Машина баз данных Скала^р МБД.П

Программно-аппаратный комплекс на основе СУБД Postgres для оперативной обработки транзакций в высоконагруженных системах

Технический обзор

версия 2.11 от 14.01.2026

skala-r.ru info@skala-r.ru

ОГЛАВЛЕНИЕ

УВЕДОМЛЕНИЕ

Информация, представленная в документе, носит исключительно информационный характер, является актуальной на дату размещения.

Технические характеристики, приведенные в документе — справочные и не могут служить основанием для претензий.

Технические характеристики могут отличаться от приведенных вследствие модификации изделий, и могут быть изменены производителем без уведомления.

Документ не является публичной офертой и не содержит каких-либо обязательств ООО «СКАЛА-Р».

ПЕРЕЧЕНЬ ТЕРМИНОВ И СОКРАЩЕНИЙ

Термин, сокращение	Определение
HDD	(англ. Hard disk drive) Твердотельный накопитель данных без подвижных частей, использующий ферромагнитные пластины для хранения информации
IP-адрес	(от англ. Internet Protocol) Уникальный числовой идентификатор, присваиваемый устройству в сети для его идентификации и адресации
IPMI	(англ. Intelligent Platform Management Interface) Интеллектуальный интерфейс управления платформой для автономного мониторинга и управления функциями, встроенными непосредственно в аппаратное и микропрограммное обеспечение серверных платформ
Postgres	Название СУБД (системы управления базами данных)
SSD	(англ. Solid-State Drive) Запоминающее устройство на основе микросхем памяти
БД	База данных
ЕРРП	Единый реестр российских программ (Минцифры)
ЕРРРП	Единый реестр российской радиоэлектронной продукции Минпромторга РФ
ОС	Операционная система
ПО	Программное обеспечение
РЭП МПТ	Единый реестр российской радиоэлектронной продукции Минпромторга РФ
СПО	Специальное программное обеспечение
СУБД	Система управления базами данных, сокр.
ЦОД	Центр обработки данных

1 ПРЕДИСЛОВИЕ

Описание документа

Этот технический обзор дает концептуальный и архитектурный обзоры **Машины баз данных Скала^р МБД.П**.

Брошюра описывает то, как оптимизированные программно-аппаратные комплексы отвечают современным вызовам, и фокусируется на **Машине баз данных Скала^р МБД.П** как одном из лидирующих решений в этом сегменте.

Аудитория

Эта брошюра предназначена для сотрудников компании **Скала^р**, партнеров и заказчиков, перед которыми ставятся задачи разработки решения, закупки, управления или эксплуатации **Машины баз данных Скала^р МБД.П**.

Обратная связь

Скала^р и авторы этого документа будут рады обратной связи по нему. Свяжитесь с командой **Скала^р** по электронной почте support_mbd@skala-r.ru.

2 ВВЕДЕНИЕ

Машина баз данных Скала^р МБД.П — это модульный программно-аппаратный комплекс для обработки и хранения данных, специально предназначенный для ‘эксплуатации СУБД Postgres Pro в составе высоконагруженных информационных систем.

Машина баз данных Скала^р МБД.П позволяет обеспечить повышенную производительность и отказоустойчивость информационных систем, в сочетании со снижением затрат за счет проработанной интеграции аппаратного и программного обеспечения, применения отказоустойчивых архитектур и эксклюзивных моделей лицензирования ПО используемого в составе комплекса.

Машина баз данных Скала^р МБД.П предназначена для размещения высоконагруженных баз данных объемом от 10 до 160 Тбайт, в зависимости от выбранного приоритета производительности или объема.

Машина баз данных Скала^р МБД.П — комплексное законченное решение, включающее в себя все необходимое от Модулей вычисления и хранения данных, включая Модуль резервного копирования, до телекоммуникационных модулей, обеспечивающих сверхскоростную сетевую среду и интерфейсы интеграции, а также систему интеллектуального управления.

Высокая производительность решения достигается в том числе применением оптимальных по производительности комплектующих и современных стандартов, накопителей SSD, сетевых протоколов 100 Gigabit Ethernet.

Помимо стандартных методов обеспечения отказоустойчивости, дублирование, избыточность, балансировка, комплектующие Машины проходят строгий отбор и валидацию, используются специализированные версии ПО, обеспечивающие возможности кластеризации (к примеру специализированная версия СУБД Postgres Pro Enterprise), а также обязательное резервирование критических компонентов и устойчивые сетевые протоколы.

Машина баз данных Скала^р МБД.П содержит все необходимые элементы для эксплуатации высоконагруженной СУБД Postgres. Физическое подключение к внешним сетям передачи данных осуществляется с помощью стандартного интерфейса Ethernet.

Реализованы функции мониторинга состояния как аппаратных, так и программных компонентов решения, а также необходимые функции управления.

Машина баз данных Скала^р МБД.П допускает размещение в одном Модуле сразу нескольких баз данных, предоставляя возможности для их консолидации и снижения стоимости эксплуатации.

Машина баз данных Скала^р МБД.П впервые была представлена в 2015 году как продукт в линейке ПАК СКАЛА-Р СР/П. С тех пор, на основании накопленного опыта эксплуатации, комплекс был значительно усовершенствован и переработан.

Решение внедрено в крупных корпоративных и государственных организациях, инсталляционная база составляет более 2000 узлов.

Программно-аппаратные комплексы **Машина баз данных Скала^р МБД.П** и составляющие их Модули включены в ЕРРРП и работают на ПО, включенном в ЕРРП. **Машина баз данных Скала^р МБД.П** также находится в ЕРРП.

3 ОТЛИЧИТЕЛЬНЫЕ ЧЕРТЫ

1. Надежное хранение и высокопроизводительная обработка больших объемов данных

- Объем баз данных до 80 Тбайт
- Объем баз данных до 160 Тбайт при средних и низких нагрузках
- Производительность более 135 ktpsB по тестам pgbench
- Формирование катастрофоустойчивых решений

2. Высокая производительность

- Сбалансированный комплект оборудования
- Оптимизация архитектуры с прицелом на обеспечение высокой производительности
- Оптимизированная локальная система хранения
- Специализированные настройки используемого программного обеспечения
- Особые алгоритмы резервного копирования и восстановления
- Проработанные варианты для типовых применений

3. Отказоустойчивость на всех уровнях

- Надежные комплектующие
- Резервирование значимых компонентов на аппаратном уровне
- Отказоустойчивая архитектура СУБД и резервного копирования
- Оперативная восстанавливаемость при сбоях (минимальные значения RTO и RPO)

4. Приоритет сохранности данных

- Полные и инкрементальные копии баз данных
- Хранение архивных журналов
- Защита кэша при сбое питания на аппаратных RAID

5. Обеспечение качества при развертывании

- Оптимальность настроек проверена тестами
- Автоматизированное развертывание исключает человеческие ошибки
- Стандартизация развертывания гарантирует соответствие решения заявленным характеристикам

6. Непрерывный контроль состояния

- Встроенная система мониторинга **Скала^р Визион** специально разработана с учетом опыта эксплуатации на стороне заказчиков и особенностей поставляемых программно-аппаратных комплексов
- Мониторинг работоспособности СУБД и оборудования
- Преднастроенные пороговые значения критичных параметров
- Различные каналы информирования о достижении пороговых значений ключевых метрик и отклонениях в работе

7. Улучшенные возможности администрирования

- Встроенная система автоматизации администрирования и обслуживания Машины **Скала^р Геном,** специально разработанная с учетом опыта эксплуатации на стороне заказчиков и особенностей поставляемых программно-аппаратных комплексов

- Автоматизированные действия по выполнению сложных операций с кластером и узлами
- В целях снижения трудозатрат по внедрению в существующие решения сохранены все стандартные механизмы управления Postgres

8. Обеспечение эксплуатации

- Централизованная поддержка решения
- Единая точка ответственности за весь комплекс
- Выпуск исправлений и рекомендаций
- Паспорт Машины в комплекте и в системе **Скала^р Геном**
- Обучение персонала заказчика
- Автоматизация управления жизненным циклом изделия
- Продвинутое управление быстрым резервным копированием и восстановлением баз данных

9. Экономическая эффективность

- Специальные условия по лицензированию СУБД Postgres Pro Enterprise
- Сокращенные сроки ввода в эксплуатацию
- Только необходимые для корпоративного решения компоненты

10. Альтернатива Oracle Exadata для транзакционных и гибридных нагрузок

- Готовая, сбалансированная, отказоустойчивая и полностью отлаженная серийная Машина баз данных для СУБД Postgres
- Высокие надежность и производительность
- Качество, подтвержденное опытом практического применения

4 ПОДТВЕРЖДЕННАЯ БЕЗОПАСНОСТЬ

Машина баз данных Скала^р МБД.П поставляется с сертифицированной ОС Альт СП (сертификат ФСТЭК 3866 от 10.08.2018, действует до 10.08.2028), которая:

Может применяться для защиты информации:

- В значимых объектах критической информационной инфраструктуры 1 категории, в государственных информационных системах 1 класса защищенности
- В автоматизированных системах управления производственными и технологическими процессами 1 класса защищенности
- В информационных системах персональных данных при установленном УЗ-1 (первый уровень защищенности персональных данных)
- В информационных системах общего пользования II класса

Соответствует требованиям следующих нормативных документов:

- «Требования безопасности информации к операционным системам» (ФСТЭК России, 2016) и «Профиль защиты операционных систем типа А четвертого класса защиты. ИТ.ОС.А4.ПЗ» (ФСТЭК России, 2017) по 4 классу защиты
- «Требования по безопасности информации к средствам контейнеризации» (ФСТЭК России, 2022, приказ № 118) по 4 классу защиты
- «Требования по безопасности информации к средствам виртуализации» (ФСТЭК России, 2022, приказ № 187) по 4 классу защиты
- «Требования по безопасности информации, устанавливающие уровни доверия к средствам технической защиты информации и средствам обеспечения безопасности информационных технологий» (ФСТЭК России, 2020, приказ № 76) по 4 уровню доверия
- «Требования по безопасности информации к системам управления базами данных» (ФСТЭК России, 2023) – по 4 классу защиты и техническим условиям 643.20663116.00002-12 ТУ

В **Машине баз данных Скала^р МБД.П** используется сертифицированная СУБД Postgres Pro Enterprise (сертификат ФСТЭК 4063 от 16.01.2019), которая:

Может применяться для защиты информации:

- В значимых объектах критической информационной инфраструктуры 1 категории, в государственных информационных системах 1 класса защищенности
- В автоматизированных системах управления производственными и технологическими процессами 1 класса защищенности
- В информационных системах персональных данных при необходимости обеспечения 1 уровня защищенности персональных данных
- В информационных системах общего пользования II класса

Соответствует требованиям следующих нормативных документов:

- «Требования по безопасности информации, устанавливающие уровни доверия к средствам технической защиты информации и средствам обеспечения

безопасности информационных технологий» (ФСТЭК России, 2020) — по 4 уровню доверия

- «Требования по безопасности информации к системам управления базами данных» (ФСТЭК России, 2023) – по 4 классу защиты

Протестирована совместимость с наложенными средствами защиты:

1. Сертифицированное антивирусное средство защиты Kaspersky Endpoint Security для Linux (сертификат ФСТЭК 2534 от 27.12.2011, действует до 27.12.2030):

Применяется для защиты информации:

- В государственных информационных системах I класса защищённости
- В информационных системах персональных данных при необходимости обеспечения 1 уровня защищённости персональных данных
- В значимых объектах критической информационной инфраструктуры 1 категории
- В автоматизированных системах управления производственными и технологическими процессами I класса защищённости
- В информационных системах общего пользования II класса

Соответствует требованиям нормативных документов:

- «Требования по безопасности информации, устанавливающие уровни доверия к средствам технической защиты информации и средствам обеспечения безопасности информационных технологий» (ФСТЭК России, 2020) - по 2 уровню доверия
- «Требования к средствам антивирусной защиты» (ФСТЭК России, 2012)
- «Профиль защиты средств антивирусной защиты типа Б второго класса защиты. ИТ.САВ3.Б2.П3» (ФСТЭК России, 2012)
- «Профиль защиты средств антивирусной защиты типа В второго класса защиты. ИТ.САВ3.Б2.П3» (ФСТЭК России, 2012)
- «Профиль защиты средств антивирусной защиты типа Г второго класса защиты. ИТ.САВ3.Б2.П3» (ФСТЭК России, 2012)
- «Требования к средствам контроля сменных машинных носителей информации» (ФСТЭК России, 2014)
- «Профиль защиты средств контроля подключения сменных машинных носителей информации второго класса защиты. ИТ.СКН.П2.П3» (ФСТЭК России, 2014)

2. Сертифицированное средство доверенной загрузки ПАК «Соболь» версия 4 (сертификат ФСТЭК 4043 от 05.12.2018, действует до 05.12.2028)

Применяется для защиты информации:

- В государственных информационных системах I класса защищённости
- В информационных системах персональных данных при необходимости обеспечения 1 уровня защищённости персональных данных
- В значимых объектах критической информационной инфраструктуры 1 категории
- В автоматизированных системах управления производственными и технологическими процессами I класса защищённости
- В информационных системах общего пользования II класса

Соответствует требованиям нормативных документов:

- «Требования по безопасности информации, устанавливающие уровни доверия к средствам технической защиты информации и средствам обеспечения безопасности информационных технологий» (ФСТЭК России, 2020) - по 2 уровню доверия
- «Требования к средствам доверенной загрузки» (ФСТЭК России, 2013), «Профиль защиты средства доверенной загрузки уровня платы расширения второго класса защиты. ИТ.СДЗ.ПР2.ПЗ» (ФСТЭК России, 2013).»

3. Сертифицированная система единой аутентификации Avanpost FAM (сертификат ФСТЭК 4492 от 13.12.2021, действует до 13.12.2026)

Применяется для защиты информации:

- в значимых объектах критической информационной инфраструктуры 1 категории
- в государственных информационных системах 1 класса защищенности
- в автоматизированных системах управления производственными и технологическими процессами 1 класса защищенности
- в информационных системах персональных данных при необходимости обеспечения 1 и 2 уровня защищенности персональных данных

Соответствует требованиям руководящих документов:

«Требования по безопасности информации, устанавливающие уровни доверия к средствам технической защиты информации и средствам обеспечения безопасности информационных технологий» (ФСТЭК России, 2020, приказ № 76) по 4 уровню доверия

5 ПРОИЗВОДСТВО В РОССИЙСКОЙ ФЕДЕРАЦИИ

Специалистами компании **Скала^р** была проведена существенная работа по созданию схем и конструктивного исполнения **Машины баз данных Скала^р МБД.П**, основанного на принципе модульности. Результаты проведенной работы на сегодняшний день не имеют аналогов на рынке РФ.

Машина баз данных Скала^р МБД.П включена в реестр российской радиоэлектронной продукции и в реестр российской промышленной продукции.

Машина баз данных Скала^р МБД.П поставляется единым комплексом, одной номенклатурной позицией как Программно-аппаратный комплекс (ПАК). При этом Машина состоит из набора отдельных Модулей (каждый из которых также является изделием в реестре РЭП МПТ), что обеспечивает гибкость комплектации и модернизации товарными позициями из реестра.

Машина баз данных Скала^р МБД.П признана произведенным в РФ товаром, в соответствии с Правилами выдачи заключения о подтверждении производства промышленной продукции на территории Российской Федерации, утвержденными постановлением Правительства от 17 июля 2015 г. № 719.

Машина баз данных Скала^р МБД.П соответствует постановлению Правительства РФ № 616 от 30 апреля 2020 г. о запрете на закупку импортной радиоэлектронной продукции и постановлению Правительства РФ № 925 от 16 сентября 2016 г. о приоритете российской радиоэлектронной продукции в 30%.

Машина баз данных Скала^р МБД.П соответствует постановлению Правительства РФ № 2013 и № 2014 от 03 декабря 2020 г. о минимальной доле закупок товаров российского происхождения.

ВНИМАНИЕ! Реестровое написание наименования **Машины баз данных СКАЛА-Р МБД.П** отличается от маркетингового написания с применением товарного знака **Скала^р**.

Товарные позиции Машин и Модулей **Скала^р** имеют один код по ОКПД 2:

- **26.20.14.160** Программно-аппаратные комплексы, созданные на серверах или устройствах, содержащие в своем составе один или более вычислительных узлов
- **26.20.14.160** Машины вычислительные электронные цифровые, поставляемые в виде систем для автоматической обработки данных

Наличие **Машины баз данных Скала^р МБД.П** на сайте государственной информационной системы промышленности (Рисунок 1).

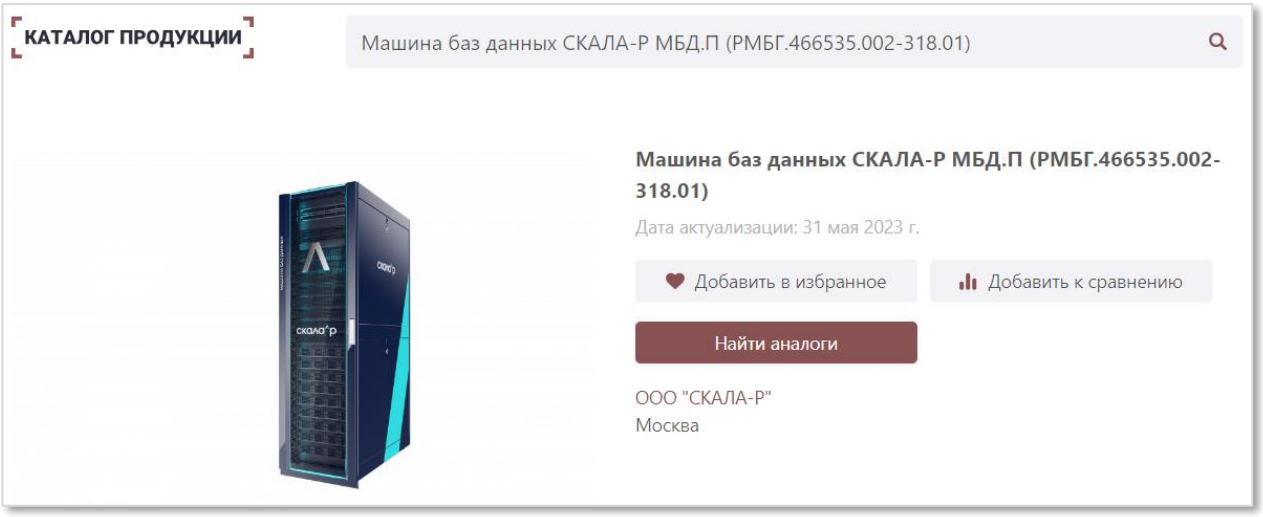

Рисунок 1 — Машина баз данных СКАЛА-Р МБД.П на сайте государственной информационной системы промышленности (ГИСП). Фрагмент страницы https://gisp.gov.ru/goods/#/product/3738080

Примеры информации о **Машине баз данных Скала^р МБД.П**, включенной в ЕРРРП, и Модулях, включенных в РЭП МПТ, представлены в таблицах 1 и 2 соответственно.

Машины и Модули различаются исполнением и платформой (материнской платой производства РФ).

Таблица 1 — Информация о Машинах Скала^р МБД.П, включенных в ЕРРРП

Наименование Машины (разработан согласно Техническим условиям РМБГ.466535.002ТУ)	**Код изделия по ОКПД2**
Машина баз данных СКАЛА-Р МБД.П (РМБГ.466535.002-518)	26.20.14.160

Таблица 2 — Информация об основных Модулях Машин Скала^р МБД.П, включенных в РЭП МПТ

Наименование Модуля (разработан согласно Техническим условиям РМБГ.466535.003ТУ)	**Код изделия по ОКПД2**
СКАЛА-Р Базовый модуль (РМБГ.466535.003-210)	26.20.14.160
СКАЛА-Р Модуль баз данных (РМБГ.466535.003-260)	26.20.14.160
СКАЛА-Р Модуль резервного копирования (РМБГ.466535.003-24)	26.20.14.160

6 ПРИНЦИПЫ ПРОЕКТИРОВАНИЯ

Чтобы лучше понять устройство **Машины баз данных Скала^р МБД.П**, можно сравнить его с традиционно используемым подходом к размещению СУБД на некотором наборе из различных аппаратных и программных компонентов.

Традиционный подход универсален

В состав оборудования, как правило, входит вычислительный узел, подключенный по сети к системе хранения данных. Узел используется для размещения программного обеспечения СУБД, сами данные хранятся в массиве и по мере необходимости передаются по сети, используя стандартные протоколы взаимодействия. Ориентация на стандартные компоненты и протоколы позволяет обеспечить предельную вариативность применения решения, а также возможность подбора компонентов для широкого спектра нагрузок. В то же время такой подход не обеспечивает оптимальности получившегося решения для конкретной задачи, что является обратной стороной универсальности.

Скала^р МБД.П создана для СУБД Postgres

Целью разработки **Машины баз данных Скала^р МБД.П** было создание полного комплекта аппаратного и программного обеспечения, адаптированного под СУБД Postgres для обработки запросов в оптимальной среде. Это позволяет использовать преимущества тонкой настройки всех уровней решения именно под функции и потребности СУБД Postgres и тем самым обеспечивает максимум ее производительности.

Комплексное размещение компонентов, применение высокопроизводительных протоколов и устройств хранения также способствуют достижению этой цели. За исключением Модуля резервного копирования, в **Машине баз данных Скала^р МБД.П** используются накопители SSD.

Быстродействие и емкость современных твердотельных накопителей позволили отказаться от использования отдельной системы хранения в **Машине баз данных Скала^р МБД.П**. Примененный подход позволяет вычислительным ресурсам непосредственно обращаться к данным, исключая необходимость их выборки на стороне системы хранения и пересылки по сети, что также положительно сказывается на производительности решения.

Машина баз данных Скала^р МБД.П поддерживает возможность переноса читающей нагрузки на реплики, высвобождая ресурсы ведущего узла под транзакционную нагрузку, выполняя ресурсоемкие выборки, с синхронной или с асинхронной реплики.

Проработанность всех программных компонентов

Основные программные элементы **Машины баз данных Скала^р МБД.П** включают в себя ПО СУБД Postgres, ПО мониторинга и администрирования, ПО управления кластером СУБД, ПО резервного копирования, но не ограничиваясь представленным списком.

В **Машине баз данных Скала^р МБД.П** обеспечена оптимизация, тонкая настройка ОС и доработка перечисленных компонентов для обеспечения их большей производительности и функционального соответствия потребностям решения в целом.

Интеллектуальное ПО Скала^р МБД.П

Практическое применение первых экземпляров **Машин баз данных Скала^р МБД.П** продемонстрировало высокую производительность решения, в то же время был выявлен ряд направлений для улучшения архитектуры аппаратной конфигурации с собственной схемотехникой.

Преодоление границ аппаратных возможностей оборудования в составе **Машины баз данных Скала^р МБД.П**, помимо масштабирования, возможно за счет развития собственного ПО Скала^р для оптимизации работы СУБД в среде ОС без балансировщика и в условиях высоких нагрузок автоматизированных банковских систем с большим количеством конкурентных клиентских соединений.

В ходе развития **Машины баз данных Скала^р МБД.П** были оптимизированы **настройки ядра операционной системы** узлов БД под конкретный вариант ее применения.

Оптимизация функционирования СУБД Postgres достигается путем изменения настраиваемых параметров для обеспечения лучшего соответствия архитектуре решения в целом, без внесения изменений во внутренние алгоритмы СУБД, что гарантирует совместимость решения с прикладным ПО, ориентированным на соответствующую версию СУБД.

Отказоустойчивость СУБД Postgres в Машине баз данных Скала^р МБД.П обеспечивается путем размещения экземпляров СУБД на трех различных узлах БД, образующих кластер. При возникновении отказа осуществляется автоматическое переключение роли мастер-СУБД на одну из реплик. При этом поддержание полной консистентной копии мастер-базы данных на репликах реализуется механизмом потоковой репликации, который позволяет передавать все изменения с ведущего узла БД на ведомые.

Применяемое в **Машине баз данных Скала^р МБД.П** ПО **Скала^р Спектр** позволяет обеспечить защиту от различных отказов, в том числе от сбоев по питанию; от сбоев процессов СУБД Postgres, связанных с нехваткой памяти, недостатком файловых дескрипторов, превышением максимального числа открытых файлов; от потерь сетевой связности между узлами кластера и других.

При тех или иных отказах и нестандартных ситуациях ПО управления кластером применяет соответствующий алгоритм реагирования. В критичных ситуациях кластер может быть остановлен для обеспечения сохранности данных.

В целом это одна из наиболее сложных задач, эффективное решение которой зависит от конкретных требований заказчика, особенностей прикладного программного обеспечения, информационно-технологической и сетевой среды инфраструктуры заказчика. В указанных условиях ряд настроек осуществляется непосредственно при развертывании решения.

Внедрение современных аппаратных RAID-контроллеров привело к значительному росту производительности и позволило повысить надежность RAID-массивов.

В **Машине баз данных Скала^р МБД.П** была существенно модернизирована подсистема управления резервным копированием. Улучшения затронули ключевые алгоритмы и настройки ПО, что позволило радикально сократить как время создания резервных копий, так и время восстановления узлов кластера. Эти показатели являются критически важными для обеспечения непрерывности бизнес-процессов не только в случае серьёзных сбоев, но и при проведении плановых работ и устранении незначительных инцидентов.

В состав **Машины** и Модулей входят собственные программные разработки ООО «СКАЛА-Р» — ПО управления кластером **Скала^р Спектр**, собственный мониторинг **Скала^р Визион** и средство администрирования **Скала^р Геном**.

В комплексе все перечисленные направления формируют целостную систему, формирующую интеллектуальную составляющую **Машины баз данных Скала^р МБД.П**.

Сопровождение и поддержка

Важным дополнением ко всему перечисленному является полная ответственность производителя за решение в целом, включая все его программные и аппаратные компоненты. Это означает не только уверенность в работоспособности изделия в целом, но и последующую поддержку от единого поставщика в режиме единого окна, а не от нескольких разных поставщиков, как бывает при самостоятельном подборе, развертывании и настройке компонентов в случае традиционного подхода.

7 СОСТАВ РЕШЕНИЯ

Термины, используемые для комплектации **Машины баз данных Скала^р МБД.П**:

Машина — это набор аппаратного и программного обеспечения в виде Модулей **Скала^р**, соединенных вместе для обеспечения определенного метода обработки данных или предоставления ИТ сервиса с заданными характеристиками.

Модуль — это структурный элемент **Машины баз данных Скала^р МБД.П**, выполняющий определенные функции в соответствии с их назначением. Он является единым и неделимым элементом спецификации, содержит набор аппаратных узлов и программного обеспечения (ПО).

Узел — это элемент Модуля, выполняющий определенную задачу в составе Модуля.

Секция (Стойка) — набор Модулей **Машины**, объединенных в один серверный шкаф.

Формирование решения основано на принципе разделения на Модули.

Машина баз данных Скала^р МБД.П состоит из Модулей, показанных на рисунке 2:

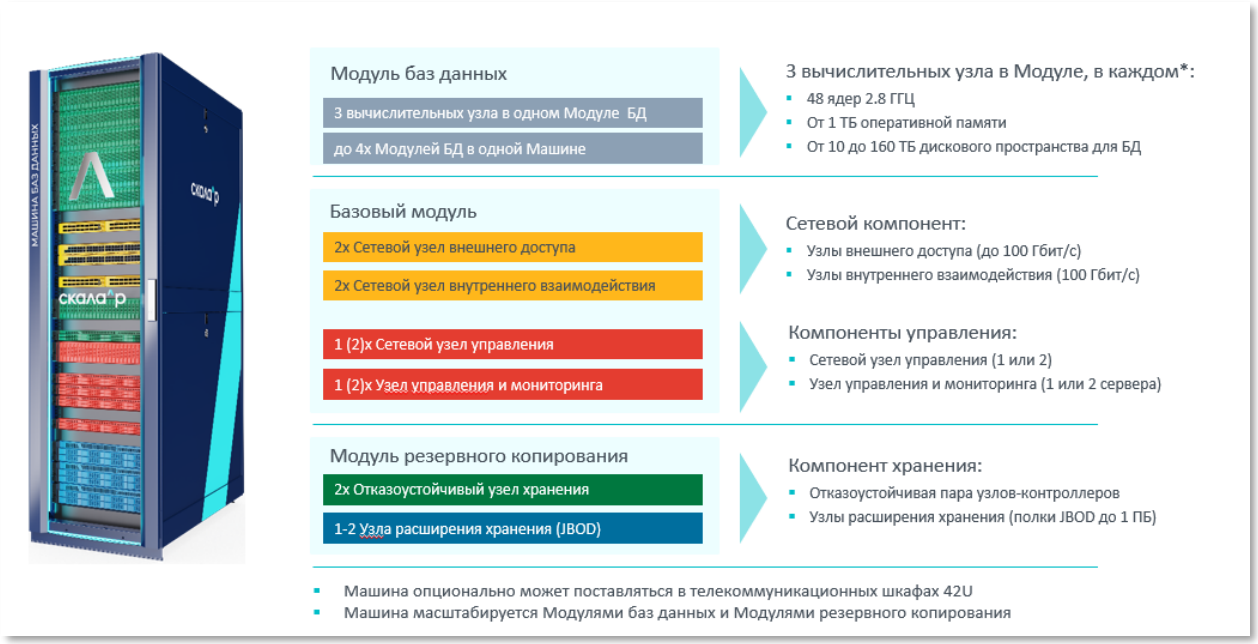

Рисунок 2 — Модули Машины Скала^р МБД.П

Машина баз данных Скала^р МБД.П основана на принципе модульности. Компонуется из набора стандартных Модулей, чем обеспечивается универсальный подход, более высокий уровень технологичности и надежности эксплуатации.

Машины баз данных Скала^р МБД.П могут поставляться в различных комплектациях и исполнениях с разными Модулями. В зависимости от требований к производительности и емкости хранения, состав Машины и Модулей подбирается под целевые показатели заказчика.

Машина баз данных Скала^р МБД.П поставляется как готовый преднастроенный комплекс, однако в процессе эксплуатации состав Машины и Модулей может расширяться для увеличения объема хранимых данных или производительности.

Для обеспечения отказоустойчивости и высокой производительности при проектировании **Машины баз данных Скала^р МБД.П** были заложены следующие технологические принципы*:

T P — дублирование критичных компонентов

P — равномерное распределение нагрузки на доступные ресурсы

T — сохранение работоспособности при отказе отдельных элементов системы (в отдельных случаях — со снижением производительности)

***Примечание** — здесь и далее по тексту отдельные перечисляемые характеристики помечены символом T в случае, если они ориентированы на обеспечение отказоустойчивости (Fault Tolerance), и символом P, если они ориентированы на обеспечение производительности (Performance).*

Базовый модуль содержит узел управления и мониторинга, в котором находится ПО, обеспечивающее развертывание/обновление системы эксплуатации и отвечающее за управление кластерами системы, выполнение резервного копирования и восстановления системы, а также за систему мониторинга Машины (контроль параметров, сбор и хранение объектов управления, метрик, визуализаций параметров).

Модуль баз данных содержит кластер из трех узлов, в котором находятся базы данных и журналы WAL.

Взаимодействие между узлами кластера **Машины баз данных Скала^р МБД.П** осуществляется с помощью сетевой подсистемы Базового модуля, которая обеспечивает внутренний интерконнект на высокой скорости, имеет выделенную сеть для управления и мониторинга, а также возможность подключения к внешним сетям.

Модуль резервного копирования хранит резервные копии баз данных и архивы WAL.

Схема внутренней коммутации **Машины баз данных Скала^р МБД.П** (для одного служебного узла) представлена на рисунке 3.

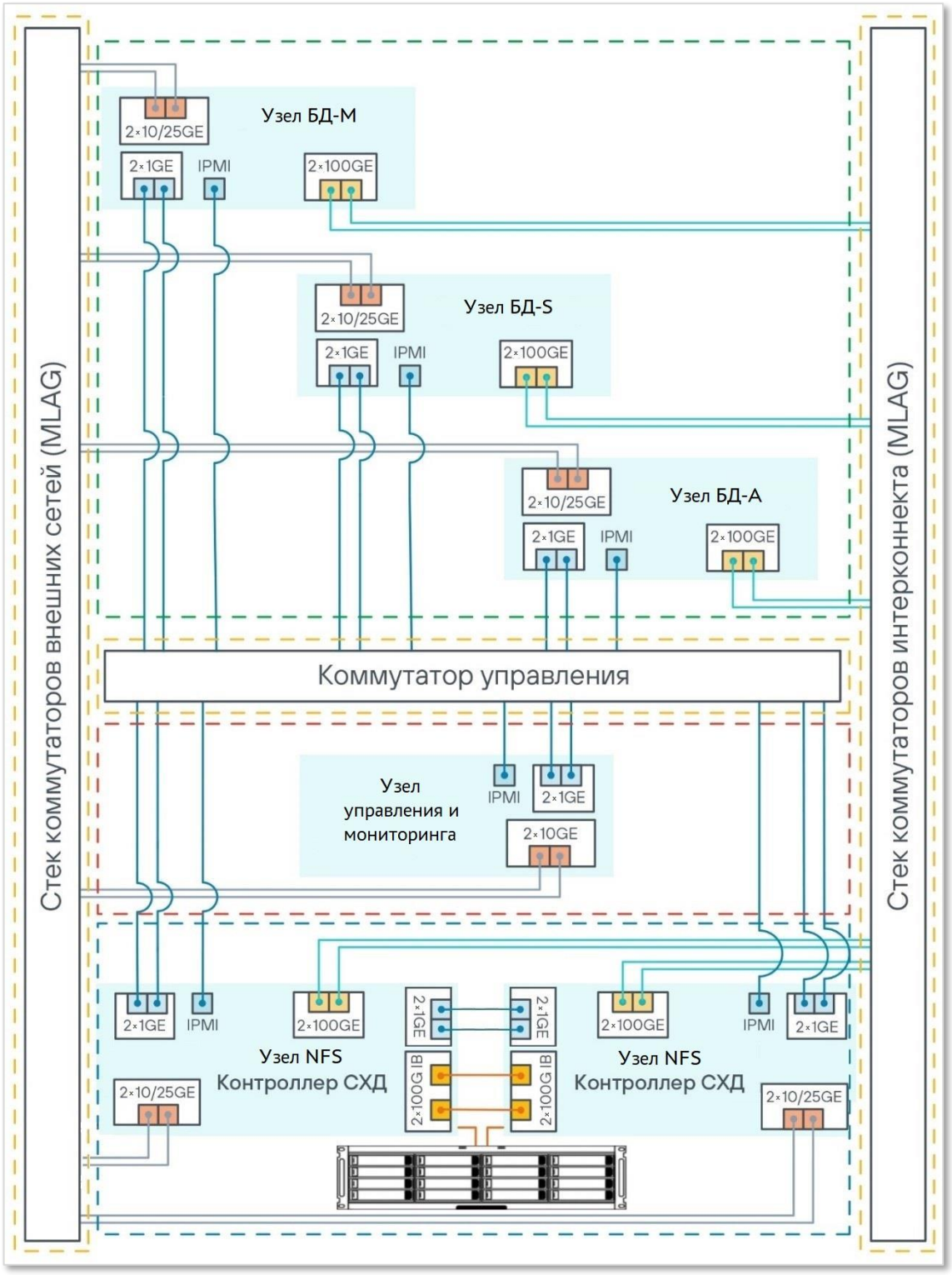

Рисунок 3 — Схема внутренней коммутации Машины баз данных Скала^р МБД.П

Базовый модуль

Базовый модуль состоит из:

- Подсистемы управления и мониторинга
- Сетевой подсистемы

Подсистема управления и мониторинга

Подсистема управления и мониторинга реализована на одном или двух физических узлах, на которые устанавливаются программные продукты разработки **Скала^р**:

- управление эксплуатацией **Машины баз данных Скала^р МБД.П** и ее компонентов: **Скала^р Геном**
- управление кластером узлов базы данных: **Скала^р Спектр**
- система собственного мониторинга: **Скала^р Визион**.

Скала^р Геном хранит репозиторий необходимых пакетов для операционных систем, оборудования и СУБД Postgres Pro.

Для установки операционной системы используются два локальных накопителя, собранных в RAID 1. В качестве операционной системы узла мониторинга используется Альт Сервер 8.4 SP. Для хранения данных мониторинга и репозитория используются локальные диски совместно с операционной системой.

Программный продукт **Скала^р Визион** предназначен для визуализации и мониторинга работы сети и оборудования, входящего в состав комплекса. Объектом мониторинга может быть любой физический или логический объект, например, память, процессор, файловая система, процесс или программа, количество пользователей, очередь файлов на обработку, объем обработанного трафика, значение температуры и другие.

Отличительной особенностью **Скала^р Визион** являются возможности мониторинга за специфичными параметрами Машины, обеспечивающими ее надежность и производительность, что позволяет производить быстрый и качественный анализ ситуаций, строить прогнозы развития ситуации в будущем.

Сбор данных с узлов производится по протоколу IPMI, на уровне операционной системы и СУБД, через установленный агент на узлах, сбор данных с активного сетевого оборудования обеспечивается протоколом SNMP.

Основной функционал:

T P — Управления кластером баз данных (**Скала^р Спектр**)

T — Управления эксплуатацией **Машины баз данных Скала^р МБД.П**, также репозиторий пакетов образов и обновлений (**Скала^р Геном**)

T — Собственного мониторинга и визуализации работы сети и оборудования, входящего в состав **Машины баз данных Скала^р МБД.П** (**Скала^р Визион**).

Пример интерфейса системы управления эксплуатацией представлен на рисунке 4, системы управления кластером - на рисунке 5, системы мониторинга - на рисунке 6.

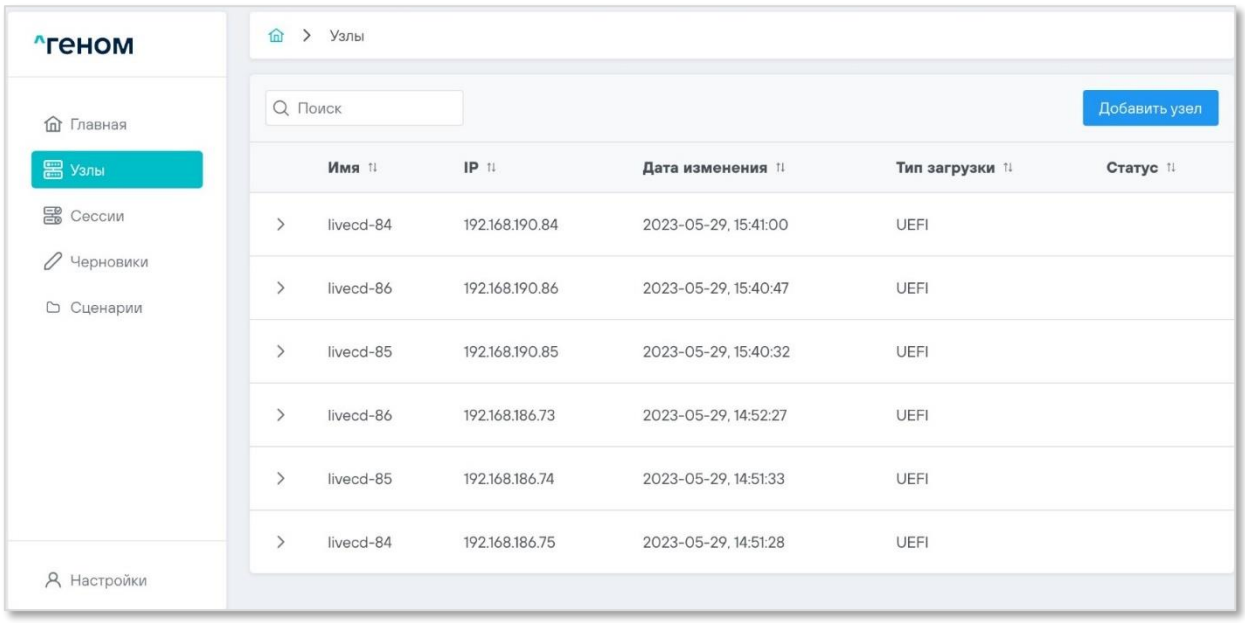

Рисунок 4 — Пример интерфейса системы управления эксплуатацией Скала^р Геном

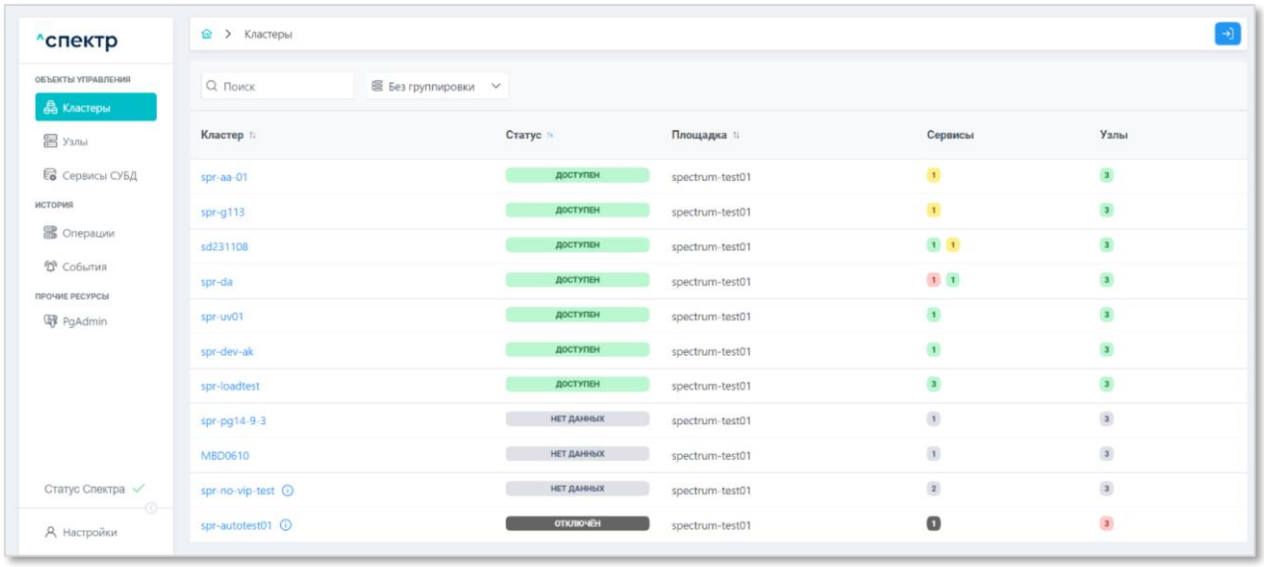

Рисунок 5 — Пример интерфейса системы управления кластерами Скала^р Спектр

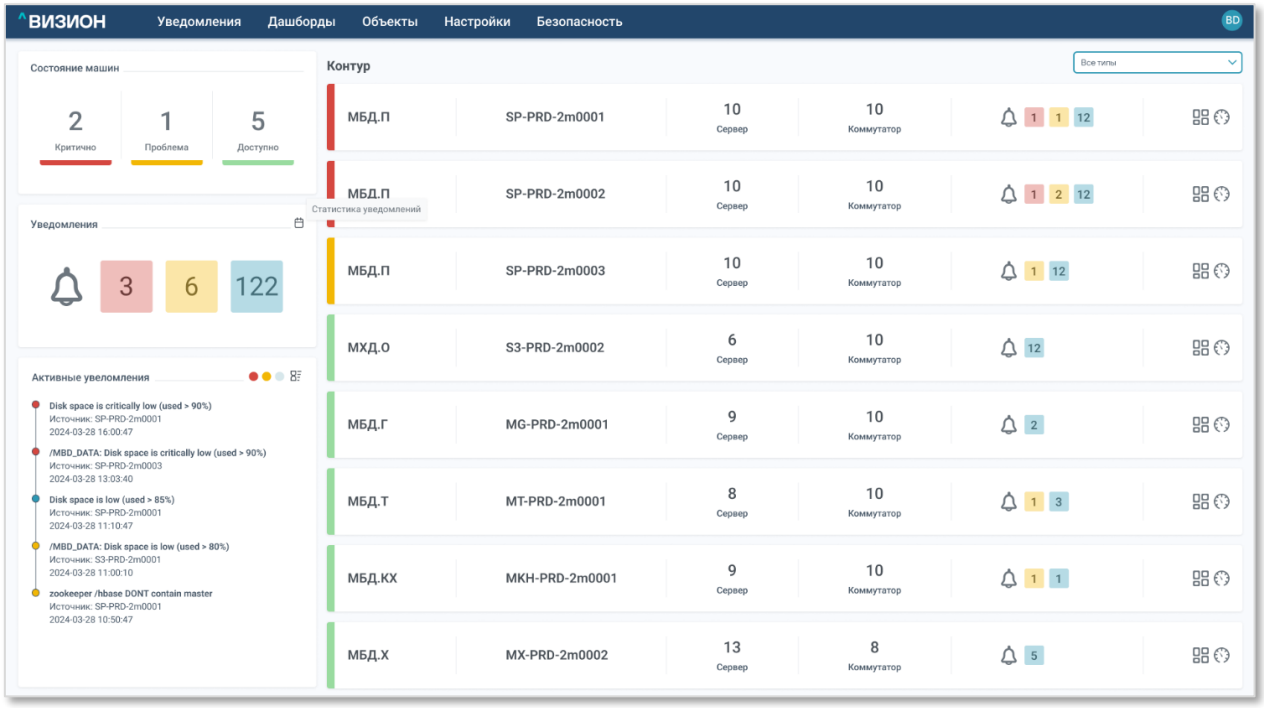

Рисунок 6 — Пример экрана системы мониторинга Скала^р Визион

В системе мониторинга используются всесторонние методы и протоколы сбора информации с объектов контроля, отражающей текущее состояние и значения следующих параметров (если соответствующие датчики установлены в оборудовании):

- по состоянию физических компонентов:
  - температура
  - скорость вращения вентиляторов
  - состояние питания
- по конфигурации:
  - количество CPU
  - объем памяти
  - имя объекта мониторинга
  - список сетевых адаптеров, их MAC- и IP-адреса
  - подключенные ресурсы хранения
- по утилизации ресурсов (счетчики):
  - загрузка CPU (общая)
  - использование памяти (занято/свободно)
  - загрузка подсистемы ввода-вывода (в IOPs или kbytes/sec)
  - использование дисковой подсистемы (свободное/занятое место)
  - утилизация сетевых интерфейсов.

Сбор данных из сетевого оборудования осуществляется через протокол SNMP с использованием частных MIB от производителей оборудования и обеспечивает мониторинг следующих параметров:

- загрузка CPU
- загрузка памяти (абсолютная и в процентах)
- состояние и значения датчиков температуры, состояние PS
- типы интерфейсов устройств
- ошибки на интерфейсах
- контроль загрузки портов сетевых устройств
- контроль пороговых значений SNMP-доступных величин.

Для приведенных параметров объектов мониторинга система позволяет выполнять следующие функции настройки и действия:

- Управление связями между объектами мониторинга
- Настройка условий перехода состояний как для одиночных объектов, так и для групп
- Создание инцидентов и условия генерации оповещений об авариях
- Хранение исторической информации об инцидентах для анализа и предсказания сбоев
- Формирование табличных и графических форм отчетности
- Выбор способов оповещения
- Добавление документации по объекту мониторинга
- Мониторинг корреляции параметров.

Выделенный производительный узел:

P — использование SSD для обеспечения высокой производительности

T — выделенные накопители (RAID 1) для загрузки ОС — обеспечение отказоустойчивости

T P — выделенные накопители для хранения служебных данных (мониторинг, ПО для развертывания, ПО управления кластером и др.)

T P — внешние интерфейсы данных дублированы (стандарт IEEE 802.3ad LACP) — повышение производительности, отказоустойчивость (в случае отказа одного из интерфейсов возможно снижение производительности)

P — 10/25 Gigabit Ethernet — для связи с внешними сетями

T — два блока питания в режиме резервирования по схеме (1 + 1)

P — 2×CPU Xeon (или аналогичный).

Узел мониторинга укомплектован двумя портами Ethernet 1 Гбит/с и двумя портами Ethernet 10/25 Гбит/с.

На базе портов 10/25 Гбит/с создается группа агрегации в режиме 802.3ad LACP, которая представляет собой на уровне операционной системы один логический bond-интерфейс. Данный bond-интерфейс предназначен для пользовательского взаимодействия с узлом мониторинга и предоставляемым им сервисом.

Порты Ethernet 1 Гбит/с используются для служебного трафика автоматизированных процессов установки, управления конфигурациями ОС, СПО. Эта сеть изолирована от любых сторонних сетей.

Схема сетевого взаимодействия узла управления и мониторинга приведена на рисунке 7.

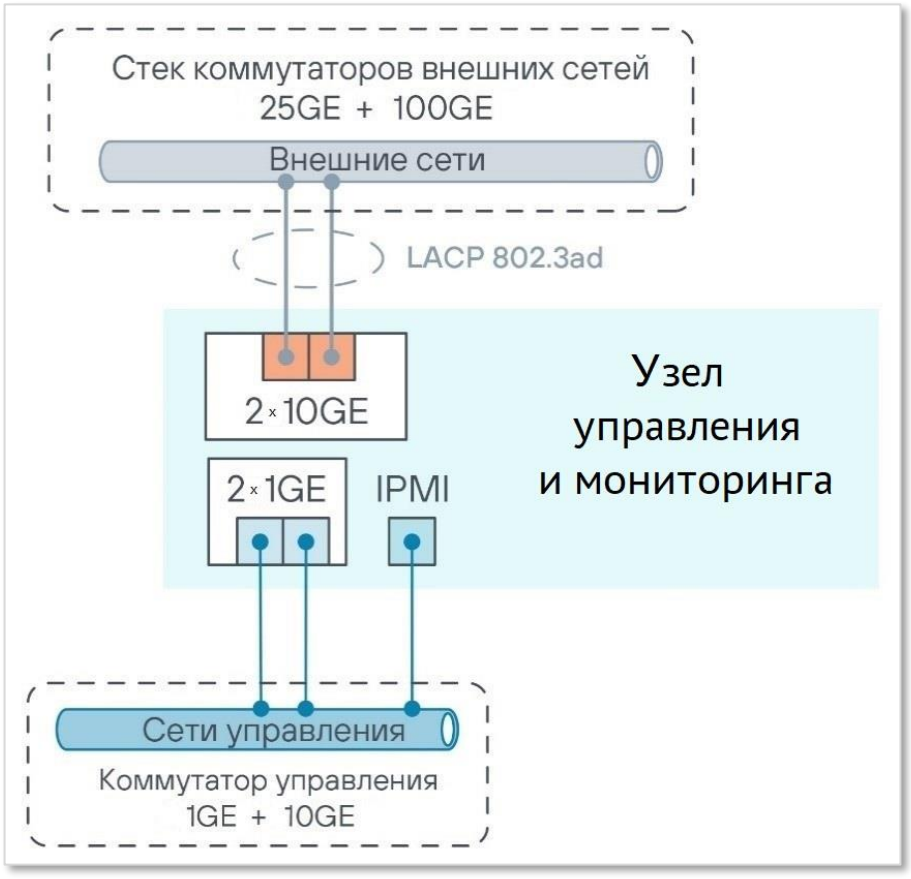

Рисунок 7 — Схема сетевого взаимодействия узла управления

Применяемое программное обеспечение:

T — ОС: Альт СП, RedOS (как варианты) — все с виртуализацией для ОС хоста и без для гостевых ОС управляющих виртуальных машин

T — KVM для управления виртуальными машинами

T P — Сервер управления кластером узлов базы данных: **Скала^р Спектр**

T — Сервер управления жизненным циклом **Машины баз данных Скала^р МБД.П** и ее компонентов: **Скала^р Геном**

T — Сервер системы собственного мониторинга: **Скала^р Визион**.

Сетевая подсистема

Сетевая подсистема состоит из коммутатора управления, коммутаторов внешних сетей и внутреннего сетевого взаимодействия (интерконнекта).

Для подключения к внешним сетям используются два коммутатора с портами 10/25 Гбит/с и uplink-портами 100 Гбит/с. Коммутаторы собраны в один виртуальный коммутатор (стек) по технологии MLAG, что позволяет подключать к паре коммутаторов узлы и другие устройства, используя протокол LACP. На стеке коммутаторов внешних сетей реализован сетевой сегмент External VLAN.

Для реализации сети интерконнекта используются два коммутатора с портами 100 Гбит/с. Коммутаторы собраны в один виртуальный коммутатор (стек) по технологии MLAG, что позволяет подключать к паре коммутаторов узлы и другие устройства, используя протокол LACP. На стеке коммутаторов внешних сетей реализован сетевой сегмент Internal VLAN.

Для реализации сетей управления используется один коммутатор с портами 1 Гбит/с и портами 10 Гбит/с. Коммутатор управления подключен к виртуальному коммутатору внешних сетей двумя портами 10 Гбит/с, собранными в транк по протоколу LACP. На коммутаторе управления реализованы сегменты IPMI, PXE и Ring VLAN.

Основные функции:

T P — передача данных между элементами **Машины баз данных Скала^р МБД.П** (интерконнект)

T P — обеспечение информационного обмена с внешними устройствами

P — обмен служебными данными, данными для мониторинга и управления.

Общая схема сетевого взаимодействия приведена на рисунке 8.

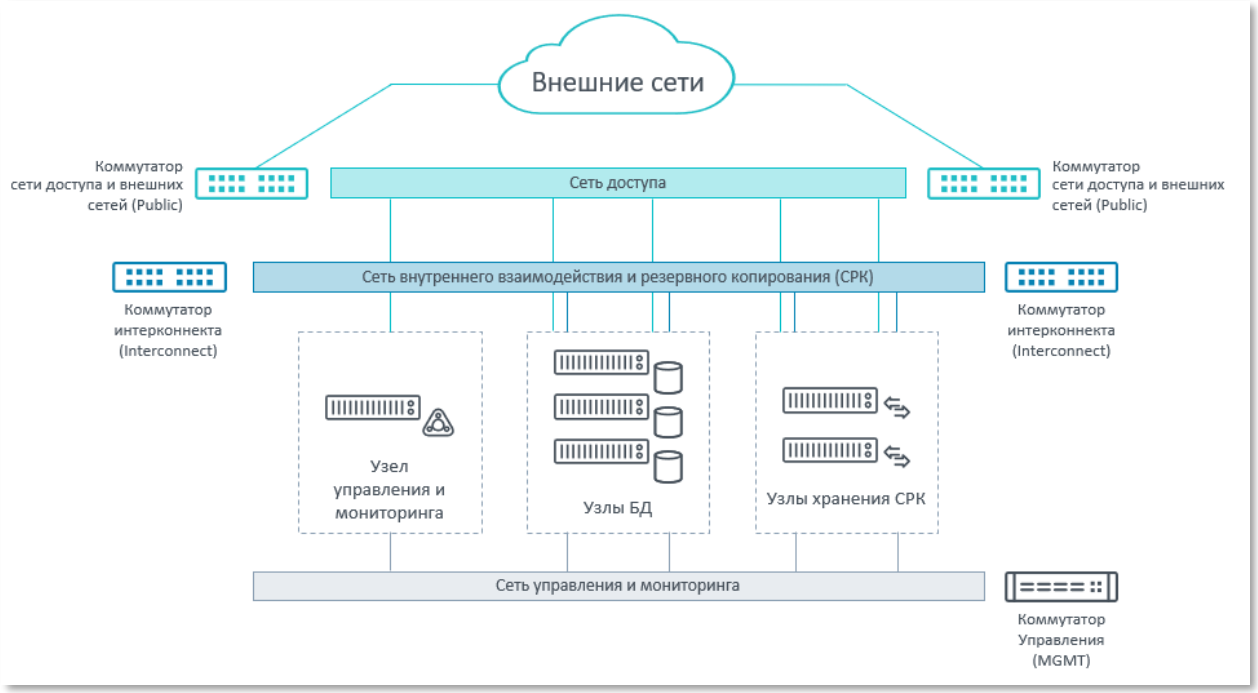

Рисунок 8 — Общая схема сетевого взаимодействия

Состав оборудования сетевой подсистемы:

T P — виртуальный коммутатор (стек), собранный по технологии MLAG из двух коммутаторов 10/25 Гбит/с +100 Гбит/с для подключения к внешним сетям

T P — виртуальный коммутатор (стек) по технологии MLAG из двух коммутаторов 100 Гбит/с для интерконнекта

P — коммутатор с 1 Гбит/с +10 Гбит/с для мониторинга, управления и служебного обмена.

Реализованные подсети:

T P — External VLAN — сеть для подключения к сервисам БД внешних пользователей и прикладных систем, подключение к узлу управления

T P — Internal VLAN — сеть для внутреннего взаимодействия между узлами БД, сеть резервного копирования, сеть кластерного взаимодействия

T — PXE VLAN — сеть для развертывания операционной системы по PXE, платформы МБД, мониторинга

T — Ring VLAN — резервная сеть кластерного взаимодействия, доступ к IPMI

T — IPMI VLAN — сеть управления оборудованием через интерфейсы удаленного управления.

Модуль баз данных

В каждом Модуле размещены 3 вычислительных узла баз данных (далее — узел БД), сконфигурированные в отказоустойчивые кластеры. В кластере могут быть размещены от 1 до 3 независимых экземпляров баз данных.

В состав **Машины баз данных Скала^р МБД.П** могут входить до 4 Модулей баз данных. Каждый Модуль состоит из отказоустойчивого трехузлового кластера (мастер, синхронная реплика, асинхронная реплика).

T — обеспечение отказоустойчивости: в случае отказа мастера его функция исполняется синхронной репликой

P — при «интеллектуальном» прикладном ПО возможно и повышение производительности (если настроить команды записи на мастер реплику, а команды чтения — на синхронную реплику).

Реализация кластера из трех узлов представлена на рисунке 9.

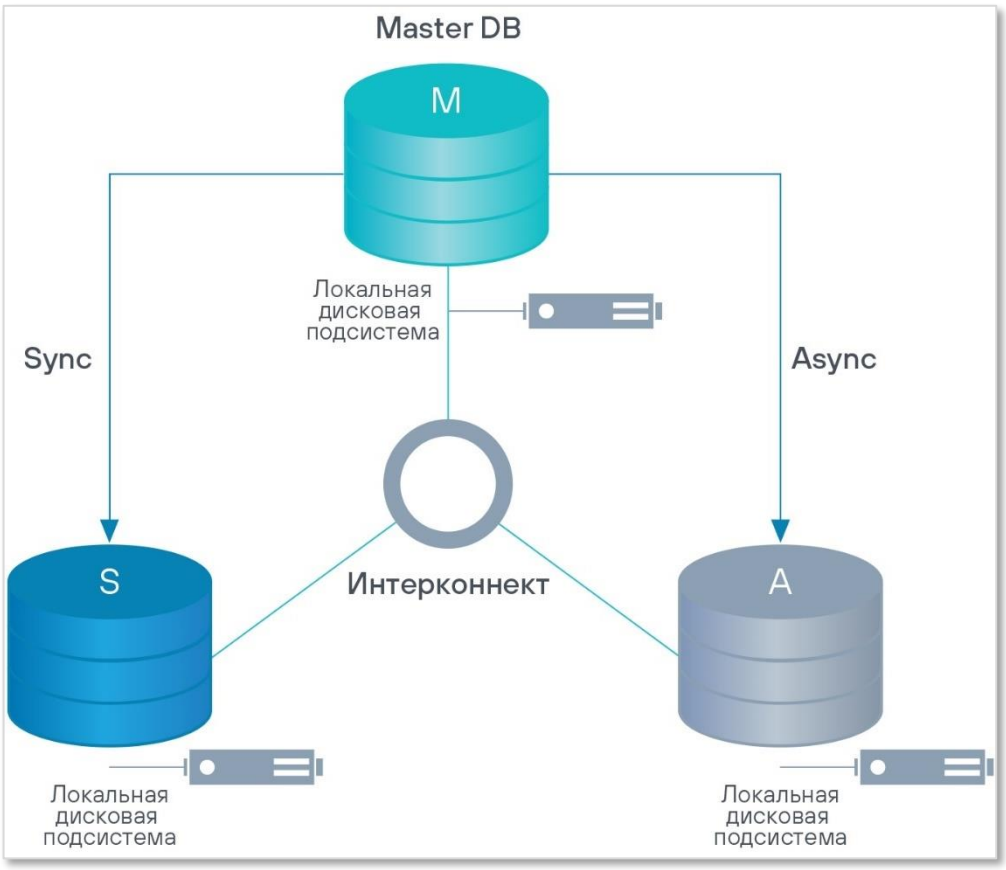

Рисунок 9 — Реализация трехузлового кластера

Каждый отдельный узел БД:

P — использование SAS SSD для томов данных и томов журналов предзаписи для обеспечения оптимальной производительности

T — выделенные накопители (RAID 1) для загрузки ОС — обеспечение отказоустойчивости

T P — локальные накопители для размещения данных (RAID 10 или 50) — исключение лишних элементов и повышение производительности (нет необходимости дополнительного внешнего обмена с системой хранения)

T P — локальные накопители для размещения WAL (RAID 10)

T P — все интерфейсы данных дублированы (стандарт IEEE 802.3ad LACP) — повышение производительности, отказоустойчивость (в случае отказа одного из интерфейсов возможно снижение производительности)

P — 10/25 Gigabit Ethernet — для связи с внешними сетями

P — 100 Gigabit Ethernet — для интерконнекта в рамках Машины

T — два блока питания в режиме резервирования по схеме (1 + 1)

P — 2×CPU Xeon (или аналогичный).

Узел БД укомплектован двумя портами Ethernet 1 Гбит/с, двумя портами Ethernet 10/25 Гбит/с и двумя портами Ethernet 100 Гбит/с.

На базе портов 10/25 Гбит/с создается группа агрегации в режиме 802.3ad LACP, которая представляет собой на уровне операционной системы один логический bond-интерфейс. Данный bond-интерфейс предназначен для пользовательского взаимодействия с узлами управления БД и предоставляемым сервисам.

На базе портов 100 Гбит/с создается группа агрегации в режиме 802.3ad LACP, которая представляет собой на уровне операционной системы один логический bond-интерфейс. Данный bond-интерфейс предназначен для служебного взаимодействия между узлами БД и доступа к подсистеме резервного копирования.

Один из портов Ethernet 1 Гбит/с используется для служебного трафика автоматизированных процессов установки, управления конфигурациями ОС, СПО. Эта сеть изолирована от любых сторонних сетей. Другой порт Ethernet 1 Гбит/с используется для служебного трафика кластера.

Схема сетевого взаимодействия вычислительного узла представлена на рисунке 10.

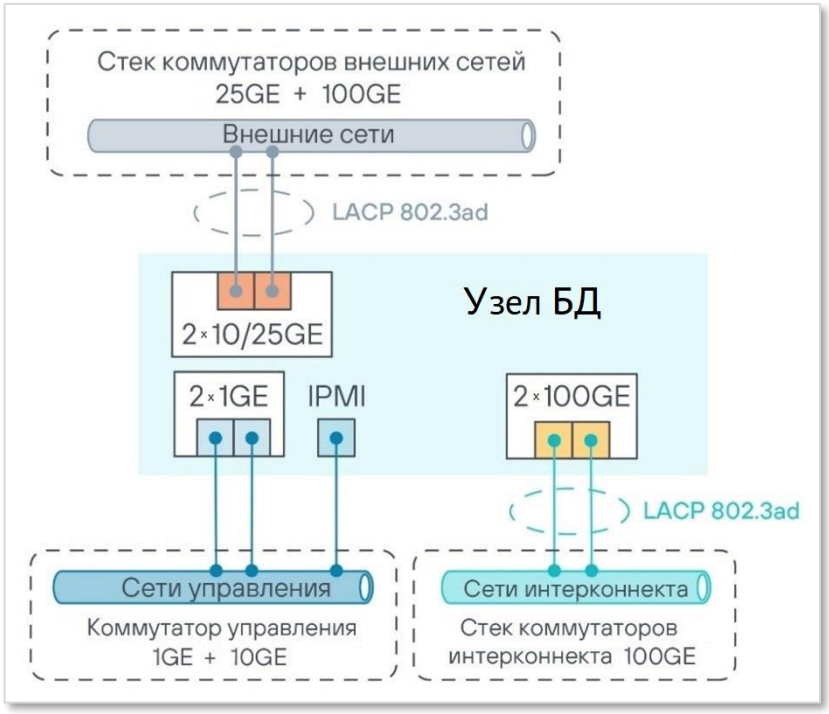

Рисунок 10 — Схема сетевого взаимодействия вычислительного узла

Применяемое программное обеспечение:

T — ОС: Альт СП или RedOS

P — СУБД: Postgres Pro Enterprise

T P — управление резервным копированием: pg_probackup

T P — управление кластером узлов базы данных: **Скала^р Спектр**

T — управление жизненным циклом **Машины баз данных Скала^р МБД.П** и ее компонентов: **Скала^р Геном**

T — система собственного мониторинга: **Скала^р Визион**

T P — аппаратные RAID-контроллеры нового поколения, оснащенные функцией защиты кэша от сбоев электропитания.

Модуль резервного копирования

Модуль резервного копирования предназначен для хранения архивных файлов журналов предзаписи WAL, создания, хранения и восстановления резервных копий баз данных и архивных копий файловых журналов предзаписи WAL.

Поддерживает проверку целостности данных резервных копий и управление политиками хранения.

Для выполнения указанных выше операций используется утилита pg_probackup. Для управления резервными копиями pg_probackup создает каталог резервных копий. В этом каталоге сохраняются все файлы резервных копий с дополнительной метаинформацией, а также архивы WAL, необходимые для восстановления на момент времени. Резервные копии разных экземпляров БД хранятся в отдельных подкаталогах каталога резервных копий.

Реализация сервиса управления дисками СРК представлена на рисунке 11.

Кластер из двух контроллеров системы резервного копирования (первичный, вторичный)

P — в нормальных условиях диски «распределены» на оба контроллера (режим «несимметричный Active-Active»), что способствует высокой производительности

T — в случае отказа одного из контроллеров функция продолжает исполняться вторым.

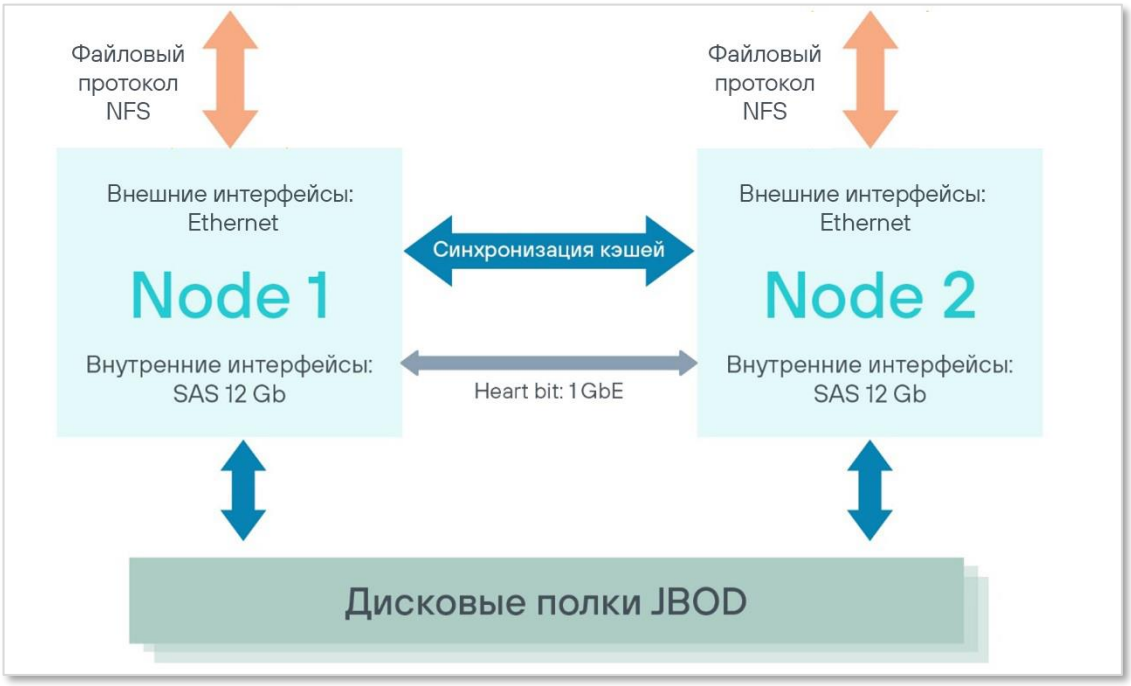

Рисунок 11 — Реализация сервиса управления дисками СРК

Функционал каждого отдельного контроллера:

P — использование общей дисковой полки с NL-SAS-дисками для обеспечения высокой производительности

T — выделенные накопители (RAID 1) для загрузки ОС — обеспечение отказоустойчивости

T P — все интерфейсы данных дублированы (стандарт IEEE 802.3ad LACP) — повышение производительности, отказоустойчивость (в случае отказа одного из интерфейсов возможно снижение производительности)

P — 10/25 Gigabit Ethernet — для связи с внешними сетями

P — 100 Gigabit Ethernet — для интерконнекта в рамках МБД

P — 2×100 Gigabit Infiniband (/Ethernet) — для синхронизации кэша контроллеров

T — два блока питания в режиме резервирования по схеме (1 + 1)

P — 2×CPU Xeon (или аналогичный)

T P — одна или две внешние полки с дисками с интерфейсом SAS подключаются к контроллерам по интерфейсу SAS 12G.

В Модуле резервного копирования для архивирования файлов WAL используется локальный режим работы pg_probackup с записью в каталог, смонтированный по NFS. Для резервного копирования используется удаленный режим работы pg_probackup с узла, предоставляющего сервис NFS.

Для реализации сервиса СРК используется программно-определяемая СХД RAIDIX в двухконтроллерной конфигурации. СХД состоит из двух контроллеров и одной или двух внешних полок с дисками с интерфейсом NL-SAS. По умолчанию во внешних полках используются жесткие диски (SAS HDD) большой емкости. Дисковые полки подключаются к контроллерам по интерфейсу SAS 12G.

Контроллер СХД укомплектован двумя портами Ethernet 10/25 Гбит/с, двумя портами Ethernet 100 Гбит/с, двумя портами VPI 100 Гбит/с (Infiniband/Ethernet 100 Гбит/с), а также двумя служебными портами Ethernet 1 Гбит/с (включая IPMI) и двумя портами Ethernet 1 Гбит/с для межкластерной связи (Heartbeat).

На базе портов 10/25 Гбит/с создается группа агрегации в режиме 802.3ad LACP, которая представляет собой на уровне операционной системы один логический агрегированный интерфейс. Данный интерфейс предназначен для подключения внешних систем (в том числе удаленных систем резервного копирования) к сервису NFS.

На базе двух портов 100 Гбит/с в режиме Ethernet создается группа агрегации в режиме 802.3ad LACP, которая представляет собой на уровне операционной системы один логический bond-интерфейс. Данный bond-интерфейс предназначен для служебного взаимодействия между узлами БД и СХД подсистемы резервного копирования, в том числе по протоколу NFS.

Два порта Infiniband 100 Гбит/с используются для прямой синхронизации кэша контроллеров СХД.

На базе двух портов Ethernet 1 Гбит/с создается группа агрегации в режиме 802.3ad LACP, которая представляет собой на уровне операционной системы один логический агрегированный интерфейс. Этот интерфейс используется для служебного трафика кластера (Heartbeat).

Схема сетевого взаимодействия контроллеров СХД представлена на рисунке 12.

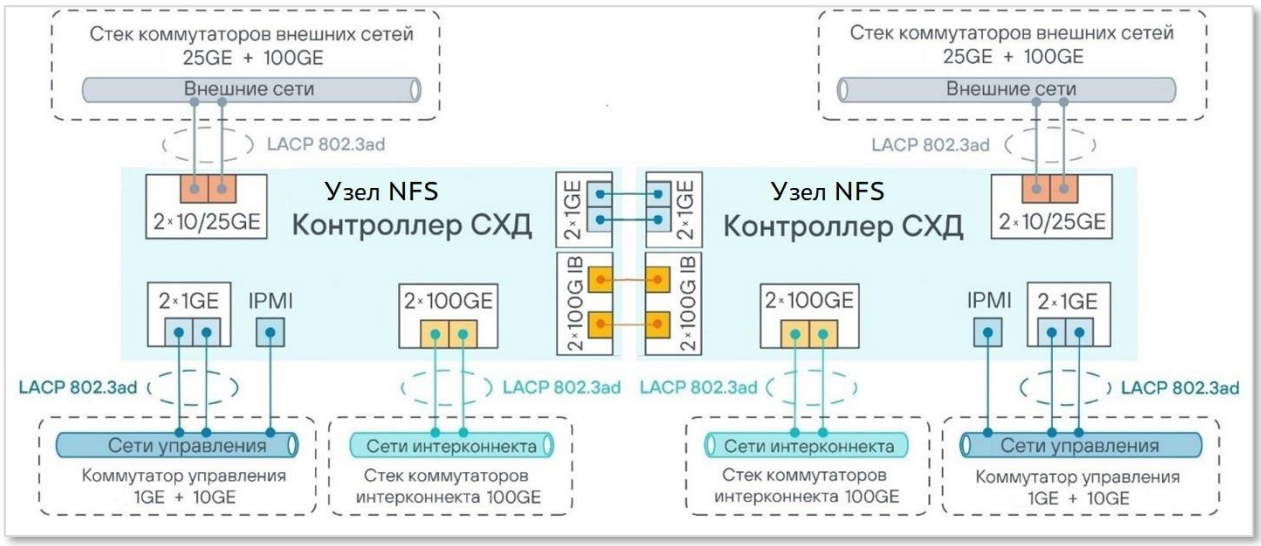

Рисунок 12 — Схема сетевого взаимодействия контроллеров СХД

Применяемое программное обеспечение:

TP — программное обеспечение для организации RAID-массивов на СХД: Raidix 5, двухконтроллерная редакция.

8 СПЕЦИФИЧНЫЕ ЧЕРТЫ

Проектирование и реализация **Машины баз данных Скала^р МБД.П** осуществлялись с учетом ряда выбранных приоритетов, оказывающих непосредственное влияние на функциональные и эксплуатационные показатели. Наиболее значимые из них следующие:

Приоритет обеспечения сохранности данных перед повышенной доступностью

Эффект:

- Гарантия сохранности данных при любых отказах
- Быстрое восстановление из резервных копий в случае сбоев

Отказ от использования виртуальной среды для реализации вычислительного узла в пользу аппаратного решения

Эффект:

- Максимум производительности на данном оборудовании (нет потерь на среду виртуализации, прочие сведены к минимуму)
- Повышение надежности решения (нет дополнительного программного уровня)

Отказ от использования выделенной системы хранения для размещения данных в пользу локальных дисков

Эффект:

- Повышение производительности: сокращение издержек на доставку блочного трафика за счет исключения максимального количества компонентов путем переноса системы хранения непосредственно на вычислительные узлы. (СУБД Postgres Pro Enterprise лучше работает с локальными томами данных)
- Повышение надежности решения (нет дополнительного сложного элемента в виде системы хранения)
- Снижение стоимости решения (нет расходов на систему хранения в целом, только на накопители SSD для данных и журналов. Для резервных копий в СРК типично применяются HDD).

Отказ от применения уникальных аппаратных разработок в пользу стандартного высоконадежного и производительного оборудования в качестве платформы для размещения компонентов решения

Эффект:

- Обеспечение стабильного уровня производительности (компоненты проверены временем)
- Повышение надежности решения (нет уникальных элементов)

- Снижение стоимости сопровождения (доступность элементов при выходе из строя)

Отказ от применения программных RAID в пользу аппаратных RAID

Эффект:

- Обеспечение более высокой производительности
- Защита кэша от сбоя электропитания, что снижает риски потери данных
- Снижение зависимости от производителей программных RAID

Отказ от использования проприетарных иностранных программных решений в пользу ПО с открытым кодом и отечественных разработок

Эффект:

- Повышение производительности за счет доработки ПО (силами **Скала^р** и партнеров)
- Повышение надежности решения (снижение рисков недоступности поддержки)

Возможность применения типовых и сторонних решений для мониторинга и управления в дополнение к предустановленным

Эффект

- Сохранение ранее сделанных инвестиций в системы управления ИТ-инфраструктурой
- Возможность построения сквозных систем управления, с включением **Машины баз данных Скала^р МБД.П** — в инфраструктуру заказчика.

9 ГАРАНТИРОВАННОЕ КАЧЕСТВО

Качественные показатели **Машины баз данных Скала^р МБД.П** обеспечиваются ее соответствием проверенному стандартному варианту, соблюдением установленных норм и требований по формированию, реализацией работ высококвалифицированными специалистами на всех этапах жизненного цикла.

Производство (комплектование, развертывание ПО и предварительная настройка)

- При производстве используются высококачественные комплектующие
- Сборка продукции осуществляется строго в соответствии с утвержденным планом размещения компонентов
- Первичное развертывание ПО осуществляется в автоматическом режиме
- Дополнительные настройки ПО осуществляются в соответствии с утвержденной пошаговой инструкцией
- Осуществляется тестирование сформированной **Машины баз данных Скала^р МБД.П**

Передача в эксплуатацию

- **Машина баз данных Скала^р МБД.П** полностью сформирована, протестирована, готова к размещению в сети заказчика и подключению прикладного ПО
- В комплекте с **Машиной баз данных Скала^р МБД.П** передается паспорт решения, эксплуатационная документация, сертификат на поддержку
- Проводится обучение специалистов заказчика работе с **Машиной баз данных Скала^р МБД.П** (опция по запросу)

Поддержка

- **Машина баз данных Скала^р МБД.П** поставляется с годовой поддержкой (может быть предоставлена также на 2, 3 и 5 лет), которая включает в себя решение всех вопросов, связанных с нарушениями работоспособности как комплекса в целом, так и его отдельных аппаратных компонентов и программного обеспечения
- Поддержка всех компонентов осуществляется через единое окно обращений в режиме 9x5 или 24x7
- Поддержка предоставляется непосредственно производителем или сертифицированным партнером
- У заказчика есть возможность выбора варианта поддержки из актуальных на момент поставки, а также дополнительных опций

В сложных случаях к решению проблем привлекаются архитекторы и инженеры, непосредственно участвовавшие в разработке **Машины баз данных Скала^р МБД.П**.

10 РЕАКЦИЯ НА ВОЗМОЖНЫЕ ОТКАЗЫ

Отказы, связанные со стандартными элементами Машины баз данных Скала^р МБД.П

В рамках **Машины баз данных Скала^р МБД.П** обеспечена отказоустойчивость основных аппаратных элементов, в том числе:

- узлов (дублирование процессоров, источников питания и др.)
- накопителей (применение в составе RAID с требуемым уровнем отказоустойчивости)
- внешних интерфейсов и внутренних элементов сети и интерконнекта (полное дублирование)
- элементов системы резервного копирования (дублирование контроллеров, интерфейсов, и применение отказоустойчивых конфигураций RAID)

Отказы перечисленных элементов отрабатываются стандартными алгоритмами в соответствии с произведенными настройками. Любой единичный отказ не повлияет на доступность системы в целом, хотя по конкретному сервису возможно некоторое снижение производительности. После устранения неисправности полная производительность **Машины баз данных Скала^р МБД.П** также восстановится.

Отказы, связанные с узлами кластера баз данных

Для обеспечения бесперебойности доступа и сохранности данных в решении реализован трехузловой кластер, состоящий из мастера СУБД, а также синхронной и асинхронной реплик. В случае отказа любого из перечисленных узлов кластера (или остановки узла для проведения обслуживания) работоспособность **Машины баз данных Скала^р МБД.П** для пользователей будет сохранена в полном объеме в автоматическом режиме средствами ПО управления кластером.

При этом при необходимости будут переназначены роли узлов кластера (актуально в случае отказа узла с мастером СУБД и узла с синхронной репликой).

После завершения обслуживания или устранения причины отказа и восстановления узла необходимые данные будут восстановлены (в зависимости от степени «отставания») из резервных копий и/или архивов WAL.

Детальные алгоритмы обеспечения отказоустойчивости кластера баз данных и рекомендации по действиям администратора в той или иной конкретной ситуации приведены в документации, передаваемой заказчику совместно с **Машиной баз данных Скала^р МБД.П**.

Поскольку для **Машины баз данных Скала^р МБД.П** избран приоритет обеспечения сохранности данных, одновременный или последовательный отказ двух узлов кластера приводит к полной остановке **Машины баз данных Скала^р МБД.П** ввиду того, что в этих условиях продолжение работы СУБД может привести к частичной или полной потере данных.

11 ТИПОВЫЕ КОМПЛЕКТЫ РЕШЕНИЯ

Типовые комплекты поставки **Машины баз данных Скала^р МБД.П** приведены на рисунке 13.

Примечание — *возможна поставка без СХД контроллеров (без контроллерной пары и полок)*

Машина баз данных Скала^р МБД.П М-1	Машина баз данных Скала^р МБД.П М-2	Машина баз данных Скала^р МБД.П М-3/М-4

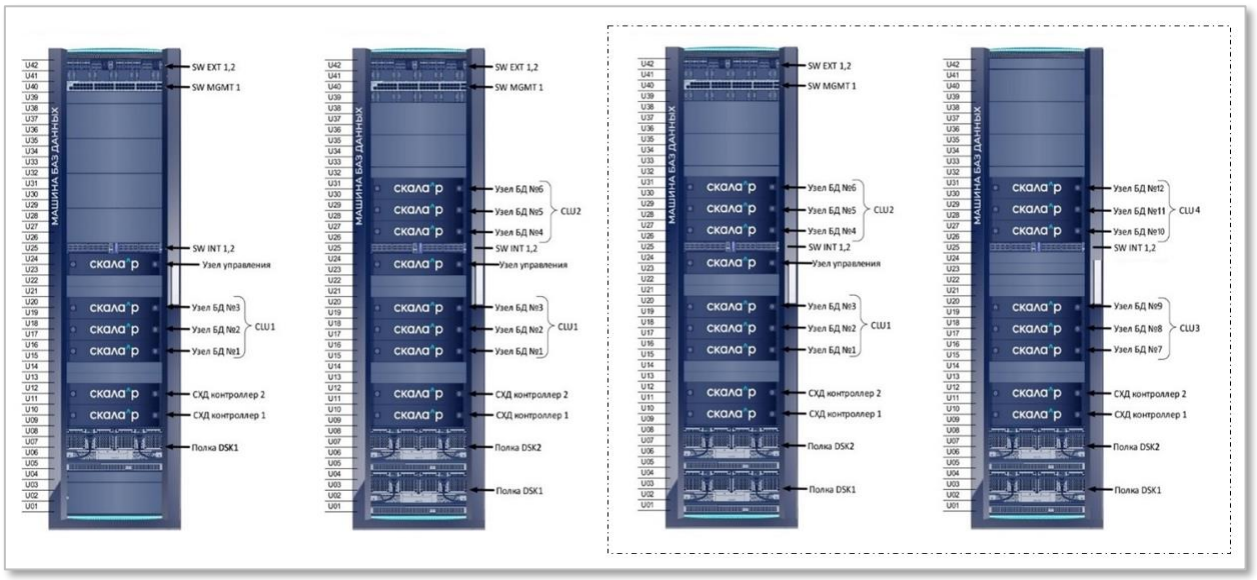

Рисунок 13 — Типовые комплекты поставки Машины баз данных Скала^р МБД

Машина Баз Данных поставляется комплектами по Модулям таким образом, чтобы можно было подобрать оптимальную конфигурацию от количества размещаемых баз данных на узлах БД. Ниже (Таблица 3) приведен пример размещения отдельных экземпляров СУБД в Модулях баз данных таким образом, чтобы разместить до 4 экземпляров отдельных СУБД в каждом Модуле. Возможна конфигурация шахматного размещения в одном Модуле до 3 разных экземпляров баз данных Postgres. Коммунальное использование одного Модуля возможно в случае разрешения размещения таких баз данных в одном секторе безопасности.

Таблица 3 — Параметры поставляемых моделей (производительный профиль, до 160 Тбайт)

Параметры / Модель	М-1	М-2	М-3	М-4
Количество узлов БД	3	2x3	3x3	4x3
Количество экземпляров СУБД	до 3	до 6	до 9	до 12
Общий полезный объем всех БД, Тбайт	до 160	до 2x160	до 3x160	до 4x160

Параметры / Модель	М-1	М-2	М-3	М-4
Полезный объем хранения системы резервного копирования (БД+WAL), на одну полку, Тбайт	190-1024	2x	3x	4x
Размещение в стойках	1	1	2	2

Лицензирование необходимого программного обеспечения осуществляется в соответствии с количеством узлов и сокетов Модулей баз данных. Стоимость лицензий учтена в стоимости решения.

12 ВАРИАТИВНОСТЬ РЕШЕНИЯ

Малый или тестовый ландшафт

Вариант решения:

- достаточно памяти RAM (512 Гбайт)
- высокопроизводительные накопители SSD (8 × 1,92 TB + 2x 1,92 для WAL)
- один аппаратный RAID контроллер
- RAID 5, RAID 1

Высокопроизводительный продуктивный ландшафт OLTP

Вариант решения:

- больше памяти RAM (2 Тбайт)
- твердотельные накопители повышенного объема (16x 7,68 TB + 4x 7,68 TB)
- два современных аппаратных RAID контроллера
- RAID 50, RAID 10

Геокластер (как опция)

Вариант решения:

- дополнительные сетевые карты в узлах БД
- специальные настройки кластерного ПО и ПО резервного копирования

Размещение нескольких экземпляров кластеров СУБД в одной Машине

Вариант решения:

- увеличение количества узлов баз данных в составе Машины; размещение нескольких экземпляров СУБД на каждом из узлов баз данных
- размещение реплик в «шахматном порядке»
- разные кластеры БД могут быть настроены под разные ландшафты (тестовый или продуктивный)

Тонкая настройка для повышения производительности (опция)

Вариант решения:

- может использоваться в комплексе с любым из вариантов
- требуется участие разработчиков прикладных систем
- достигается направлением чтения и записи на разные узлы кластера путем внесения соответствующих настроек в прикладных системах

13 ТРЕБОВАНИЯ К РАЗМЕЩЕНИЮ РЕШЕНИЯ

Решение поставляется в виде отдельного серверного монтажного шкафа 19”, высота 42U.

Наполнение шкафа оборудованием и совокупный вес зависит от выбранного варианта решения и может составлять от 400 до 800 кг.

Для подключения шкафа к системе электроснабжения должны быть предусмотрены два независимых входа электропитания.

Потребляемая мощность шкафа составляет от 6 до 11 кВт.

Должны быть предусмотрены соответствующие мощности по отводу тепла.

Для подключения к локальной сети необходим резервированный канал 4×100 Gigabit Ethernet или до 8×10/25 Gigabit Ethernet.

При развертывании решения на нем будут осуществлены настройки сетевых адресов в соответствии со структурой сети заказчика. Заказчик должен предоставить необходимые данные в соответствии с номенклатурой компонентов решения.

В сети заказчика должны быть настроены соответствующие маршруты и права доступа.

Дальнейшие мероприятия по вводу в эксплуатацию осуществляются заказчиком путем проведения настройки прикладных программных систем.

14 ПРИМЕРЫ РАБОТАЮЩИХ РЕШЕНИЙ

Пример: размещение 3х сервисов на одном кластере

Решение – **Машина баз данных Скала^р МБД.П.**

Консолидация трех сервисов баз данных на едином кластере с целью оптимизации использования вычислительных ресурсов.

- 1 аппаратный кластер (3 узла)
- 3 сервиса БД
- СУБД Postgres Pro Enterprise Certified
- объем БД № 1 ~18 Тбайт
- объем БД № 2 ~ 17,5 Тбайт
- объем БД № 3 ~ 10 Тбайт

«Шахматное» размещение БД по узлам кластера (Рисунок 14).

Специфика

- Повышение утилизации вычислительных ресурсов
- «Шахматное» размещение БД по узлам кластера:
  каждый из узлов БД является мастером для одной из баз,
  синхронной копией — для другой и асинхронной копией — для третьей

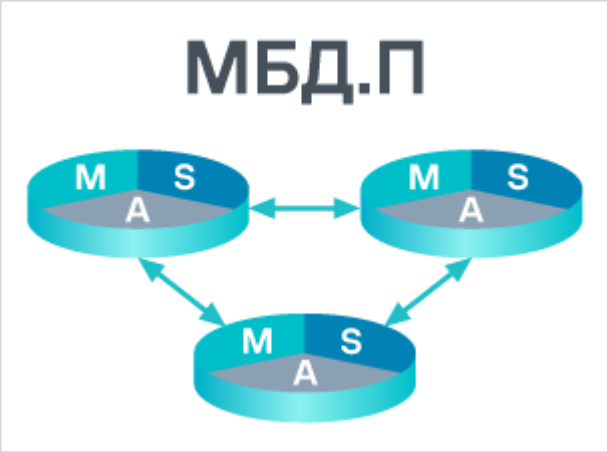

Рисунок 14 — «Шахматное» размещение БД по узлам кластера

Пример: метрокластер

Решение – две связанные **Машины баз данных Скала^р МБД.П**

Каждая СУБД имеет высокодоступную конфигурацию (мастер + синхронная реплика) на своей основной площадке, при этом обеспечивается асинхронное (или синхронное) реплицирование на соседнюю площадку для восстановления в случае сбоя ЦОД.

В каждом ЦОД развернут собственный Модуль резервного копирования. Архивация WAL и создание резервных копий выполняются в каждом ЦОД независимо друг от друга.

Переключение роли в случае сбоя узла или ЦОДа происходит автоматически.

В составе каждой **Машины баз данных Скала^р МБД.П**:

- 2 сервиса БД
- 2 узла кластера одного сервиса и 1 узел второго
- СУБД Postgres Pro Enterprise Certified

Специфика

- Обеспечение отказоустойчивости в пределах одного региона реализовано за счет асинхронной (синхронной) реплики на удаленной площадке
- Автоматическое восстановление при сбое узла или площадки
- Централизованное управление метрокластерами осуществляется с помощью ПО **Скала^р Спектр**.

Реализация метрокластера

- Удаленность порядка 20 км.
- Канал 100 Гбит/с
- На каждой площадке есть реплика основного кластера

Пример реализации метрокластера приведен на рисунке 15.

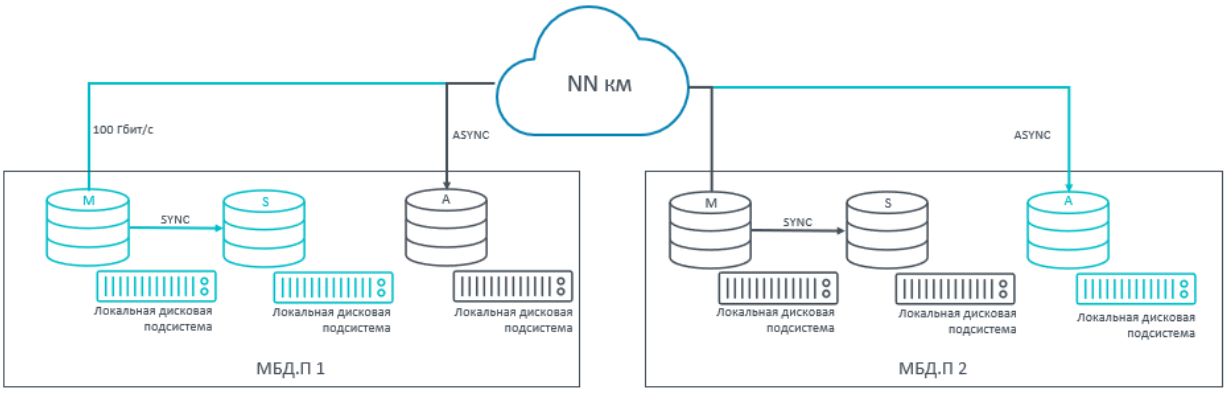

Рисунок 15 — Пример организации метрокластера

Пример: геораспределенный кластер

Решение – две связанные **Машины баз данных Скала^р МБД.П**

Элементами обеспечения катастрофоустойчивости системы на уровне баз данных является размещение СУБД на двух площадках. Одна из площадок является активной, вторая – горячим резервом.

Синхронизация данных между двумя площадками, а также между серверами внутри площадки осуществляется средствами встроенной потоковой репликации системы управления базами данных Postgres Pro. В рамках одной площадки синхронизация данных осуществляется через сеть интерконнект, между площадками – через сеть георепликации.

Также сеть интерконнект используется для резервного копирования и архивации WAL журналов. В каждом ЦОД развернут собственный Модуль резервного копирования. Архивация WAL и создание резервных копий выполняются в каждом ЦОД независимо друг от друга.

В любой момент времени активной является СУБД только на одной из площадок. Переключение или смена роли площадки может осуществляться либо в плановом режиме, либо в случае катастрофы.

В составе каждой **Машины баз данных Скала^р МБД.П**:

- 1 аппаратный кластер (3 узла)
- 1 сервис БД
- СУБД Postgres Pro Enterprise Certified

Специфика

- Обеспечение отказоустойчивости на уровне катастрофы реализовано за счет полной копии кластера на резервной площадке.
- Повышение производительности СУБД за счет реализации на уровне приложения распределения чтения по репликам базы данных.
- Централизованное управление геораспределенным кластером осуществляется с помощью ПО **Скала^р Спектр**.

Реализация геораспределенного кластера

- Вторая, взаимодействующая **«Машина баз данных Скала^р МБД #2»** (полная копия основной)
- Удаленность порядка 500 км.
- Канал 25 Гбит/с
- Каскадная репликация

Пример реализации геокластера приведен на рисунке 16.

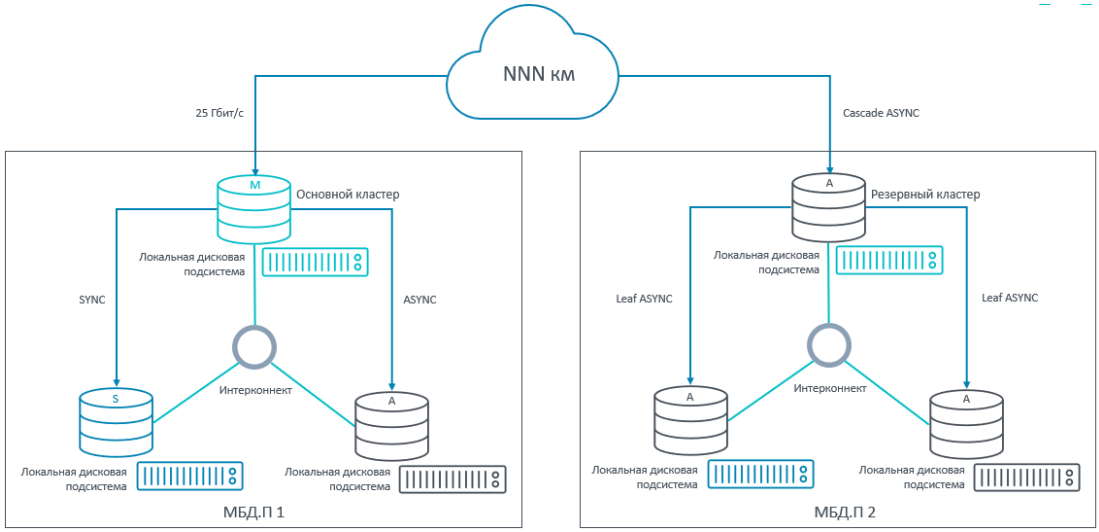

Рисунок 16 — Пример реализации геокластера

Результаты тестирования

Тесты проводились с помощью ПО pgbench, обеспечивающего TPC-B подобную нагрузку на базу данных. Стандартный встроенный скрипт выдает семь команд в транзакции со случайно выбранными aid, tid, bid и delta (режим RW):

1. **BEGIN;**
2. **UPDATE pgbench_accounts SET abalance = abalance + :delta WHERE aid = :aid;**
3. **SELECT abalance FROM pgbench_accounts WHERE aid = :aid;**
4. **UPDATE pgbench_tellers SET tbalance = tbalance + :delta WHERE tid = :tid;**
5. **UPDATE pgbench_branches SET bbalance = bbalance + :delta WHERE bid = :bid;**
6. **INSERT INTO pgbench_history (tid, bid, aid, delta, mtime) VALUES (:tid, :bid, :aid, :delta, CURRENT_TIMESTAMP);**
7. **END;**

При выборе встроенного теста «select-only» из указанных выше команд выполняется только «SELECT» (режим «SO»).

Аппаратные характеристики:

- 2xCPU Xeon 24 ядра
- 1024 Гбайт оперативной памяти
- Накопители SAS SSD

Параметр pgbench sf = 75 000 приводит к созданию БД объемом ~1 Тбайт; функция объединения в пулы pgbench не использовалась, для каждого пользователя использовалось выделенное подключение к СУБД. Тесты проводились на RedOS8 для сертифицированной версии Postgres Pro Enterprise. С 2023 года мы усложнили тестирования и начали использовать БД объемом 3 Тбайта. На рисунке 17 показаны результаты теста Машины 2025 года.

Рисунок 17 — Производительность Машины баз данных Скала^р МБД.П по результатам теста pgbench

15 О РЕЗУЛЬТАТАХ РАСЧЕТА НАДЕЖНОСТИ

Машина баз данных Скала^р МБД.П ориентирована на обеспечение отказоустойчивого и высокопроизводительного функционирования СУБД Postgres. При реализации проектов с применением **Машины баз данных Скала^р МБД.П** возникает потребность в знании реальных показателей надежности, обеспечиваемых этим решением.

Значения показателей зависят от конкретной конфигурации решения и используемого набора оборудования.

Специалистами **Скала^р** в соответствии с требованиями «ГОСТ 27.301-95 Надежность в технике. Расчет надежности. Основные положения» разработана специальная математическая модель, позволяющая оценить основные показатели надежности решения.

Модель была применена к «среднему» типовому варианту конфигурации **Машины баз данных Скала^р МБД.П,** который включает в себя два трехузловых кластера и две дисковых полки в составе СХД. Более полно структура выбранного для расчета варианта решения отражена на рисунке 18.

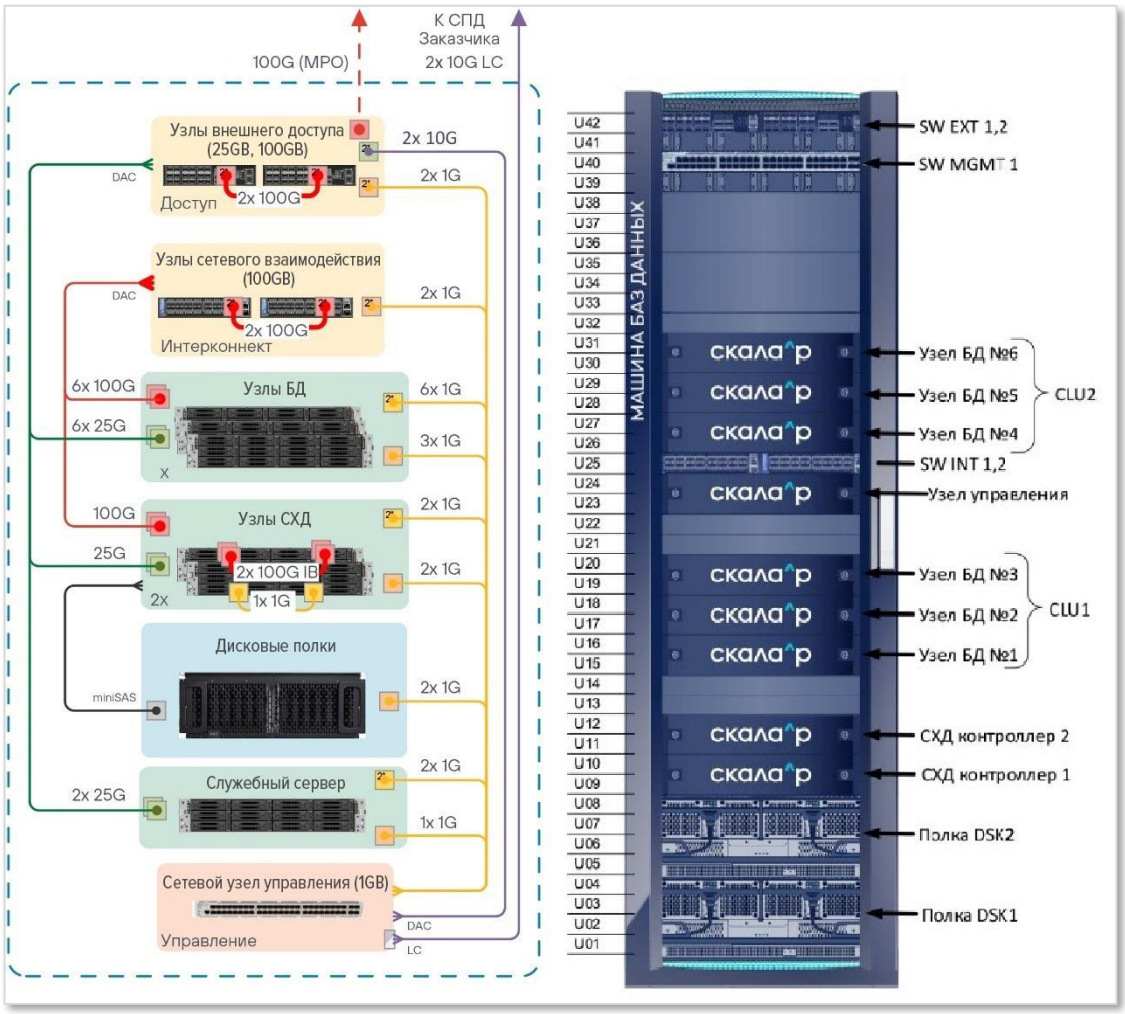

Рисунок 18 — Схема взаимодействия для расчета надежности Машины баз данных Скала^р МБД.П

ЗАКЛЮЧЕНИЕ

Машина баз данных Скала^р МБД.П — программн0-аппаратный комплекс для обработки и хранения данных с использованием СУБД Postgres в высоконагруженных системах.

Основные черты **Машины баз данных Скала^р МБД.П:**

- Отказоустойчивость
- Высокая производительность
- Приоритет сохранности данных
- Готовность к быстрому развертыванию
- Удобная эксплуатация
- Экономическая эффективность

Машина баз данных Скала^р МБД.П и ее Модули произведены в РФ и внесены в реестры РЭП МПТ (в том числе как ПАК) и реестры ПАК Минцифры

Структурно в **Машину баз данных Скала^р МБД.П** входят:

- Базовый модуль
- Модуль баз данных
- Модуль резервного копирования

Каждый из Модулей — это специально подобранный комплект оборудования, а также предустановленное и настроенное программное обеспечение, адаптированное для обеспечения функционала решения в целом и простоты его модернизации.

Надежность, производительность **Машины баз данных Скала^р МБД.П** подтверждается проведенными тестами, специальными расчетными данными, практическим использованием решений в течение ряда лет.

Дополнительная информация по **Машине баз данных Скала^р МБД.П** предоставляется по запросу info@skala-r.ru.

О КОМПАНИИ

Скала^р — лидер российского рынка ПАК по версии CNews Analytics, 2024.

Программно-аппаратные комплексы (Машины) **Скала^р** выпускаются с 2015 года и представляют широкий технологический стек для построения динамических инфраструктур и инфраструктур управления данными высоконагруженных информационных систем.

Продукты **Скала^р** включены в ЕРРРП, произведенной на территории Российской Федерации, и в ЕРРП для ЭВМ и БД. Соответствует критериям доверенности и использованию для объектов критической информационной инфраструктуры (КИИ).

Машины **Скала^р** являются серийно выпускаемыми преднастроенными комплексами, которые быстро развертываются и вводятся в эксплуатацию. Глубокая интеграция технических средств и программного обеспечения в ПАК **Скала^р** позволяет получить расширенные возможности и функциональность, которые недоступны при использовании отдельных компонентов.

Модульный принцип обеспечивает интеграцию разнородных компонентов ИТ-инфраструктуры в единую платформу предприятий, корпораций и ведомств. Единые поддержка и сервисное обслуживание для всех продуктов линейки **Скала^р** от производителя обеспечивают оперативное разрешение инцидентов на стыке технологий.