УДК 519.6

ОСОБЕННОСТИ РЕАЛИЗАЦИИ КОММУНИКАЦИОННОГО ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ КС-ЭВМ ПРИ ИСПОЛЬЗОВАНИИ БЕСКОММУТАТОРНОЙ ТЕХНОЛОГИИ

Д. А. Жуков, В. М. Вялухин
(РФЯЦ-ВНИИЭФ)

Описываются особенности реализации коммуникационного программного обеспечения для построения компактных высокопроизводительных вычислительных комплексов с применением бескоммутаторной технологии на базе архитектуры InfiniBand. Приводятся описания настройки коммуникационного программного обеспечения, алгоритмов межпроцессорного взаимодействия в стандарте MPI в бескоммутаторной коммуникационной среде. Данные программные решения рассмотрены применительно к компактному высокопроизводительному вычислительному комплексу, разработанному в РФЯЦ-ВНИИЭФ.

Ключевые слова: бескоммутаторная технология, коммуникационное программное обеспечение, система межпроцессорных обменов, архитектура InfiniBand, MPI-процесс, КС-ЭВМ (компактная суперЭВМ).

Введение

Особенности разработки мультипроцессорных вычислительных комплексов для параллельных вычислений связаны с применением высокоскоростных средств коммуникации между отдельными вычислительными модулями (ВМ), объединенными между собой в единую коммуникационную сеть. Такие вычислительные комплексы, как правило, представляют достаточно сложные программно-аппаратные реализации, позволяющие объединять в единое вычислительное поле несколько тысяч ВМ, обеспечивающих большой объем параллельных вычислений. Достаточно высокая стоимость таких систем на стадиях проектирования и дальнейшей эксплуатации делает их недоступными для использования небольшими предприятиями и коммерческими организациями.

Для решения сравнительно небольших задач с применением параллельных вычислений, как правило, достаточно использовать компактные рабочие станции (с терафлопной производительностью), имеющие в своем составе несколько ВМ, объединенных в единую коммуникационную сеть.

Широко распространенной на сегодняшний день коммуникационной технологией является технология InfiniBand [1], которая позволяет разрабатывать с использованием высокоскоростного коммутируемого межпроцессорного взаимодействия высокопроизводительные вычислительные системы с различными топологиями.

Особенностью вычислительного комплекса, разработанного в РФЯЦ-ВНИИЭФ, является бескоммутаторная технология объединения ВМ в единую коммуникационную инфраструктуру на базе архитектуры InfiniBand. По сравнению с коммутаторными системами данная система имеет более низкую стоимость, меньшие габаритные размеры и не требует больших затрат на обслуживание.

В данной работе рассматриваются различные варианты бескоммутаторных топологий вычислительной среды, а также особенности реализации коммуникационного программного обеспечения (ПО) для параллельных вычислений в стандарте MPI (Message Passing Interface) [2]. Приводится практическое решение по настройке параметров коммуникационной среды.

Особенности бескоммутаторной топологии

В основу бескоммутаторной технологии объединения ВМ положена возможность стандарта InfiniBand передавать данные между абонентами без использования коммутируемого соединения. Такая технология может эффективно использоваться при небольшом количестве ВМ.

Основными элементами коммуникационной среды при использовании бескоммутаторной технологии являются InfiniBand-адаптеры, обеспечивающие соединение *точка-точка* между ВМ, в результате чего всю вычислительную систему можно представить в виде совокупности нескольких независимых InfiniBand-подсетей. Такой подход дает возможность представить топологию всей вычислительной системы в виде полного графа, количество вершин которого соответствует количеству ВМ, и реализовать несколько вариантов топологий.

На рис. 1 приведены возможные варианты бескоммутаторных топологий с использованием InfiniBand-адаптеров. Данные схемы соединения ВМ предполагают межпроцессорное взаимодействие между абонентами по принципу *каждый с каждым* и не предусматривают транзитных пересылок пакетов внутри ВМ.

Разработанный в РФЯЦ-ВНИИЭФ вычислительный комплекс состоит из трех ВМ, объединенных между собой с помощью двухпортовых InfiniBand-адаптеров. Структурная схема вычислительного комплекса приведена на рис. 1 слева. Результатом такого соединения является наличие трех независимых InfiniBand-подсетей IB_{S1}, IB_{S2}, IB_{S3}. Каждая подсеть управляется собственным менеджером подсети OpenSM [1] пакета OFED [3], который является стандартным коммуникационным ПО для InfiniBand-сетей.

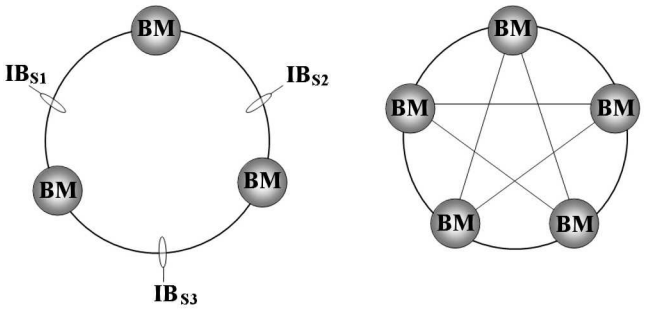

Рис. 1. Возможные варианты топологий для реализации бескоммутаторной коммуникационной среды

Коммуникационное ПО КС-ЭВМ

Вычислительный комплекс РФЯЦ-ВНИИЭФ работает под управлением ОС Unix/Linux и включает в себя два основных программных компонента:

- системное ПО транспортного уровня;
- коммуникационное ПО уровня пользовательского процесса.

В качестве системного ПО в данном вычислительном комплексе используется пакет OFED, который представляет собой программную реализацию поддержки высокоскоростного коммуникационного оборудования для InfiniBand-сетей. Данное ПО включает в себя набор драйверов различных уровней для реализации высокоскоростных коммуникационных протоколов. В состав пакета также входят набор тестов для измерения коммуникационных характеристик InfiniBand-фабрики, утилиты мониторинга и управления коммуникационной средой InfiniBand.

Состав коммуникационного ПО для реализации параллельных вычислений в стандарте MPI представлен пакетами OpenMPI [4] и MVAPICH [5]. Эти пакеты являются переносимыми реализациями библиотек функций MPI для разработки и выполнения параллельных приложений в вычислительной среде многопроцессорных комплексов и систем массового параллелизма. В частности, для реализации межпроцессорных обменов в бескоммутаторной коммуникационной среде были выбраны реализации MPI-библиотек с поддержкой многоканальной передачи данных.

Проблемы использования библиотек MPI в бескоммутаторной коммуникационной среде

Бескоммутаторная технология объединения ВМ в коммуникационную сеть накладывает ряд ограничений на использование стандартного коммуникационного ПО. В частности, библиотека MPI, представленная пакетом MVAPICH, рассчитана на межпроцессорное взаимодействие внутри единой коммуникационной сети и не предназначена для использования в вычислительных системах, состоящих из нескольких независимых InfiniBand-подсетей. Основной проблемой при передаче сообщения удаленному абоненту в случае использования бескоммутаторной

мутаторной технологии является выбор нужной InfiniBand-подсети в соответствии с *ранком* (rank) удаленного MPI-процесса, т. е. выбор нужного подканала передачи данных.

С точки зрения архитектуры InfiniBand каждый MPI-процесс в пределах одной InfiniBand-подсети может адресоваться уникальной парой очередей QP (Queue Pair), которая является своего рода виртуальным каналом, и локальным идентификатором LID (Local Identifier). Локальный идентификатор присваивается менеджером подсети SM (Subnet Manager) [1] каждому порту или портам адаптера. В пределах одной InfiniBand-подсети LID, так же как и QP, являются уникальными. Еще один немаловажный параметр при отправке сообщения — подканал передачи данных, который фактически является логическим номером физического порта InfiniBand-адаптера. Количество подканалов передачи данных на каждом из ВМ определяется общим количеством физических портов HCA (Host Channel Adapter).

При стандартном использовании пакета MVAPICH внутри единой коммуникационной среды основная настройка подканалов передачи данных в соответствии с ранками MPI-процессов происходит на стадии загрузки и инициализации MPI-процессов, в результате чего модификация каждой QP выполняется с использованием циклического алгоритма, начиная с нулевого подканала. Механизм взаимодействия MPI-процессов внутри единой коммуникационной среды с использованием двухпортовых InfiniBand-адаптеров приведен на рис. 2, где цифрой 1 обозначена фаза инициализации подканалов передачи данных, цифрой 2 — передача данных в канал; P1 и P2 — порты InfiniBand-адаптера.

Дальнейшее коммутирование подканалов передачи данных при обмене сообщениями между абонентами сети происходит с использованием алгоритма Round Robin. При использовании бескоммутаторной технологии объединения ВМ данный алгоритм приводит к неправильному выбору подканала передачи данных, в результате чего пакет отправляется не тому абоненту.

Основные этапы по модификации пакета MVAPICH для работы в бескоммутаторной коммуникационной среде связаны с доработкой механизма инициализации MPI-приложений и механизма коммутирования подканалов передачи данных с учетом нескольких InfiniBand-подсетей.

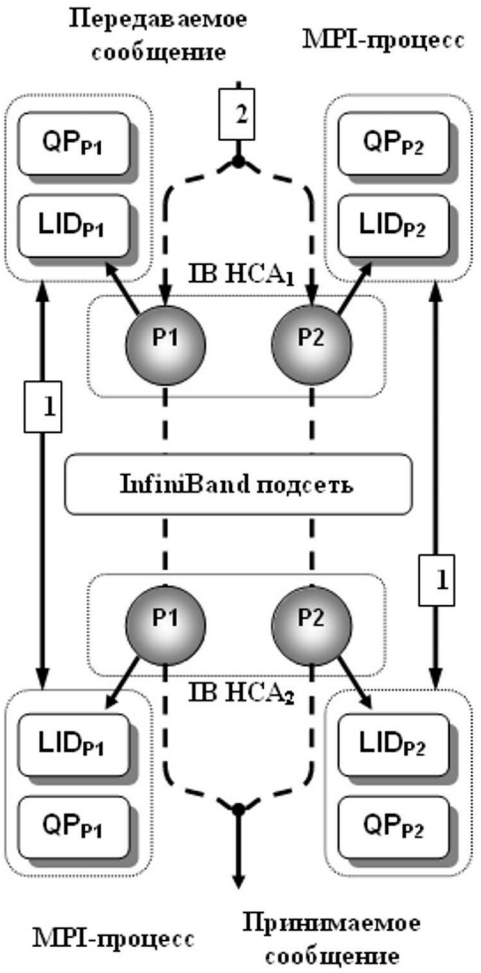

Рис. 2. Механизм инициализации MPI-процессов и передачи сообщений внутри единой коммуникационной среды

Подсистема запуска и инициализации MPI-приложений

Основу механизма транспортного уровня библиотеки MPI составляют виртуальные соединения VC (Virtual Connection), которые создаются на стадии загрузки и инициализации MPI-процессов. Каждое виртуальное соединение описывает взаимодействие MPI-процесса с абстрактным устройством, которое может быть представлено программным интерфейсом библиотеки низкоуровневых примитивов (функции инициализации InfiniBand-адаптера, функции взаимодействия между абонентами и др.). Виртуальное соединение для каждого MPI-процесса включает в себя локальную коммуникационную таблицу, описывающую логические подканалы передачи данных.

Для бескоммутаторной коммуникационной среды был введен дополнительный параметр, с помощью которого можно идентифицировать абонентов на принадлежность одной из InfiniBand-подсетей. Роль такого идентификатора играет префикс SP (Subnet Prefix) InfiniBand-подсети. Данный параметр настраивается с помощью SM для каждой InfiniBand-подсети индивидуально на стадии ее инициализации.

В результате выполненной авторами работы локальная коммуникационная таблица для каждого MPI-процесса в бескоммутаторной коммуникационной среде в случае использования двухпортовых InfiniBand-адаптеров была дополнена следующими параметрами:

1) локальные идентификаторы портов: $\mathrm{LID_{P1}}$, $\mathrm{LID_{P2}}$;
2) пары очередей для каждого MPI-процесса: $\mathrm{QP_{P1}}$, $\mathrm{QP_{P2}}$
3) префиксы InfiniBand-подсетей: $\mathrm{SP_{P1}}$, $\mathrm{SP_{P2}}$.

Стадия загрузки и инициализации MPI-процессов включает в себя обмен локальными коммуникационными таблицами между абонентами разных подсетей с последующим формированием для каждого виртуального соединения глобальной коммуникационной таблицы.

Механизм обмена локальными таблицами осуществляется по сети Ethernet с помощью менеджера процессов PM (Process Manager), который является составной частью библиотеки MPI. Дальнейшая модификация QP для каждого MPI-процесса происходит с учетом совпадения префиксов InfiniBand-подсетей. Такой механизм инициализации позволяет значительно снизить накладные расходы при дальнейшем межпроцессорном взаимодействии.

В каждую глобальную коммуникационную таблицу MPI-процесса включается необходимая информация обо всех процессах, участвующих в межпроцессорных обменах в рамках одной задачи. Таблицы атрибутов располагаются в соответствии с ранками MPI-процессов, образуя хэш-таблицу с прямой адресацией.

Межпроцессорное взаимодействие в бескоммутаторной коммуникационной среде

В бескоммутаторной коммуникационной среде основным критерием наличия физической связи между двумя абонентами (MPI-процессами) является совпадение префиксов InfiniBand-подсетей. Алгоритм выбора нужной InfiniBand-подсети (подканала передачи данных) предусматривает выбор атрибутов удаленного соединения из глобальной коммуникационной таблицы локального MPI-процесса (инициатора обменов) в соответствии с ранком MPI-процесса приемника. Механизм коммутирования подканалов передачи данных в бескоммутаторной коммуникационной среде, реализованный в модифицированной библиотеке MPI, приведен на рис. 3.

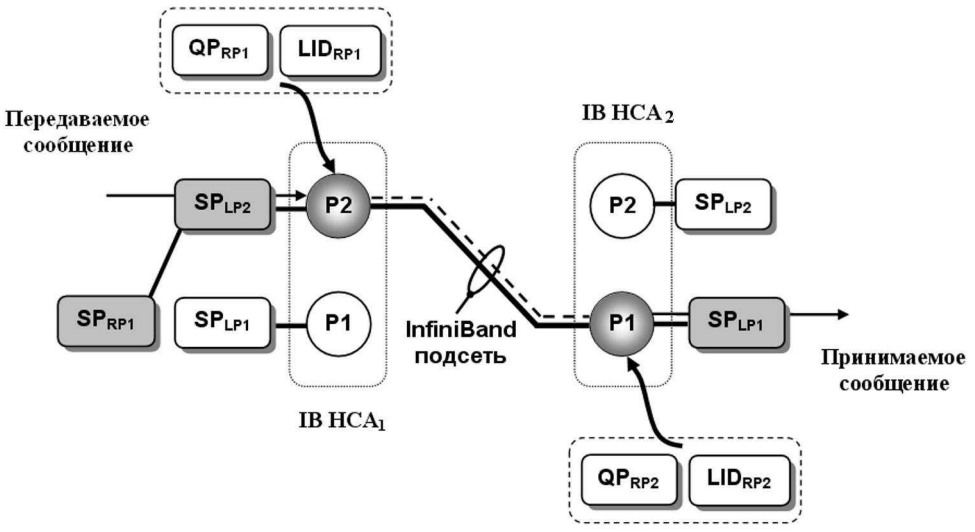

Рис. 3. Механизм коммутирования подканалов передачи данных в бескоммутаторной коммуникационной среде: $\mathrm{SP_{LP}}$, $\mathrm{SP_{RP}}$ — префиксы InfiniBand-подсети соответственно для локального и удаленного MPI-процессов; $\mathrm{QP_{RP}}$, $\mathrm{LID_{RP}}$ — атрибуты удаленного подканала передачи данных

Функция анализа префиксов InfiniBand-подсетей возвращает номер логического подканала передачи данных для MPI-процесса источника. В дальнейшем номер подканала играет роль индекса, с помощью которого выбирается уже смодифицированная QP MPI-процесса приемника и LID порта приемника InfiniBand-адаптера. Передача сообщения удаленному абоненту в данном случае происходит с задействованием одного конкретно выбранного канала (порта) передачи данных в отличие от стандартной реализации пакета MVAPICH с поддержкой многоканальной передачи данных, где задействуются все каналы передачи данных.

Заключение

Результатом работы являются модифицированная библиотека MPI на базе пакета MVAPICH, ориентированная на применение в вычислительных системах, реализованных с использованием бескоммутаторной технологии, а также практические решения по настройке параметров и конфигурации коммуникационной подсистемы. Разработанное ПО используется на компактных вычислительных комплексах, производимых в РФЯЦ-ВНИИЭФ.

Дальнейшее направление работ предусматривает оптимизацию протоколов межпроцессорного взаимодействия с целью повышения эффективности коммуникационных характеристик вычислительного комплекса.

Список литературы

1. *Shanley T.* InfiniBand Network Architecture. Addisson-Wesley, 2006.
2. *Gropp W., Lusk E., Skjellum A.* Using MPI: Portable Parallel Programming with the Message-Passing Interface. Second edition. MIT Press, 1999.
3. OFED. http://www.openfabrics. org.
4. OpenMPI. http://www.open-mpi.org.
5. MVAPICH. http://mvapich.cse.ohio-state.edu.

Статья поступила в редакцию 09.02.11.