InfiniBand

IT-инфраструктуры: реальность и перспективы

Публикация дает представление об основных особенностях и преимуществах построения современных IT-инфраструктур на базе infiniband-протокола, а также областях их применения и перспективах развития.

Введение

С момента появления infiniband-протокола прошло более 6 лет. Изначально он создавался для поддержки связности инфраструктуры Internet и интерконнекта серверов и, прежде всего, как расширение пропускной способности PCI-шины. Однако по мере эволюции IT-архитектуры – повышения производительности процессоров и пропускной способности основных интерфейсов (Ethernet, FC и с переходом с PCI на PCI-Express) – потребность в использовании infiniband стала возрастать и расширяться в качестве универсального транспорта данных) (рис. 1). Сразу заметим, что в настоящий момент производительности полного 4 Гбит/с двухконтроллерного дискового массива достаточно для обслуживания не менее 10–12 серверов (с усредненным набором бизнес-приложений), поэтому речь не идет о тотальном переходе на infiniband-интерфейс (в условиях еще не поддерживаемой полной функциональной совместимости). Рассматриваются условия расширения его применения от междоменного серверного интерконнекта и HPC-приложений. Последнему обстоятельству способствовал ряд изменений на рынке, в частности, в конце 2005 г. появились первые системы хранения с infiniband-интерфейсом; а к середине 2006 г. значительно расширился парк infiniband-коммутаторов и infiniband-маршрутизаторов с поддержкой практически всех основных сетевых протоколов; infiniband стал поддерживать расстояния до 200 м; анонсированы и уже доступны системы управления всей консолидированной сетевой инфраструктурой (включая infiniband).*

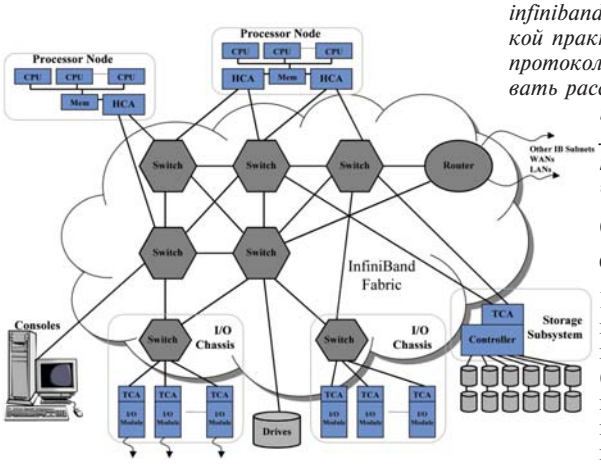

Рис. 1. IT-инфраструктура на основе infiniband-фабрики *(HCA – Host Channel Adapter, TCA – Target Channel Adapter).*

Определение и особенности InfiniBand (IB)

IB – один из немногих протоколов, разработанный для высокоскоростной передачи данных (high bandwidth) с минимальными задержками (low latency) в качестве некой альтернативы высоколатентной архитектуре Ethernet и уже с самого начала применяемый для организации высокопроизводительных кластеров для межузлового соединения.

IB разрабатывался как протокол для передачи данных между вычислительными системами. В задачи таких протоколов входят управление целостностью данных, защита данных от сбоев, использование функционала операционных систем. Одним из главных требований к таким протоколам является низкая задержка при передаче блоков данных. Это требование пришло также из HPC-решений (high-performance computing/clustering), т.к. большие задержки на Ethernet серьезно ограничивали производительность вычислительных комплексов.

По назначению и распространенности можно выделить 3 основных типа протоколов (или 3 способа организации обмена данными), используемых при построении IT-инфраструктуры:

- *блок-ориентированный протокол* – в основном используется для организации взаимодействия серверов на основе блоков данных с дисковыми системами хранения;
- *сетевой протокол* – предназначен для взаимодействия между различными системами (чаще всего это IP-ориентированные протоколы);
- *протокол передачи информации между процессами (IPC – Inter Process Com-*

*) "IBM e-server BladeCenter and Topspin InfiniBand Switch Technology", Redpaper, апрель 2005.

munication), выполняемыми в рамках одной или нескольких систем.

Примеры протоколов, их классификация и основные характеристики представлены в табл. 1.

Табл. 1. Основные характеристики протоколов

	Block	Network	IPC
Задержка	несколько миллисекунд	сотни миллисекунд	несколько микросекунд
Размер блока	очень большой	от мелкого до большого	от мелкого до большого
Использование	хранение данных	универсальный протокол	вычислительные кластеры
Протоколы	FC	Ethernet, TCP/IP	InfiniBand

IPC-протоколы, в отличие от двух других, используют свои (отличные) механизмы для организации передачи данных. В частности, IB имеет:

- возможность описания всех аспектов ввода/вывода;
- возможность использования разделяемой памяти вместо разделяемых шин;
- варьируемую полосу пропускания – от 2,5 (1X) до 60 (12X) Gbps;
- полную прозрачность для операционных систем;
- возможность организации виртуальных потоков.

Именно последние два пункта позволяют серьезно расширить применение IB и вывести его за рамки построения высокоскоростных сетей для HPC. Разработанные технологии и протоколы позволяют осуществлять одновременную передачу пакетов FC, Ethernet, SCSI поверх существующей IB-инфраструктуры (рис. 2). Виртуализация канала передачи и консолидация протоколов обеспечивают гибкость в построении решений и экономию средств. Рассмотрим более подробно основные особенности IB-интерфейса.

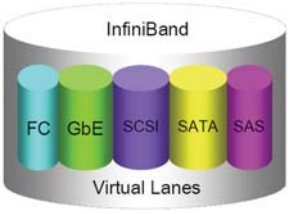

Рис. 2. IB-протокол за счет организации виртуальных потоков позволяет осуществлять одновременную передачу пакетов FC, Ethernet, SCSI поверх существующей инфраструктуры.

Инкапсуляция протоколов в IB

Понимание того, какая совместимость и на каком уровне обеспечивается при инкапсуляции других протоколов в IB, формируется из логики и реализации ее механизмов. InfiniBand поддерживает множество протоколов верхнего уровня, которые дают возможность эксплуатировать InfiniBand различными типами программного обеспечения с различными требованиями и целями.

Один из наиболее востребованных – это *IPoIB-протокол* (Internet Protocol over InfiniBand, или "IP-протокол поверх InfiniBand"). IPoIB поддерживает большой спектр приложений промышленного стандарта, микропрограммных средств и операционных систем. Любой протокол связи, который "лежит" поверх IP (типа TCP/IP или UDP/IP), может транспортироваться по InfiniBand через IPoIB-интерфейс.

Поддержка на InfiniBand *SDP-протокола* (Sockets Direct Protocol), обеспечивающего обмен данными несколько выше стека протокола, дает возможность приложениям, написанным для TCP-sockets-интерфейса, функционировать без осуществления полной поддержки всех TCP/IP уровней. SDP-протокол обеспечивает асинхронный-sockets-интерфейс, через который приложения и микропрограммные средства могут связаться, без необходимости понимания нижних уровней TCP- или IP-протокола.

Различие между синхронными и асинхронными sockets небольшая, но это критически важно для SDP. Запросы от приложений, написанных для синхронных sockets, требуют возвращения запроса (получение "конца" об его успешном завершении). После чего любые используемые буфера данных могут стираться и использоваться повторно. Для асинхронных sockets все запросы ставятся в очередь, а буфера освобождаются по мере их завершения.

Использование асинхронных sockets является стандартным для приложений под ОС Windows, что требует отсутствия каких-либо модификаций "поверх" SDP-протокола. Исторически, UNIX- и Linux-приложения использовали синхронные sockets и поэтому "прозрачно" не отображались в SDP-протокол. Для устранения этих несоответствий поддерживается процесс конверсии sockets синхронный-к-асинхронному, который может быть обеспечен между sockets интерфейса к приложению (синхронный) и к SDP-интерфейсу (асинхронный). Это преобразование может включать задержку "конца" от запроса, копирование буферов данных или других методов для достижения полной унаследованной совместимости.

Промышленным стандартом для инкапсуляции storage-протоколов в InfiniBand является SRP-протокол (Storage RDMA Protocol, RDMA – remote direct memory access). SCSI-команды, генерируемые или приложением, или файловой системой (это может быть часть операционной системы) непосредственно работают на SRP-интерфейсе, который обеспечивает полную унаследованную SCSI-поддержку. SCSI-протокол и структуры данных, используемые в SCSI, транспортируются без модификации по SRP поверх InfiniBand, точно так же, как SCSI транспортируется по Fibre Channel (FCP-протокол) и по TCP/IP (iSCSI-протокол).

Области, в которых FC, iSCSI и SRP отличаются, связаны с сетевыми функциями типа открытие устройства (device discovery); обозначение (enumeration) сетевых устройств; поддержка многодоступности. Есть 2 технических решения, чтобы отобразить эти сетевые аспекты на "вершине" SRP: первый – эмуляция Fibre Channel с его сетью и обозначениями; второй – эмуляция iSCSI и его сети и обозначений.

И, наконец, еще один из "массовых" протоколов, эмулируемых в InfiniBand, это MPI-протокол (Message Passing Interface), используемый в приложениях для технических и научных расчетов (HPC-решения).

Необходимо помнить, что наиболее полно совместимость InfiniBand осуществляется на уровне приложений. Вопросы сетевого дизайна, учета существующей функциональности (которая в ряде случаев может значительно отличаться у вендоров) и совместимости на сетевом уровне полностью ложатся на разработчика проекта. При этом, чем проще существующая инфраструктура, тем проще интегрировать в нее InfiniBand. По мере ее усложнения растет и сложность интеграции IB в существующую инфраструктуру.

Варьируемая пропускная способность IB

Одной из привлекательных сторон InfiniBand является его варьируемая пропускная способность (4/16/48 Gbps – для данных), реализуемая на основе медного или оптоволоконного кабеля (табл. 2).

Табл. 2. Три уровня пропускной способности IB

InfiniBand Link	Signal Pairs	Signaling Rate	Data Rate	Full-Duplex Data Rate
1X	2	2.5 Gbps	2.0 Gbps	4.0 Gbps
4X	8	10 Gbps (4*2.5 Gbps)	8 Gbps (4*2 Gbps)	16 Gbps
12X	24	30 Gbps (12*2.5 Gbps)	24 Gbps (12*2 Gbps)	48 Gbps

С появлением в прошлом году оптоволокна для IB также резко возросли поддерживаемые расстояния на его основе (табл. 3). В настоящее время проводятся исследования его использования на расстоянии нескольких сотен километров.

Табл. 3. Поддерживаемые расстояния на основе IB

Cable Type	Link Rate	Distance*)
CX-4 Copper	1X	0-20 m
	4X	0-15 m
	12X	0-10 m
Optical Fiber:		
62.5 micron multimode	4X	2-75 m
50 micron @ 500 MHz/Km	4X	2-125 m
50 micron @ 2000 MHz/Km	4X	2-200 m

*) хотя в спецификации на кабель могут быть указаны и большие длины, поставщик оборудования их может ограничивать, например, из-за деградации функции восстановления от ошибок (bit-error ratio – BER).

Реализации систем хранения на основе IB

В настоящее время у ряда производителей появились системы хранения, в которых в качестве внешнего интерфейса используется IB. В качестве примера можно привести систему производства LSI Logic – Engenio 6498 storage system, которая использует IB-интерфейс, а также FC- и SATA-диски. Данная система доступна через OEM-партнеров LSI Logic.

Другим производителем, представившим IB системы хранения данных, является компания DataDirectNetwork, продукция которой ориентирована на рынок высокопроизводительных вычислений. В настоящий момент этот производитель предлагает две системы, которые могут поставляться как с интерфейсом FC, так и с IB. Это модели S2A9500 и S2A9550.

Использование параллельных и разделяемых файловых систем, традиционно применявшихся в HPC для увеличения скорости работы, позволяет получить дополнительный уровень виртуализации с IB-решениями. Файловая система ADIC StorNext FX обеспечивает одновременный доступ к данным с различных аппаратных платформ и операционных систем. В настоящий момент поддерживают-

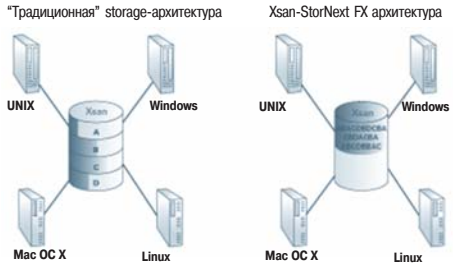

Рис. 3. При организации хранилища на базе Xsan-StorNext FX архитектуры управление томами для разных платформ осуществляется из общего пула.

ся клиенты для AIX, HP-UX, IRIX, Red Hat Linux, SuSE Linux, Mac OS X, Solaris, UNICOS/mp и Windows (рис. 3).

Особенности реализации инфраструктур на основе IB

Привлекательными сторонами IB в сравнении с базовыми характеристиками других протоколов (iSCSI, GbE, 10GbE, FC) являются его основные параметры и низкая цена оборудования. IB является открытым стандартом, что обеспечивает конкуренцию на рынке между производителями оборудования и большое количество разработчиков (в том числе независимых) продуктов, использующих IB.

Производительность одного порта IB 4X (наиболее распространенного в настоящий момент) можно сравнить с двумя портами 4Gb FC. Учитывая меньшую стоимость адаптеров и вдвое меньшую потребность в портах фабрики, получаем экономию более чем в 2 раза. Но, конечно, не следует забывать о том, что при этом не обеспечивается полная совместимость с существующей инфраструктурной функциональностью, и в каждом конкретном случае необходимо считать все преимущества и возможные потери. Кроме традиционных адаптеров PCI-X, которые ограничивают полосу пропускания одним гигабайтом, в настоящее время производители оборудования IB предлагают адаптеры PCI-Express DDR, позволяющие достичь скорости передачи данных 20Gb/s.

Помимо решения вопросов, связанных с высокоскоростным подключением серверов и систем хранения данных, использование адаптеров IB позволяет значительно упростить резервирование и повышение надежности. Для построения высоконадежного решения требуется, чтобы в системе отсутствовали единые точки отказа (SPOF), т.е. все компоненты должны быть дублированы. При использовании традиционной схемы необходимо иметь избыточные сетевые адаптеры и активное сетевое оборудование, адаптеры подключения к SAN и оборудование организации SAN. При использовании IB-архитектуры минимум в 2 раза сокращается количество используемого оборудования и подключений, благодаря высокой производительности и возможности организации виртуальных потоков (GbE, FC) внутри одного канала IB.

Благодаря появлению в последнее время новых продуктов, поддерживающих IB, появилась возможность использовать IB-технологии для построения решений, традиционно базирующихся на технологии FC, т.е. организации сетей хранения данных. Обобщенно, информационная архитектура с использованием систем хранения данных строится по следующей схеме. Серверы через сеть хранения данных (SAN) подключаются к системам хранения данных, включающие в себя дисковые системы и устройства для резервного копирования и долговременного хранения данных (ленточные и оптические накопители и библиотеки).

В настоящее время большинство SAN построены с использованием оборудования, основанного на FC-протоколе. Благодаря появлению IB-FC gateway модулей, IB систем хранения, уже сейчас можно предложить внедрение новых систем на базе IB, которые позволяют не только увеличить полосу пропускания в 5 раз, но и обеспечить совместимость с используемым FC-оборудованием.

В дополнение к высокоскоростной сети хранения данных, в этом случае получаем высокоскоростные каналы для передачи информации между серверами, например, с помощью организации IP-сети. Контроль за всей инфраструктурой осуществляется с помощью единой системы управления, что позволяет снизить затраты на администрирование.

Компания Cisco предлагает спектр продуктов для построения решений, основанных на использовании единого интерконнекта на технологии IB. В частности, коммутаторы Cisco SFS 3012 помимо 24 10-Gbps InfiniBand портов имеют возможность подключать до 12 плат расширения для подключения Cisco InfiniBand-to-Ethernet или InfiniBand-to-Fibre Channel gateway модулей.

Помимо экономии аппаратных ресурсов, возникает экономия человеческих ресурсов, которые тратятся на управление инфраструктуры передачи данных. Компания Cisco представляет расширение своих систем управления, таких, как CiscoWorks LMS, Resource Manager Essentials и Dynamic Fault Manager для управления IB-коммутаторами, в частности, серией Cisco SFS 7000. Таким образом, появляется возможность управления из единого центра как сетевой инфраструктурой, так и инфраструктурой высокоскоростной передачи данных.

Пример построения IB-инфраструктуры

Построение системы ввода/вывода на базе IB позволяет создать высокопроизводительную систему с минимальными затратами на оборудование. Типичным примером является построение системы высокой доступности для баз данных и серверов приложений.

Как видно из схемы (рис. 4), для удаления единой точки отказа необходимо установить в серверное оборудование минимум 4 адаптера (2 FC и 2 GbE). При этом требуются различные системы для контроля

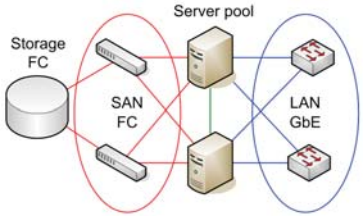

Рис. 4. "Типовая" IT-инфраструктура, как правило, имеет две сети – LAN и SAN.

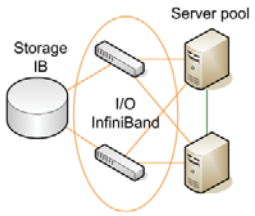

Рис. 5. Пример агрегирования сетевой инфраструктуры на базе IB.

и управления сетями LAN и SAN. При использовании IB-технологии "картина" упрощается до одной сети (рис. 5).

Упрощение инфраструктуры, необходимой для обеспечения высокой доступности, ведет к упрощению администрирования, повышению надежности и увеличению утилизации ресурсов. Высокая пропускная способность IB позволяет использовать меньшее количество адаптеров и портов для получения требуемой производительности и обеспечивает высокую масштабируемость и гибкость решения.

Высокая степень интеграции и возможность подключения оборудования FC в фабрику IB позволяет говорить об эффективности внедрения IB решений в существующих ЦОД (рис. 6). Использование таких коммутаторов, как Cisco SFS 3000 серии, позволяет объединить

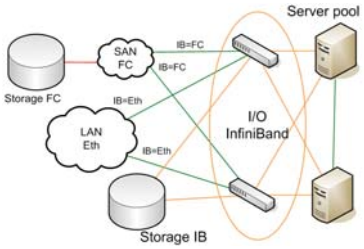

Рис. 6. Пример внедрения IB-решений в существующую ЦОД.

существующие системы FC с новым оборудованием IB. Технологии дают возможность объединить до 4-х каналов FC в один канал IB. Единая система управления CiscoWorks дополнительно снижает затраты на управление решением.

Появление продуктов и наличие открытых стандартов позволяет строить решения по хранению данных на технологии InfiniBand. Тем не менее, конечному пользователю, не имеющему специально подготовленный персонал, внедрить такое решение не под силу. В данном случае следует прибегать к помощи системных интеграторов, имеющих опыт и знания в области внедрения подобных решений. Высокая квалификация сотрудников ЛАНИТ позволяет гарантировать выполнение подобных проектов на самом высоком уровне.

Заключение

В настоящее время IB-инфраструктуры, в основном, предназначаются для HPC-решений. Однако развитие семейства IB-продуктов, а также растущие требования к скорости передачи данных и эффективности IT-инфраструктур могут послужить импульсом к использованию IB-решений в центрах обработки данных и для решения бизнес-задач.

Юрий Барабанщиков,
Департамент сетевой интеграции группы компаний ЛАНИТ