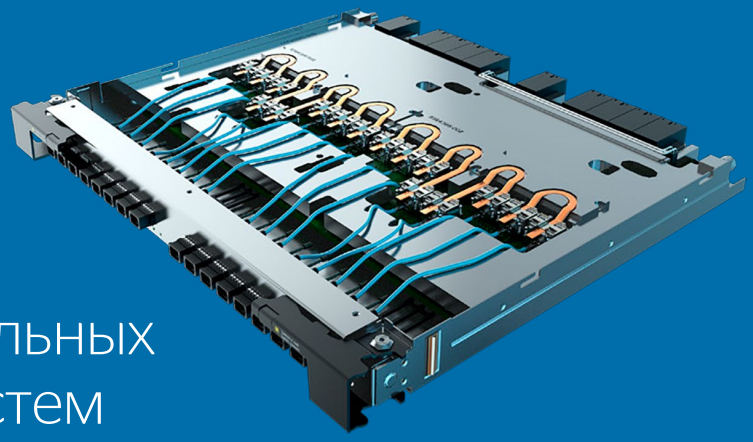
Технология BXI для суперкомпьютеров Bull Sequana X1000

Высокоскоростные коммутационные технологии для экстремальных высокопроизводительных вычислительных систем



Развитие эксафлопсных вычислений приводит к повышению производительности, увеличению количества поддерживаемых узлов и вычислительных ядер и росту объема данных. При таком уровне масштабирования вычислительных технологий оптимизация сетей, составляющей основу любой системы, является наивысшим приоритетом для поддержания высокой скорости работы. Коммутационные технологии с большой долей вероятности станут ключевым решением для создания эксафлопсных систем. Именно поэтому одним из основных элементов программы развития эксафлопсных вычислений Bull стала разработка собственной коммутационной технологии следующего поколения. Технология Bull eXascale Interconnect (BXI) обеспечивает принципиально новые показатели производительности, масштабируемости, эффективности, надежности и качества сервиса для экстремальных рабочих нагрузок.

Технология BXI отличается исключительной масштабируемостью (до 64 000 узлов для первой версии технологии). Основные отличительные особенности:

- Высокоскоростные каналы (100 Гбит/с/с)
- Высокая скорость передачи сообщений (>100 млн сообщений/с)
- Минимальные требования к памяти и использование компонентов с низким значением задержки.

Сокращение потери пропускной способности

Основной отличительной особенностью технологии BXI является система управления обменом данными с полным аппаратным шифрованием, что позволяет полностью выделить ресурсы центральных процессоров для компьютерных вычислений, а процессами обмена данных управляет технология BXI. В результате, в отличие от других широко используемых сетей, технология BXI обеспечивает высокую пропускную способность операций обмена данными даже тогда, когда система находится под большой вычислительной нагрузкой. Аппаратные компоненты BXI соотносятся напрямую с коммуникационными библиотеками, включая MPI (Message Passing Interface) и PGAS (Partitioned Global Address Space). Благодаря этому аппаратному ускорению технология BXI обеспечивает высочайший уровень производительности при обмене данными для высокопроизводительных компьютерных приложений. Кроме того, она отличается высокой пропускной способностью, низкой задержкой и высокой скоростью передачи сообщений. Архитектура технологии BXI основана на базе коммуникационной библиотеки Portals 4. Это позволяет выполнить полную оптимизацию для всех типов коммуникаций MPI, включая последние расширения MPI-2 и MPI-3 и PGAS. Несообщающийся протокол Portals 4 гарантирует минимальные постоянные требования к памяти, независимо от размеров системы.

Качество сервиса

Высочайшее качество сервиса (QoS) технологии BXI позволяет определить несколько виртуальных сетей и гарантировать, например, что объемные сообщения подсистемы ввода/вывода не будут ограничивать скорость передачи небольших сообщений данных. Кроме того, функции адаптивной маршрутизации технологии BXI позволяет в динамичном режиме предотвращать возникновение аппаратных ограничений пропускной способности.

Надежность и отказоустойчивость

Для обеспечения высокой надежности работы в технологии BXI используется как сквозная ошибка проверок и перенаправление данных, так и на уровне отдельных соединений. Кроме того, во всех микросхемах ASIC используются схемы кода коррекции ошибок (ECC) для обнаружения и устранения ошибок. Эти механизмы гарантируют непрерывность работы в случае неустойчивого или устойчивого отказа (на уровне соединения или коммутатора).

Компоненты технологии BXI

Технология BXI основана на двух типах микросхем ASICs (сетевой контроллер Network Interface Controller (NIC) и коммутатор) и поставляется полным набором программного обеспечения. Коммутаторы BXI управляются с помощью распределенного и внеполосного пакета управления, что позволят выполнять масштабирование до 64 000 узлов. Внеполосное управление исключает возникновение помех для трафика управления и трафика приложений. Более подробное описание компонентов технологии BXI приводится на следующей странице.

Использование технологии BXI в суперкомпьютерах Bull Sequana

Bull Sequana – это новая платформа для высокопроизводительных компьютерных вычислений с высокой эффективность энергопотребления благодаря технологии Direct Liquid Cooling (DLC). Bull Sequana включает 3 стойки с 288 (макс.) вычислительными узлами, подключенными между собой с помощью двухуровневой технологии BXI с топологией Fat-Tree. Предлагаются 2 версии: недорогая система с 1 контроллером NIC на 1 узел и версия с 2 контроллерами NIC на 1 узел для повышения скорости работы. Для создания крупномасштабных высокопроизводительных вычислительных систем элементы Bull Sequana подключены между собой с помощью внешних коммутаторов BXI.

Bull
atos technologies

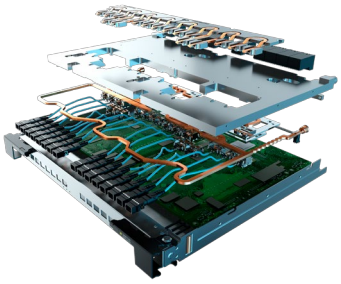
Сетевой контроллер Network Interface Controller (NIC) технологии BXI

Контроллер BXI NIC доступен в виде мезонинной платы для модулей Bull Sequana и в виде стандартных карт PCIe для других узлов. Плата BXI NIC соединяет каждый узел с межсоединением BXI.

- PCIe, 3 поколение, 16 каналов
  - Специализированный разъем NIC 4x BXI для передачи данных в одну сторону со скоростью 12,5 ГБ/с
- Реализация в примитивах взаимодействия Portals 4
  - Снижение нагрузки при обмене данными и вычислениях за счет переноса рабочих задач на NIC
  - Двухсторонняя отправка сообщений MPI
  - Односторонняя отправка сообщений PGAS/MPI
- Обход ОС и приложений
  - Прием контролируется сетевым контроллером NIC без прерываний и участия ОС
  - Ответ на операцию «положить» или «получить» не требует активности приложения
- Коллективные операции разгружают аппаратный ускоритель
- Сквозной механизм восстановления надежности для неустойчивых и устойчивых отказов
- Резервирование виртуальных каналов: разделение различных типов сообщений во избежание зависаний и для оптимизации использования ресурсов сети (балансировка нагрузки и обеспечение высокого качества работы)
- Счетчики производительности и ошибок для анализа работы приложений

Мезонинная плата BXI для блейд-модуля Bull Sequana

Коммутатор технологии BXI



Коммутатор с прямым жидкостным охлаждением BXI L1 для суперкомпьютера Bull Sequana

Коммутатор BXI предлагается в виде коммутационных модулей L1/L2 внутри Bull Sequana с медными соединителями и в виде внешнего коммутатора с 48 портами.

Коммутационные модули Bull Sequana охлаждаются с помощью усовершенствованной версии запатентованного Bull решения Direct Liquid Cooling (DLC) и проверенной временем технологии охлаждения, которая максимально сокращает объем энергопотребления за счет использования тепловой воды с температурой до 40°C.

Микросхема ASIC коммутатора BXI представляет собой неблокирующийся координатный соединитель с низким уровнем задержки и 48 портами. Он поддерживает 48 портов BXI со скоростью 100 Гбит/с. Совокупная пропускная способность микросхемы составляет 1200 ГБ/с (48 портов * 12,5 ГБ/с/ в 1 направлении * 2 направления).

Крупные коммутаторы (288 или 576 портов) создаются путем объединения 2 уровней микросхем ASICs.

16 виртуальных каналов, доступных в BXI, могут использоваться для предотвращения возникновения блокировок и для повышения эффективности обмена данными.

Для баланса трафика в случае перегрузки каждый порт имеет выделенную таблицу маршрутизации, а три порта вывода определены для каждого адресата: один предварительно определенный детерминированный путь и два адаптивных пути выбираются из доступных вычислительных путей, покрывая максимум физических вариантов использования. За счет адаптивной маршрутизации входящие сообщения могут быть направлены на менее загруженные порты вывода.

Для максимальной точности при назначении временных меток, что может иметь важное значение для коммуникационного анализа, коммутатор BXI предлагает тактовый генератор для синхронизации всех узлов в системе с допускаемым предельным значением не более 1 с для 32 000 узлов.

Программный пакет технологии BXI

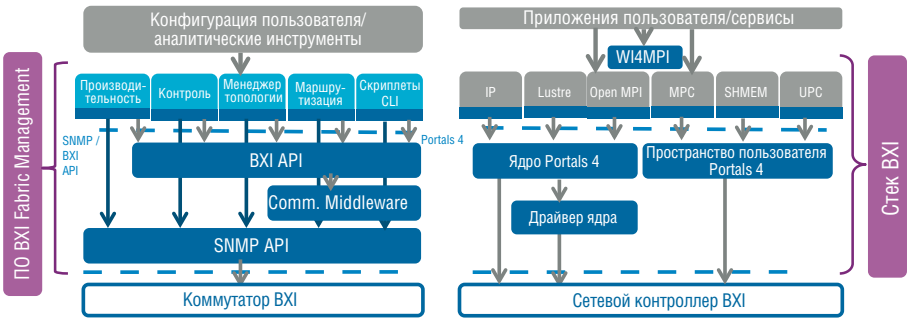
Технология BXI предлагается со специальным программным пакетом, который включает стек BXI Stack, разрешающий пользователям осуществлять программирование и запускать приложения на системах с технологией BXI, и ПО BXI Fabric Management для установки, настройки конфигурации и контроля работы систем с технологией BXI.

ПО BXI Fabric Management поддерживает следующие функциональные возможности:

- Маршрутизирующие решения для быстрых реакций (<5 секунд) на неполадки и предоставление стабильной платформы для запуска приложений (надежность на уровне до 25% ошибок при подключении для 64 000 узлов).
- Автономное определение топологии и проверка неправильных подключений.
- Получение информации и управление локальными и глобальными событиями коммутирующей сети позволяет оперативно реагировать на события.
- Высокочастотные счетчики производительности (включая гистограммы) предоставляют точную информацию о трафике. Четыре фильтра предусмотрены для определения определенных событий, например, сообщения, созданные определенным пользователем, набором узлов или определенным типом трафика.
- Многофункциональный интерфейс командной строки позволяет легко и просто настраивать конфигурацию и контролировать коммутирующую сеть или выполнять мониторинг и диагностику возможных неполадок.

Параллельные приложения могут использовать все преимущества функциональных возможностей сети BXI с помощью библиотек MPI, SHMEM или UPC.

- Все компоненты реализованы напрямую, используя интерфейс программирования приложений Portals 4.
- Сервисы уровня ядра также реализованы с помощью Portals 4.
- Portals 4 LND (Lustre Network Driver) предоставляет параллельную файловую систему Lustre с прямым доступом к Portals 4.
- Компонент IPoPtl (IP over Portals) позволяет реализовать масштабную, эффективную и надежную IP-коммуникацию для унаследованного ПО.



hpc@atos.net