Организация баз данных
2002. №2(4)

УДК 519.683

© 2002 г. **В.А. Анненков**,
Е.А. Нурминский, д-р физ.-мат. наук,
С.В. Смирнов, канд. физ.-мат. наук
(Институт автоматики и процессов управления ДВО РАН, Владивосток)

АНАЛИЗ ПРОИЗВОДИТЕЛЬНОСТИ МЕЖПРОЦЕССОРНОГО ОБМЕНА В МВС-1000М/17[1]

Анализируются возможности повышения скорости межпроцессорного обмена в МВС-1000М/17. На примере MPI/LAM изучается влияние системных параметров, режимов обмена и типов данных на производительность межпроцессорного обмена.

Введение

Одним из направлений развития высокопроизводительной вычислительной техники являются массивно-параллельные компьютеры с распределенной памятью [1,2]. К данному классу относятся Intel Paragon, IBM SP1, Parsytec, МВС-100/1000, вычислительные кластеры. Такие компьютеры состоят из большого числа типовых однородных вычислительных модулей, соединенных некоторой коммуникационной средой. Вычислительные модули представляют собой серийные процессоры с локальной памятью. Известными преимуществами вычислителей данного класса являются возможность наращивать вычислительную мощность за счет добавления вычислительных модулей и дешевизна, позволяющая в определенных классах задач служить выгодной альтернативой крайне дорогим суперкомпьютерам.

Основным ограничивающим фактором вычислительного процесса в компьютерах с распределенной памятью является относительно медленное межпроцессорное взаимодействие по сравнению со скоростью локальной обработки данных самими процессорами. В нашей статье анализируются возможности повышения скорости межпроцессорного обмена в МВС-1000, которая является одной из распространенных реализаций этой архитектуры, в связи с чем следует особо подчеркнуть возможности поэтапной модерни-

[1] Работа выполнена при финансовой поддержке Российского фонда фундаментальных исследований (код проекта 01-07-90225) и федеральной целевой программы «Интеграция» (код проекта В-0009).

зации и настройки аппаратно-программного комплекса вычислителя на определенные классы задач.

В МВС-1000 межпроцессорный обмен ведется по сети Fast Ethernet. Существует реальная возможность практически кратного ускорения обмена по этой сети путем ”связывания каналов” (channel bonding) [3], т.е. объединения нескольких каналов Fast Ethernet. В такой конфигурации вычислительный модуль соединяется с коммутатором более чем одним каналом. Связывание каналов в узлах под управлением ОС Linux позволяет организовать равномерное распределение нагрузки приема/пере-дачи между соответствующими каналами. Аппаратная модернизация требует относительно небольших затрат на дополнительные сетевые карты и коммутаторы Fast Ethernet.

В настоящее время наблюдается тенденция удешевления компонент, необходимых для построения сети Gigabit Ethernet. Тем не менее скачкообразный переход сразу всего комплекса на Gigabit Ethernet может быть затруднен по финансовым соображениям, поэтому следует рассматривать некоторые промежуточные варианты, когда дальнейшее наращивание вычислительной мощности комплекса производится путем добавления группы процессорных элементов, соединенных с вычислительной сетью системы через гигабитный интерфейс.

В такой ситуации исключительно важной является эффективная организация взаимодействия между процессорами с разными сетевыми интерфейсами. В частности, процессор с гигабитной сетевой картой может взаимодействовать через гигабитный аплинк (uplink) коммутатора Fast Ethernet сразу с несколькими процессорами с сетевыми картами Fast Ethernet. В таком смешанном варианте удается с минимальными затратами получить эффективную реализацию схемы обмена одного выделенного процессора со многими, что существенно расширяет круг задач, которые могут быть столь жеэффективно реализованы на вычислителе.

Модернизированный таким образом МВС-1000 можно применять, например, при решении задач многоканальной обработки спутниковой информации высокого разрешения, в задачах высококачественной визуализации динамики сложных пространственных объектов, в задачах расчета динамики океана с применением моделей, в которых необходимо обращать двумерный эллиптический оператор для всей счетной области [4].

В данной работе исследовалась пропускная способность при межпроцессорном обмене данными через общую память, через гигабитный аплинк коммутатора и обмен по сети Fast Ethernet при связывании пар каналов. Работа направлена на отработку составляющих модернизации вычислителя. Кроме того, изучение особенностей различных режимов передачи данных в вычислительном комплексе необходимо для решения проблемы отображения алгоритмов на архитектуру вычислительной системы.

Вычислительный комплекс МВС-1000М/17-МВС-1000/16

Система МВС-1000, установленная в Институте автоматики и процессов управления ДВО РАН (http://www.dvo.ru/bbc/hardware/mbc1000/), представляет собой многомашинный параллельный вычислитель, разработанный совместно ИПМ им.М.В.Келдыша и НПО ”Квант”. Ведущей машиной служит МВС-1000М/17, ведомой – МВС-1000/16.

Основу вычислительной мощности МВС-1000М/17 составляют семнадцать вычислительных узлов на базе микропроцессоров Intel Pentium III. В состав вычислительного узла входят 2 процессора Pentium III 1ГГц, оперативная память DIMM SDRAM 1Гбайт, PC-133, жесткий диск 20Гбайт. Шестнадцать вычислительных узлов (ВУ) объединены в кластер двумя сетями Fast Ethernet, семнадцатый узел – управляющий (УВУ).

Коммуникационная среда МВС1000М/17 включает 24-портовый коммутатор Fast Ethernet для служебных пересылок (Compex SRX2224), 16-портовый коммутатор Fast Ethernet с Gigabit Ethernet uplink (Intel Express 460T Standalone Switch) и 8-портовый коммутатор Gigabit Ethernet (Compex GSC1008). Каждый ВУ входит в две независимо коммутируемые сети Fast Ethernet, а УВУ входит в сеть для служебных пересылок и локальную сеть пользователей.

Первая сеть Fast Ethernet предназначена для служебных пересылок, она обеспечивает функции управления ВУ и доступ из них к общим периферийным ресурсам. Вторая сеть Fast Ethernet служит для передачи данных внутри параллельных программ, работающих на группе ВУ.

Для межмашинного обмена данными служит сеть Gigabit Ethernet. К ней подключены гигабитные аплинки коммутаторов сетей внутрипрограммных пересылок ведущей машины МВС-1000М/17 и ведомой МВС-1000/16. Межмашинный обмен служебными сообщениями производится по сети Fast Ethernet.

Все узлы вычислительной системы работают под управлением операционной системы Red Hat Linux 7.2. Вычислительная система реализована на модульной основе, это облегчает проведение модернизации и позволяет вносить изменения в конструкцию вычислителя.

На рис.1 представлена структурная схема из проекта модернизации МВС-1000М/17. С помощью связывания каналов планируется увеличить пропускную способность сети для передачи сообщений внутри программ. Все ВУ построены на базе серверной платформы GS-SR101. Материнская плата включает двухпортовый встроенный сетевой интерфейс Fast Ethernet, и есть возможность подключения дополнительной сетевой карты. Имеющийся коммутатор Intel Express 460T Standalone Switch поддерживает режим связывания каналов. Предполагается, что дополнительные вычислительные узлы, напрямую выходящие на гигабитный коммутатор, должны иметь высокие показатели производительности как процессоров, так и ши-

ны данных и шины PCI, чтобы выполнять обмены по сети одновременно с обработкой информации. По существующей терминологии эти дополнительные узлы можно классифицировать как станции (СТ).

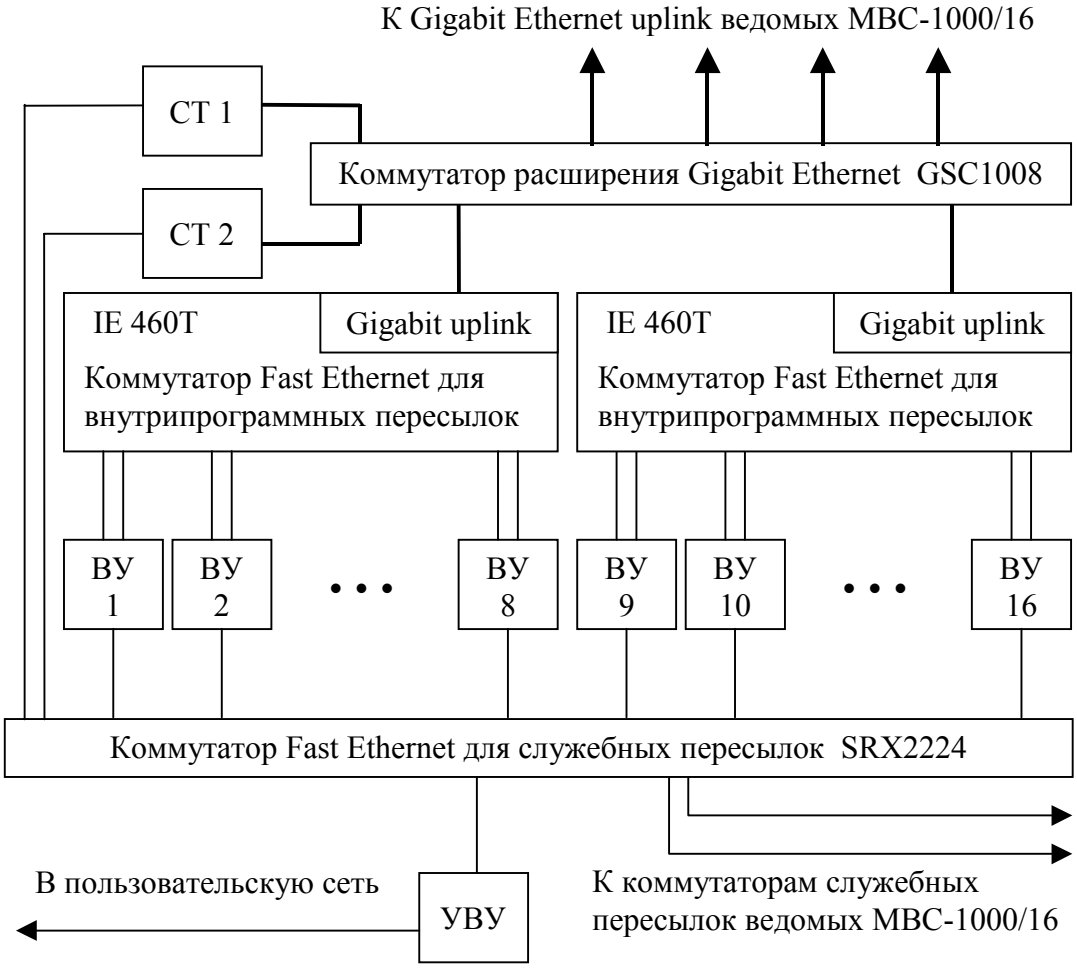

Рис. 1. Структура МВС-1000М/17 (проект).

Для целей предварительного тестирования в качестве станции использовался двухпроцессорный компьютер на базе процессора Intel Pentium III с частотой 800МГц. Операционная система Linux SuSe 8.0 на базе ядра версии 2.4.18. В качестве сетевого интерфейса применялась плата NetGear GA620T, установленная в 64-разрядный разъем PCI (частота 66МГц). Для межпроцессорного обмена была установлена реализация MPI/LAM [5] версии 6.5.7 (http://www.lam-mpi.org/6.5/).

Количественные характеристики межпроцессорного взаимодействия в МВС-1000М/17

Для определения количественных характеристик межпроцессорного обмена применялись средства MPI/LAM. Таймирование операций обмена производилось с помощью высокоточной функции MPI_Wtime. Для уменьшения накладных расходов на организацию вычислительной работы все из-

мерения проводились многократно в цикле. На графиках в качестве длины сообщений, представляющих собой одномерные массивы двойных слов (8 байт или 64 двоичных разряда), представлены размеры массивов (длины векторов).

Вычислительный узел комплекса МВС-1000/17 представляет собой двухпроцессорную SMP (symmetrical multiprocessor) систему, в связи с этим необходимо исследовать количественные характеристики межпроцессорного взаимодействия через общую память на уровне SMP-узла. Отметим, что наиболее высокие результаты достигаются при использовании коммуникационного протокола usysv (устанавливается при конфигурировании пакета LAM с параметром --with-rpi=usysv). Важным параметром системы является размер буфера в разделяемой оперативной памяти. При малом размере буфера суммарные накладные расходы при передаче сегментов длинного сообщения значительно снижают пропускную способность межпроцессорного обмена. Увеличение буфера в разделяемой памяти достигается увеличением параметра LAM_SHMSHORTMSGLEN, задающего максимальный размер короткого сообщения. Представленные результаты получены при двух значениях этого порогового параметра: 8 Кбайт, выставляемого по умолчанию в LAM, и 80 Кбайт. На графиках (см. рис. 2) первому значению соответствуют штриховые линии, второму – сплошные. Таймирование выполнялось на односторонних операциях приема и передачи. Процессоры по очереди принимают и отправляют сообщения (ping-pong). На рис.2 представлены показатели производительности межпроцессорного обмена, выполненного на блокирующих операциях MPI_Send и MPI_Recv.

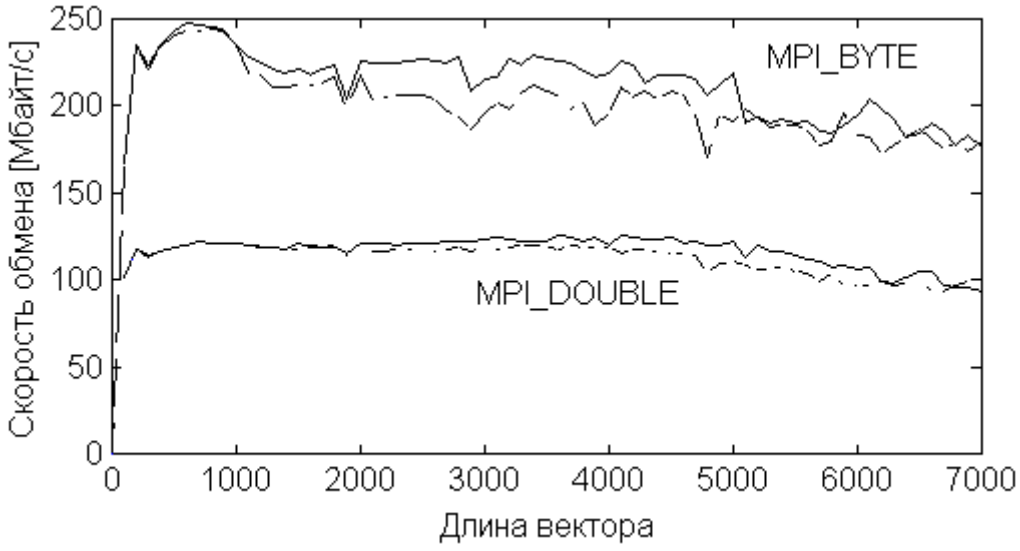

Рис. 2. Производительность межпроцессорного обмена в SMP-узле с применением блокирующих операций.

Значения пропускной способности приближаются к половине максимальной пропускной способности обмена с оперативной памятью, что озна-

чает высокую эффективность межпроцессорного обмена на данных операциях.

На рис.3 представлены показатели производительности межпроцессорного обмена, выполненного на неблокирующих операциях MPI_Isend и MPI_Irecv.

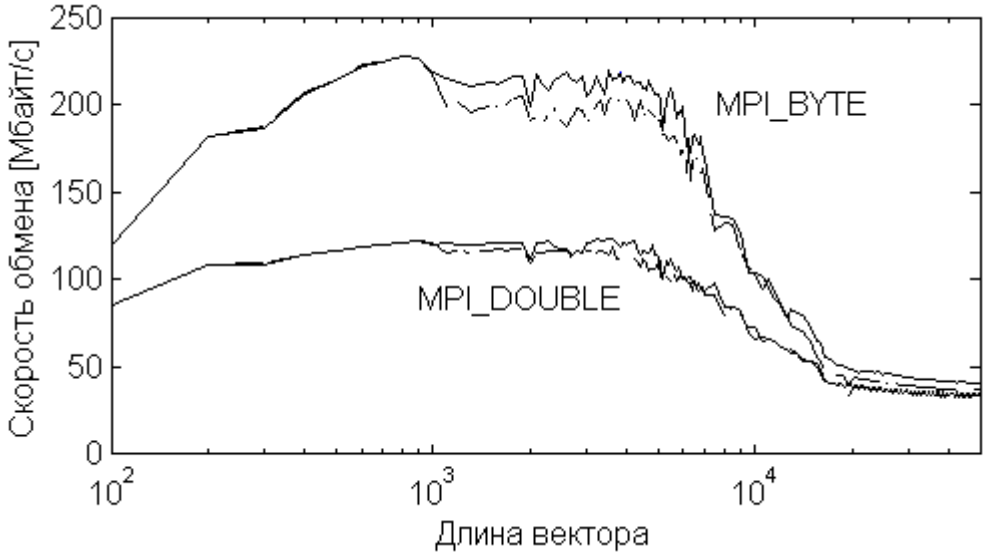

Рис.3. Производительность межпроцессорного обмена в SMP-узле с применением неблокирующих операций.

В этом случае дополнительные накладные расходы на запись сообщения в промежуточный буфер приводят к снижению производительности межпроцессорного обмена. Отметим катастрофическое падение производительности на передаче длинных сообщений, когда суммарный размер буферов с данными превышает размер кэша L2. В этой ситуации при всех перемещениях данных задействуется внешняя системная шина.

Существенной особенностью данной реализации MPI/LAM является зависимость скорости обмена от типа передаваемых данных: передача массивов данных с типом MPI_DOUBLE (двойное слово) производится почти в два раза медленнее, чем массивов данных с типом MPI_BYTE. Для построения графиков использовались максимальные значения из полученных в нескольких замерах производительности.

Пример результатов одиночного замера пропускной способности представлен на рис. 4. На кривых хорошо заметны острые отрицательные пики. Эти пики при следующих замерах смещаются, поэтому они отсутствуют на рис. 2 и рис. 3. Таймирование в схеме обмена одного выделенного процессора со многими выполнено следующим образом. Сначала замеряется скорость отправки, когда выделенный процессор многократно рассылает сообщения всем 16 процессорам. Затем замеряется скорость приема, когда выделенный процессор многократно получает сообщения от 16 процессоров.

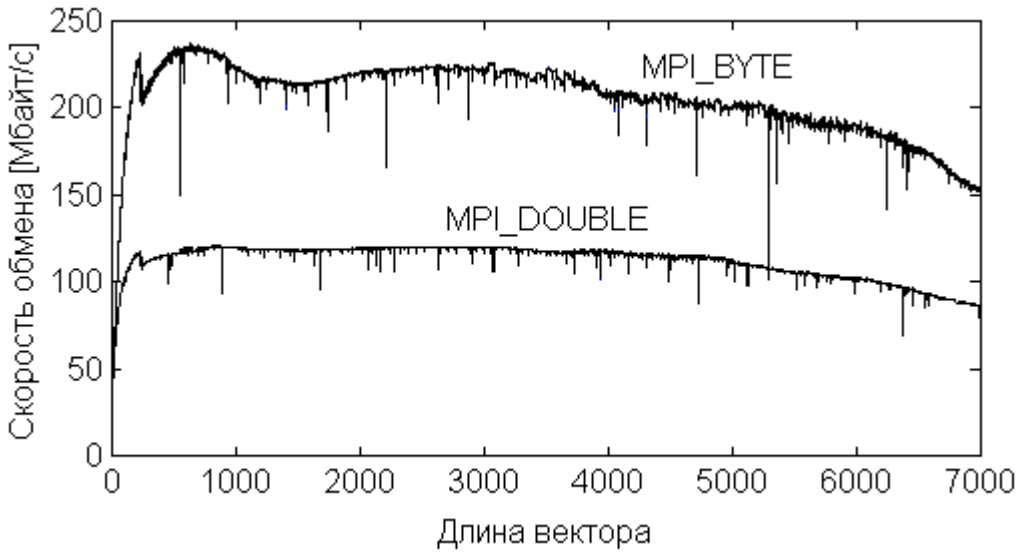

Рис. 4. Пример одиночного измерения производительность межпроцессорного обмена в SMP-узле с применением блокирующих операций.

Целью таймирования является получение показателей максимальной пропускной способности, именно на такой режим по умолчанию настроены параметры драйвера acenic сетевой карты NetGear GA620T. Для достижения оптимальных характеристик при передаче коротких одиночных сообщений необходимо изменение настроек драйвера. Результаты измерений представлены на рис. 5.

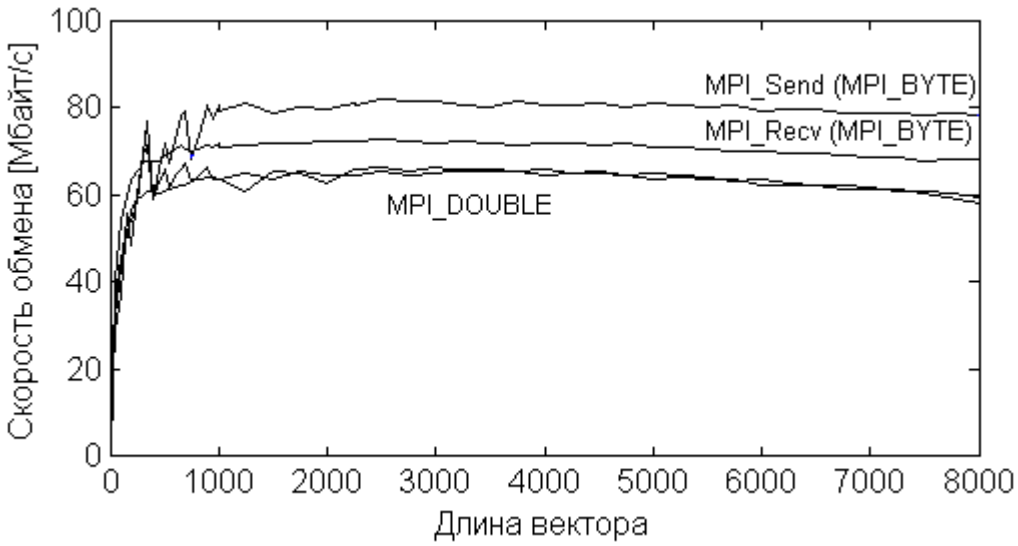

Рис.5. Производительность межпроцессорного обмена через гигабитный аплинк.

При длине сообщений более 8Кбайт (длина вектора более 1000) кривые показателей производительности упорядочиваются следующим образом. Наиболее быстро, со скоростью порядка 80 Мбайт/с, происходит рассылка байтовых массивов. Со скоростью до 71 Мбайт/с производится прием байтовых массивов. Прием и рассылка массивов двойных слов производятся примерно с одинаковой скоростью до 64 Мбайт/с.

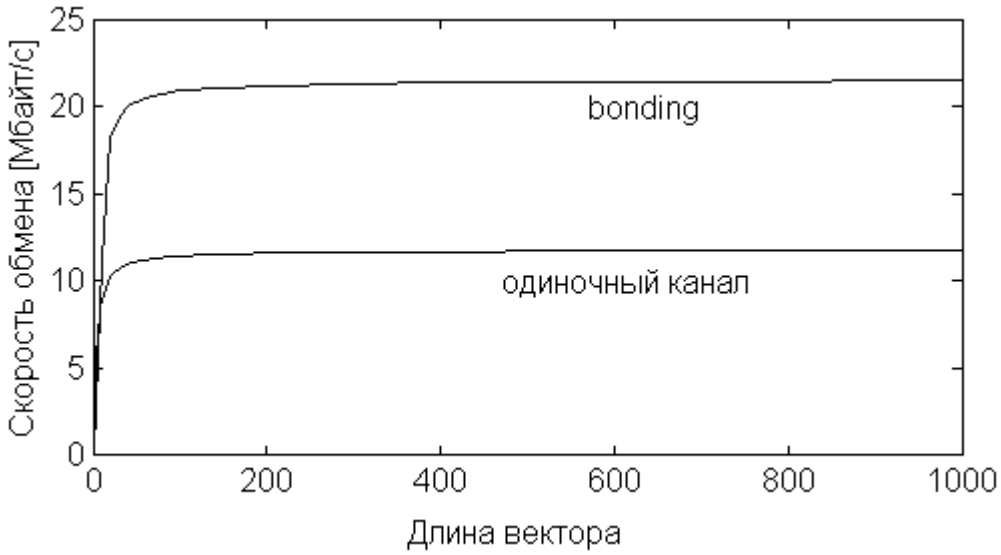

Рис. 6. Производительность межпроцессорного обмена для одиночного и двух связанных каналов Fast Ethernet.

Результаты измерения пропускной способности при связывании пар каналов Fast Ethernet (bonding) демонстрируют эффективность этого метода. Верхняя кривая на рис.6 построена по данным измерений для связанных пар каналов, нижняя – по данным для одиночного канала. Пропускная способность повышается с 12Мбайт/с до 22Мбайт/с, что примерно соответствует 92% от двукратного ускорения.

Заключение

В данном исследовании изучались характеристики обмена данными между процессорами вычислительного комплекса МВС1000М/17 в нескольких коммуникационных средах. Эта информации необходима для предварительного планирования постановки различных проблемных приложений на вычислительный комплекс. Дальнейшее тестирование необходимо проводить уже с учетом специфики конкретного приложения.

При анализе прохождения параллельной программы в случае интенсивного межпроцессорного обмена данными необходимо учитывать ограничивающие факторы всех промежуточных звеньев.

При межпроцессорном обмене через общую память ограничивающим фактором является скорость обмена процессоров с оперативной памятью. В этом случае совмещение межпроцессорного обмена принципиально возможно только с такой вычислительной работой, которая почти не требует обращения к оперативной памяти. При обработке больших потоков данных, по-видимому, оптимальным решением проблемы обмена является применение блокирующих операций. Такой подход требует некоторой синхронизации работы процессоров, но позволяет избежать лишних операций записи в промежуточные буферы.

При межпроцессорном обмене по сети Gigabit Ethernet известным решением проблемы повышения пропускной способности до заявленного значения порядка 1Гбит/с является применение увеличенного размера сегмента передаваемых по TCP/IP данных (Jumbo box). Этот размер превышает максимальный размер сегмента в стандарте Fast Ethernet. Таким образом, чтобы достичь эффективной пропускной способности при прохождении данных и по сети Gigabit Ethernet и по сети Fast Ethernet, необходимо отойти от стандарта Fast Ethernet. В используемом участке сети Fast Ethernet все устройства и драйверы должны поддерживать опцию Jumbo box. Следует отметить, что для повышения пропускной способности гигабитного соединения применяется методика накопления сегментов сообщений. Это означает, что обработка короткого сообщения может начаться не сразу после приема. Если принципиальной чертой проблемной задачи является обмен короткими сообщениями, требуется соответствующая настройка драйвера сетевой гигабитной карты, направленная на уменьшение времени задержки.

ЛИТЕРАТУРА

1. *Корнеев В.В.* Параллельные вычислительные системы. М.:Нолидж, 1999.
2. *Zabrodin A.V., Levin V.K., Korneev V.V.* The massively parallel computer system MBC-100. Lecture Notes in Computer Science, № 964. //Parallel Computing Technologies. Third International Conference,PaCT-95, St.Petersburg, Russia, Sept.1995, Springer. P.341-355.
3. *Becker D.J., Sterling T., Savarese D., Dorband J.E., Ranawak U.A., Packer C.V.* BEOWULF: A Parallel Workstation for Scientific Computation, in Proceedings, International Conference on Parallel Processing, 1995.
4. *Левин В.А., Смирнов С.В.* Вычислительные алгоритмы численной модели динамики океана для параллельной ЭВМ //Сб. научн. статей. Владивосток: ИАПУ ДВО РАН. 2001, С.180-191.
5. *Burns G., Daoud R., Vaigl J.* LAM: An open cluster environment for MPI, in Proceedings of Supercomputing Symposium ’94 (J.W. Ross, ed.): University of Toronto, 1994. P.379–386.