Software Engineering Conference Russia 2018
October 12-13, Moscow

Оптимизация интерконнекта кластерного решения при помощи **InfiniBand RDMA**

на примере доработки MySQL Cluster

Михаил Купчук
руководитель группы исследования технологий обработки информации

MEDIO TRIBE

МТС : ИТ

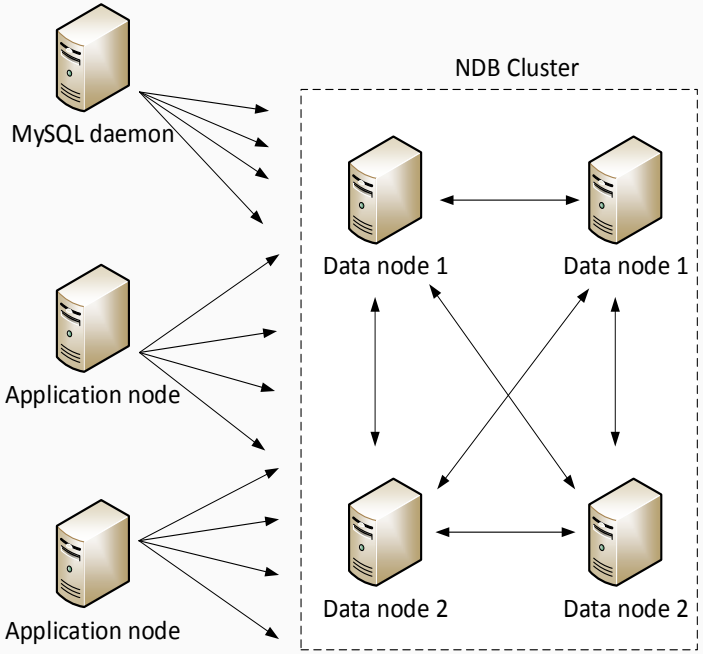
MySQL Cluster и наш продукт

- В нашем продукте используется 4 года в «боевой» среде для хранения балансов абонентов:
  5 независимых региональных площадок с нагрузкой по 70k бизнес-операций в сек на каждой
- На данный момент production стенды справляются с задачами

- Требования к количеству обрабатываемых запросов в единицу времени регулярно растёт
- Горизонтальное масштабирование даёт результат, но наверняка есть альтернативы
- Уменьшение времени обработки запроса никогда не бывает лишним

MySQL Cluster

- Кластерная in-memory БД
- Ядро – NoSQL БД NDBCLUSTER
- Резервирование с избыточных хранением копий данных на нескольких узлах
- Online масштабирование
- Для разработки приложений предоставляется свой API
- SQL интерфейс через MySQL daemon



InfiniBand

- Высокая пропускная способность (начиная с 10Gb/s и до 100Gb/s)
- Низкое время задержки – производителем декларируется < 1µs end to end
- RDMA (Remote Direct Memory Access) – технология доступа к памяти удалённого компьютера по сети, в которой данные передаются минуя ОС и ЦП

Разработанное решение

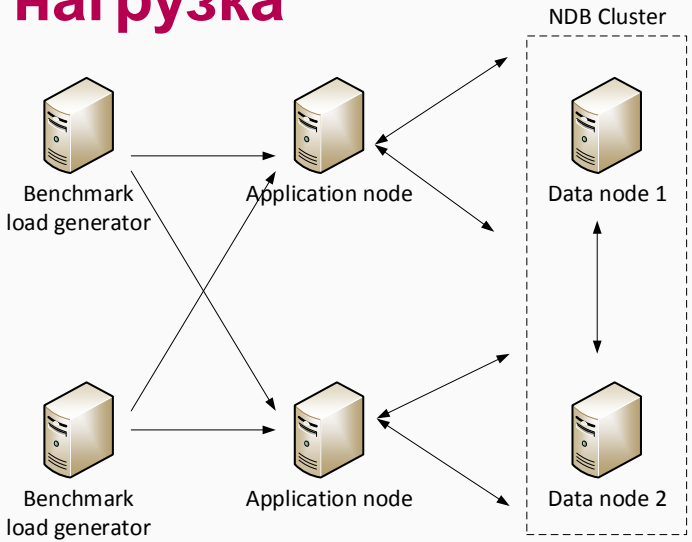
Сделано

- Разработан новый тип transporter’a, использующего InfiniBand RDMA, в клиентской и серверной частях NDB
- Использован «родной» API InfiniBand – IB Verbs

Не сделано

- Поддержка полноценного zero-copy при сетевом взаимодействии
- Кастомизация кода менеджеров соединений для максимально эффективного использования RDMA transporter’a

Тестовый стенд и нагрузка



Сервера

2 * Intel Xeon E5-2690 v4 @ 2.60GHz

8 * 16GB RAM

10GE Intel 82599ES PCIE Card

56Gbps IB Card Mellanox MT27500 Family ConnectX-3

Нагрузка

- 3 типа бизнес-операций
  - 2 бизнес-операций на чтение данных, 1 на обновление
  - Каждая бизнес-операция разворачивается в операции с 2-3 таблицами, с доступом к 30 записями в каждой таблице
- Подача осуществляется в многопоточном режиме

Результаты

TCP

- ~10100 бизнес-операций в секунду
- Среднее время отклика от 2.1 до 4.4 мс в зависимости от типа бизнес-операции

InfiniBand:

- ~12500 бизнес-операций в секунду
- Среднее время отклика от 1.15 до 3.08 мс в зависимости от типа бизнес-операции
- Экстраполяция на текущие стенды, используемые в продуктиве, даёт ~87500 бизнес-операций в секунду

Ссылки и контакты

- E-mail: mgkupchu@mts.ru

- Портал по программированию RDMA: https://www.rdmamojo.com/2012/05/18/libibverbs/
- Серия обучающих статей по основам использования IB Verbs: https://thegeekinthecorner.wordpress.com/2010/08/13/building-an-rdma-capable-application-with-ib-verbs-part-1-basics/

MEDIO TRIBE

МТС : ИТ