Высокоскоростная сеть Альфа

РФЯЦ-ВНИИТФ

РОСАТОМ

Сеть Альфа: введение

РФЯЦ-ВНИИТФ
РОСАТОМ

Назначение

Высокоскоростная коммуникационная сеть Альфа предназначена для передачи данных между узлами вычислительной системы с высокой скоростью и малой коммуникационной задержкой

Области применения

Большие объемы данных

- Системы хранения данных
- Системы управления базами данных (СУБД)

Расчетно-информационные задачи

- В качестве коммуникационной сети в суперкомпьютерах и кластерах, в том числе с применением GPU

Специализированное исполнение

- Отказоустойчивая сеть в специализированных и бортовых вычислительных комплексах

Общие сведения

РФЯЦ-ВНИИТФ
РОСАТОМ

Особенности

- Чип сети на базе ПЛИС
- Интерфейс с хостом: PCIe Gen 3 x16
- Поддержка медных и оптических кабелей
- Прямой доступ в память удаленного узла (RDMA)
- Отправка данных происходит без участия ядра ОС (в пространстве пользователя)
- Инжекция данных в сеть в режиме PIO и DMA
- Поддержка счетчиков производительности и исключительных ситуаций

Характеристики

- Задержка между двумя соседними узлами:
  - MPI: 1,9 мкс
  - SHMEM: 1,7 мкс
- Задержка на хоп: 0,5 мкс
- Пропускная способность порта: 80 Гбит/с
- Пропускная способность на MPI: 72,5 Гбит/с
- Масштабирование: до 4096 узлов
- Энергопотребление адаптера: до 29,5 Вт
- Энергопотребление коммутатора: до 300 Вт

Сравнение основных характеристик

РФЯЦ-ВНИИТФ
РОСАТОМ

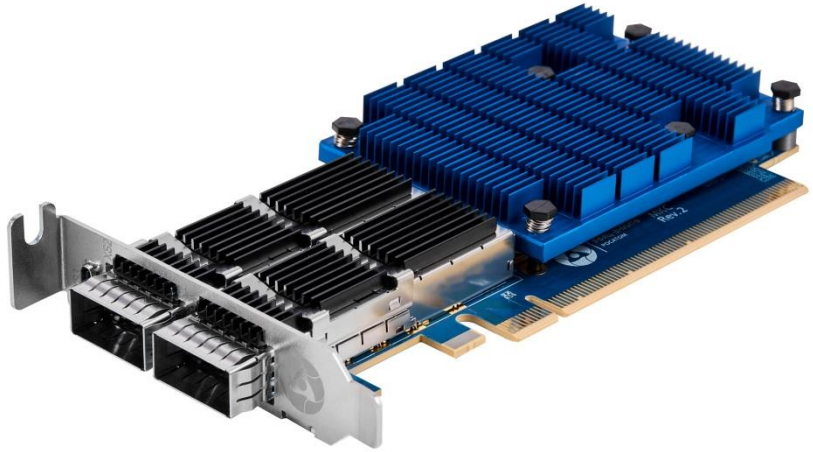
Сеть	Ангара	Ethernet 100G	IB EDR	Intel OmniPath	IB HDR	Slingshot	IB NDR	Альфа
Год выпуска	2013	2015	2015	2015	2019	2019	2022	2025
Топология	4D-тор	дерево, другие	Fat Tree, kD-тор	Fat Tree	Fat Tree, kD-тор, Dragonfly	Dragonfly	Fat Tree, kD-тор, Dragonfly	5D-тор, Fat Tree, Dragonfly +
Линк MPI, Гбит/с	50	100	100	100	200	200	400	72.5
Задержка MPI, мкс	0.85	~1-3	0.9	0.9	0.9	1.8	1	2.4
Темп выдачи сообщений, МТ/с	13	–	165	110	180	–	330	50
RDMA	+	+–	+	–	+	+	+	+

Исполнение

Варианты сетевого оборудования

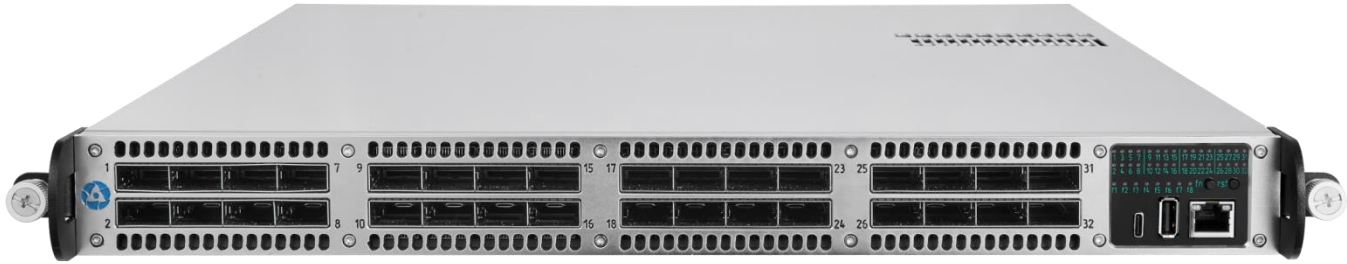
Сетевой Адаптер

- 2 порта **QSFP-DD**
- Габаритные размеры (Д×Ш×В), мм **142,25 × 68,9 × 17,25**
- Форм-фактор: **low-profile PCI Express card**



Коммутатор

- 32 порта **QSFP-DD**
- Габаритные размеры (Д×Ш×В), мм **650 × 440 × 43,6**



Сетевой адаптер

РФЯЦ-ВНИИТФ
РОСАТОМ

Подключение к хост-системе

Адаптер сети Альфа подключается к хост-системе через интерфейс PCI Express 3.0 x16

Данный интерфейс используется для:

- настройки и мониторинга адаптера
- посылки пакетов
- выполнения адаптером DMA обращений (чтение, запись, атомарные операции) к памяти хост-системы

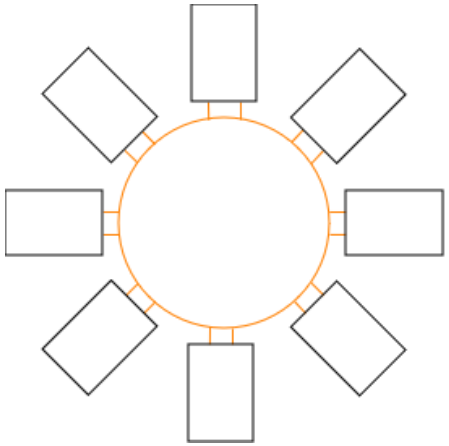
BAR-регионы адаптера:

- регистровый файл
- инжекционные конвейеры для посылки пакетов

Использование

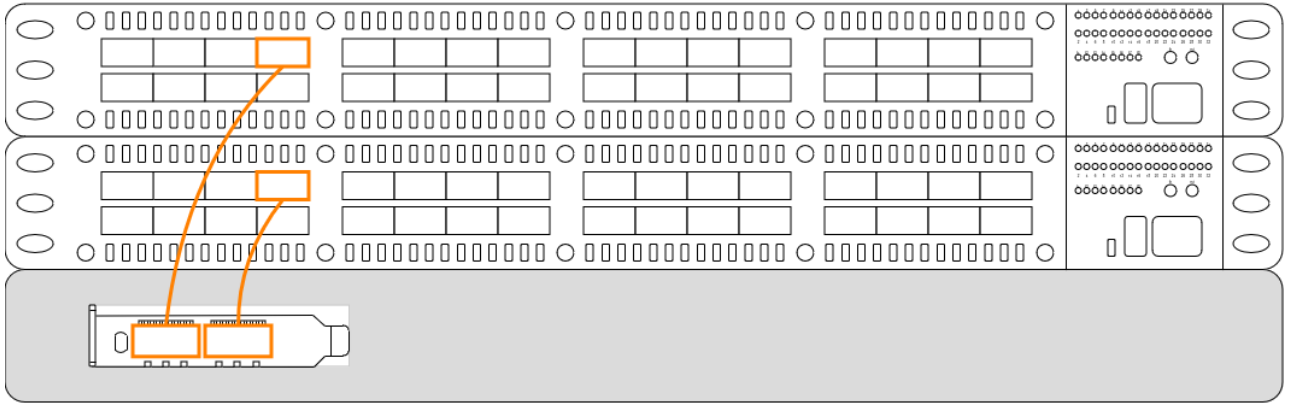
Без коммутатора

- До 8 адаптеров в кольце



С коммутатором

- Используется 1 порт
- Используется 2 порта (обеспечение отказоустойчивости)

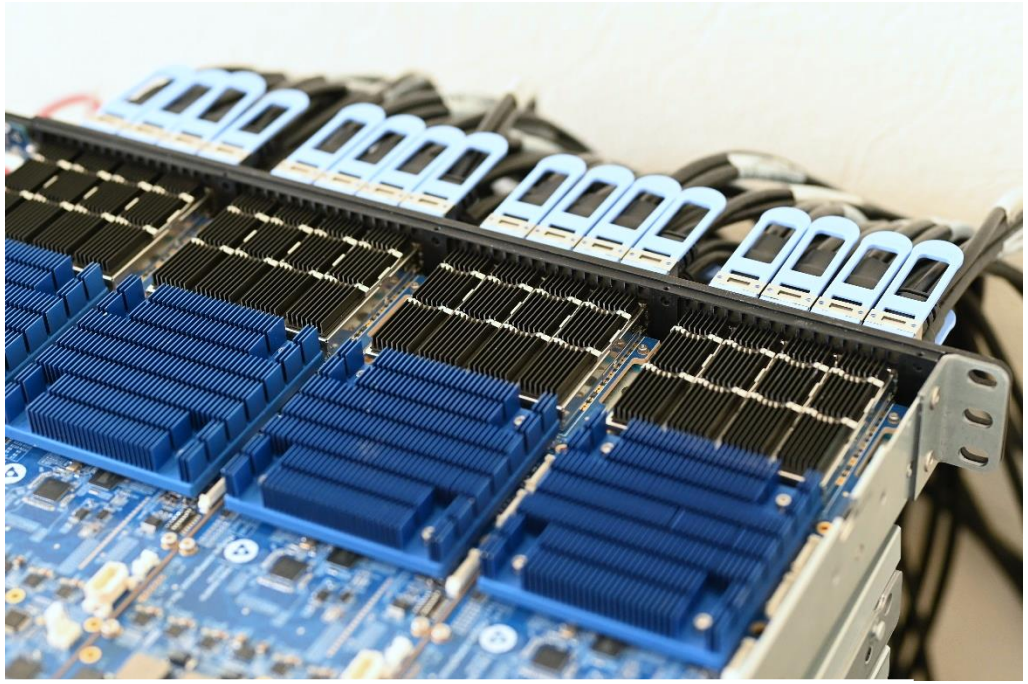


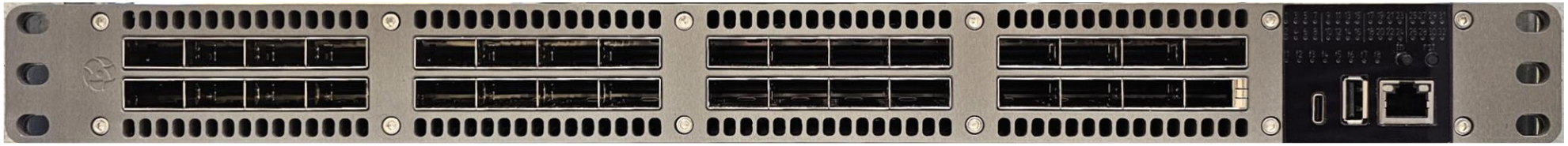
Сетевой коммутатор

РФЯЦ-ВНИИТФ
РОСАТОМ

Особенности

- Монтажная единица: 1U
- 32 порта QSFP-DD
- Представляет собой 4 модуля коммутации, соединенных в кольцо
- Наличие индикации
- Блок питания с резервированием





Аппаратные возможности

Используется детерминированная маршрутизация

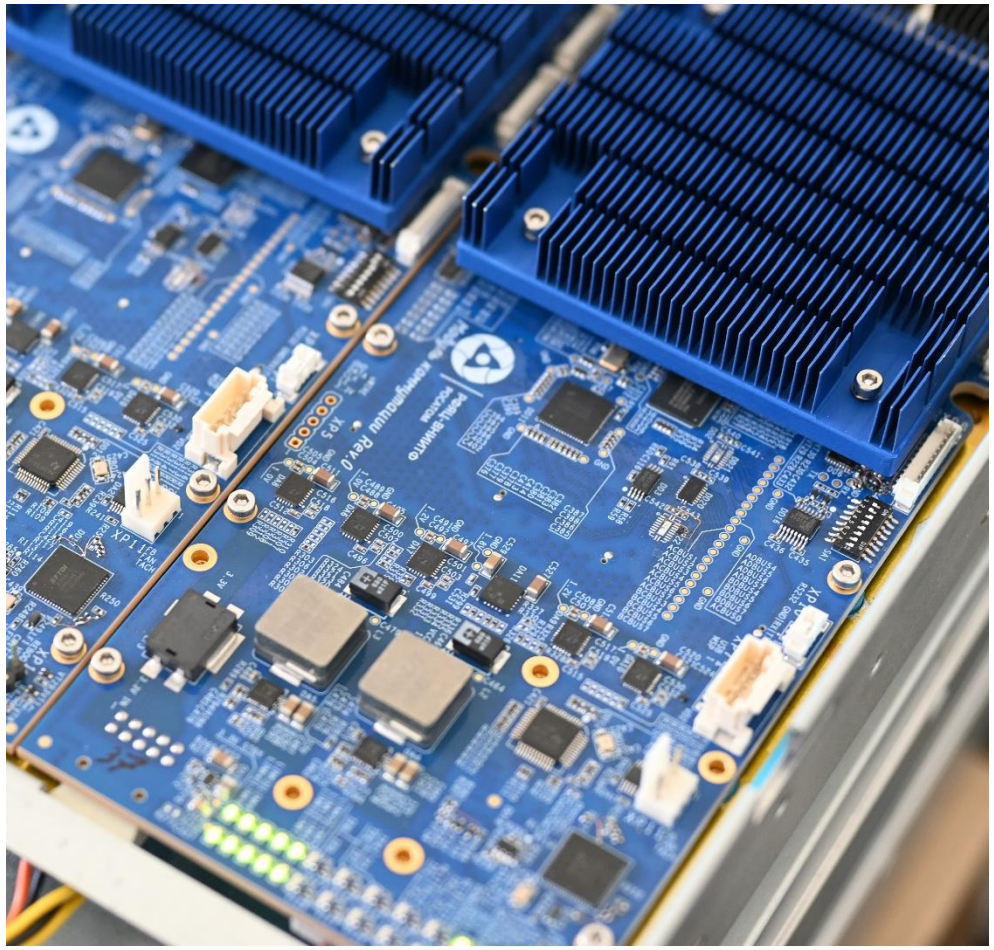
Прямой доступ к памяти удаленного узла

Поддерживаются следующие типы операций:

- PUT – запись данных в память удаленного узла
- GET – чтение данных из памяти удаленного узла
- ADD/CAS – атомарная операция с ячейкой памяти удаленного узла
- FADD/FCAS – атомарная операция с ячейкой памяти удаленного узла с возвратом значения
- INTERRUPTS – вызов прерывания на удаленном узле (поддерживается до 512 прерываний вида MSI-X)
- *_SUFFIX – отправка уведомления о завершении операции на удаленный узел внутри основного пакета

Аппаратная поддержка кольцевого буфера

Гарантированная и негарантированная доставка пакетов



Передача данных

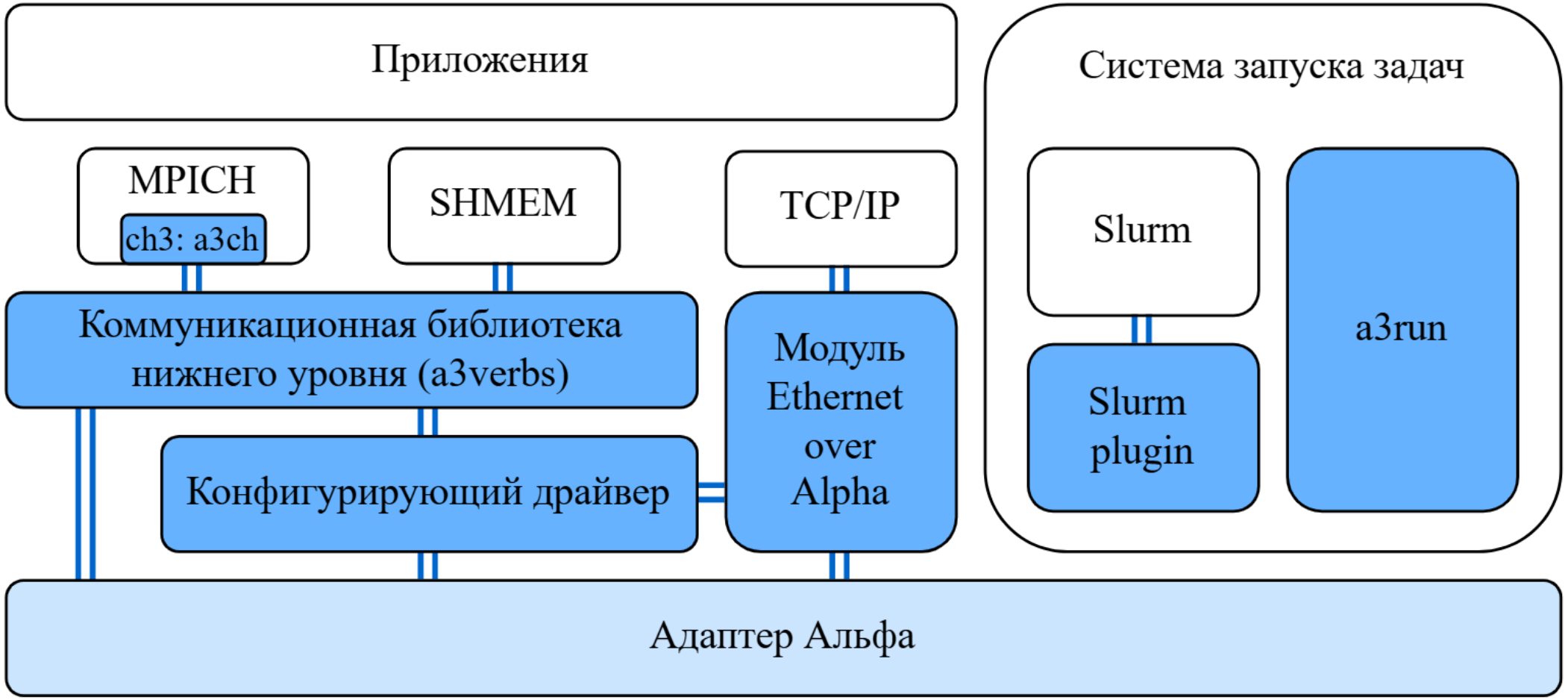
Адресация памяти

- На каждом узле выделяется набор регионов DRAM-памяти, доступных для сетевого адаптера на чтение/запись. Можно задать до 8 разных регионов, размер каждого может достигать 1 ТБ.
- Удалённая память адресуется через номер региона и смещение (OFFSET) относительно его начала, которые указываются в заголовке пакета. Гранулярность задания смещения – 32-битное слово.

Отправка данных

- Сообщение нарезается на пакеты размером до 512 байт
- Для каждого пакета формируется заголовок, в котором указывается:
  - тип операции
  - номер DMA региона
  - смещение (OFFSET) внутри DMA региона
  - размер данных
  - другая информация
- Готовый пакет копируется в инжекционный конвейер адаптера (кольцевой буфер)
- Поддерживается два сценария отправки данных:
  - PIO – готовый пакет копируется в инжекционный конвейер адаптера
  - DMA – адаптер самостоятельно работает с памятью хоста (формирует, отправляет и получает пакеты)

Архитектура ПО

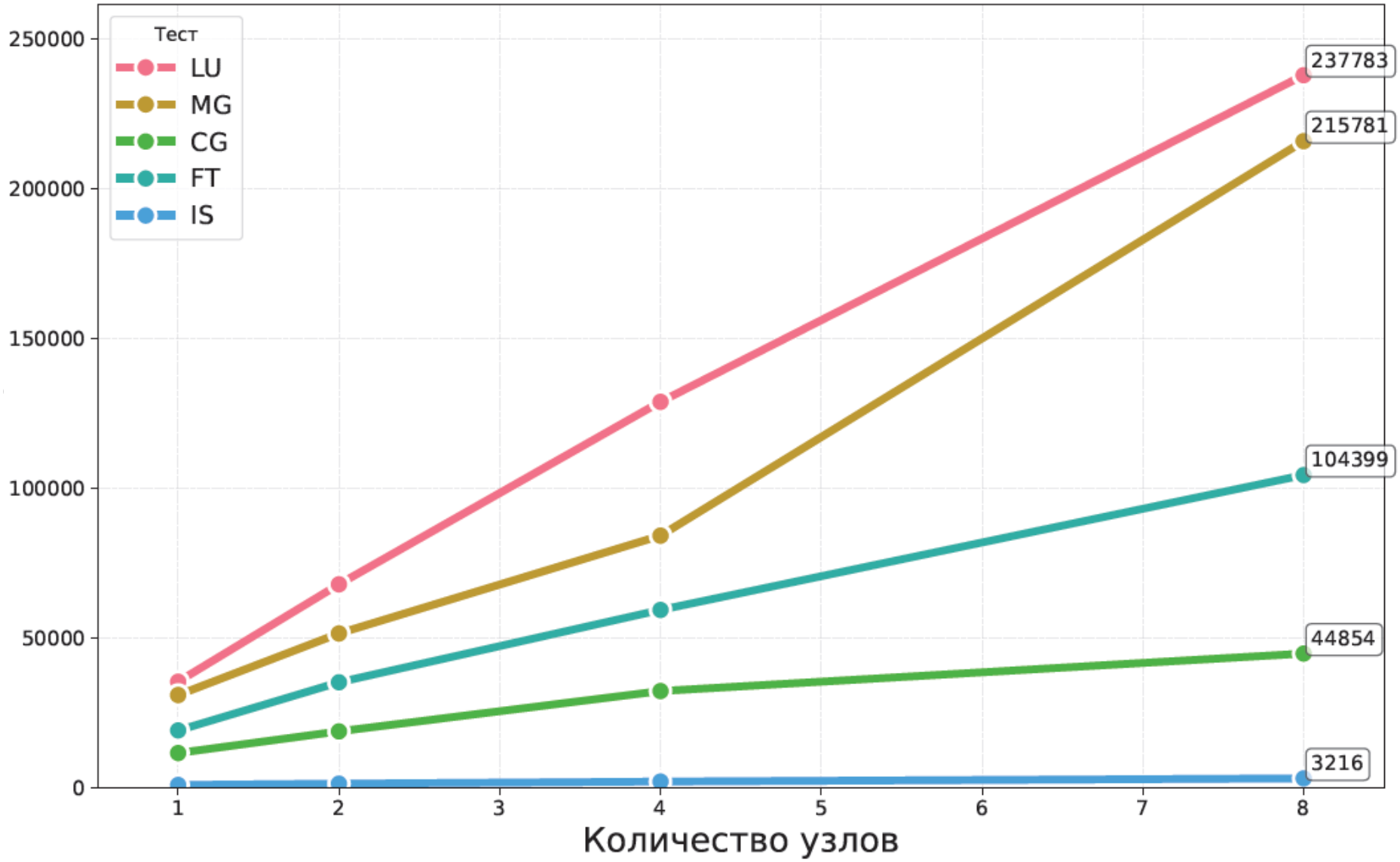


Результаты тестирования, NPB

РФЯЦ-ВНИИТФ
РОСАТОМ

Кластер и ПО

- 8 узлов:
  2x Intel Xeon Gold 6330,
  256 ГБ
- NPB 3.3.1
- MPICH 4.3.0
- 8 процессов на узел



Тестирование и совместимость

РФЯЦ-ВНИИТФ
РОСАТОМ

Аппаратная совместимость

Успешное совместное тестирование

РСК

ЦИАМ

kraftway
РОССИЙСКИЕ ТЕХНОЛОГИИ

ИВК

ТеСИС

Архитектура ПАК СУБД

ПАК СУБД

РФЯЦ-ВНИИТФ
РОСАТОМ

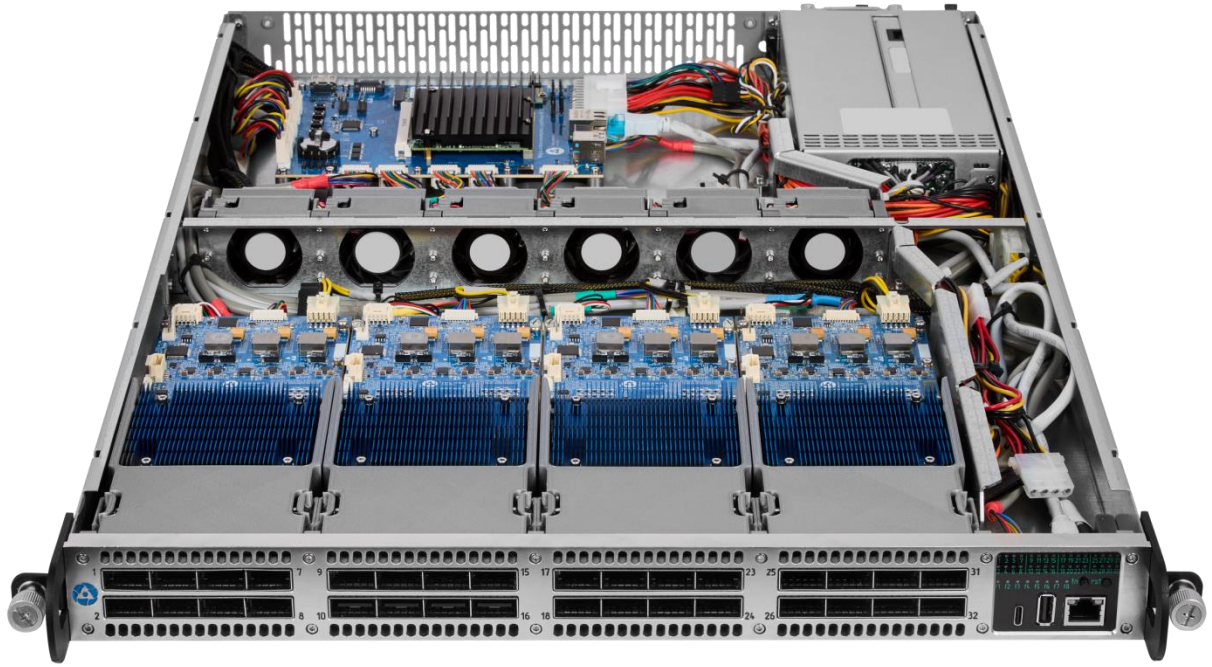
Преимущества

- Горизонтальное масштабирование путём увеличения размера кластера серверов БД вплоть до 100 серверов
- Сверхмалая коммуникационная задержка сети Альфа позволяет сократить время на выполнение синхронного commit, что обеспечивает повышение темпа обработки зависимых по данным транзакций

Результаты

- Повышение темпа обработки транзакций PostgresSQL более, чем в 2 раза на сети Альфа по сравнению с сетью 100 Гбит/с Ethernet
- Увеличение допустимого размера БД до 1 ПБ

Контакты

Симонов Алексей Сергеевич

AleSeSimonov@rosatom.ru