10 тысяч долларов в расчете на процессорное ядро, однако и это на порядок больше, чем в "среднем" современном ПК.

Напомним, однако, что для корректного сопоставления отношения стоимость/производительность (даже если забыть об использовании TCO - полной стоимости владения) необходимо сопоставлять стоимости конкретных конфигураций. Так, если стоимость более высокопроизводительных МП велика, но составляет небольшую часть полной стоимости компьютера (например, с большой дорогой ОП, дорогой системной платой, дорогой подсистемой НЖМД), то отношение стоимость/производительность компьютера на базе этих МП будут лучше, чем при использовании более дешевых и более медленных МП.

Наконец, следует еще раз обратить внимание на то, что ПК-кластеры оказались эффективными из-за определенного соотношения между техническими характеристиками ПК (ПК-серверов) и их стоимостью, которые были рассмотрены выше. Эта ситуация обусловлена, в частности, потребностями массового рынка ПК в таких высоких технических характеристиках. С нашей точки зрения, сегодняшние потребности большинства приложений ПК (особенно офисных) к аппаратным ресурсам полностью удовлетворяются.

В случае, если в дальнейшем для ПК не появятся новые более ресурсоемкие приложения, произойдет диверсификация рынка, в т.ч. рынка МП, ПК будут резче отставать от высокопроизводительных компьютерных систем, даже одно-двухпроцессорных, и хотя последние будут еще дороже, общая картина с уровнем производительности и отношением стоимость/производительность может поменяться. Таким образом, сегодняшняя эффективность Beowulf-кластеров не обязана сохраняться в более отдаленном будущем.

2. Использование кластеров Beowulf в химических приложениях

Выше мы говорили о производительности процессоров и компьютерных, в т.ч. кластерных систем, на универсальных тестах. Учитывая разнообразие расчетных методов, используемых в квантовой химии, молекулярной механике и молекулярной динамике, такой подход в целом оправдан. Кроме рассмотренных выше тестов, имеются многие

другие [17], которые мы по ряду обстоятельств считаем менее интересными для наших задач (упомянем лишь смесь ScienceMark, включающую, в частности, тестовую задачу молекулярной динамики, см. www.sciencemark.org).

В [8] была найдена корреляция между SPECfp_base2000 и производительностью ряда микропроцессоров (МП) на тестах Gaussian-98. Однако применение, например, "in-core" методов квантовой химии для ускорения расчетов требует большой емкости ОП (см. выше) и соответственно применения 64-разрядных МП. Но тогда сопоставление производительности на основе SPECfp2000 становится некорректным.

Имеются также попытки составить тесты (смеси задач), специализированные для области вычислительной химии (см., например, [20]). Такие тесты требуют оперативного обновления результатов по мере появления новых аппаратных средств (что на практике выполняется не всегда), а методики составления смесей дискуссионны. С нашей точки зрения, более интересны смеси задач для конкретных приложений, например, Gaussian (см. табл. 5, [8]), а еще лучше - сопоставление времен выполнения конкретных типовых расчетов [21].

Однако нас интересует производительность не однопроцессорных компьютеров, а Beowulf-кластеров (в т.ч. в сопоставлении с многопроцессорными компьютерами традиционных архитектур). Здесь на первый план выходит ускорение, достигаемое при распараллеливании.

В таблице 6 приведены программные средства и модели распараллеливания, используемые в ряде популярных программных продуктов в области квантовой химии и молекулярной динамики. Данные этой таблицы показывают что большинство программ в

Таблица 5. Производительность 32- и 64-разрядных микропроцессоров на наборе стандартных тестов Gaussian 98 [8]

Микропроцессор	Частота, ГГц	Время, мин.
Intel Pentium 4	1,8	1120
AMD Athlon	1,4	1358
Compaq Alpha 21264	0,67	2005

этой области распараллелено в модели обмена сообщениями с использованием средств типа MPI или PVM [1, 22], причем применение MPI, являющегося стандартом (www.mpi-forum.org), преобладает.

Модели распараллеливания, специфические для компьютерных систем с общим полем ОП, применяются в некоторых очень популярных программах (например, Gaussian и Amber). Эти модели недоступны для распараллеливания в кластерах, а попытки обеспечить в кластере общее поле памяти программным путем, например на уровне ОС, оказываются неэффективными (в качестве примера можно указать на Mosix, //open-mosix.sourceforge.net).

Как указано в [1], большинство программистов предпочитают иметь хотя бы

Таблица 6. Применяемые средства распараллеливания программ вычислительной химии

Программа	Метод распараллеливания	
	Модель общего поля памяти	Модели обмена сообщениями (*)
ADF2000	-	MPI/PVM
Amber 7.0	Директивы компилятора, OpenMP	MPI
Gaussian-03	Автоматическое распараллеливание компилятором, директивы компилятора, ручное распараллеливание (fork)	Linda
Gamess-US	-	TCGMSG/MPI/DDI
Jaguar 4.2	-	MPI
Mopac 2002	-	MPI
NWChem 3.4.1	-	TCGMSG/MPI
Turbomole 5.3	-	MPI
AMMP	-	PVM
GROMACS	-	MPI

Примечания: (*) Включая также программные средства более высокого уровня, работающие в кластерах и MPP-система

иллюзию совместно используемой (общей) ОП. Однако нам известно только одно подобное программное средство (более высокого уровня, чем обычные обмены сообщениями типа MPI/PVM) - Linda [1,17], которое используется только в одном программном комплексе (Gaussian, см. ниже). В Linda создается некая иллюзия структурированной совместно используемой распределенной ОП, для доступа к которой применяется очень небольшой набор примитивных операций (как и в случае с MPI/PVM, это - набор подпрограмм, вызываемых из Фортрана или Си).

В соответствии с законом Амдала [17,22], достигаемый уровень распараллеливания с ростом числа процессоров не возрастает неограниченно, а стремится к некой константе, зависящей от доли последовательных (не распараллеливаемых) вычислений в общем времени расчета. Практика распараллеливания задач вычислительной химии подтверждает подобное уменьшение эффективности добавления новых МП (узлов), когда их число становится достаточно большим (рис.3-7). Эта зависимость в реальности определяется очень многими параметрами - производительностью МП, характеристиками межсоединения, используемым методом вычислительной химии, конкретной программной реализацией, особенностями рассчитываемого химического объекта и проч.

Так, на рисунке 3 продемонстрирована зависимость уровня распараллеливания от характеристик межсоединения для задач молекулярной динамики (напомним, что для

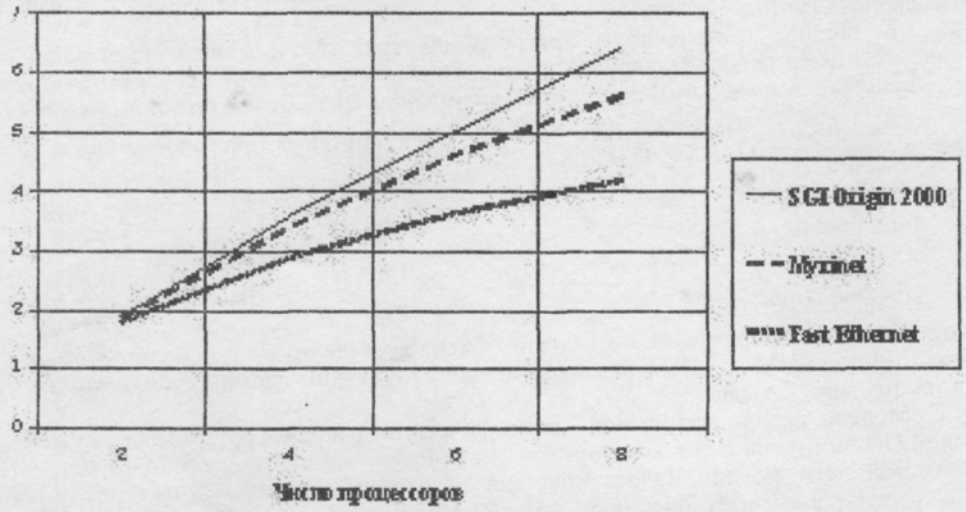

Рисунок 3.
Влияние межсоединения на распараллеливание. Сравнение распараллеливания Amber 6.0: кластер пластоцианина в воде (11500 атомов) [23]

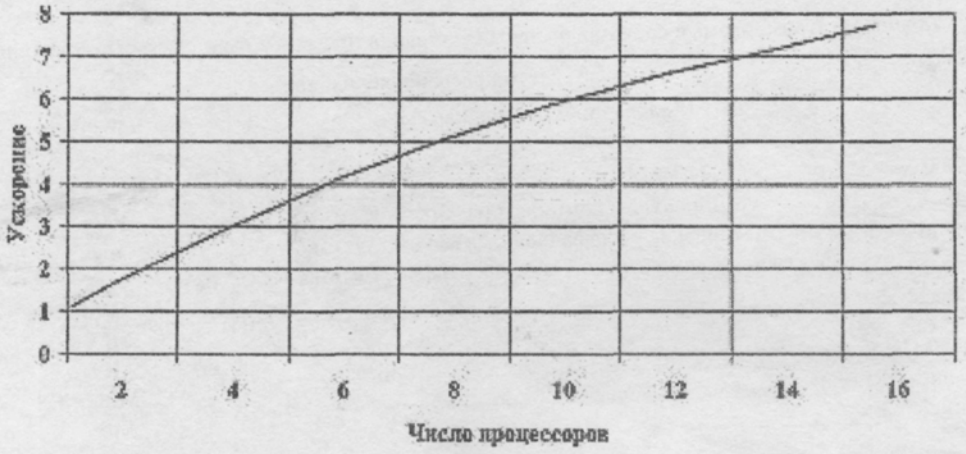

Рисунок 4.
Распараллеливание полуэмпирической программы MOPAC2002 (C960) на SGI Origin 300 (R14000-500МГц) [21]

этого случая важна и ПС, и задержки). Современные полуэмпирические квантовохимические программы распараллеливаются не слишком хорошо (рис. 4). Что касается неэмпирических методов квантовой химии, то здесь эффективность распараллеливания сильно зависит от метода расчета.

В кластерах хорошо распараллеливается метод DFT (почти линейно до 8 МП даже для Fast Ethernet), см., например, наши результаты на рисунке 5. При небольшом числе МП эффективно распараллеливается метод ССП (рис. 5), но при большом числе МП эффективность распараллеливания резко падает (рис. 6). Распараллеливание метода MP2 иллюстрирует рисунок 7. Однако во всех случаях сильна зависимость от конкретной программной реализации, рассчитываемого химического объекта и других факторов, указанных выше. Поэтому приведенные данные служат лишь некоторой небольшой иллюстрацией к утверждению о том, что уровень распараллеливания является принципиально важным вопросом при обсуждении эффективности применения кластеров для задач вычислительной химии.

3. Применение новых 64-разрядных процессоров в узлах кластеров

По причинам, детали которых мы здесь не обсуждаем (см., однако, табл.1 и 2), с нашей точки зрения, в кластерах имеет смысл применять лишь МП IBM Power4, Intel Itanium 2/Madison /Deerfield (64-разрядные МП) или x86-совместимые МП Pentium 4/Xeon, AMD Athlon или Opteron. Первые два доступны в слишком дорогих для массового

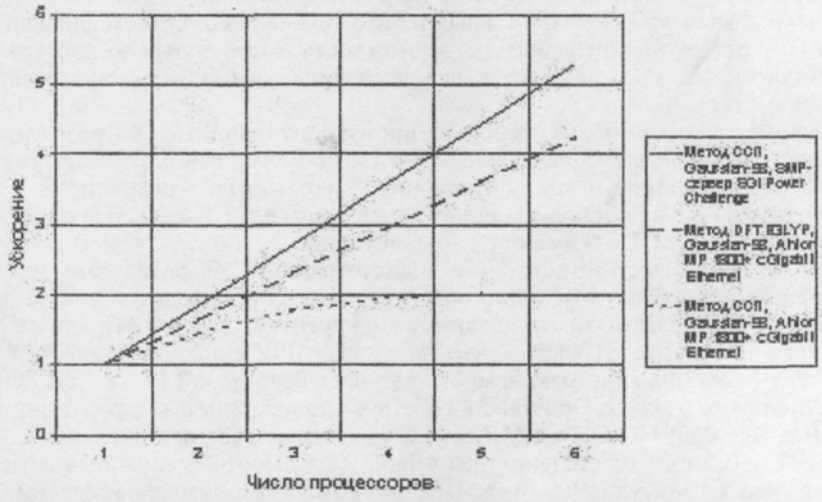

Рисунок 5.

Сравнение распараллеливания в программах Gaussian: данные тестов, проведенных в ЦКОХИ Молекулы: C6(NO2)3(NH2)3, 6-31G**, 300 б.ф., (ССП test178, nosymm) Валиномицин, 3-21G (DFT test397)

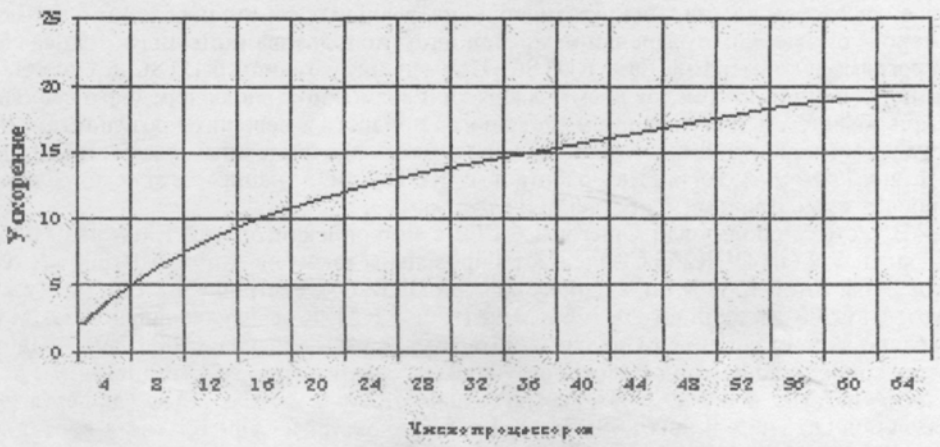

Рисунок 6.

Распараллеливание метода ССП в Gaussian-94 (Linda) на суперЭВМ Cray-T3E/600 МГц. Молекула: фуллерен (C60, 6-31G*, 900 б.ф.) [24]

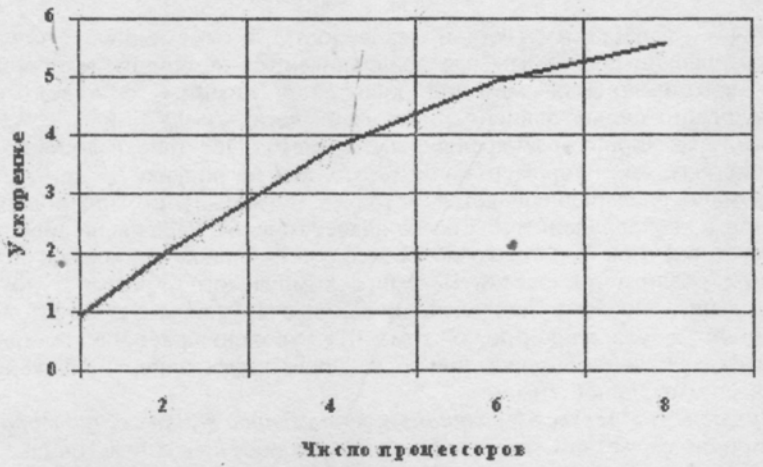

Рисунок 7.
Ускорение метода MP2 на кластере Fast Ethernet (узлы 2 x Pentium II/400 МГц). Молекула: бензол, 6-31G**. Программа GAMESS-US [25]

рынка кластеров конфигурациях (хотя Itanium 2 формально относится к 64-разрядным и x86-совместимым МП), а из остальных только появившиеся совсем недавно AMD Opteron имеют 64-разрядную адресацию, т.е. возможность иметь большую доступную задаче ОП (потребности в этом для ряда задач вычислительной химии были рассмотрены нами ранее в данной статье).

Кроме возможного ускорения производительности, 64-разрядная адресация позволяет рассчитывать молекулярные системы с большими размерностями.

Поэтому ниже мы приводим некоторые результаты тестирования производительности (в основном полученные нами) для ПК-серверов на базе Opteron. Эти 64-разрядные МП имеют существенно более высокое отношение производительность/стоимость, чем альтернативные 64-разрядные платформы. При построении кластеров МП Opteron по уровню производительности и отношению стоимость/производительность являются конкурентами как 64-разрядных МП, так и 32-разрядных Intel Xeon. Известно, что на тестах SPECcpu2000 Opteron/1.8 ГГц немного опережает Xeon/3.06 ГГц, хотя немного уступает Pentium 4/3 ГГц с FSB 800 МГц.

Микроархитектура Opteron способствует пониженным задержкам при обращении в ОП и повышенной ПС ОП в SMP-конфигурациях за счет независимых каналов доступа МП в ОП, в отличие от общей шины в Intel Xeon. Поэтому оценки производительности Opteron весьма актуальны, и в Центре компьютерного обеспечения химических исследований РАН (ЦКОХИ) были проведены измерения производительности 2-процессорной SMP-системы на базе Opteron как на тестах универсального характера (Linpack, n=100 и 1000; STREAM и lmbench (2.0.4)), так и на конкретных прикладных задачах квантовой химии. Для указанных универсальных тестов необходимо применение таймеров с высоким разрешением, подобно использованной нами ранее [10,14] подпрограммы-таймера на базе RDTSC. Применение команды RDTSC в Opteron имеет своими особенностями малое время задержки и возможность внеочередного выполнения, что при обычном использовании может иногда привести к неверным результатам. В связи с этим в соответствующих случаях мы воспользовались системным вызовом gettimeofday ОС Linux, который корректно работает с RDTSC и в наших тестах на Opteron не привносит существенной "собственной задержки".

В тестах использован Opteron/1.6 ГГц с материнской платой RioWorks HDAMA с двухканальной ОП DDR266 CL2.5. Тесты проведены также на Athlon MP1800+ (с близкой тактовой частотой 1533 МГц, с одноканальной DDR266), Pentium 4/1.8 ГГц (кэш L2 256K, RDRAM PC800) и Pentium III 1266 МГц (кэш L2 512K, с двухканальной ОП PC133); проведено сопоставление с другими литературными данными (Табл.7-9). На тестах Linpack сопоставлены компиляторы pgf90 5.0 (бета-версия), g77-3.3 и Intel ifc 7.1, а также библиотеки AMD acml-0.9 (бета-версия), Intel MKL-6.0 и Atlas 3.5.1. Сервер с Opteron работал под ОС UnitedLinux/SuSE Linux Enterprise Server for Opteron (бета-версия).

Измеренная нами задержка доступа в ОП на тестах lmbench составила 109 нс (табл. 7). Это лучше, чем в Athlon MP, но всего лишь близко к результатам Xeon/3.06 ГГц. Для Itanium 2 с набором микросхем HP zx1 задержка равна 156 нс. На тестах ПС ОП (STREAM)

удалось достигнуть ПС 2.3-2.5 Гбайт/с, что более чем в 2 раза превосходит задержку, наблюдаемую у Athlon MP1800+ с одноканальной DDR266. Конечно, Pentium 4/3 ГГц с FSB 800 МГц и двухканальной DDR400 будут далеко впереди. Однако известно, что у Pentium 4 в двухпроцессорных системах из-за конкуренции на системной шине параллельное выполнение двух нитей тестов STREAM немного снижает суммарную ПС, в то время как на Opteron нами обнаружен рост суммарной ПС примерно в 1,8 раза. Известные приложения квантовой химии - 32-разрядные программы Gaussian-98 и Gamess-US, отличающиеся высокими требованиями к производительности с плавающей запятой и различным поведением с точки зрения локализации в кэше, требованиями к ПС ОП и т. д. в зависимости от метода расчета, на Opteron/1.6 ГГц по процессорному времени ускорились в 1,5-1,9 раза по сравнению с Athlon MP1800+ (на одном процессоре). При этом ускорение при распараллеливании программ на 2 процессора SMP-сервера для Opteron оказалось заметно выше, чем для Athlon MP (табл. 8). Однако ускорение Gamess-US в Opteron по сравнению с Pentium 4/1.8 ГГц оказалось всего около 1,2-1,4. Это может быть связано с худшей по сравнению с Gaussian-98 оптимизированностью исходных текстов. Интересно, что ускорение Gamess-US при оптимизации под Pentium 4 по сравнению с оптимизацией под Pentium III составило 4-12% при выполнении на Pentium 4/1.8 Гц (табл. 9). Неожиданно высокие результаты для Pentium III Tualatin/1266 МГц (табл. 8) связаны, вероятно, с более высокой емкостью кэш-памяти второго уровня в этом МП (512 Кбайт, в 2 раза больше, чем в Athlon MP).

Приведенные результаты показывают, что хотя определенная корреляция SPECfp_base2000 и производительности МП на задачах квантовой химии наблюдается,

Таблица 7. Сопоставление процессоров узлов кластера на универсальных тестах

Процессоры	Linpack (MFLOPS)		Пиковое значение, GFLOPS	Задержка обращения в память, нс
	n=100	n=1000		
Xeon/3.06 (1)	1190	2355	6.1	102 (2)
Pentium 4/1.8	761	1831	3.6	-
Athlon MP1800+	732	1623	3.1	191 (2)
Opteron/1.6	818	2103	3.2	109
Itanium 2/1.0	1102	3534	4.0	156 [4]

Примечания.(1) с DDR266/CL2.0. Данные Linpack приведены для Pentium 4/2.53 ГГц. Данные для Athlon, Opteron и Pentium 4/1.8 ГГц получены авторами, остальные взяты из официальной таблицы тестов на сайте //netlib2.cs.utk.edu (наши данные для Athlon опубликованы ранее [10, 14] и помещены в официальную таблицу). Для Pentium 4 использован ifc с ключами -O3 -tpp7 -xW -ip и библиотека MKL, для Opteron - g77 с ключами -O3 -m32 -mfpmath=sse -malign -march=athlon-xp -funroll-loops -fomit-frame- pointer и библиотека Atlas. Для n=1000 при использовании на Opteron старых кодов, оптимизированных под Athlon, ускорение по сравнению с AthlonMP составило 1.3-1.4 (зависит от библиотеки). (2)Данные //www.tecchannel.de/hardware/1164/15.html: для Xeon - с DDR266/CL2.0; для Athlon - приведены для Athlon MP2600+, с DDR266/CL2.0 .

Таблица 8. Эффективность применения некоторых 64- и 32-разрядных микропроцессоров (в SMP-конфигурациях) на задачах квантовой химии с небольшим объемом ввода-вывода

Тесты Gaussian-98	N CPU	Процессорное время расчета, сек				
		AMD		Intel		SGI Origin 3000
		Opteron 1.6 ГГц	Athlon MP1800+ 1533 МГц	Pentium III 1266 МГц	Itanium 733 МГц	R14000A 600МГц
test178(RHF)	1	72	131	117	80	94
(симметрия D3H)	2	43	92	84	-	-
test178 (без	1	358	580	535	-	-
симметрии)	2	185	342	341	-	-
$C_3H_{11}O_4$/MP2	1	170	311	348	189	247
CH_6N_2/MP4	1	1977	3030	3759	2070	3158
test397(DFT)	1	21763	36027	-	27160	29298
	2	11077	19702	-	-	-

Примечания. Для Opteron, Athlon и Pentium III нами использована двоичная версия Gaussian-98 Rev.A11 [26], транслированная с помощью pgf77-4.2 с оптимизацией под Pentium III. Данные для Itanium и R14000 взяты из [21]. Значения SPECfp_base2000 для Opteron, Athlon MP, Itanium и R14000A равны соответственно 1029, 587, 623 и 499 (см. //www.specbench.org). По данным SPECfp2000 и наших тестов прикладных программ, производительность Opteron с плавающей запятой существенно выше, чем в Athlon XP. Но пиковая производительность Opteron меньше, чем у Pentium 4, из-за гораздо более низкой тактовой частоты, и на тестах, имеющих высокую долю операций с плавающей запятой (например, тестах Linpack), Opteron может сильно уступать Xeon/Pentium 4 (Табл.1).

Таблица 9. Эффективность применения 64- и 32-разрядных x86-совместимых микропроцессоров на задачах квантовой химии с большим объемом ввода-вывода

Тесты Gamess-US (conventional) [27]	N CPU	Процессорное время расчета, сек		
		Opteron 1.6 ГГц	Athlon MP1800+ 1533 МГц	Pentium 4 1.8 ГГц
test178(RHF)	1 2	49/57 30/47	96/102 -	68/71(1) 41/53(2)
test397(DFT)	1	11446/28049	-	13996
$C_3H_{11}O_4$/MP2	1 2	83,6/84,3 69/108	132/102 -	93(3) -

Примечания. Через слэш приведено старт-стопное время, которое в conventional- методах из-за большого ввода-вывода гораздо больше (лимитирует время ожидания ввода-вывода). (1) Все данные для Athlon получены при трансляции GAMESS- US с помощью ifc 7.1 с ключами -O3 -tpp6 -ip, а для остальных процессоров - с ключами -O3 -tpp7 -xW -ip. При трансляции с ключами -O3 -tpp6 с последующим прогоном на Pentium 4 время выполнения равно 71/75 сек. (2) Для Pentium 4 приведены данные не для SMP-сервера, а для двух однопроцессорных узлов кластера, соединенных Fast Ethernet со связыванием двух каналов. (3) При трансляции с ключами -O3 -tpp6 -ip время выполнения составляет 104/113 сек.

имеется целый ряд исключений, зависящих от специфики задач и микроархитектуры МП (что и следовало ожидать, учитывая разнообразие расчетных методов). Соотношение производительностей Opteron и Pentium 4/Xeon достаточно сильно зависит от теста.

ВЫВОДЫ. 1) Дан краткий обзор факторов, влияющих на эффективность применения кластеров Beowulf в расчетах методами квантовой химии и молекулярной динамики.

2) Показано, что при большом числе процессоров (свыше 8-16) эффективность распараллеливания в типовых режимах расчетов (в частности, для полуэмпирических и неэмпирических квантовохимических методов ССП и теории возмущений Меллера-Плессета) существенно снижается.

3) Анализ технических характеристик и полученные авторами результаты измерений производительности серверов на базе Opteron показывают перспективность их применения в узлах кластеров при проведении квантовохимических расчетов неэмпирическими методами квантовой химии (ССП, MP2, MP4, DFT).

Работа поддержана РФФИ, проект 01-07-90072.

ЛИТЕРАТУРА

1. *Танненбаум Э.*(2002) Архитектура компьютера (пер. с англ.), "Питер", М.
2. *Столлингс У.* (2002) Структурная организация и архитектура компьютерных систем (пер. с англ.), "Вильямс", М.
3. *Кузьминский М., Волков Д.* (1995) Открытые системы, №6, 33 - 40.
4. *Кузьминский М.*(1999) Открытые системы, №3, 16 - 20.
5. *Кузьминский М.*(2003) Открытые системы, №7 - 8, 12 - 15.
6. *Кузьминский М.*(2000) Открытые системы, №11, 9 - 13.
7. *Кузьминский М.*(2001) Открытые системы, №1, 16 - 19.
8. *Kenny Yu. J.-S., Chin-Hui Yu.* (2003) J. Chem. Inf. Comput. Sci., **42**, 673 - 679.
9. *Касперски К* (2003) Техника оптимизации программ. Эффективное использование памяти, БХВ-Петербург, СПб.
10. *Кузьминский М.*(2002) Открытые системы, №9, 10 - 18.
11. "Gaussian 94 Workshop Notes. Mountain View, Feb. 27 - March 1 1996" (1996), SGI, Inc. and NCSA, Mountain View.
12. *Кузьминский М.*(1998) Открытые системы, №3, 19-23.
13. *Кузьминский М., Мускатин А.*(2001) Открытые системы, №7 - 8, 17-22.
14. *Кузьминский М.Б., Мендкович А.С. и др.* (2002) в кн.: Высокопроизводительные параллельные вычисления на кластерных системах", Сб. материалов II Международного научно-практического семинара, 26-29 ноября 2002 г., Изд-во ННГУ, Н. Новгород, с.169 - 174.
15. *Корнеев В.В.* (1999) Параллельные вычислительные системы, Нолидж, М.
16. *Андреев А., Воеводин В., Жуматий С.* (2000) Открытые системы, №5 - 6, 9 - 14.

17. *Воеводин В.В., Воеводин Вл.В.* (2002) Параллельные вычисления, БХВ-Петербург, СПб.
18. *Митрофанов В., Слуцкий А., Ларионов К., .Эйсымонт М.* (2003) Открытые системы, №5, 29 - 35.
19. *Кузьминский М.* (2003) Открытые системы, №6, 10 - 17.
20. *Guest M.F.* (2000) "Performance of Various Computers in Computational Chemistry" Technical Report, Daresburgy Lab., Cheshire, UK.
21. SGI Computational Chemistry Applications Performance Report.Spring 2002 (2002), Silicon Graphics, Inc.
22. *Немнюгин С., Стесик О.* (2002) Параллельное программирование для многопроцесорных вычислительных систем, БХВ-Петербург, СПб.
23. SGI Computational Chemistry Applications Performance Report.Fall 2000 (2000) Silicon Graphics, Inc.
24. *Gorb L.,Yanov I., Leszczynski J.* (2000) Parallel Computing, **26**, 1043 - 1060.
25. *Hawick K.A., Grove D.A., Coddington P.D., Buntine M.A.* (2000) Commodity Cluster Computing for Computational Chemistry, Technical Report DHPC-073, Adelaide Univ. (Australia).
26. Gaussian 98, Revision A.11.3, Frisch M. J., Trucks G. W., Schlegel H. B., Scuseria G. E., Robb M. A., Cheeseman J. R., Zakrzewski V. G., Montgomery J. A., Stratmann R. E., Burant J. C., Dapprich S., Millam J. M., Daniels A. D., Kudin K. N., Strain M. C., Farkas O., Tomasi J., Barone V., Cossi M., Cammi R., Mennucci B., Pomelli C., Adamo C., Clifford S., Ochterski J., Petersson G. A., Ayala P. Y., Cui Q., Morokuma K., Malick D. K., Rabuck A. D., Raghavachari K., Foresman J. B., Cioslowski J., Ortiz J. V., Baboul A. G., Stefanov B. B., Liu G., Liashenko A., Piskorz P., Komaromi I., Gomperts R., Martin R. L., Fox D. J., Keith T., Al-Laham M. A., Peng C. Y., Nanayakkara A., Gonzalez C., Challacombe M., Gill P. M. W., Johnson B., Chen W., Wong M. W., Andres J. L., Gonzalez C., Head-Gordon M., Replogle E. S., and Pople J. A., Gaussian, Inc., Pittsburgh PA, 1998.
27. *Schmidt M.W., Baldridge K.K., Boatz J.A., Elbert S.T., Gordon M.S., Jensen J.H., Koseki S., Matsunaga N., Nguyen K.A., Su S., Windus T.L., Dupuis M., Montgomery J.A.*(1993), J. Comput. Chem., **14**, 1347 - 1363.

COMPUTATIONAL RESOURCES OF CLUSTERS AND THEIR USE IN CHEMICAL CALCULATIONS

M.B. Kuzminskiy, A.S. Mendkovich

N.D. Zelinsky Institute of Organic Chemistry RAS (IOC RAS);
47, Leninsky pr., Moscow, 119991 Russia
tel: (095) 135-63-88, (095) 135-63-77; fax: (095) 135-53-88; e-mail: kus@free.net

An overview of main technical characteristics of Beowulf clusters (both their nodes and interconnect) is presented. Performance data, in particular about parallelization of quantum chemical and molecular dynamics calculations in this clusters, are presented. The tests of servers based on 64-bit x86-compatible new generation microprocessors - AMD Opteron - are performed, and promise of their using in clusters is demonstrated.

Key words: clusters, parallelization, quantum chemistry, molecular dynamics