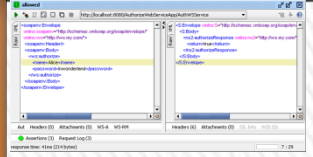
ФЕДЕРАЛЬНОЕ АГЕНТСТВО СВЯЗИ
Федеральное государственное бюджетное образовательное учреждение высшего образования
«Санкт-Петербургский государственный университет телекоммуникаций им. проф. М.А. Бонч-Бруевича»

Технология InfiniBand

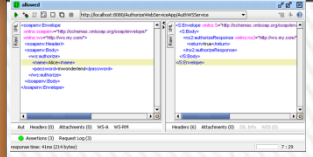
Научный руководитель
доктор технических наук профессор Лариса Константиновна Птицына

Санкт-Петербург, 2015



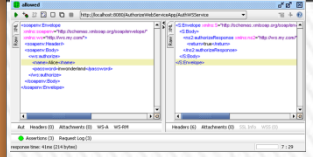
Приоритеты технологии Infiniband

- Иерархическая приоритизация трафика;
- Масштабируемость;
- Возможность резервирования;
- Низкая латентность;



Infiniband обеспечивает

- Возможность описания всех аспектов ввода-вывода;
- Возможность использования разделяемой памяти вместо разделяемых шин;
- Варьируемую полосу пропускания от 1X до 12X;
- **Полную прозрачность для операционных систем;**
- **Возможность организации виртуальных потоков.**

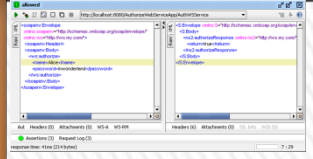


Преимущества Infiniband

- Единственный порт для всего трафика сервера - сети, системы хранения данных и трафика кластера

Эффект преимущества:

1. Множество разнообразных кабелей, подключаемых к шасси в стойке, заменяются единственным общим кабелем – по одному на каждое устройство сервер/система хранения.
2. Совместно используемая система ввода-вывода создает общую точку соединения, уменьшая число прямых соединений (кабелей), что уменьшает требуемое число адаптеров и снижает общие издержки.

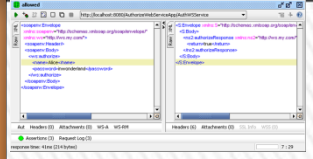


Преимущества Infiniband

- Серверу более не требуется множество разнородных адаптеров

Эффект преимущества:

1. Изготовители серверов могут исключить устаревшие (унаследованные) компоненты, снизив стоимость системы и разместив ее в более компактном корпусе.
2. Более компактный корпус снижает стоимость оборудования информационного центра благодаря оптимизации расположения оборудования в стойке.
3. Предприятие может заменить или модернизировать аппаратное обеспечение сервера, не нарушая конфигурации системы ввода-вывода.

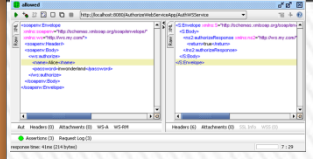


Преимущества Infiniband

- Масштабируемые характеристики подсистемы ввода-вывода

Эффект преимущества:

1. Сервер InfiniBand может иметь несколько высокоскоростных каналов ввода-вывода для использования с системами хранения или в коммуникациях

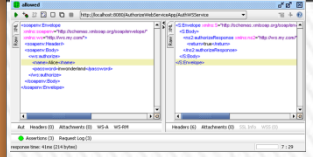


Преимущества Infiniband

- Отделение процессора сервера от устройств ввода-вывода и увеличение расстояния между процессором и контролерами ввода-вывода до нескольких километров

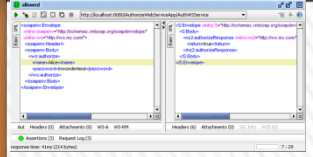
Эффект преимущества:

1. Гибкость в размещении серверов и контроллеров ввода-вывода в пределах центра данных и способность добавлять вычислительные мощности отдельно от расширения системы хранения.
2. Сетевой протокол InfiniBand позволяет реализовать возможности глобальной маршрутизации в пределах пакетов данных, поддерживая различные местоположения систем хранения WAN.
3. Производительность сервера возрастает, поскольку задачи ввода-вывода переносятся в структуру InfiniBand .



Режимы Infiniband – режимы передачи данных по шинам для портов

SDR – Single Data Rate;
DDR – Double Data Rate;
QDR – Quad Data Rate;
FDR – Fourteen Data Rate;
EDR – Enhanced Data Rate

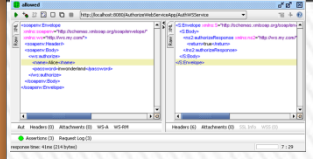


Шины Infiniband

Подобно многим современным шинам, например, PCI Express, SATA, USB 3.0, в Infiniband используются дифференциальные пары для передачи последовательных сигналов. Две пары вместе составляют одну базовую двунаправленную последовательную шину (англ. *lane*), обозначаемую 1x. Базовая скорость — 2,5 Гбит/с в каждом направлении. Порты Infiniband состоят из одной шины или агрегированных групп 4x или 12x базовых двунаправленных шин. Чаще всего применяются порты 4x.

Для портов существует несколько режимов передачи данных по шинам.

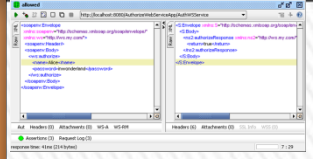
Основное назначение Infiniband — межсерверные соединения, в том числе и для организации RDMA (Remote Direct Memory Access).



Симплексная пропускная способность Infiniband в пересчете на полезный трафик для различных режимов и количества линий

	SDR	DDR	QDR	FDR	EDR
1X	2 Gbit/s	4 Gbit/s	8 Gbit/s	13.64 Gbit/s	25 Gbit/s
4X	8 Gbit/s	16 Gbit/s	32 Gbit/s	54.54 Gbit/s	100 Gbit/s
12X	24 Gbit/s	48 Gbit/s	96 Gbit/s	163.64 Gbit/s	300 Gbit/s

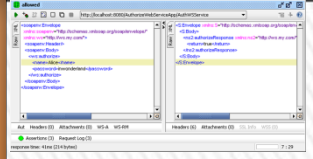
Кодирование (бит)	**8/10**	**8/10**	**8/10**	**64/66**	**64/66**	**64/66**
Типичные задержки, мкс[8]	**5**	**2.5**	**1.3**	**0.7**	**0.7**	**0.5**
Год появления[9]	**2001, 2003**	**2005**	**2007**		**2011**	**2014**[7]



Адаптеры технологии InfiniBand

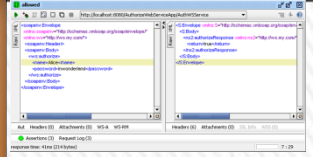
InfiniBand использует коммутируемую среду с соединениями точка-точка, в отличие от ранних вариантов сетей Ethernet, которые использовали общую среду и, изначально, шинное соединение. Все передачи начинаются и заканчиваются на адаптере канала. Каждый вычислительный узел содержит *HCA*-адаптер (host channel adapter), подключаемый к процессору по интерфейсу PCI Express (ранее через PCI-X). Между адаптерами пересылаются данные и управляющая информация, в том числе необходимая для реализации QoS (quality of service).

Для периферийных устройств предполагалось использование TCA-адаптеров (target channel adapter), но они не получили распространения, а такие периферийные устройства создаются на базе стандартных материнских плат.



Адаптеры HCA технологии InfiniBand

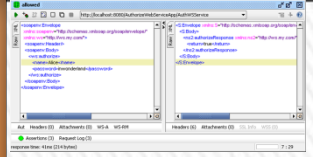
HCA-адаптеры обычно имеют один или два порта 4x, которые могут подключаться либо к таким же портам HCA и TCA, либо к коммутаторам (свитчам). Коммутаторы могут быть организованы в сети с топологиями типа утолщенное дерево (Fat Tree), Сеть Клоза, реже — многомерный тор, двойная звезда, и в различных гибридных комбинациях



Порты и кабели InfiniBand 4 X

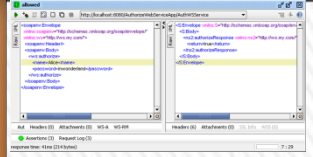
- CX4 (SFF-8470, например, Fujitsu MicroGiGaCN), только до скоростей DDR (иногда до QDR)
- QSFP (SFF-8435, SFF-8436, 40Гбит/с)
- QSFP+ (QSFP14, SFF-8685, 56 Гбит/с)
- zQSFP+ (QSFP28, SFF-8665, 100 Гбит/с)

›



Порты и кабели InfiniBand 12 X

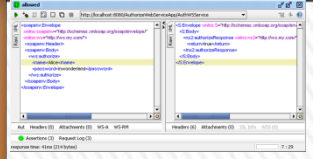
- 12x MicroGiGaCN (Fujitsu FCN-260C024)
- CXP



Кабели InfiniBand

- Пассивные электрические кабели (медные многожильные провода), длиной обычно в единицы метров, иногда до 30 м. Для более длинных кабелей доступны меньшие скорости (7 м для QDR).
- Активные электрические кабели (то же, но с усилителями, позволяют немного увеличить максимальную длину кабеля для данной скорости).
- Активные оптические кабели с интегрированным оптоволоконным кабелем длиной от единиц до десятков и сотен метров.
- Активные оптические модули с оптическим коннектором MTP/MTO для подключения оптоволоконных кабелей типа OM3/OM4 (8 волокон), либо SR4, либо LC/LC.

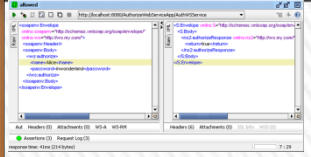
Сигналы Infiniband могут передаваться на несколько дюймов по печатным платам.



Адаптер HCA QLogic QLE7340 (QDR, 40 Гбит/с).

Фрагмент коммутатора Voltaire ISR-6000 с портами SDR 4x

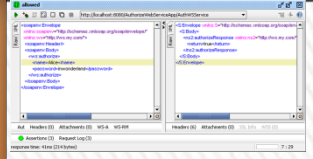
Ранние версии Infiniband использовали электрические кабели 4x с разъёмами CX4 (SFF 8470).



Физический уровень InfiniBand

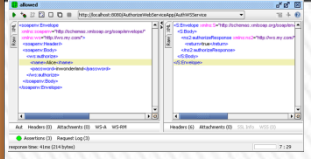
На физическом уровне InfiniBand передает данные в виде пакетов длиной до 4 КБ (килобайт), которые после объединения формируют сообщение. Некоторые устройства поддерживают меньший максимальный размер пакетов, например, 2 КБ

›



Типы сообщений InfiniBand

- операция доступа к памяти — чтение или запись в память получателя (RDMA).
- канальные операции пересылки сообщений (отправитель посылает сообщение с данными, получатель принимает его в заранее выделенный буфер)
- транзакционная операция
- передача нескольким получателям (multicast, поддерживается не всеми коммутаторами)
- атомарная операция в память удаленного узла (атомарное сложение и сравнение-с-обменом для 64-битных целых).



Разделение сообщений InfiniBand на сервисы

Сообщения Infiniband разделяются на сервисы в зависимости от гарантий доставки и необходимости инициализации соединения перед обменом:

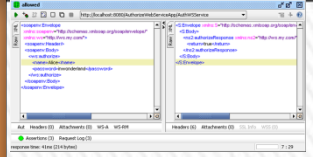
•Reliable Connected (RC) — надежная доставка, необходима инициализация соединения между получателем и отправителем.

•Unreliable Connected (UC) — ненадежная доставка, необходима инициализация.

•Reliable Datagram (RD) — опциональный сервис, реализуется редко. Надежная доставка без инициализации.

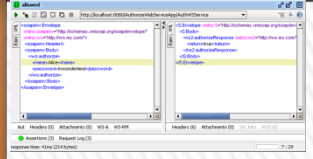
•Unreliable Datagram (UD) — ненадежная доставка, не требует инициализации.

•Позже был введен сервис XRC[13], комбинирующий в себе некоторые свойства RC и RD.



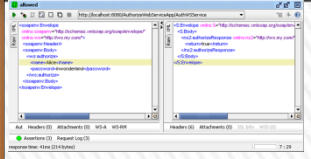
Принцип RDMA

Infiniband позволяет использовать принцип **RDMA** (англ. *Remote Direct Memory Access* — удалённый прямой доступ к памяти), при котором передача данных из памяти удаленного компьютера в локальную память инициатора запроса производится непосредственно сетевым контроллером, при этом исключается участие CPU удаленного узла. RDMA позволяет передавать данные без дополнительной буферизации и не требует активной работы ОС, библиотек или приложения на узле, к памяти которого производится обращение.



Низкоуровневые универсальные программно-аппаратные интерфейсы

Infiniband может использоваться с двумя низкоуровневыми универсальными программно-аппаратными интерфейсами (API), которые были разработаны на базе U-Net (Cornell, середина 1990-х) и VIA (*Virtual Interface Architecture*[en], конец 1990-х).



Высокоуровневые программные интерфейсы

С помощью verbs или uDAPL могут быть реализованы высокоуровневые программные интерфейсы и протоколы, в частности:

- MPI (*Message Passing Interface*) — популярный стандарт пересылки сообщений в компьютерных кластерах. Существует множество реализаций MPI, поддерживающих сети Infiniband.
- SHMEM, GASnet и другие популярные интерфейсы для работы с RDMA
- IPoIB (IP over Infiniband) — группа протоколов, описывающих передачу IP-пакетов поверх Infiniband[15]:
  - RFC 4390 Dynamic Host Configuration Protocol (DHCP) over InfiniBand
  - RFC 4391 Transmission of IP over InfiniBand (IPoIB)
  - RFC 4392 IP over InfiniBand (IPoIB) Architecture
- SRP (англ. *SCSI RDMA Protocol*) — протокол обмена данными между SCSI-устройствами с использованием RDMA[15]. Определён в стандарте ANSI INCITS 365—2002.
- DDP (англ. *Direct Data Placement*): RFC 4296 —— архитектура для реализации прямого размещения данных (DDP) и удаленного прямого доступа к памяти (RDMA) в Internet-сетях.
- SDP (англ. *Socket Direct Protocol*) — протокол установления виртуальных соединений и обмена данными между сокетами поверх Infiniband[15], передача данных не использует TCP-стек операционной системы, однако использует IP-адреса и может использовать IPoIB для их разрешения.
- '*iSER* (iSCSI Extensions for RDMA) — IETF-стандарт для адаптации iSCSI к RDMA сетям

Спасибо за внимание!