УДК 004.3

ТЕХНОЛОГИЯ INFINIBAND И ЕЕ ПРИМЕНЕНИЕ В ВЫСОКОПРОИЗВОДИТЕЛЬНЫХ ВЫЧИСЛИТЕЛЬНЫХ СИСТЕМАХ

Елисеева Т.В., Никифоров В.С.

Научный руководитель – доцент Середкин В.Г.

Сибирский федеральный университет

Определение и история

InfiniBand – высокопроизводительная коммутируемая архитектура, предназначенная для соединения серверов, систем связи, запоминающих устройств. Применяется как для внутренних, так и внешних системных соединений.

InfiniBand (IB) появилась в 1999г. путем слияния двух конкурирующих проектов: Next Generation I/O (ngio) (разработчики - Intel, Microsoft, Sun) и Future I/O, разработанным Compaq, IBM, и Hewlett-Packard. Первоначально она предполагалась как комплексная "системная сеть", которая соединит процессорные узлы (CPU) и обеспечит высокоскоростную передачу информации. В теории, это должно было сделать конструкцию кластеров намного более простой и потенциально менее дорогой, потому что все больше устройств могли иметь совместный доступ, можно было бы легко менять их взаимную конфигурацию и рабочую нагрузку. Со временем InfiniBand превратился в широко распространенное соединение для коммерческих центров обработки данных с низкими издержками, низкой задержкой, высокой пропускной способностью.

Благодаря своим техническим и технологическим преимуществам, таким как гибкий транспортный механизм или поддержка различных физических линий, архитектура InfiniBand (IBA) завоевала популярность в среде высокопроизводительных вычислительных систем, о чем может свидетельствовать статистика авторитетной международной организации TOP500.

Почти в трети представленных суперкомпьютеров используется технология InfiniBand.

Технические характеристики

В InfiniBand используется двунаправленная последовательная шина (SDR) с пропускной способностью 2,5 Гбит\сек. Т.к. в SDR передача информации осуществляется с помощью шифровки 10\8 (как в DDR и QDR), получается 2 Гбит полезной информации. В FDR и EDR уже применяется 66\64.

	SDR	DDR	QDR	FDR	EDR
X	2 Гбит\сек	4 Гбит\сек	8 Гбит\сек	14 Гбит\сек	25 Гбит\сек
4X	8 Гбит\сек	16 Гбит\сек	32 Гбит\сек	56 Гбит\сек	100 Гбит\сек
12X	24 Гбит\сек	48 Гбит\сек	96 Гбит\сек	168 Гбит\сек	300 Гбит\сек

Таблица 1. Теоретически рассчитанные показатели производительности, не включающие дополнительные служебные требования.

На данный момент максимальная пропускная способность составляет 300 Гбит\сек. К 2014г. планируется увеличить производительность до 1000 Гбит\сек.

Показатель латентности достаточно низок: колеблется от 100 наносекунд (микросхемы переключателя QDR) до 2,6 микросекунд (Mellanox InfiniHost DDR III HCAs). В целом же задержка не превышает 1 микросекунды.

Ограничения расстояния передачи информации находятся в диапазоне от 10 метров по медному кабелю при пассивном DDR-соединении до нескольких сотен метров по оптиковолоконному кабелю.

Архитектура

Архитектура InfiniBand состоит из процессорных узлов и комплексов устройств ввода\вывода, соединенные через связную архитектуру, образованную каскадом связанных между собой роутеров и коммутаторов.

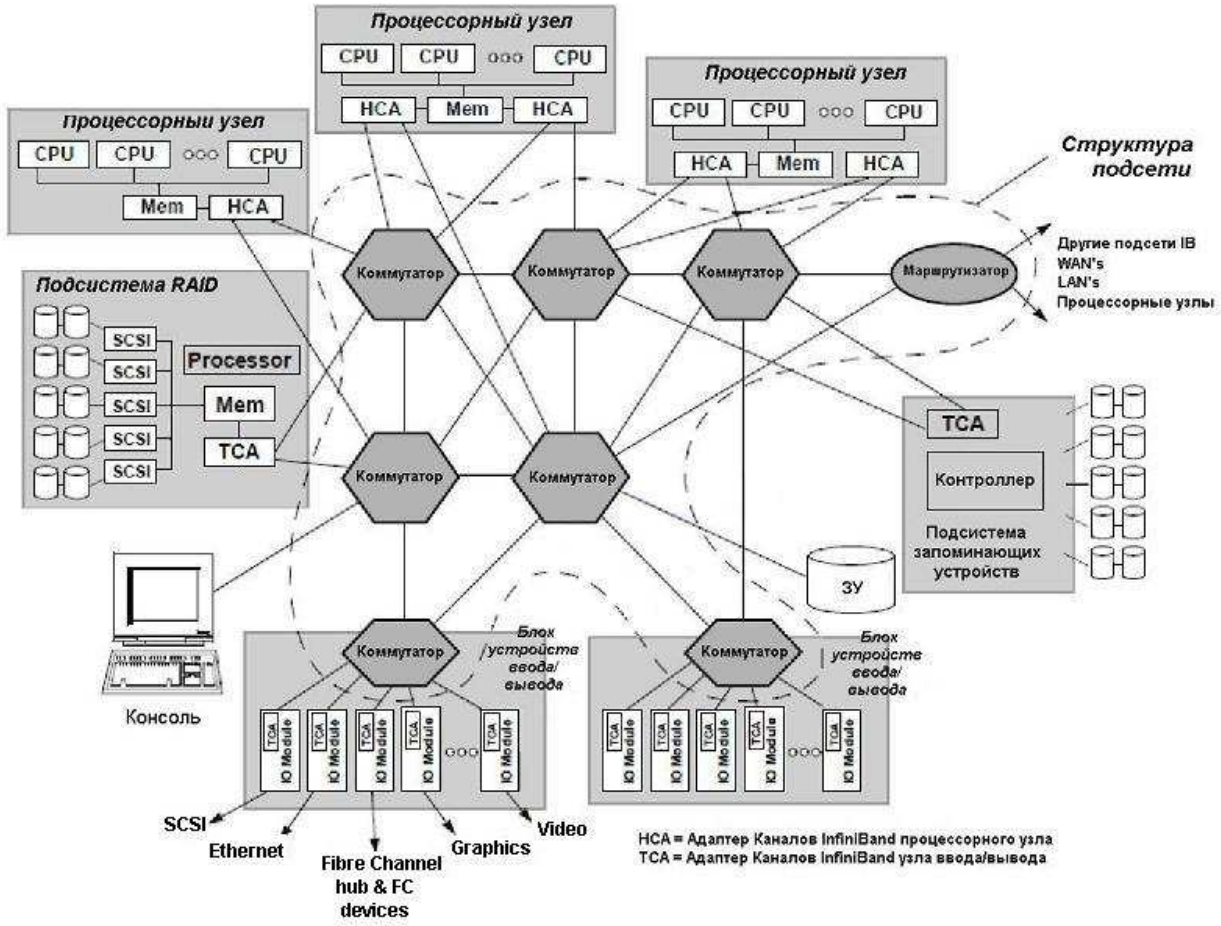

Рисунок 1. Архитектура InfiniBand

Комплексы устройств ввода могут состоять как из одиночных специализированных интегральных микросхем (ASIC) присоединяемых устройств, так и из SCSI или LAN адаптеры подсистем защиты RAID, являющиеся не менее сложными, чем процессорные узлы.

Архитектура IB может быть описана серией слоев, причем протокол каждого независим от других слоев. Каждый слой зависит от сервиса нижнего уровня и предоставляет услугу верхнему уровню.

Физический слой определяет порядок помещения битов в провод, чтобы сформировать символы и определяет символы, используемые для того, чтобы структурировать (от запуска пакета до конца пакета) символы данных, и создает заливку между пакетами (бездействие).

Канальный слой описывает пакетный формат и протоколы для пакетной работы,

например, управление потоком и как пакеты направлены в пределах подсети между источником и местом назначения.

Сетевой слой описывает протокол для направления пакетов между подсетями.

В транспортном слое описываются сетевые протоколы и протоколы канального уровня, поставляющие пакет требуемому месту назначения.

IBA поддерживает любое число протоколов верхнего уровня для различных потребительских нужд. IBA также определяет сообщения и протоколы для определенных функций управления. Эти протоколы управления разделяются на управление подсетью и службы подсети, имеющие свои уникальные свойства.

Архитектура IB не имеет топологических ограничений, причем реализация простой топологии довольно проста. Модульные коммутаторы базируются на архитектуре дерева FAT.

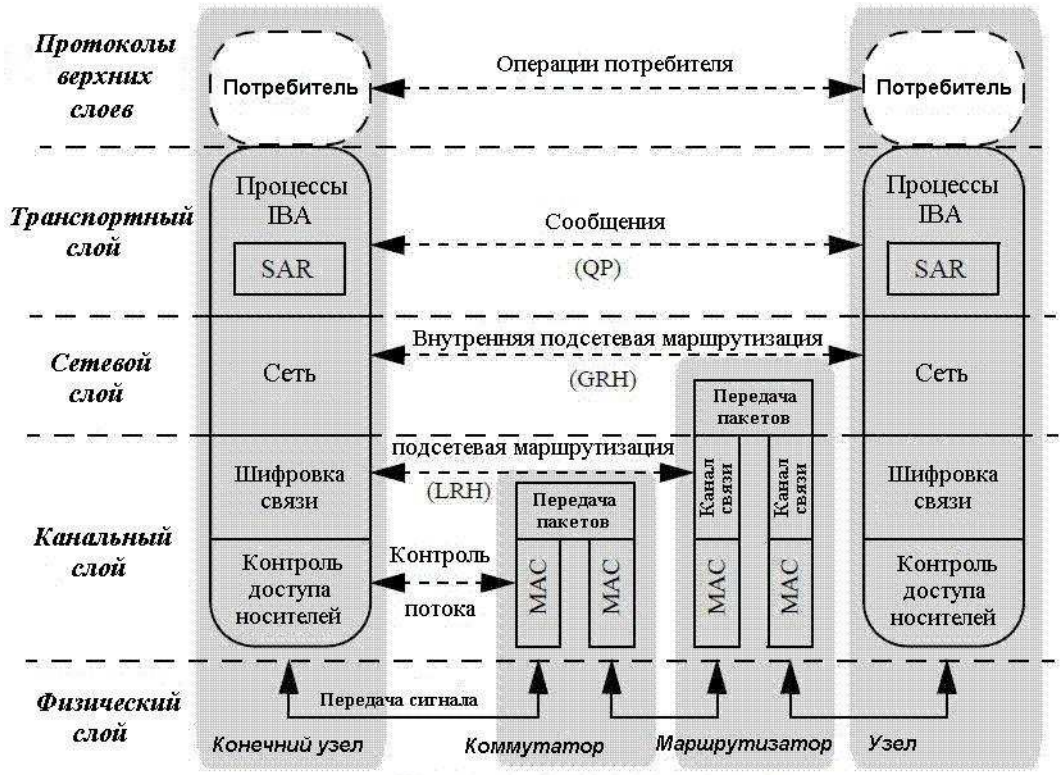

Рисунок 2. Уровни архитектуры InfiniBand

Протоколы, API и взаимодействие с операционными системами.

Кластеры, использующие традиционные приложения, такие как Oracle RAC, полагаются на собственные соединения или TCP/IP, чтобы управлять сложным характером межкластерного трафика. Каждый дополнительный узел (node) в кластере влечет за собой все больше издержек и трафика, ограничивая совокупный потенциал производительности кластера. IP-over-InfiniBand (IPoIB — группа протоколов, описывающих передачу IP-пакетов поверх Infiniband)) устраняет эти недостатки.

Межпроцессорное соединение IPoIB также позволяет кластеру работать над единственным приложением, подобно большой системе SMP, без известных проблем масштабируемости, имеющих место на системах SMP. Протокол SDP (Socket Direct Protocol) позволяет устанавливать вирутальные соединения и обмениваться данными между сокетами поверх IB, используя IP-адреса (для их разрешения может включаться IPoIB). Однако не использует TCP в качестве стека операционной системы (ОС). Группа протоколов RDMA используется для передачи данных на соответствующий сетевой контроллер, минуя ОС и CPU, что позволяет значительно выйграть в производительности. RDMA (remote direct memory access) использует для передачи информации между SCSI-устройствами протокол SRP. Также IB применяет протокол DDR.

Чтобы получить прямой доступ в удаленную память или процессорный узел, не описывая конкретный тип оборудования, используется библиотека API UDAPL (User Direct Access Programming Library).

По сути, в IB нет никакого стандартного API программирования в пределах спецификации. Стандарт только перечисляет ряд операций и функции, которые должны существовать. Синтаксисом этих функций занимаются поставщики. Фактический стандарт до настоящего времени был синтаксисом, разработанным Союзом OpenFabrics, который был принят большинством поставщиков InfiniBand, и для Linux и для Windows. Стек программного обеспечения Infiniband, разработанный Союзом OpenFabrics, выпущен как "Распределение Предприятия OpenFabrics (OFED)", при выборе лицензии GPL2 или BSD лицензируют для Linux, и как "WinOF" при выборе лицензии BSD для Windows.

Преимущества технологии InfiniBand

- Высокая пропускная способность (превышает показатели основных аналогов, таких как FibraChannel или Ethernet).
- Может поддерживать простые организации, например одиночную компьютерную систему.
- Может быть расширена, путем включения следующих моментов: репликация компонентов для повышения надежности системы, расположение каскадом компонентов коммутационной матрицы, дополнительные модули ввода\вывода для повышения производительности и масштабирования узла ввода-вывода, дополнительные хост-узлы вычислительных элементов для масштабируемых элементов, а также возможность различной комбинации всего вышеперечисленного.
- Дает возможность вычислительной системе поддерживать уровень при возрастающих требованиях к пропускной способности, масштабируемости, разгрузке ЦП, доступности и поддержки интернет-технологий.
- IBA сосредотачивается на движущихся данных в и из памяти узла и оптимизирован для отдельного управления интерфейсами памяти. Это разрешает аппаратным средствам быть близко связанными или даже интегрированными с комплексом памяти узла, удаляя любой барьер производительности.
- Будучи разработанным как сеть первого порядка, IBA достаточно гибок, чтобы быть реализованным как сеть второго порядка, реализуя разрешениея наследства и

миграции. Даже в этом случае, работа оптимизации памяти IBA разрешает максимально доступное использование пропускной способности и увеличение эффективности ЦП.

Вышеперечисленные преимущества технологии InfiniBand в том числе и в кластерной системе ИКИТ СФУ.