Сетевые ускорения в комплексе Скала-СР / Postgres Pro: настоящее и будущее

Скала-СР / Postgres Pro

IBS

Mellanox TECHNOLOGIES

Идея, системная интеграция, сопровождение

Суперкомпьютерное сетевое оборудование с поддержкой RDMA

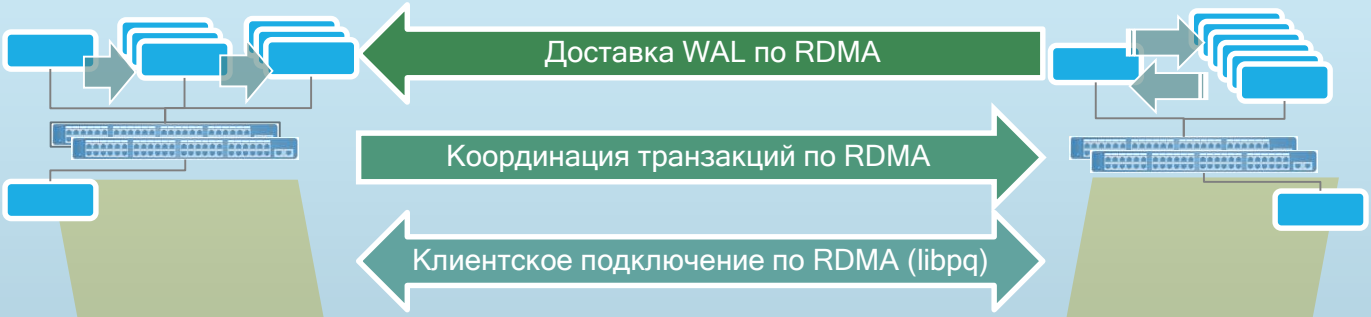
Специальная версия Postgres Pro EE с поддержкой RDMA

Российское серверное оборудование

PROFESSIONAL Postgres

DEPO [computers]

Скала-СР / Postgres Pro: 300 и 500



300

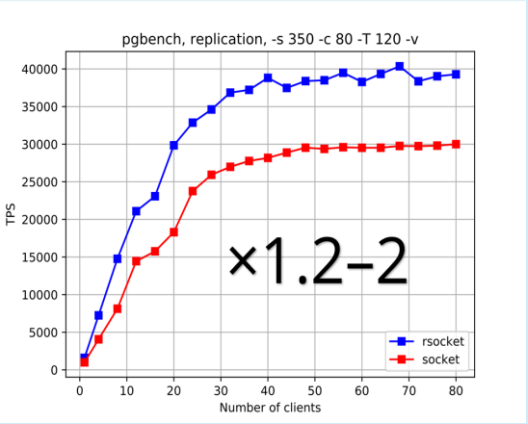
Кластер общего назначения

500

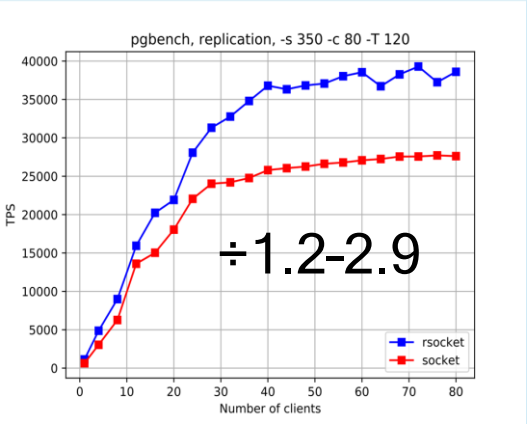
Мультимастер

Репликация по RDMA

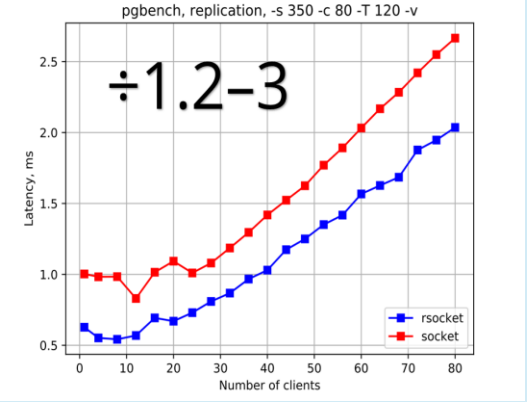
Read-write with synchronous replication, TPS



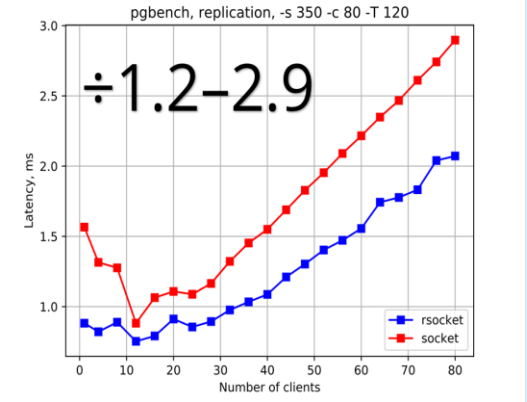
Read-write with remote apply replication, TPS



Read-write with synchronous replication, latency

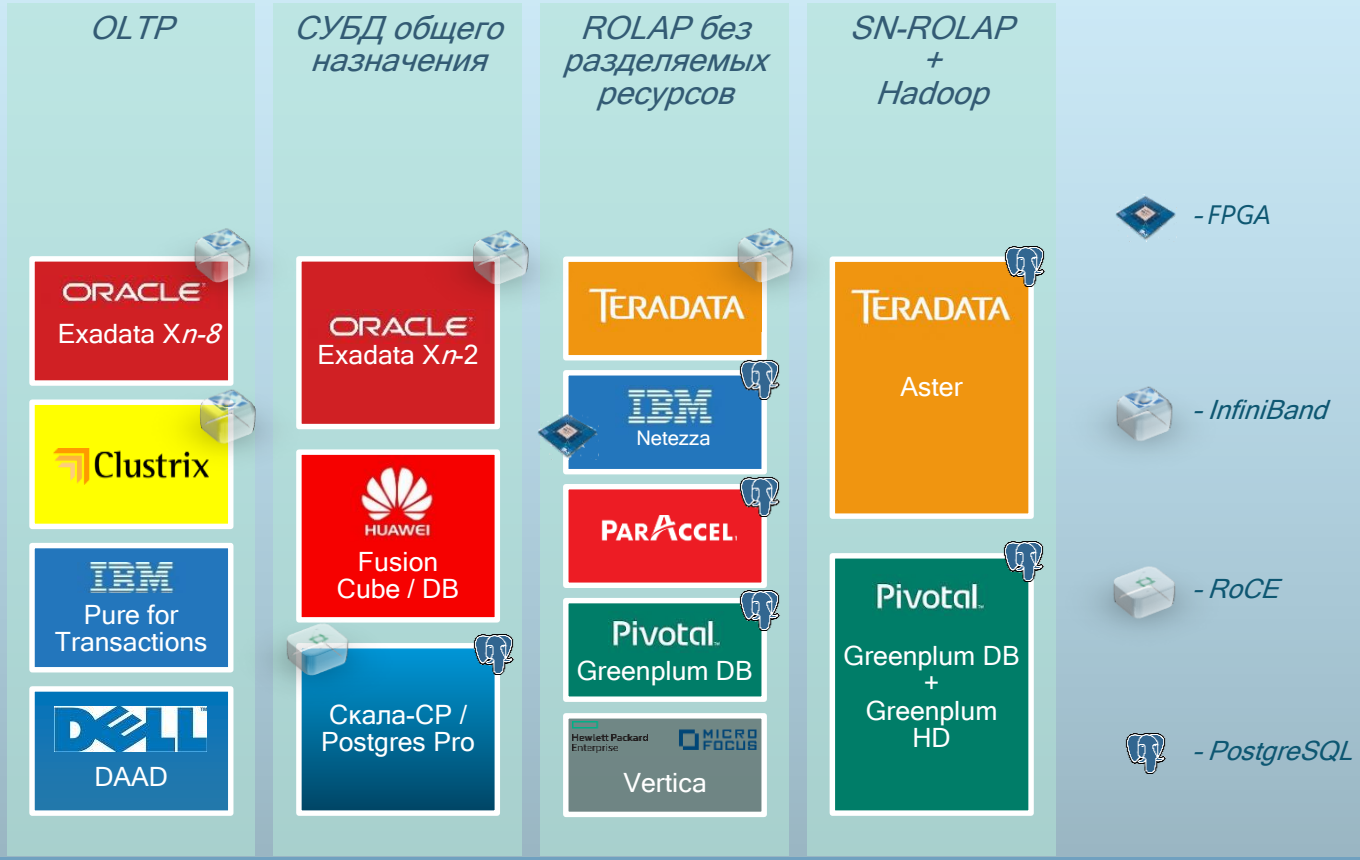


Read-write with remote apply replication, latency



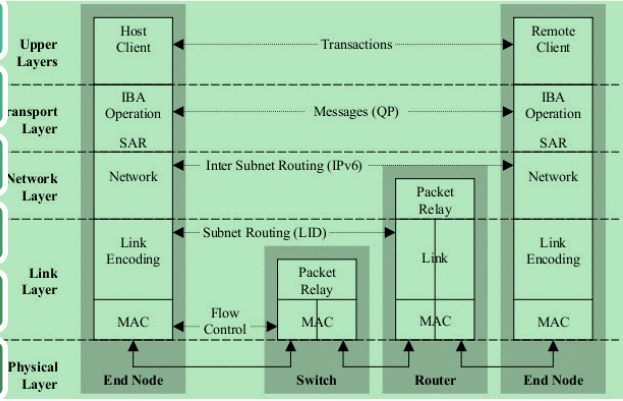
PROFESSIONAL Postgres

Аппаратные трюки в мире машин баз данных



Почему InfiniBand?

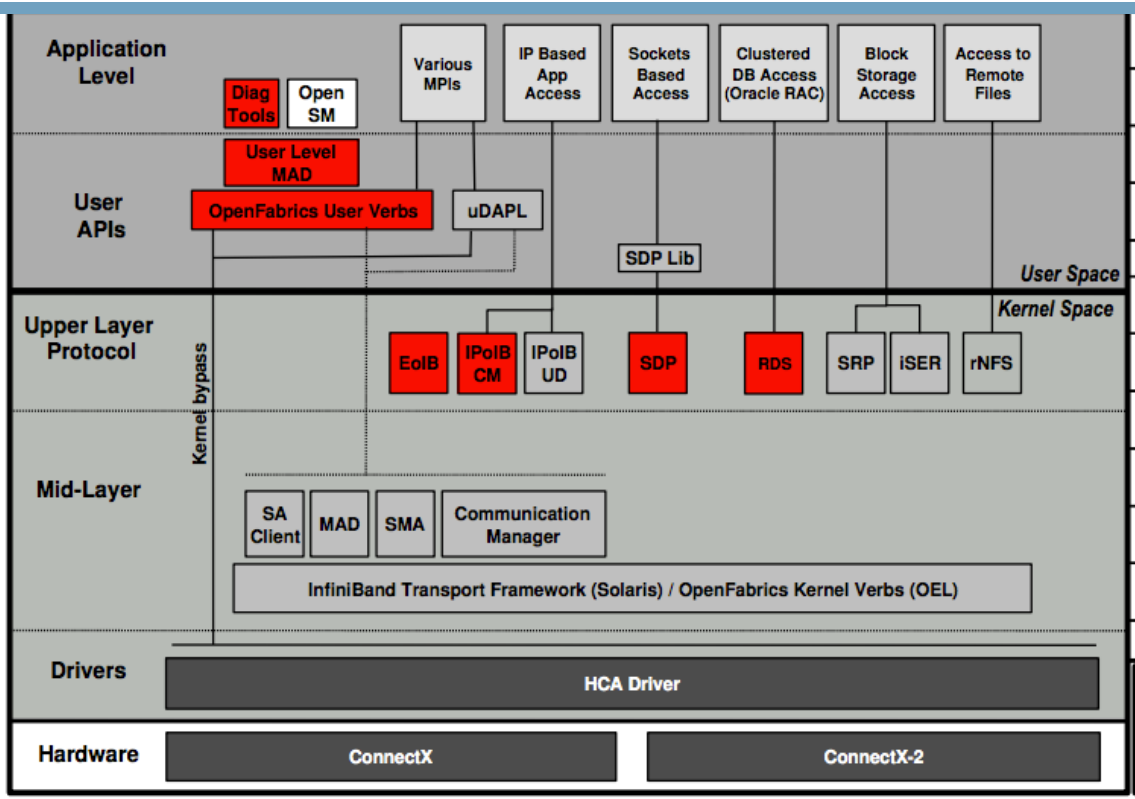
- Открытый сетевой стандарт
- Полностью переделанные все 7 сетевых уровней
- Наиболее быстрая сетевая инфраструктура - до 200Гб/с
- Самые низкие задержки- 90ns на каждом коммутаторе
- Полный HW Offload всех сетевых задач
- За счет этого, используя Verbs API, E2E задержки <0.7µs



Какой прирост производительности при переходе от 10GbE к 40Gb/s IB?

В 40 раз!

InfiniBand у «больших вендоров»: Oracle iDB и Teradata BYNet

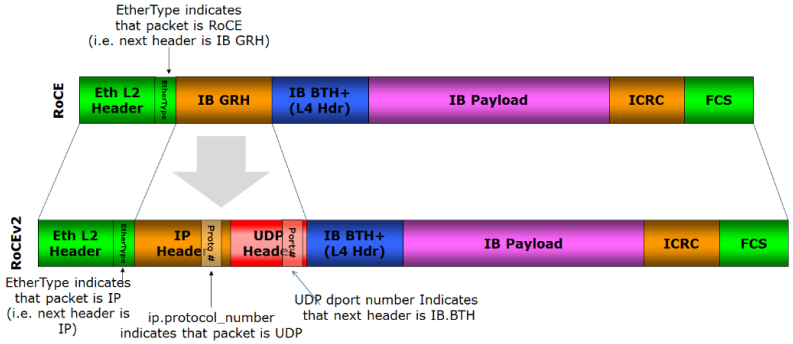


ИТ-директор: «*сеть должна быть только Ethernet!*»

RoCE: свойства Infiniband в Ethernet

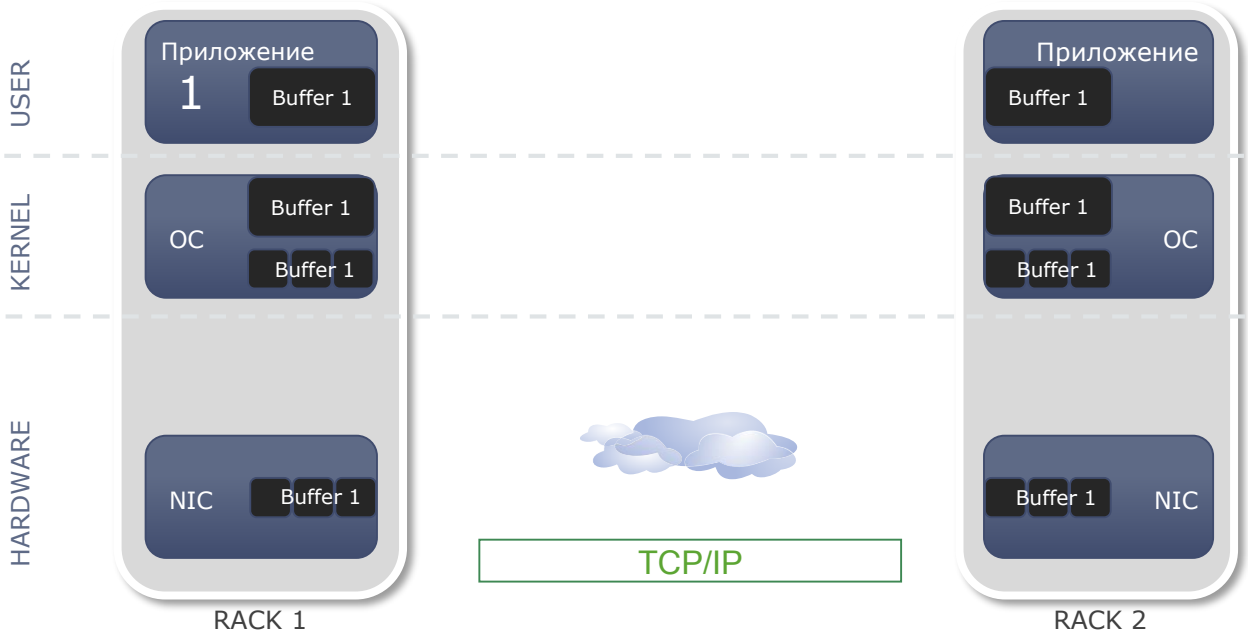
Mellanox TECHNOLOGIES

- Низкие задержки сети <300µs, 1.2µs E2E
- Гарантированная доставка пакетов (Zero Packet Loss)
- PFC+ECN для организации lossless-среды
- RoCE API на базе Verbs с небольшими отличиями

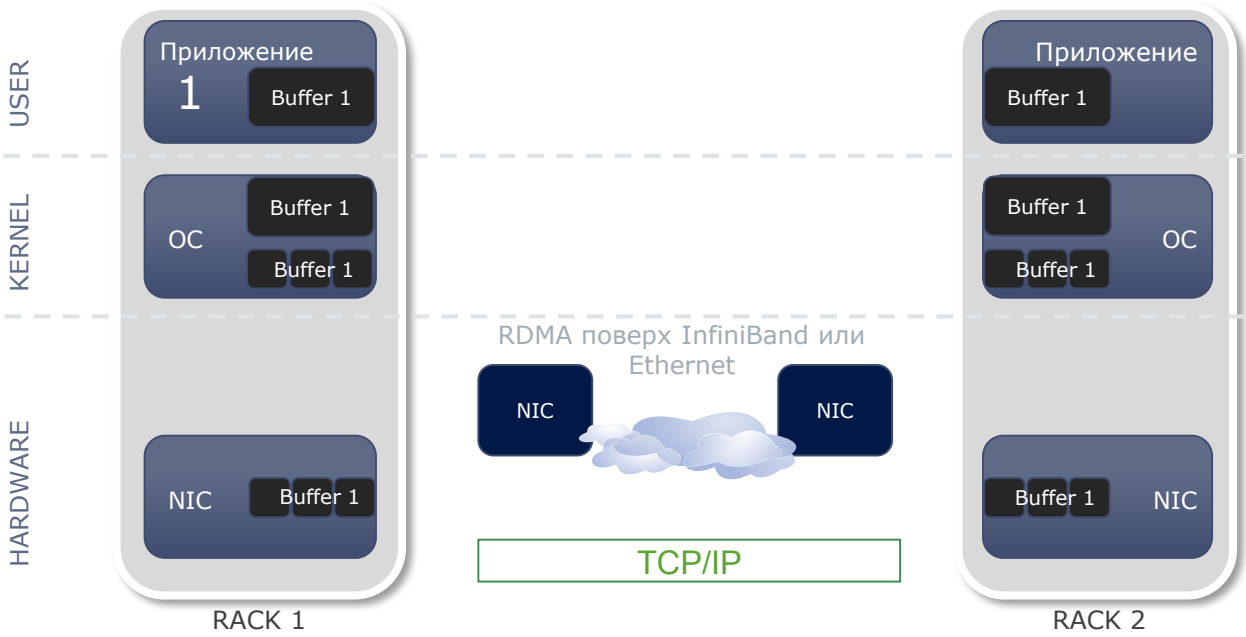


RoCE™

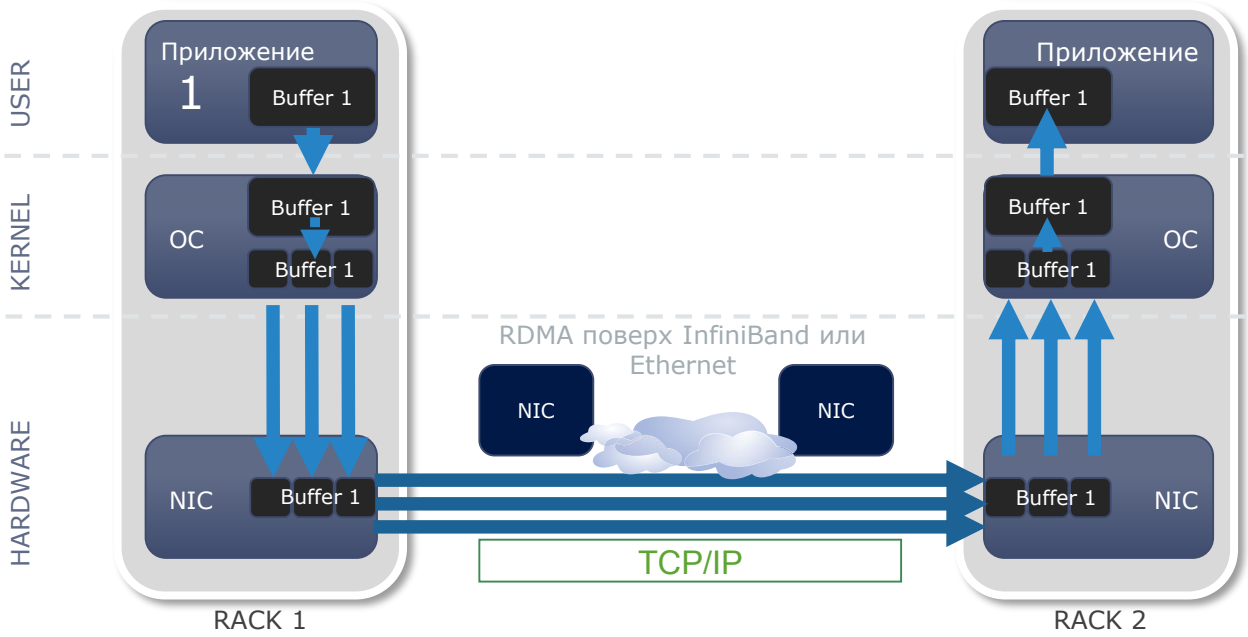
Zero-copy + CPU Offload



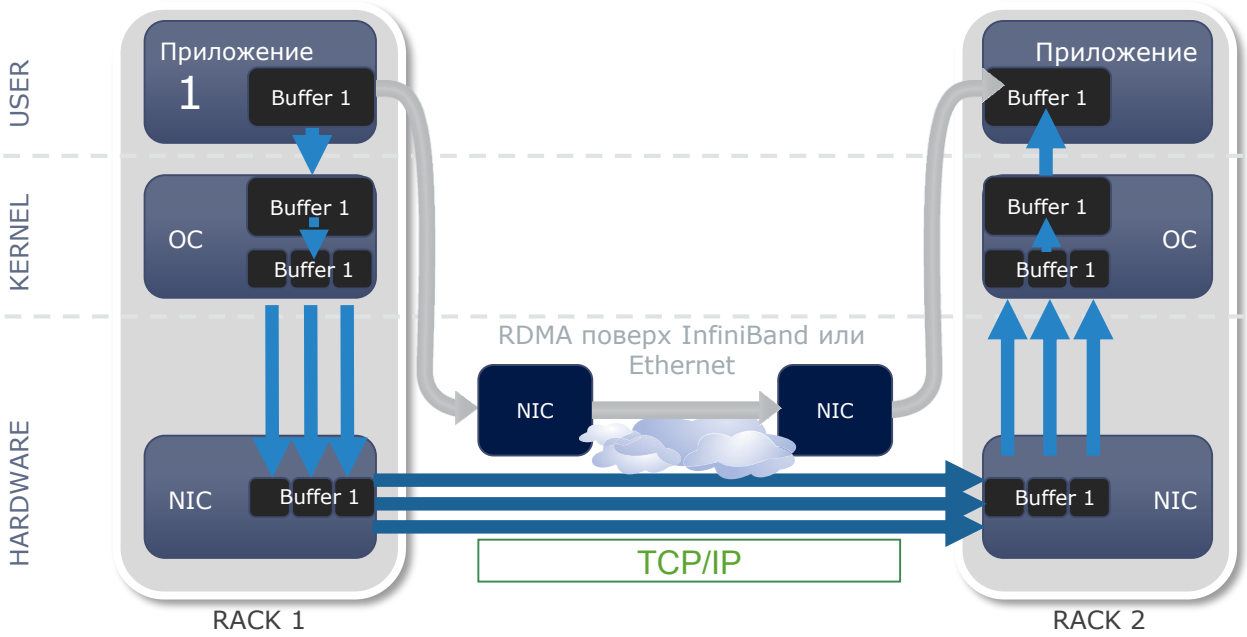
Zero-copy + CPU Offload



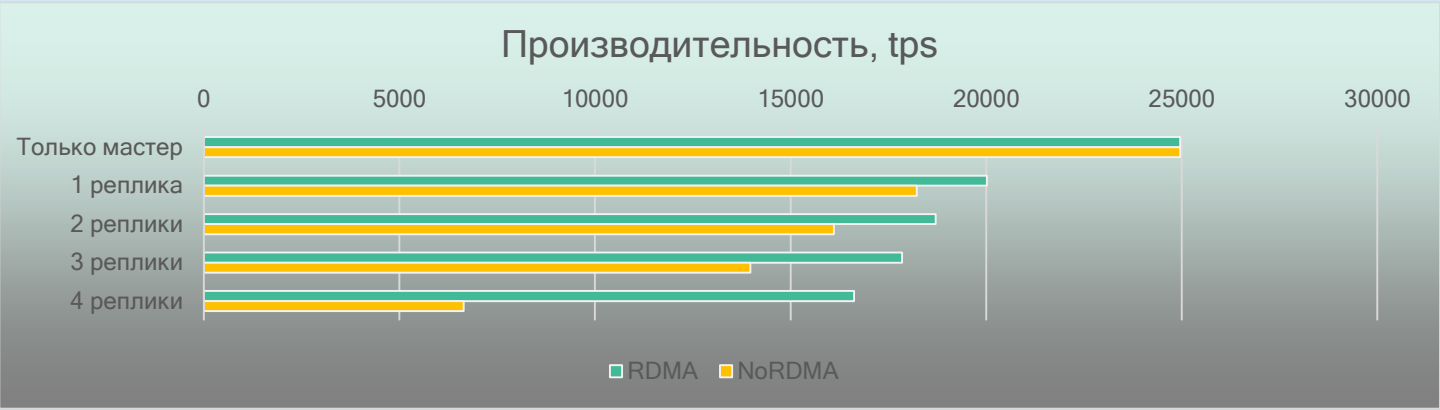
Zero-copy + CPU Offload

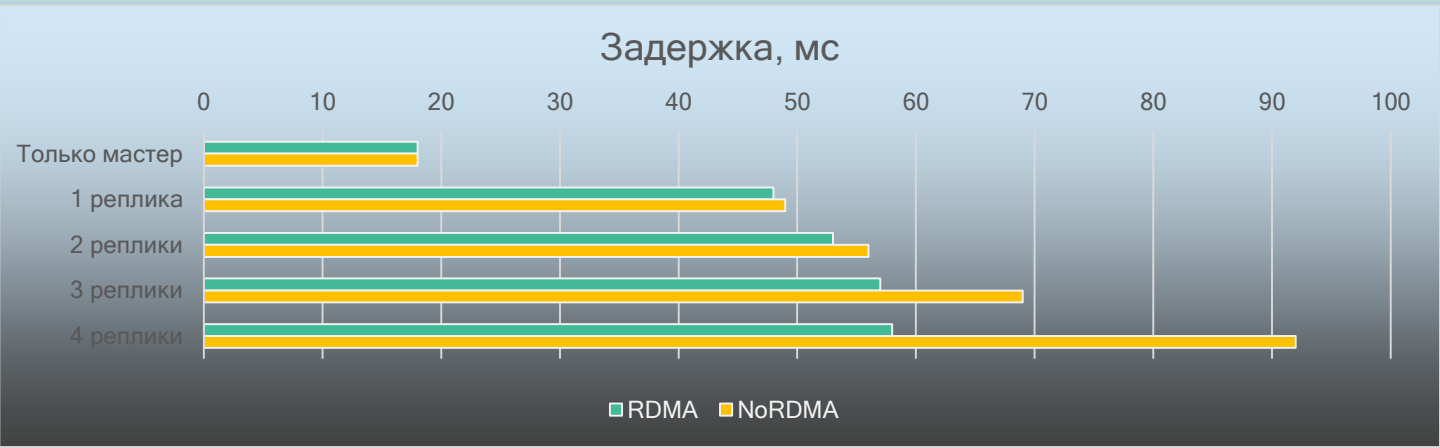


Zero-copy + CPU Offload

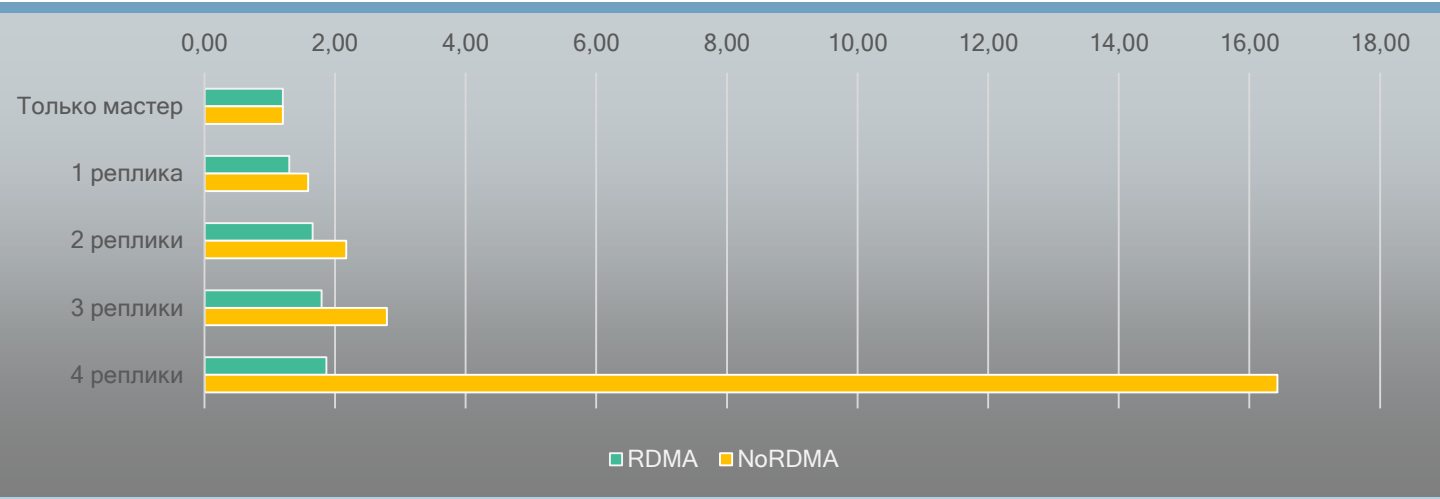


Эффект CPU offload?
Много синхронных реплик





Время ЦПУ мастера на транзакцию (мс)



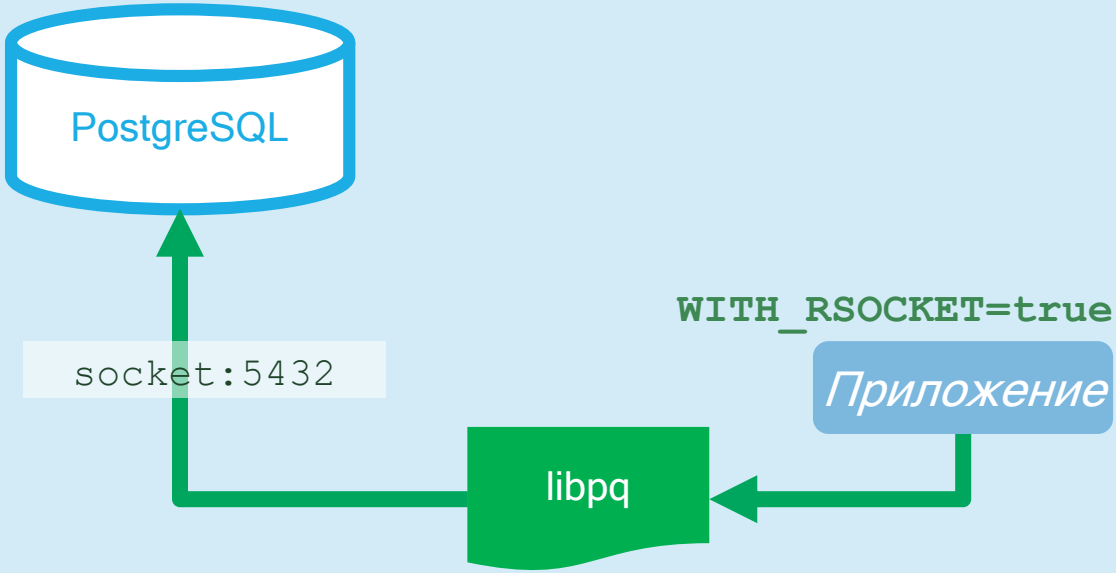
Disclaimer: поверхностный расчёт

(из Load Average на всю систему, с ожиданиями и прочими неаккутартностями)

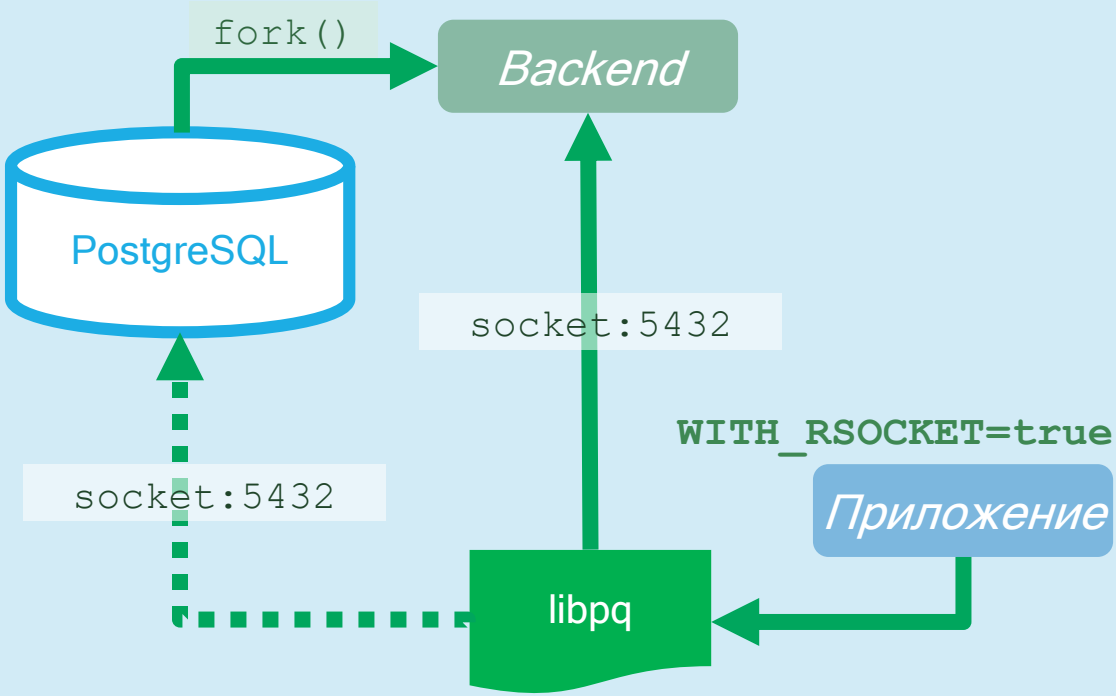
Но эффект очевидный, значительный

И это далеко не максимум возможностей!

Первая реализация: RSocket

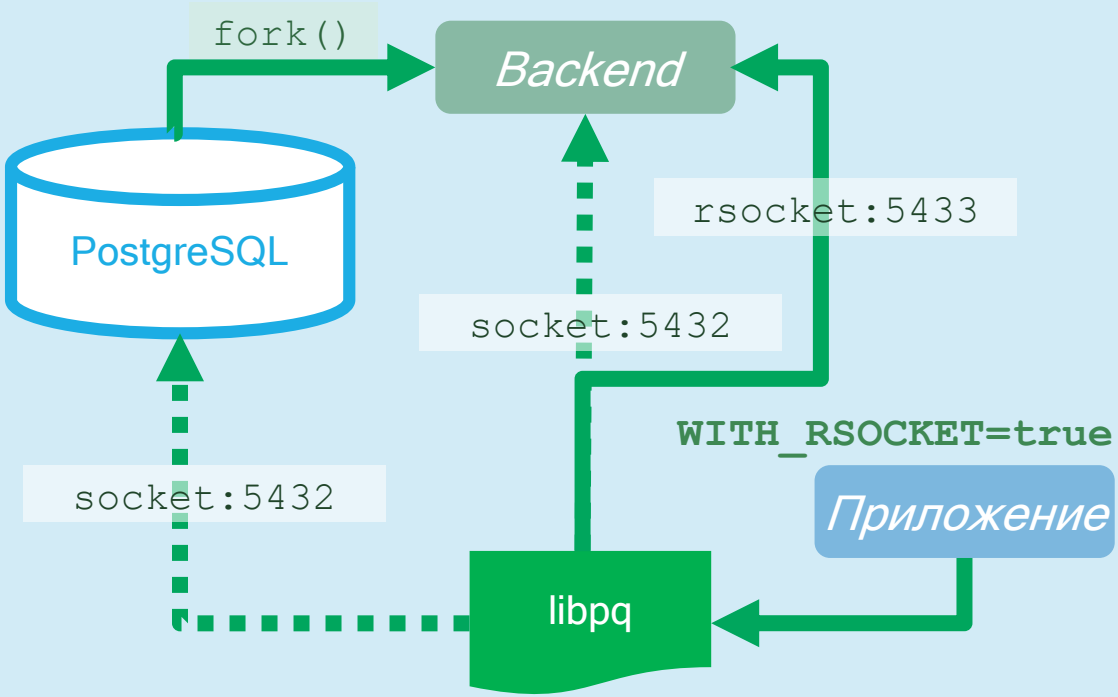


Как установить RDMA-соединение?



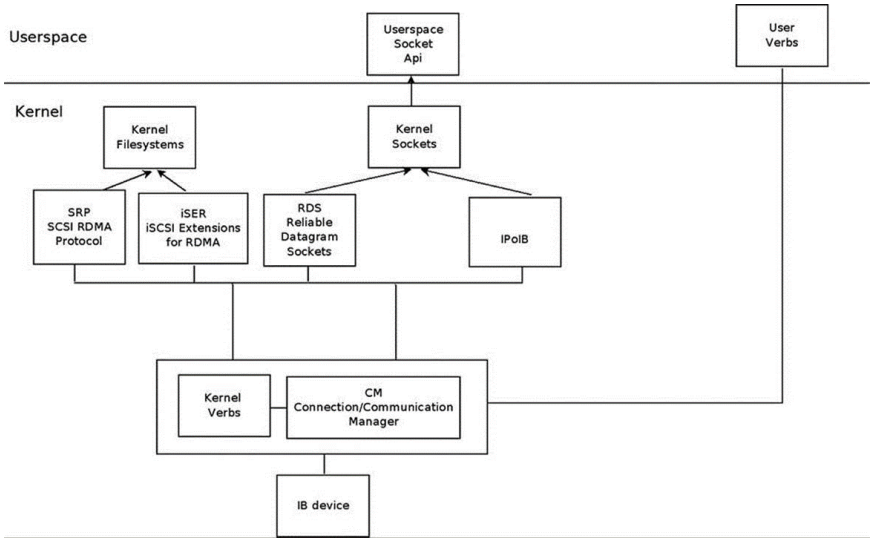
PROFESSIONAL Postgres

...и работаем через RSocket



PROFESSIONAL Postgres

Verbs: родной язык RDMA

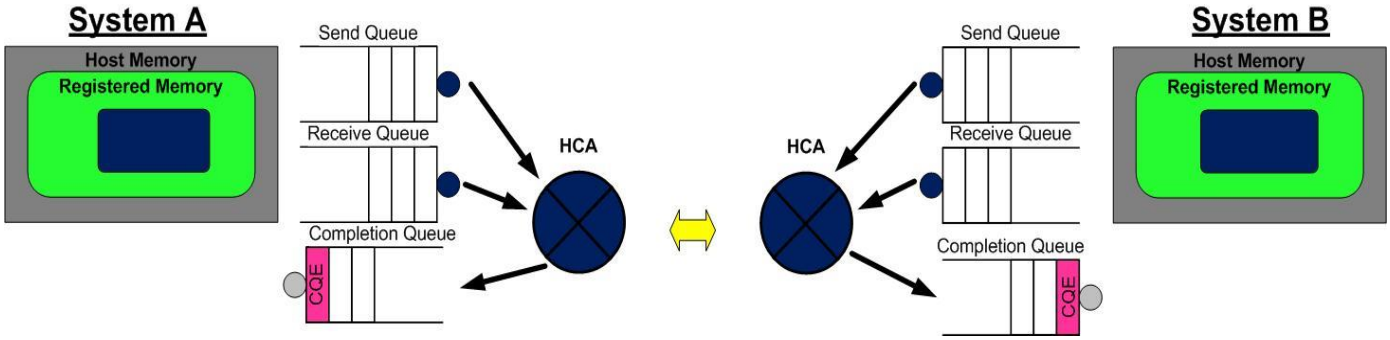


Verbs: родной язык RDMA

LID – Local Identifier (address which is assigned to any port in subnet)
GID – Global Identifier (128b address to identify endpoint of multicast group)
QP – Queue Pair
CQ – Completion Queue
PD – Protection Domain
MR – Memory Region
MW – Memory Window
AV – Address Vector
WR – Work Request
WQE – Work Queue Entry
SR – Send Request
RR – Receive Request
CQE – Completion Queue Entry

SEND: send only the data (the responder need to post RR)
RDMA Write: send the data + remote address and key
RDMA Read: send a read request to the remote side:
remote address and key + remote size
ATOMIC (Cmp&Swp, Fetch&Add) – Atomic operations
(Read-Modify-Write in one atomic operation) which are
used to implement mutual exclusives objects

Verbs: родной язык RDMA



Возможности приблизиться к «родному языку»

SDP

- Socket Direct Protocol
- Исторически первая реализация «переходника» с TCP-приложений на RDMA
- Устарел, не поддерживается в OFED

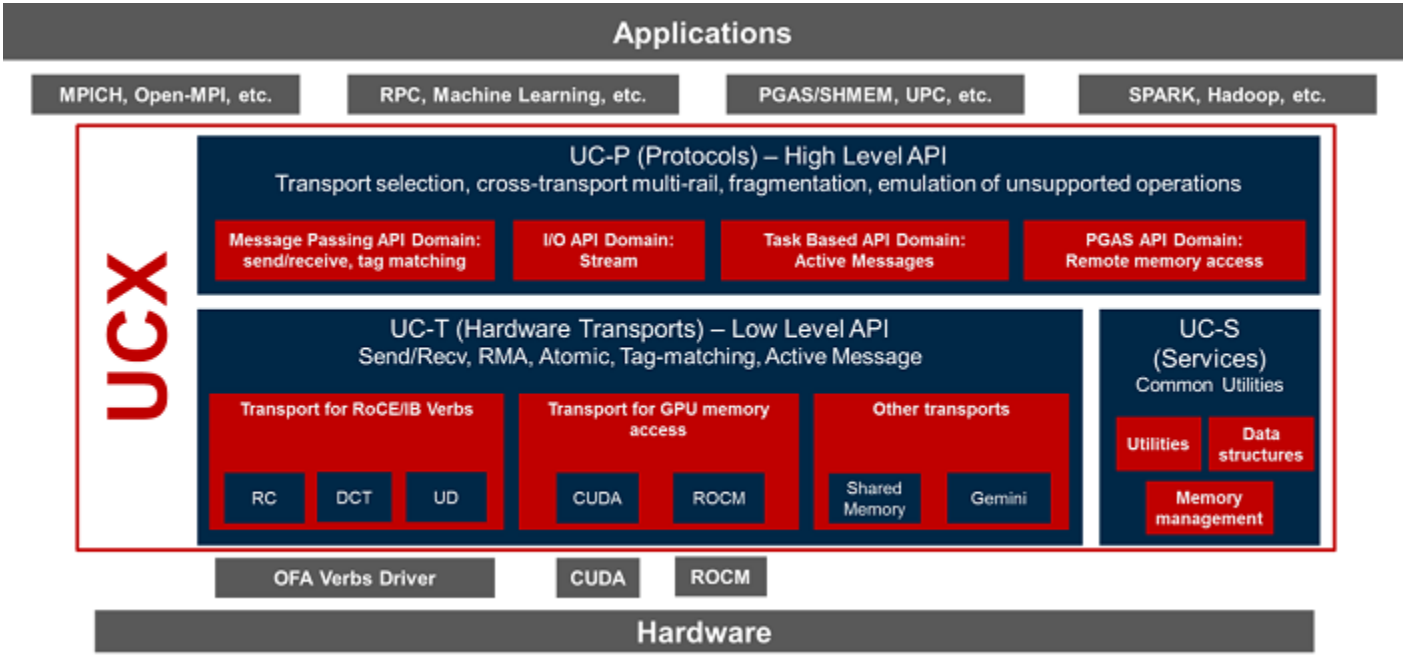
VMA

- Mellanox Voltaire Messaging Accelerator
- Та же идея, что RSocket: сокетоподобный язык для миграции приложений на RDMA
- Свободное ПО, активно развивается и поддерживается Mellanox

UCX

- Двухуровневый фреймворк, позволяющий как быстро, так и «тонко» перейти на RDMA
- Свободное ПО, активно развивается и поддерживается Mellanox

UCX


Applications
MPICH, Open-MPI, etc.
RPC, Machine Learning, etc.
PGAS/SHMEM, UPC, etc.
SPARK, Hadoop, etc.
UCX
UC-P (Protocols) – High Level API
Transport selection, cross-transport multi-rail, fragmentation, emulation of unsupported operations
Message Passing API Domain: send/receive, tag matching
I/O API Domain: Stream
Task Based API Domain: Active Messages
PGAS API Domain: Remote memory access
UC-T (Hardware Transports) – Low Level API
Send/Recv, RMA, Atomic, Tag-matching, Active Message
Transport for RoCE/IB Verbs
RC
DCT
UD
Transport for GPU memory access
CUDA
ROCM
Other transports
Shared Memory
Gemini
UC-S (Services)
Common Utilities
Utilities
Data structures
Memory management
OFA Verbs Driver
CUDA
ROCM
Hardware

Текущие работы и перспективы

Переход на UCX

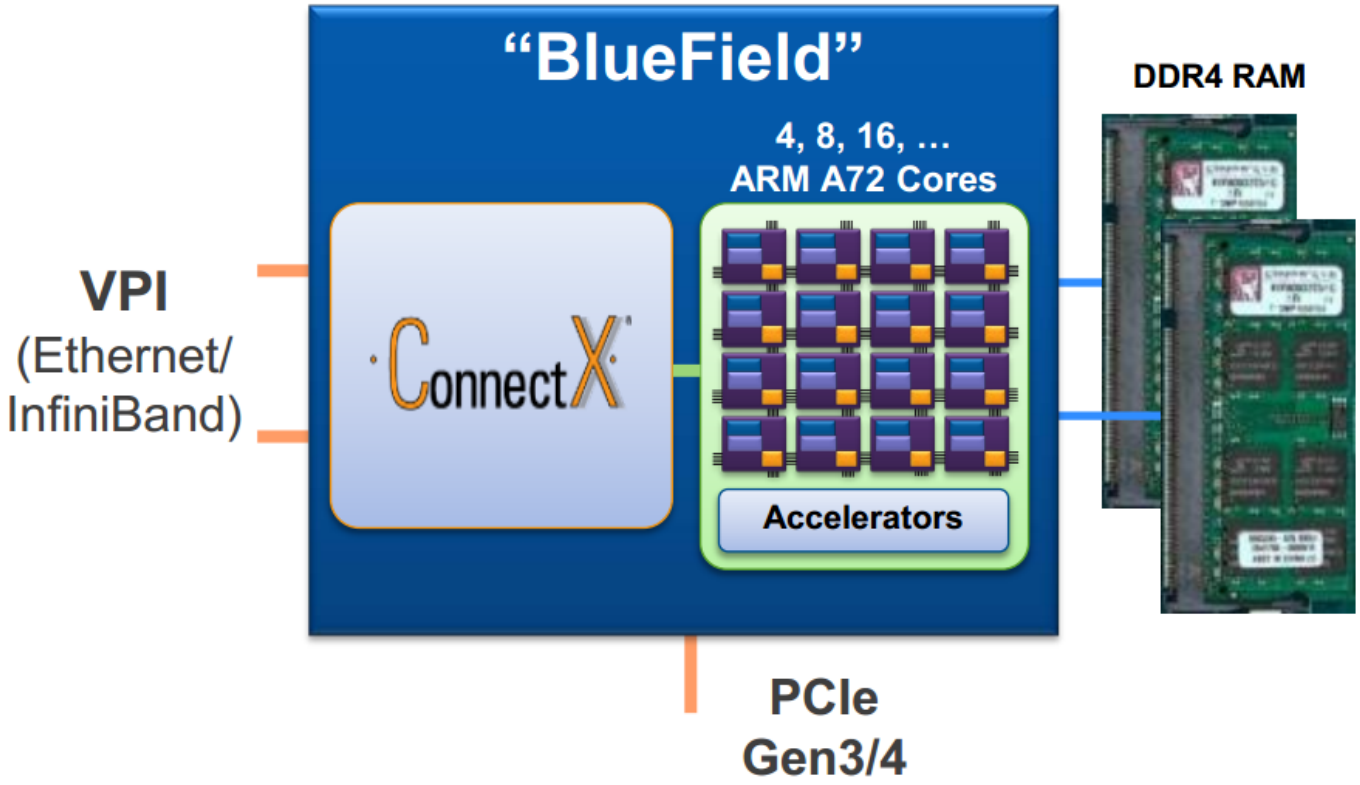
- Перспектива уйти от fork()
- Архитектурные тонкости…

RDMA для pg_shardman

- Возможность мгновенной коммуникации между сегментами

Postgres PROFESSIONAL

BlueField:
компьютер в сетевом адаптере



IBS

BlueField: прорабатываемые возможности

Система хранения для PostgreSQL

- JBOF + NVMf
- 16 NVMe SSD без PCIe-коммутатора

Система управления на сетевом адаптере

- Corosync + Pacemaker - внутри хоста, и на отдельном чипе
- Мониторинг

Спасибо за внимание!

skalar@ibs.ru