УДК 519.6

УПРАВЛЕНИЕ ЕДИНОЙ КОММУНИКАЦИОННОЙ СРЕДОЙ ИЗ ПРОГРАММНО-НЕЗАВИСИМЫХ КОММУНИКАЦИОННЫХ ПОДСИСТЕМ

В. Г. Басалов
(РФЯЦ-ВНИИЭФ)

Дается описание разработанного пакета программ Multi Cluster Subnet Manager, являющегося менеджером программно-независимых коммуникационных подсистем в составе единой коммуникационной среды. Приводится описание механизмов, реализованных в данном пакете для обнаружения, конфигурирования, активации и поддержания программно-независимых коммуникационных подсистем.

Ключевые слова: многопроцессорная вычислительная система, единая коммуникационная среда, программно-независимая коммуникационная подсистема, логическая кластеризация, менеджер подсети.

Введение

Количество высокопроизводительных вычислительных комплексов, построенных с использованием коммуникационной среды, основанной на архитектуре InfiniBand, непрерывно увеличивается. В связи с этим возникает необходимость объединить отдельные вычислительные мощности в единое целое, сохранив при этом независимость каждой составной части. Такое объединение позволит независимым вычислительным системам (ВС) получить доступ к разного рода общим ресурсам, например единой системе хранения данных.

С другой стороны, построение многопроцессорных вычислительных комплексов с большой (более 500 Тфлопс) производительностью приводит к разрастанию коммуникационной среды. Уже сегодня размеры таких систем достигают нескольких тысяч узлов. Дальнейшее увеличение количества используемых в них устройств (коммутаторов, вычислительных узлов и устройств ввода-вывода) приведет к тому, что существующее коммуникационное программное обеспечение не справится с управлением коммуникационной средой. Избежать подобной ситуации позволяет разбиение единой коммуникационной среды многопроцессорного вычислительного комплекса на программно-независимые коммуникационные подсистемы, названные логическими кластерами. Логическими кластерами управляют независимые менеджеры, которые обеспечивают выполнение задач на всем вычислительном поле комплекса.

Решение указанных проблем заложено в архитектуре InfiniBand. Она позволяет объединять отдельные подсети с помощью IB-to-IB-коммутатора, каждая подсеть содержит один или более коммутаторов и вычислительных модулей (ВМ). Пример такой коммуникационной среды представлен на рис. 1.

Каждая подсеть управляется автономно. На уровне подсети адресация между узлами происходит с помощью локальных идентификаторов LID (Local Identifier). Устройства, находящиеся в различных подсетях, адресуются через IB-to-IB-коммутаторы посредством глобальных идентификаторов GID (Global Identifier).

Основными достоинствами такой схемы являются:

- масштабируемость вычислительных комплексов;
- возможность дублирования LID в разных подсетях;
- ограничение распространения последствий отказов в работе и топологических изменений;

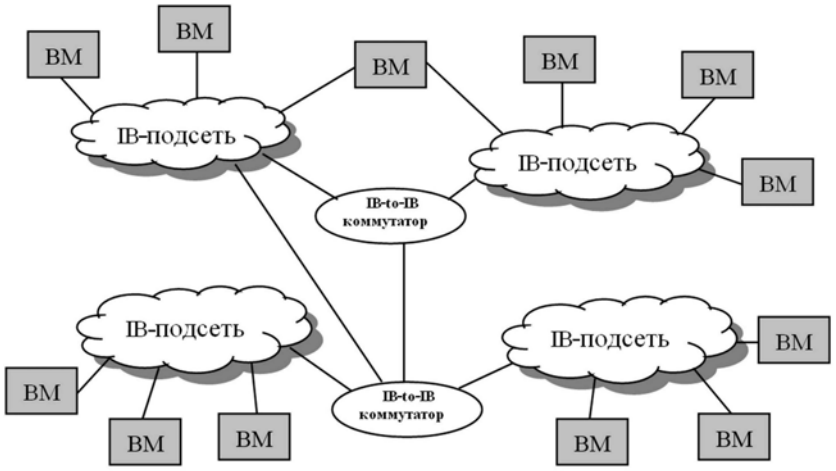

Рис. 1. Компоненты единой коммуникационной среды

– ограничение границ управления коммуникационной средой [1].

К сожалению, приведенный выше механизм в настоящее время не реализован ни в аппаратной, ни в программной части. Начало производства IB-to-IB-коммутаторов было намечено на 2008 год, но в связи с отсутствием спроса их выпуск отложен на неопределенное время. Из материалов Международного симпозиума, прошедшего весной 2008 г. в Сонома (США) [2], следует, что разработка программной поддержки для IB-to-IB-коммутаторов стартует только после начала их производства. В связи с вышесказанным стала актуальной разработка технологии по управлению единой коммуникационной средой из программно-независимых коммуникационных подсистем без использования IB-to-IB-коммутаторов.

В статье описывается созданный автором пакет программ Multi Cluster Subnet Manager (MCSM). MCSM является менеджером подсети независимого логического кластера, входящего в состав единой коммуникационной среды. На рис. 2 представлен вычислительный комплекс, единая коммуникационная среда которого образована слиянием двух ВС с целью получения доступа к общей файловой подсистеме.

Каждый независимый логический кластер управляется запущенным на нем MCSM. Функциями MCSM являются обнаружение, конфигурирование, активация и поддержание коммуникационной подсети независимого логического кластера, а также построение маршрутов передачи данных как между узлами внутри независимого логического кластера, так и между узлами разных независимых логических кластеров единой коммуникационной среды.

Независимая логическая кластеризация

Без использования IB-to-IB-коммутаторов объединение нескольких независимых логических кластеров в единую коммуникационную среду представляет собой обычную IB-подсеть, которая содержит набор конечных узлов — ВМ, соединенных между собой коммуникационной средой, состоящей из коммутаторов и линий связи, только большего размера.

Основной проблемой при создании MCSM стало обнаружение топологии независимых логических кластеров. Поскольку они не ограничены конечными портами (в архитектуре InfiniBand портами IB-to-IB-коммутаторов являются канальные адаптеры, как в вычислительных узлах), то при запуске стандартного менеджера подсети на любом логическом кластере будет обнаружена вся единая коммуникационная среда. Поэтому необходимо ограничить процесс обнаружения топологии коммуникационной среды пределами логического кластера.

Каждый логический кластер единой коммуникационной среды искусственно ограничивается коммутаторами и портами на них, названными *граничными* коммутаторами и портами. Граничными коммутаторами независимого логического кластера являются те коммутаторы, порты которых непосредственно соединены линиями связи с коммутаторами других независимых логических кластеров. На рис. 3 граничные коммутаторы выделены штриховкой. Физические номе-

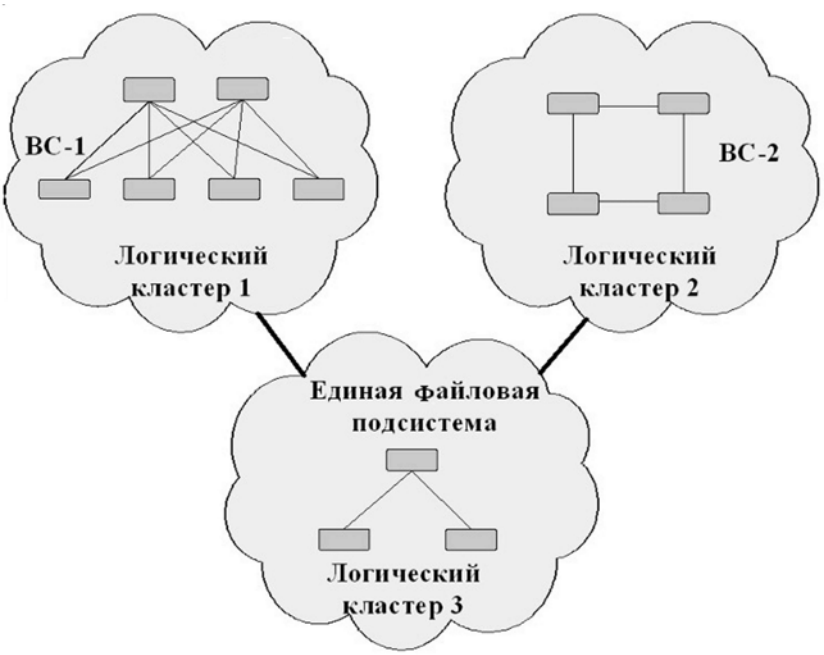

Рис. 2. Топология единой коммуникационной среды

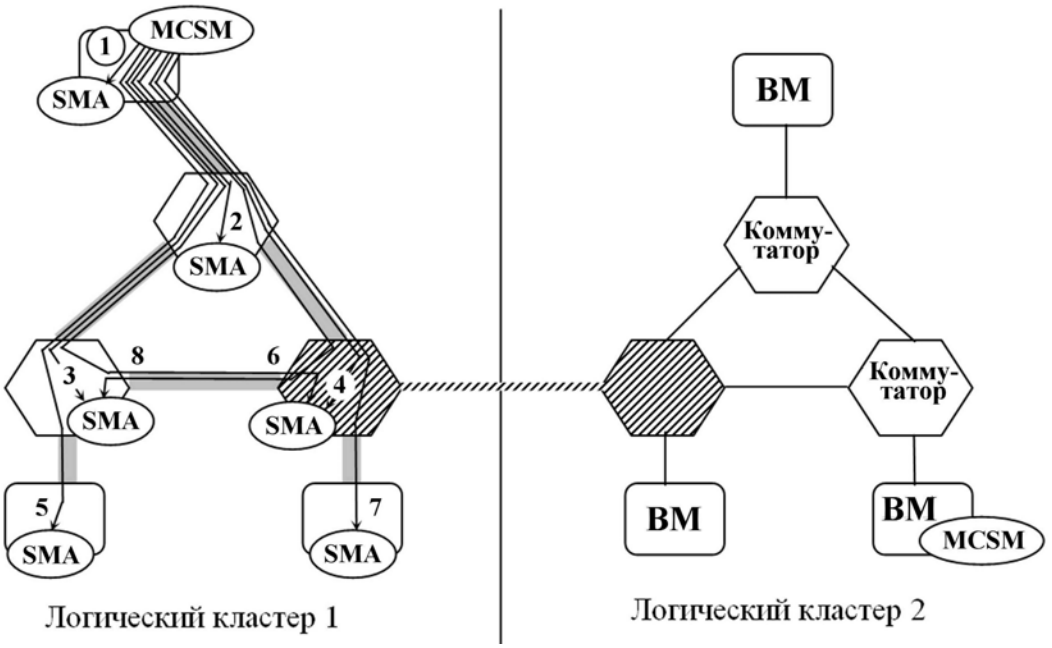

Рис. 3. Пример процесса обнаружения топологии логического кластера

ра GUID (Globally Unique Identifier) граничных коммутаторов и номера граничных портов заносятся в конфигурационный файл, для каждого логического кластера — в свой собственный.

Рассмотрим подробнее пример процесса обнаружения топологии коммуникационной среды, реализованный в MCSM и представленный на рис. 3. MCSM запускается на инструментальном ВМ логического кластера 1. На всех устройствах IB-подсети запускаются агенты менеджера подсети (SMA), в задачу которых входит обработка запросов от MCSM. Каждая стрелка на рис. 3

представляет запрос о состоянии узла, посланный MCSM. Номер, соответствующий стрелке, показывает, в каком порядке эти запросы посылались. Первый запрос посылается в канальный адаптер ВМ, на котором запущен MCSM. После ответа на запрос 2 MCSM посылает новые запросы, имеющие номера 3 и 4, через все активные порты коммутатора. Соответственно, получив ответ на запрос 3, MCSM посылает запросы 5 и 6. В свою очередь, получив ответ на запрос с номером 4, MCSM определяет, что пришел он от граничного коммутатора. Коммутатор имеет три активных порта, через которые можно послать новые запросы, но, так как один из них является граничным, MCSM сформирует только два запроса, помеченные на рисунке номерами 7 и 8. В результате процесс обнаружения топологии коммуникационной среды ограничивается и не распространяется за пределы логического кластера.

Обнаружение изменений топологии

Архитектура InfiniBand определяет два механизма обнаружения изменений топологии: опрос подсети и системное прерывание [3]. Изменение топологии подсети происходит тогда, когда изменяется состояние хотя бы одного порта в коммуникационной среде, например, вследствие отказа оборудования.

В MCSM реализованы оба вышеназванных механизма. Каждый MCSM отслеживает изменения топологии только в своем логическом кластере и ничего "не знает" о топологических изменениях, происходящих в других логических кластерах. Этим достигается ограничение распространения последствий отказов в работе и топологических изменений пределами одного логического кластера, что позволяет не допускать влияния на функционирование остальных частей единой коммуникационной среды.

Назначение локальных идентификаторов узлам единой коммуникационной среды

В процессе обнаружения логического кластера MCSM назначает LID для каждого обнаруженного узла. Адресное пространство LID позволяет задействовать в единой коммуникационной среде до 49 151 портов. Каждому порту в единой коммуникационной среде присваивается уникальный LID. MCSM имеет два механизма назначения LID: назначение случайного LID и назначение предопределенного LID.

В первом случае, зная заранее примерное число портов в каждом логическом кластере, можно разбить все адресное пространство LID на непересекающиеся подмножества путем назначения минимального LID для каждого логического кластера (рис. 4). Назначение локальных идентификаторов для каждого логического кластера начинается с минимального значения, определенного в конфигурационном файле, которое затем увеличивается на единицу для каждого обнаруженного порта. Внутри логического кластера назначение LID портам происходит случайным образом.

Достоинством этого механизма является то, что подготовительная работа проста: заранее для каждого логического кластера необходимо определить только его минимальный LID. Недостаток — невозможность гарантии того, что одному порту будет назначен один и тот же LID при разных процессах обнаружения топологии коммуникационной среды.

Этот недостаток устраняется при использовании механизма предопределенных LID. Он предусматривает большую подготовительную работу. Необходимо создать и поддерживать базу данных обо всех устройствах единой коммуникационной среды. База содержит GUID каждого порта единой коммуникационной среды и заранее присвоенный ему уникальный LID. Обладание этой информацией значительно облегчает поддержку приложений, использующих транспортные протоколы RDMA [1] и IPoIB [4].

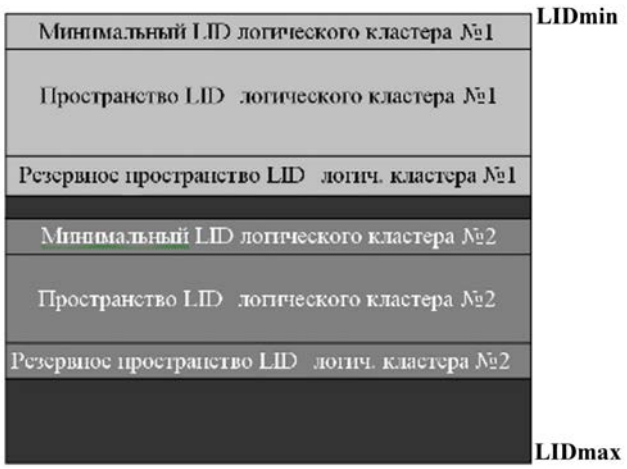

Рис. 4. Адресное пространство LID единой коммуникационной среды

Вычисление маршрутных таблиц

Используя текущую топологическую информацию, полученную в процессе обнаружения сети, MCSM должен создать пути в единой коммуникационной среде, по которым пакеты данных будут достигать пункта назначения. Другими словами, ему необходимо вычислить маршрутные таблицы для каждого коммутатора своего логического кластера. Маршрутная таблица коммутатора представляет собой одномерный массив, номерами элементов которого являются номера LID узлов, а значениями этих элементов — номера портов коммутатора.

Алгоритм вычисления путей в MCSM разделен на два этапа. На первом этапе осуществляется вычисление путей внутри логического кластера, которое выполняется с помощью стандартных алгоритмов маршрутизации, реализованных в OpenSM [4]. В настоящее время в MCSM используются алгоритмы маршрутизации Min Hop, Up/Down и DOR [5]. На каждом логическом кластере может быть запущен тот алгоритм, который лучше соответствует его топологии. Например, если топология логического кластера представляет собой гиперкуб, MCSM на нем следует запускать с алгоритмом DOR.

На втором этапе осуществляется вычисление путей между узлами логических кластеров единой коммуникационной среды. В каждом логическом кластере определяются транзитные коммутаторы и порты, через которые происходит обмен информацией с другими логическими кластерами. Все транзитные коммутаторы логического кластера обязательно будут граничными, но не все граничные коммутаторы будут транзитными. Эта информация определяется заранее и хранится в конфигурационном файле. Используя ее, MCSM дополняет маршрутные таблицы, созданные на первом этапе выполнения алгоритма маршрутизации, портами, обеспечивающими достижение узла с любым LID в единой коммуникационной среде. Если между логическими кластерами существует несколько транзитных коммутаторов и линий связи, они загружаются равномерно.

Некоторые приложения, работающие в единой коммуникационной среде, используют транспортные протоколы RDMA и IPoIB. Одним из таких приложений для мультипроцессорных систем является параллельная файловая система Lustre [6]. В основе этой системы лежит механизм *клиент-сервер*. Построение канала связи между клиентом и сервером происходит следующим образом:

- источник через широковещательную сеть (архитектура InfiniBand поддерживает коллективные обмены) посылает запрос, содержащий GUID получателя;
- узел, GUID которого совпал с присланным, направляет свой ответ в источник;
- источник посылает подтверждение об установлении канала.

Для поддержки коллективных обменов в MCSM разработан механизм создания маршрутных таблиц коллективных обменов в единой коммуникационной среде. В каждом независимом логическом кластере для каждого коммутатора вычисляются маршрутные таблицы коллективных обменов, образующие дерево связей между ВМ. Необходимо объединить локальные деревья логических кластеров в глобальное дерево единой коммуникационной среды. Для этого в каждом логическом кластере заранее выбирается коммутатор, в маршрутную таблицу коллективных обменов которого будет добавлен номер порта, ведущего в смежный логический кластер. GUID коммутатора и номер порта MCSM получает из конфигурационного файла.

Тестирование MCSM

Для проверки функционирования MCSM была собрана экспериментальная ВС, в которую вошли следующие аппаратные и программные компоненты:

- три 24-портовых InfiniBand-коммутатора;
- шесть серверов SuperMicro на базе процессоров AMD Opteron 2 ГГц, RAM — 16 Гбайт, сетевой InfiniBand-адаптер — Host Channel Adapter (HCA) Voltaire 400EX;
- сервер под управлением ОС Windows 2000 (для администрирования);
- операционная система Red Hat Enterprise Linux AS 4 (Nahant Update 1, ядро 2.6.21.6) [7];
- программное обеспечение для коммуникационной среды InfiniBand — OFED-1.2.0 [8];
- параллельная файловая система Lustre-1.6.5.

На рис. 5 представлена схема единой коммуникационной среды экспериментальной ВС, виртуально разделенной на три независимых логических кластера. Каждый логический кластер

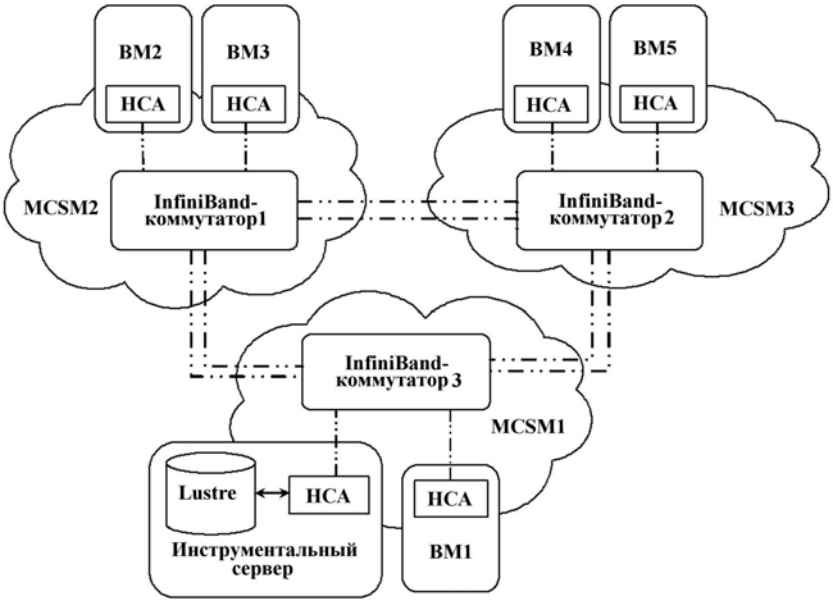

Рис. 5. Логическая кластеризация единой коммуникационной среды экспериментальной ВС (–··–·· — InfiniBand)

управляется своим MCSM с определенными заранее граничными условиями.

Проверка функционирования единой коммуникационной среды экспериментальной ВС производилась как с помощью низкоуровневых тестов, входящих в пакет OFED-1.2.0 и имеющих разные транспортные протоколы (RDMA, IPoIB), так и с помощью MPI-приложений (тест PMB). Тестовые задачи запускались на ВМ, находящихся как в одном логическом кластере, так и в разных логических кластерах, а также на всем вычислительном поле экспериментальной ВС.

При тестировании на каждом логическом кластере был запущен IB-администротор [8], представляющий собой пакет диагностических утилит. В процессе работы администратор через MCSM собирает информацию о состоянии узлов, портов и связей логического кластера и предоставляет ее пользователю.

Заключение

Разработанный пакет программ MCSM представляет собой менеджер коммуникационной среды логического кластера, являющегося частью единой коммуникационной среды вычислительного комплекса. MCSM обеспечивает обнаружение, конфигурирование, активацию и поддержание коммуникационной среды логического кластера, а также маршрутизацию сообщений между ВМ единой коммуникационной среды.

Механизмы и алгоритмы, примененные при разработке MCSM, позволяют:

- объединять различные ВС в глобальную ВС с единой коммуникационной средой;
- виртуально разбивать большие коммуникационные среды на независимо управляемые логические кластеры без использования аппаратных средств;
- ограничить распространение топологических изменений, вызванных отказами в работе аппаратуры, пределами логического кластера. При этом локальные ВС остаются автономными и не влияют на функционирование всей глобальной ВС. Это особенно актуально при профилактическом обслуживании отдельных локальных ВС;
- уменьшить количество подконтрольного оборудования, что упрощает работу MCSM, делая ее более надежной.

Применимость разработанного пакета программ MCSM была подтверждена тестированием его на экспериментальной ВС.

В дальнейшем планируется применять данный пакет для управления коммуникационной средой больших вычислительных комплексов.

Список литературы

1. InfiniBand™ Architecture Release 1.0.a. Volume 1. General Specifications. Jule 19, 2001.
2. *Khapyorsky S.* OpenSM and InfiniBand Mannagement Update // Int. Sonoma Workshop. April 7—9, 2008. http://www. openfabric.org/ archives/spring 2008 sonoma/ wednesdey/ opensm.pdf.
3. *Bermudez A., Casado R., Quiles F. J. et al.* Evaluation of a subnet management mechanism for InfiniBand networks // Int. Conf. on Parallel Proc. (ICPP'03). IEEE, 2003.
4. Multicast. http://www.ietf.org/rtc/rtc4392. txt.
5. OpenSM Release Notes. http://linux.die.net/ man/8/opensm.
6. File System Lustre. http://www.lustre.org.
7. Red Hat Enterprise Linux. http://www. redhat.com.
8. OFED. http://www.openfabrics.org.

Статья поступила в редакцию 03.12.10.