



Машина баз данных

Скала^р МБД.П

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

Технический обзор

Релиз МБД.П26RS

версия 4.6 от 09.09.2026



ОГЛАВЛЕНИЕ

Перечень терминов и сокращений.....	3
1 Предисловие	4
2 Введение	5
3 Отличительные черты	6
4 Подтвержденная безопасность.....	8
4.1 Операционная система РЕД ОС.....	8
4.2 СУБД Postgres Pro Enterprise.....	8
4.3 Средства защиты	9
5 Производство в Российской Федерации.....	11
6 Принципы проектирования	13
7 Состав решения	15
7.1 Базовый модуль	17
7.2 Модуль баз данных	24
7.3 Модуль резервного копирования.....	27
8 Специфичные черты.....	30
9 Гарантированное качество	32
10 Реакция на возможные отказы	33
11 Типовые комплекты решения	34
12 Вариативность решения	36
13 Требования к размещению решения	37
14 Примеры работающих решений.....	38
15 Техническая поддержка	42
16 Заключение	44
О Компании	45

ПЕРЕЧЕНЬ ТЕРМИНОВ И СОКРАЩЕНИЙ

Термин, сокращение	Определение
Avanpost FAM (Federated Access Manager)	ПО для единой аутентификации сотрудников в корпоративных ресурсах организации
IPMI	(англ. Intelligent Platform Management Interface) Стандартизированный интерфейс для удаленного управления и мониторинга серверного оборудования на аппаратном уровне
KVM	(англ. Kernel-based Virtual Machine) Гипервизор уровня ядра Linux, который позволяет создавать виртуальные машины и управлять ими
MIB	(англ. Management Information Base) База управляющей информации, виртуальная база данных, используемая для управления объектами в сети связи
Postgres	Название СУБД (системы управления базами данных)
RAID	(англ. Redundant Array of Independent Disks) Избыточный массив независимых дисков, технология виртуализации данных для объединения нескольких физических дисковых устройств в логический модуль для повышения отказоустойчивости и/или производительности
WAL	(англ. Write Ahead Log) Журнал предварительной записи изменений для обеспечения целостности транзакций
БД	База данных
ЕРРП	Единый реестр российских программ (Минцифры)
ЕРРРП	Единый реестр российской радиоэлектронной продукции Минпромторга РФ
Модуль транзакционной обработки	В составе МБД.П используется как синоним Модуля баз данных
Машина Скала^р (Машина СКАЛА-Р)	Набор аппаратного и программного обеспечения в виде Модулей Скала^р, соединенных вместе для обеспечения определенного метода обработки данных или предоставления ИТ-сервисов с заданными характеристиками. Зарегистрирован в реестрах Минпромторга как программно-аппаратный комплекс (код ОКПД2 26.20.14.160)
Подсистема	Логическое объединение компонентов по функциональному признаку, с целью пояснения состава и принципов действия ПАК
РЭП МПТ	Единый реестр российской радиоэлектронной продукции Минпромторга РФ
Узел	Узел вычисления (сервер) или сетевой узел (коммутатор) в составе Модуля, в зависимости от контекста

1 ПРЕДИСЛОВИЕ

Уведомление

Документ носит исключительно информационный характер и является актуальным на дату размещения.

Технические характеристики, указанные в документе, носят исключительно справочный характер и не могут служить основанием для претензий.

Технические характеристики произведенных ПАК могут отличаться от приведенных в настоящем документе вследствие модификации ПАК.

Технические характеристики и комплектация ПАК могут быть изменены производителем без уведомления.

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

Описание документа

Настоящий документ дает концептуальный и архитектурный обзоры **Машины баз данных Скала^р МБД.П** (далее **Скала^р МБД.П**, **Машина**, **МБД.П**), название в реестрах Минпромторга РФ «Машина баз данных СКАЛА-Р МБД.П».

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

Аудитория

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

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.

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
- предназначена для размещения высоконагруженных баз данных объемом от 20 и до 300 Тбайт, в зависимости от выбранного приоритета на производительность или максимальный объем

Машина баз данных Скала^р МБД.П — комплексное, законченное решение, включающее в себя все необходимое, от Модулей для обработки и хранения данных, включая Модуль резервного копирования, до телекоммуникационных модулей, обеспечивающих скоростную 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, обязательное резервирование критических компонентов, использование устойчивых сетевых протоколов.

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

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
- работают на ПО, включенном в ЕРРП

Машина также находится в ЕРРП.

3 ОТЛИЧИТЕЛЬНЫЕ ЧЕРТЫ

Надежное хранение и высокопроизводительная обработка больших объемов данных

- Объем баз данных до 300 Тбайт
- Модельный ряд включает 20 Тбайт, 40 Тбайт, 80 Тбайт, 160 Тбайт, 300 Тбайт
- Производительность более 160 kTPS по тестам rgbench
- Поддержка катастрофоустойчивых кластерных архитектур

Высокая производительность

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

Отказоустойчивость на всех уровнях

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

Приоритет сохранности данных

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

Контроль качества при производстве

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

Непрерывный контроль состояния

- Встроенная система мониторинга **Скала^{Ар} 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

Улучшенные возможности администрирования

- Встроенная система автоматизации администрирования и обслуживания **Машины** от **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

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

Экономическая эффективность

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

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

4.1 Операционная система РЕД ОС

Машина баз данных Скала^Ар МБД.П поставляется с сертифицированной ОС РЕД ОС (версия 8с, сертификат ФСТЭК 4060 от 12.01.2019, действует до 12.01.2029), которая:

может применяться для защиты информации:

- Государственных информационных систем и торговых площадок до I класса защищенности включительно
- Информационных систем для обработки персональных данных до I уровня защищенности включительно
- Компонентов критической информационной инфраструктуры до I категории значимости включительно
- Автоматизированных систем управления техническими процессами до I класса защищенности включительно
- Информационных систем общего пользования до II класса защищенности

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

4.2 СУБД Postgres Pro Enterprise

В Маchine баз данных Скала^Ар МБД.П используется сертифицированная СУБД Postgres Pro Enterprise (сертификат ФСТЭК 4063 от 16.01.2019, действует до 16.01.2029), 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

4.3 Средства защиты

*Протестирована совместимость
с наложенными средствами защиты*

4.3.1 Kaspersky Endpoint Security для Linux

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

применяется для защиты информации:

- в государственных информационных системах I класса защищенности
- в информационных системах персональных данных при необходимости обеспечения 1 уровня защищенности персональных данных
- в значимых объектах критической информационной инфраструктуры 1 категории
- в автоматизированных системах управления производственными и технологическими процессами I класса защищенности
- 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 2 уровню доверия
- «Требования к средствам антивирусной защиты» (ФСТЭК России, 2012)
- «Профиль защиты средств антивирусной защиты типа Б второго класса защиты. ИТ.САВЗ.Б2.ПЗ» (ФСТЭК России, 2012)
- «Профиль защиты средств антивирусной защиты типа В второго класса защиты. ИТ.САВЗ.Б2.ПЗ» (ФСТЭК России, 2012)
- «Профиль защиты средств антивирусной защиты типа Г второго класса защиты. ИТ.САВЗ.Б2.ПЗ» (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.ПЗ» (ФСТЭК России, 2014)

4.3.2 «Соболь» версия 4

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

применяется для защиты информации:

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

4.3.3 Avanpost FAM

Сертифицированная система единой аутентификации **Avanpost FAM** (сертификат ФСТЭК 4492 от 13.12.2021, действует до 13.12.2026), которая:

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
- 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
- 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
- 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%
- 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. Машина баз данных СКАЛА-Р МБД.П на сайте государственной информационной системы промышленности (ГИСП). Фрагмент страницы <https://gisip.gov.ru/goods/#/product/5427701>

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	Код изделия по ОКПД2
Машина баз данных Скала^р МБД.П	26.20.14.160

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

Наименование Модуля	Код изделия по ОКПД2
СКАЛА-Р Базовый модуль	26.20.14.160
СКАЛА-Р Модуль транзакционной обработки	26.20.14.160
СКАЛА-Р Модуль резервного копирования	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.

Скала^р МБД.П создана для СУБД Postgres

Целью разработки **Машины 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 **МБД.П** используются накопители SSD.

Быстродействие и емкость современных твердотельных накопителей позволили отказаться от использования отдельной системы хранения в **Машине**. 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.

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 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** 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 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**, 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 **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 **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.

В **Машине** 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 серьезных 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 **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 одного 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^Ар МБД.П.**:

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

Модуль — это структурный элемент **Машины Скала^Ар 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.

Скала^Ар МБД.П состоит из Модулей, показанных на Рисунке 2.



Рисунок 2. Модули Машины Скала^Ар МБД.П

Машина баз данных Скала^Ар МБД.П:

- основана на принципе модульности. Компонуется из набора стандартных Модулей, чем обеспечивается универсальный подход, более высокий уровень технологичности и надежности эксплуатации
- может 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
- поставляется как готовый преднастроенный комплекс, однако в процессе эксплуатации состав **Машины** и Модулей может расширяться для повышения емкости хранимых данных или увеличения производительности

Для обеспечения отказоустойчивости и высокой производительности при проектировании **Машины баз данных Скала[^]р МБД.П** были заложены технологические принципы и применен ряд технических решений. К технологическим принципам проектирования относятся*:

Т Р — дублирование критичных компонентов

Р — равномерное распределение нагрузки на доступные ресурсы

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

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

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

Модуль баз данных состоит из 3-узлового кластера, на котором размещены экземпляры БД и журналы WAL.

Взаимодействие между узлами кластера **Машины 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 БД и архивы WAL.

Схема внутренней коммутации **Скала^Ар МБД.П** (для одного служебного узла) представлена на Рисунке 3.

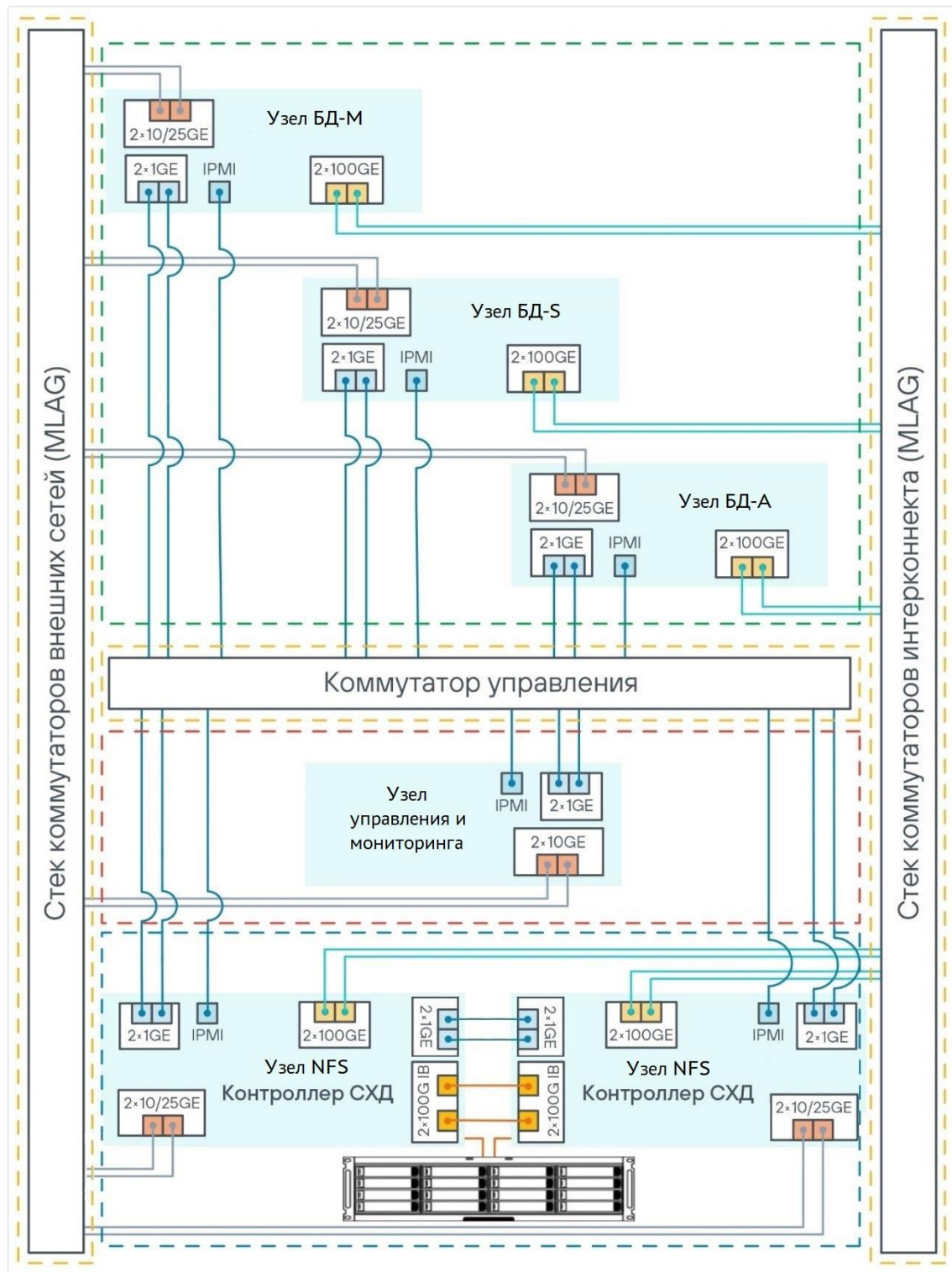


Рисунок 3. Схема внутренней коммутации Машины баз данных Скала^Ар МБД.П

7.1 Базовый модуль

Базовый модуль состоит из:

- Подсистема управления и мониторинга

- Сетевая подсистема

7.1.1 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**
- встроенная система мониторинга: **Скала^р Визион**

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

Для установки ОС используются два локальных накопителя, собранные в RAID 1. Для хранения данных мониторинга и репозитория используются локальные диски совместно с ОС.

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 ОС и СУБД, обеспечивается через установленного агента на узлах, сбор данных с активного сетевого оборудования обеспечивается протоколом SNMP.

Основной функционал:

- ТР** — Управления кластером баз данных (**Скала^р Спектр**)
- Т** — Управления эксплуатацией **Машины МБД.П**, также репозиторий пакетов образов и обновлений (**Скала^р Геном**)
- Т** — Собственного мониторинга и визуализации работы сети и оборудования, входящего в состав **МБД.П** (**Скала^р Визион**)

Пример интерфейса системы управления эксплуатацией представлен на Рисунке 4, системы управления кластером – Рисунок 5, системы мониторинга – Рисунок 6.

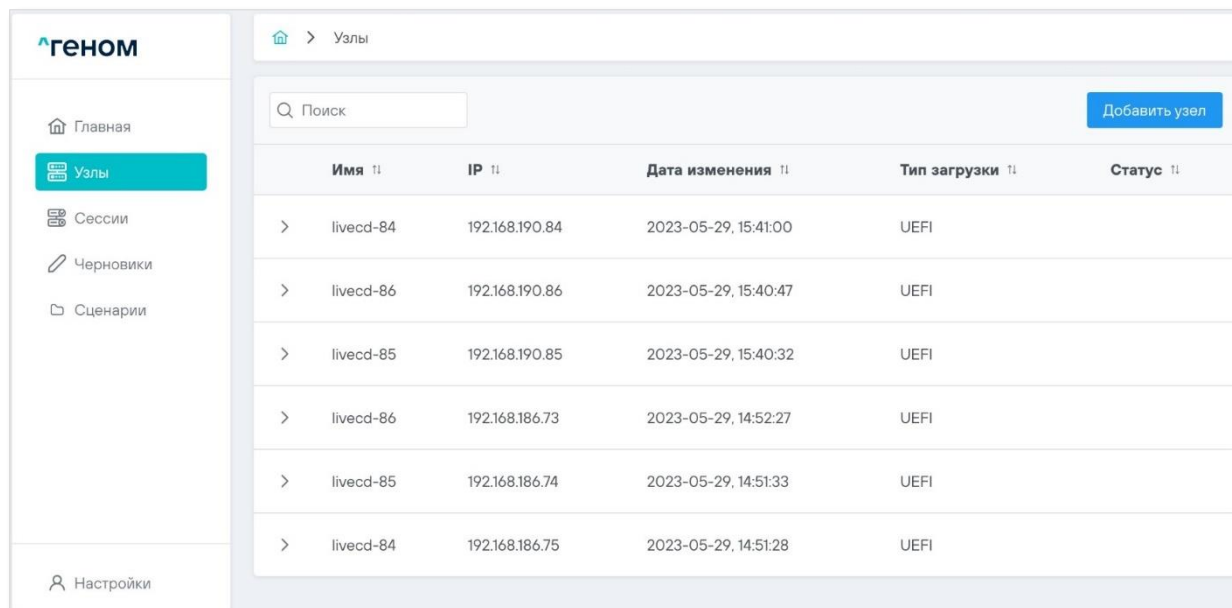


Рисунок 4. Пример интерфейса системы управления эксплуатацией Скала^р Геном

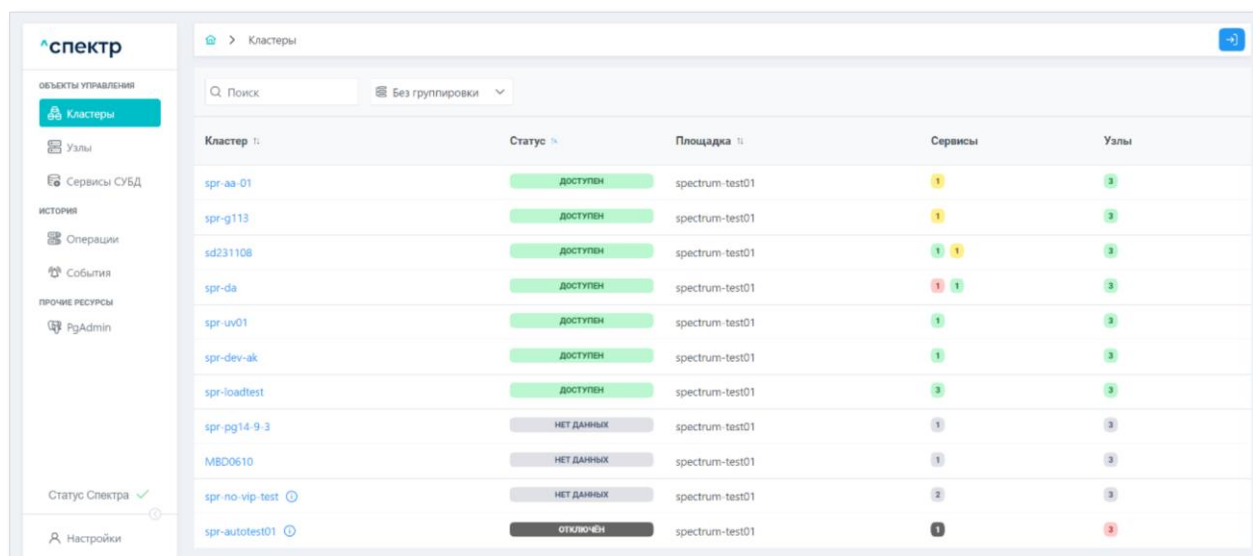


Рисунок 5. Пример интерфейса системы управления кластерами Скала^р Спектр

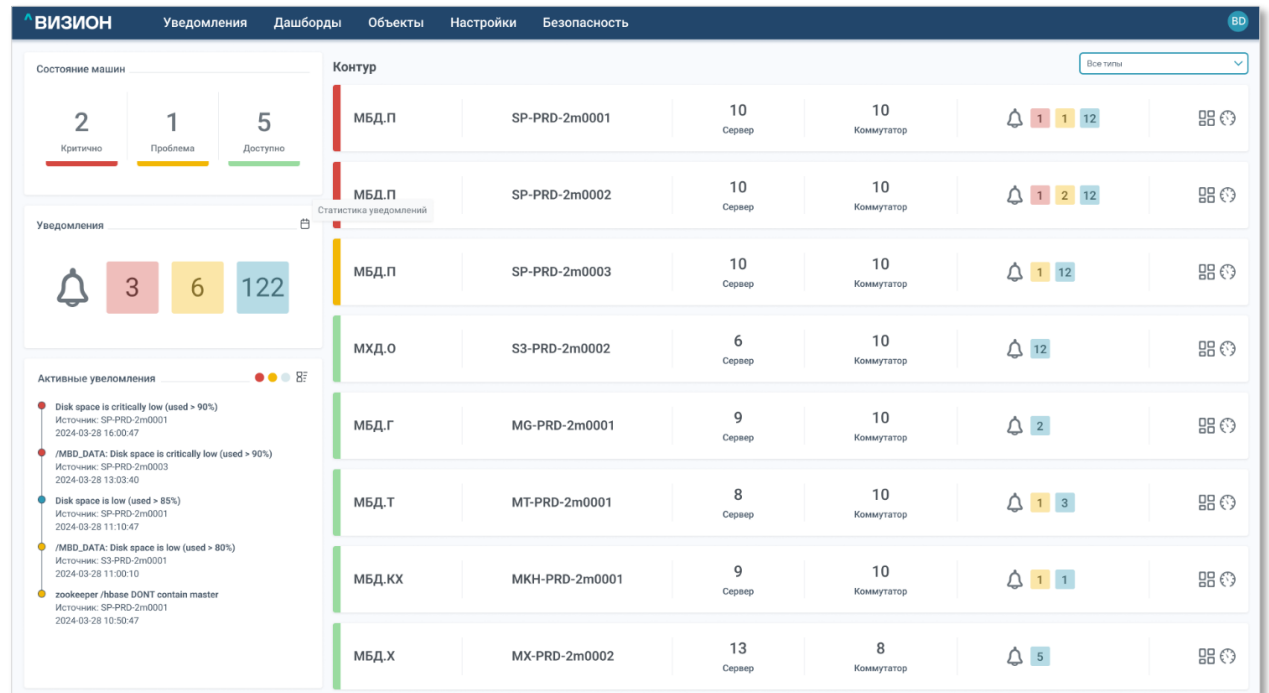


Рисунок 6. Пример экрана системы мониторинга Скала^р Визион

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
- Конфигурационные:
 - количество ЦПУ
 - объем памяти
 - имя объекта мониторинга
 - список сетевых адаптеров, их MAC- и IP-адреса
 - подключенные ресурсы хранения
- По утилизации ресурсов (счетчики):
 - загрузка ЦПУ (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 ЦПУ
- 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** — порт Ethernet 10/25 Гбит/с — для связи с внешними сетями
- T** — два блока питания в режиме резервирования по схеме (1 + 1)
- P** — 2×ЦПУ 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 ОС 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.

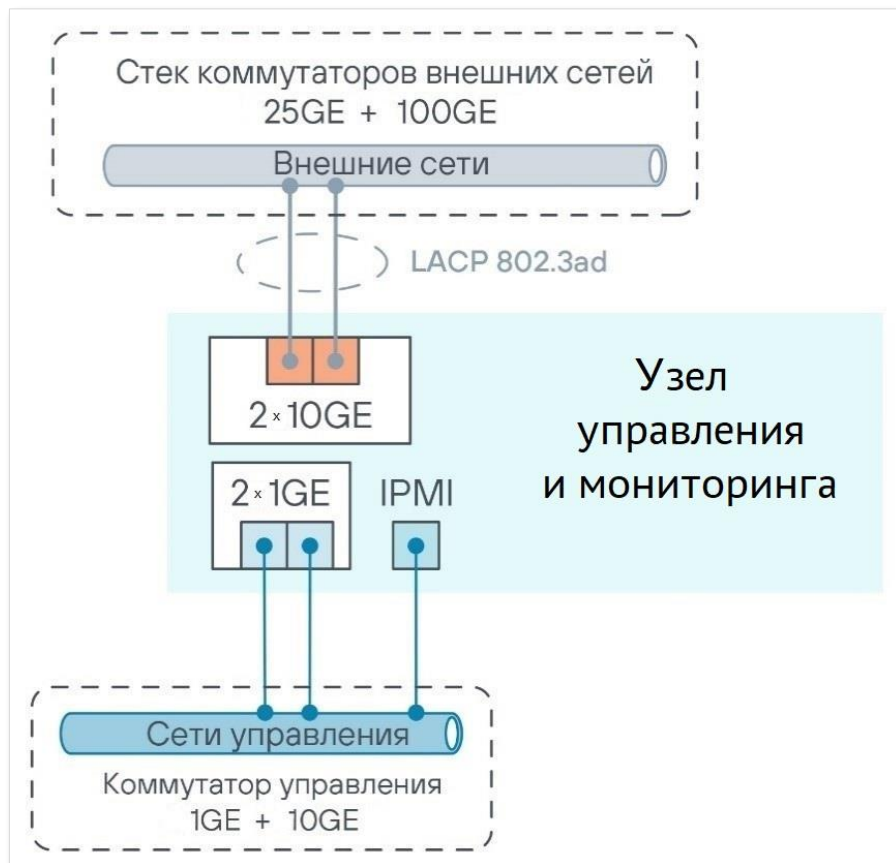


Рисунок 7. Схема сетевого взаимодействия узла управления

Применяемое программное обеспечение:

- T** — ОС: РЕД ОС с виртуализацией на служебных узлах с ПО **Скала^р Геном** и на Модуле транзакционной обработки; без виртуализации - на гостевых узлах
- T** — KVM для управления виртуальными машинами
- T P** — Сервер управления кластером узлов базы данных: **Скала^р Спектр**
- T** — Сервер управления жизненным циклом **Машины МБД.П** и ее компонентов: **Скала^р Геном**
- T** — Сервер системы собственного мониторинга: **Скала^р Визион**

7.1.2 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 объединены в один виртуальный коммутатор (стек) по технологии MLAG, что позволяет подключать узлы и другие устройства к паре коммутаторов, используя протокол LACP. На стеке коммутаторов внешних сетей реализован сетевой сегмент Public VLAN.

Для реализации сети интерконнекта используются два коммутатора с портами 100 Гбит/с. Коммутаторы объединены 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, объединенными в транк по протоколу LACP. На коммутаторе управления реализованы сегменты IPMI и PXE.

Основные функции:

- Т Р** — передача данных между элементами **Машины баз данных Скала^{Ар} МБД.П** (интерконнект)
- Т Р** — обеспечение информационного обмена с внешними устройствами
- Р** — обмен служебными данными, данными для мониторинга и управления

Общая схема сетевого взаимодействия приведена на Рисунке 8.

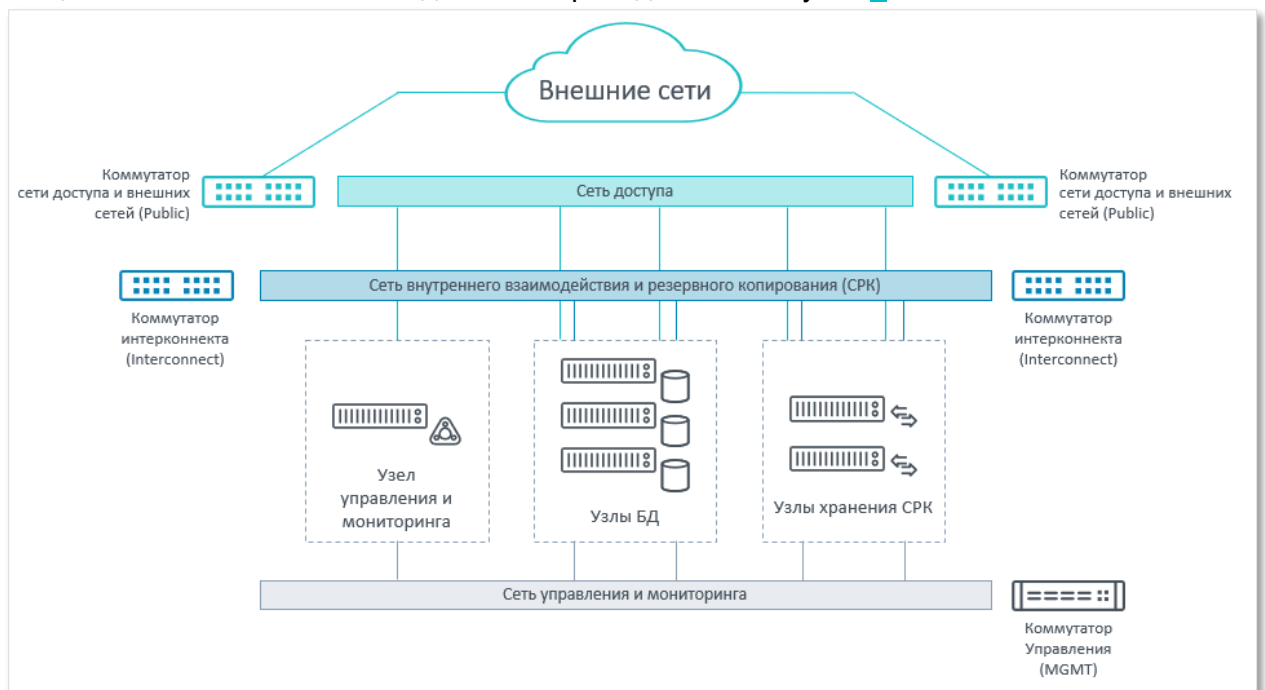


Рисунок 8. Общая схема сетевого взаимодействия

Состав оборудования сетевой подсистемы:

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

- TP** — Public VLAN — сеть для подключения к сервисам БД внешних пользователей и прикладных систем, подключение к узлу управления
- TP** — Interconnect VLAN — сеть для внутреннего взаимодействия между узлами БД, сеть резервного копирования, сеть кластерного взаимодействия
- T** — PXE VLAN — сеть для развертывания ОС по PXE, платформы МБД
- T** — Mgmt VLAN — для удаленного управления и мониторинга
- T** — IPMI VLAN — сеть управления оборудованием через интерфейсы удаленного управления

7.2 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 БД.

В состав **Машины баз данных Скала^Ар 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 выполняется 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.

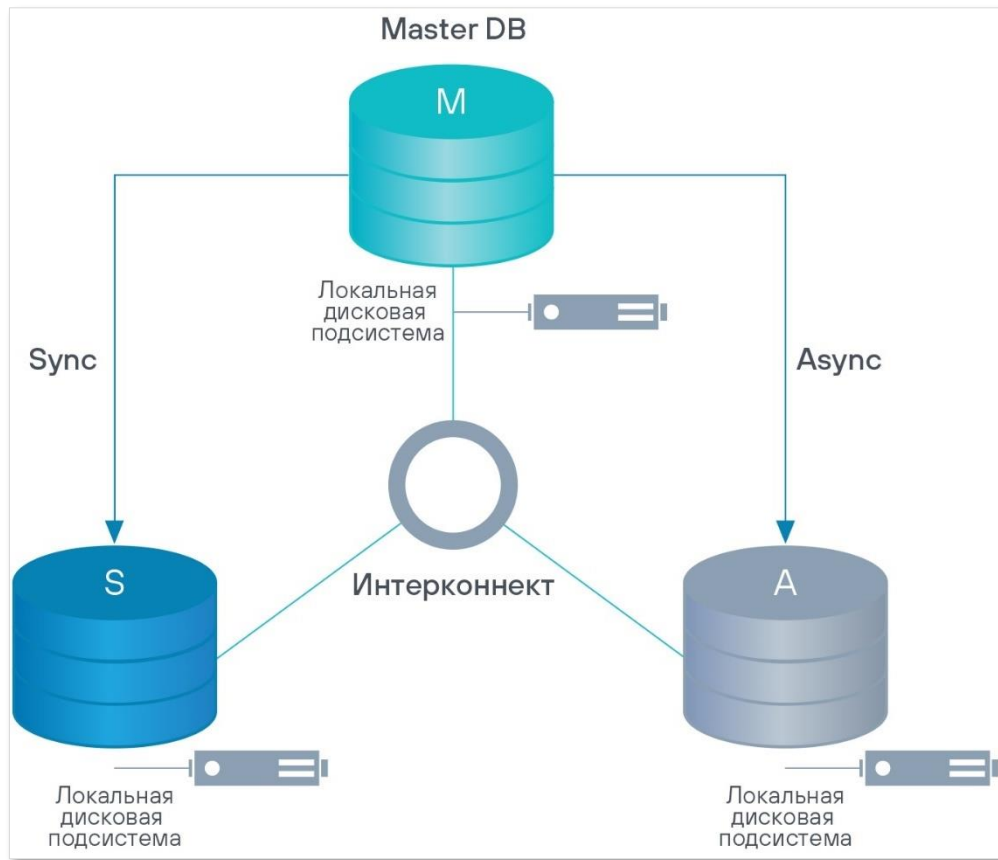


Рисунок 9. Реализация трехузлового кластера

В каждом отдельном узле 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 или RAID 1)
- T P** — все интерфейсы данных дублированы (стандарт IEEE 802.3ad LACP) — повышение производительности, отказоустойчивость (в случае отказа одного из интерфейсов возможно снижение производительности)
- P** — порт Ethernet 10/25 Гбит/с — для связи с внешними сетями
- P** — порт Ethernet 100 Гбит/с — для интерконнекта в рамках **Машины**
- T** — два блока питания в режиме резервирования по схеме (1 + 1)

P — 2×ЦПУ 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 ОС 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 ОС 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](#).

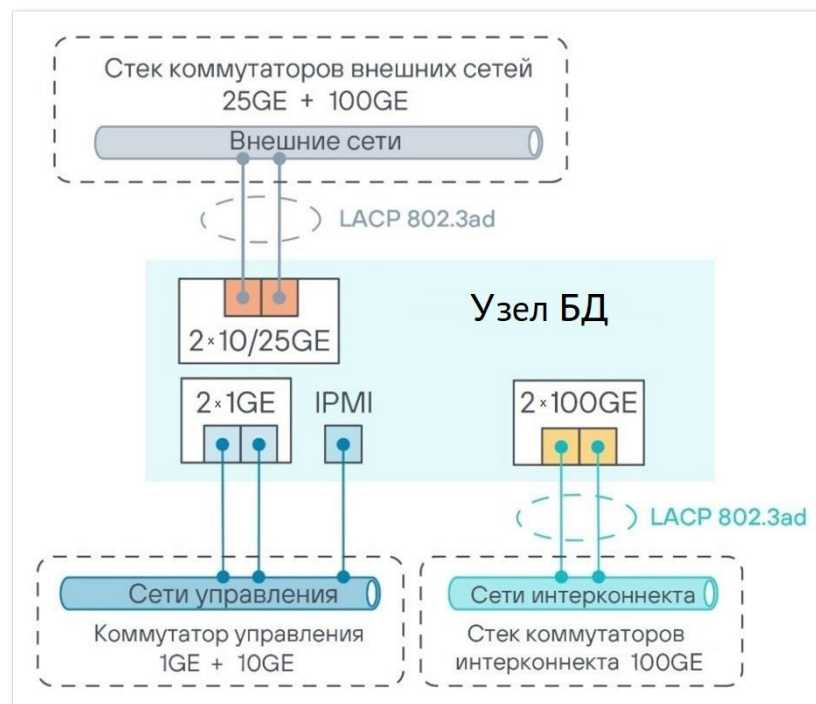


Рисунок 10. Схема сетевого взаимодействия вычислительного узла

Применяемое программное обеспечение:

T — ОС: РЕД ОС

P — СУБД: Postgres Pro Enterprise

T P — управление резервным копированием: pg_probackup

T P — управление кластером узлов БД: Скала^{Ар} 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

7.3 Модуль резервного копирования

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

Поддерживает проверку целостности данных резервных копий и управление политиками хранения.

Для выполнения указанных 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 метаданной, 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](#) ниже.

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

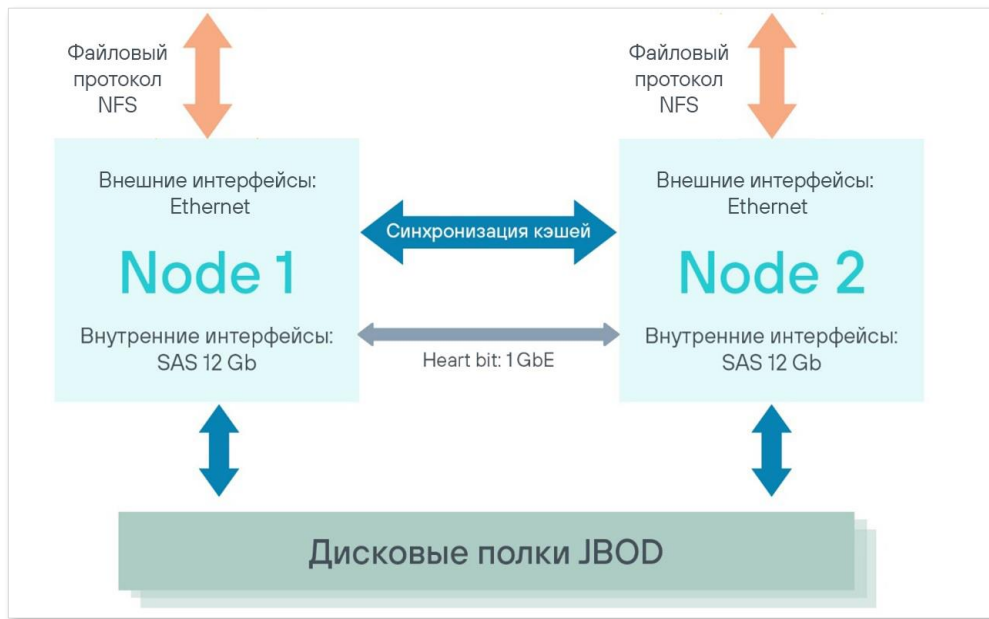


Рисунок 11. Реализация сервиса управления дисками СРК

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 дублированы (стандарт IEEE 802.3ad LACP) — повышение производительности, отказоустойчивость (в случае отказа одного из интерфейсов возможно снижение производительности)
- P** — порт Ethernet 10/25 Гбит/с — для связи с внешними сетями
- порт Ethernet 100 Гбит/с — для интерконнекта в рамках МБД
- P** — 2×100 Гбит/с Infiniband (/Ethernet) — для синхронизации кэша контроллеров
- T** — два блока питания в режиме резервирования по схеме (1 + 1)
- P** — 2×ЦПУ 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 ОС 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 ОС 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 ОС один логический агрегированный интерфейс, который используется для служебного трафика кластера (Heartbeat).

Схема сетевого взаимодействия контроллеров СХД:

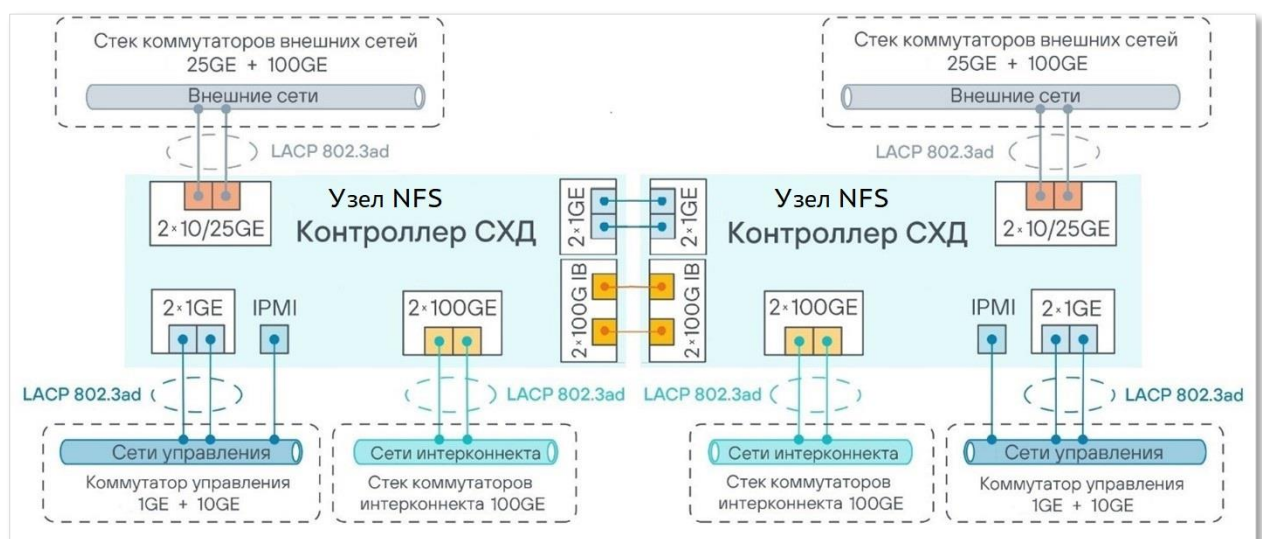


Рисунок 12. Схема сетевого взаимодействия контроллеров СХД

Применяемое программное обеспечение:



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

8 СПЕЦИФИЧНЫЕ ЧЕРТЫ

Проектирование и реализация **Машины баз данных Скала^Λр 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

Оптимальная конфигурация оборудования для максимальной производительности

Эффект:

- Максимум производительности на данном оборудовании
- 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 узлы обработки 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 для компонентов решения

Эффект:

- Обеспечение стабильной производительности: компоненты проверены временем
- Повышение уровня надежности: компоненты широко используются в индустрии

- Снижение стоимости сопровождения: доступность элементов при выходе из строя

Отказ от применения программных RAID в пользу аппаратных RAID

Эффект:

- Обеспечение более высокой производительности и эффективности
- Защита кэша RAID-контроллера от сбоя электропитания снижает риски потери данных

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

Эффект:

- Повышение качества за счет возможности доработки ПО силами **Скала^р** и партнеров
- Повышение надежности решения из-за наличия технической поддержки
- Лицензионная чистота используемых продуктов

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

Эффект:

- Возможность эксплуатации уже развернутых у заказчика решений по мониторингу и управлению ИТ-инфраструктурой в соответствии с внутренними стандартами
- Возможность построения сквозных систем управления, с включением **Машины баз данных Скала^р МБД.П** как одного из элементов инфраструктуры заказчика

9 ГАРАНТИРОВАННОЕ КАЧЕСТВО

Качественные показатели **Машины баз данных Скала^р МБД.П** обеспечиваются ее соответствием эталонному варианту конфигурации, 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**

Передача в эксплуатацию

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

Поддержка

- **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 в режиме одного окна 9×5 или 24×7 в зависимости от заказанного уровня сервиса
- 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**.

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:

- Функциональных узлов (серверов)
- Дублирование аппаратных компонентов: всех сетевых интерфейсов, источников питания и др.
- Применение дисковых RAID-массивов с требуемым уровнем отказоустойчивости
- Наличие запасных (spare) дисков в RAID-массивах
- Дублирование контроллеров подсистемы резервного копирования

Отказы перечисленных элементов обрабатываются 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 **МБД.П** также восстановится.

Отказы, связанные с узлами кластера баз данных

Для обеспечения бесперебойности доступа и сохранности данных в решении реализован 3-узловой 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 Скала^р** 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 штатной работы 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**.

Поскольку для **МБД.П** избран приоритет обеспечения сохранности данных, одновременный или последовательный отказ двух узлов кластера приводит к полной остановке **Машины** 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^Ар МБД.П** приведены на Рисунке [13](#).

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

**Машина баз данных
Скала^Ар МБД.П М-1**

**Машина баз данных
Скала^Ар МБД.П М-2**

**Машина баз данных
Скала^Ар МБД.П М-3/М-4**

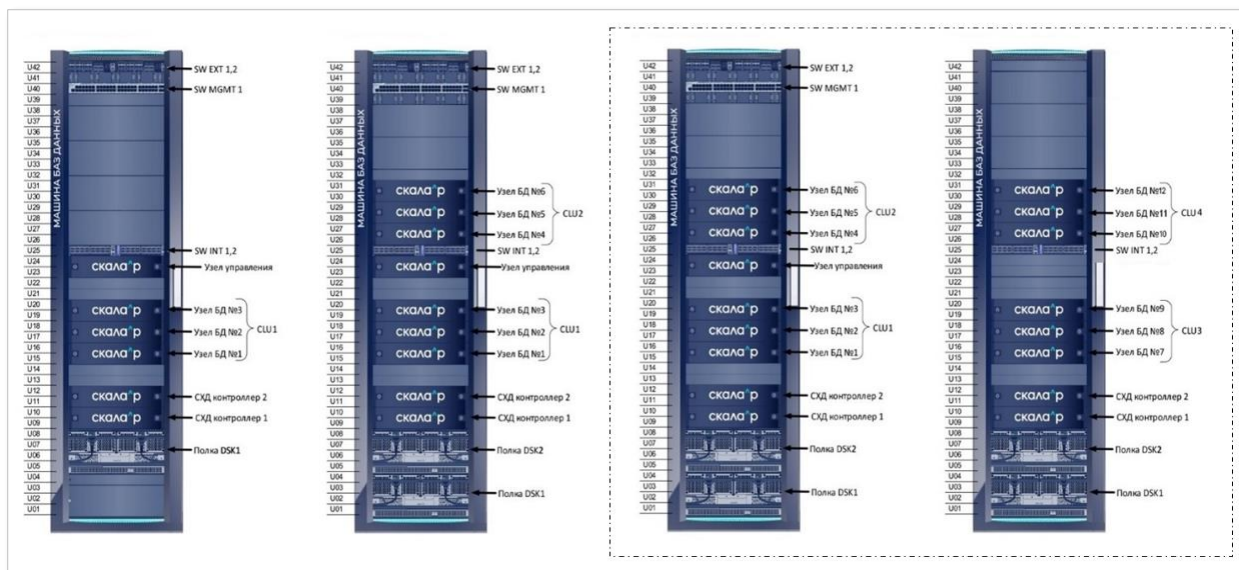


Рисунок 13. Типовые комплекты поставки Машины баз данных Скала^Ар МБД

Скала^Ар МБД.П поставляется комплектами по Модулям таким образом, чтобы можно было подобрать оптимальную конфигурацию от количества размещаемых баз данных на узлах БД. В Таблице [3](#) приведен пример размещения отказоустойчивого сервиса СУБД в Модулях баз данных, каждый из которых включает до 3 экземпляров СУБД.

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

Параметры / Модель	М-1	М-2	М-3	М-4
Количество узлов БД	3	2×3	3×3	4×3
Количество экземпляров СУБД	до 3	до 6	до 9	до 12
Общий полезный объем всех БД, Тбайт	до 300	до 2×300	до 3×300	до 4×300
Полезный объем хранения системы резервного копирования (БД+WAL), на одну полку, Тбайт	140-1360	2×	3×	4×
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

Высокопроизводительный продуктивный ландшафт OLTP

Вариант решения:

- больше памяти RAM (2 Тбайт)
- твердотельные накопители повышенного объема (16×15,36 TB + 2×15,36 TB)
- два современных аппаратных RAID контроллера
- RAID 50, RAID 1

Геокластер (как опция)

Вариант решения:

- дополнительные сетевые карты в узлах БД для связи между ЦОД
- специальные настройки кластерного ПО и ПО резервного копирования

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 Гбит/с Ethernet или до 8×10/25 Гбит/с 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: растянутый (метро-) кластер

Решение – две связанных **Машины баз данных Скала^Ар МБД.П**

Каждая СУБД имеет высокодоступную конфигурацию (мастер + синхронная реплика) на своей основной площадке, при этом обеспечивается асинхронное (или при необходимости 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 ЦОД происходит автоматически.

В составе каждой **Машины 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 растянутыми (метро-) кластерами осуществляется с помощью ПО **АСпектр**.

Реализация растянутого (метро-) кластера

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

Пример реализации растянутого (метро-) кластера:

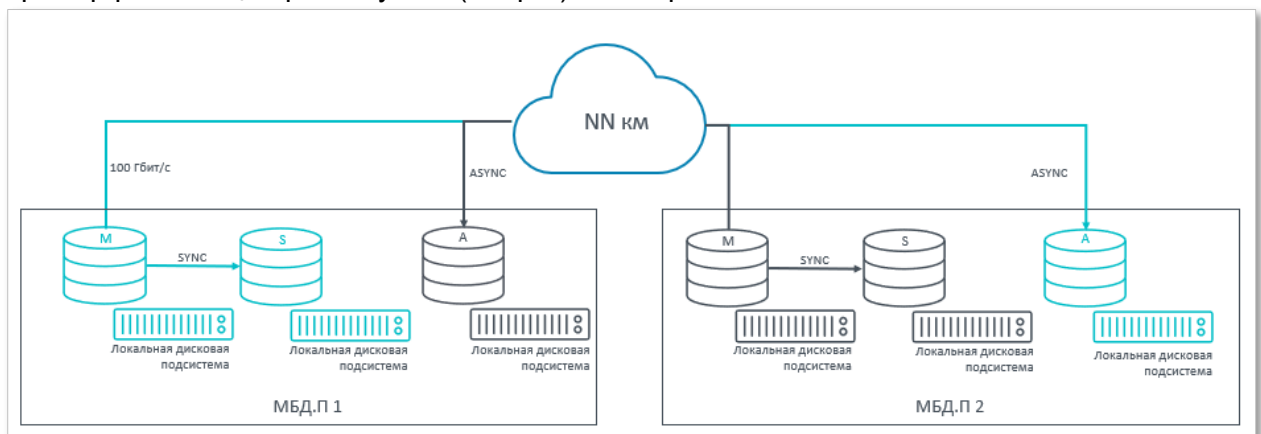


Рисунок 14. Пример организации метрокластера

Пример: катастрофоустойчивый (геораспределенный) кластер

Решение – две связанных **Машины баз данных Скала^Ар МБД.П**

Элементами обеспечения катастрофоустойчивости системы на уровне БД 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. 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 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 **^Спектр**

Реализация катастрофоустойчивого (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 15.

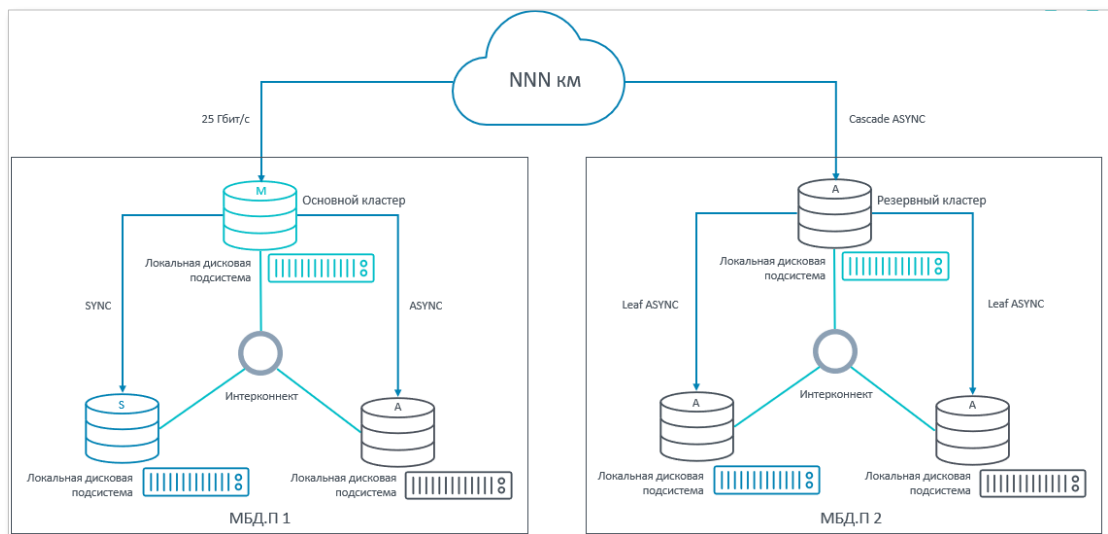


Рисунок 15. Пример реализации геокластера

Результаты тестирования

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

```
BEGIN;
UPDATE pgbench_accounts SET abalance = abalance + :delta WHERE aid = :aid;
SELECT abalance FROM pgbench_accounts WHERE aid = :aid;
UPDATE pgbench_tellers SET tbalance = tbalance + :delta WHERE tid = :tid;
UPDATE pgbench_branches SET bbalance = bbalance + :delta WHERE bid = :bid;
INSERT INTO pgbench_history (tid, bid, aid, delta, mtime) VALUES (:tid, :bid,
:aid, :delta, CURRENT_TIMESTAMP);
END;
```

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

Аппаратные характеристики:

- 2×ЦПУ Xeon 64 ядра
- 1024 Гбайт оперативной памяти
- Накопители SAS SSD

Параметр `pgbench sf = 205 000` приводит к созданию БД объемом ~3 Тбайт; функция объединения в пулы `pgbench` не использовалась, для каждого пользователя использовалось выделенное подключение к СУБД. Тесты проводились на РЕД ОС 8с для сертифицированной версии Postgres Pro Enterprise 18.3.2.

Результаты теста показаны на Рисунке 16.

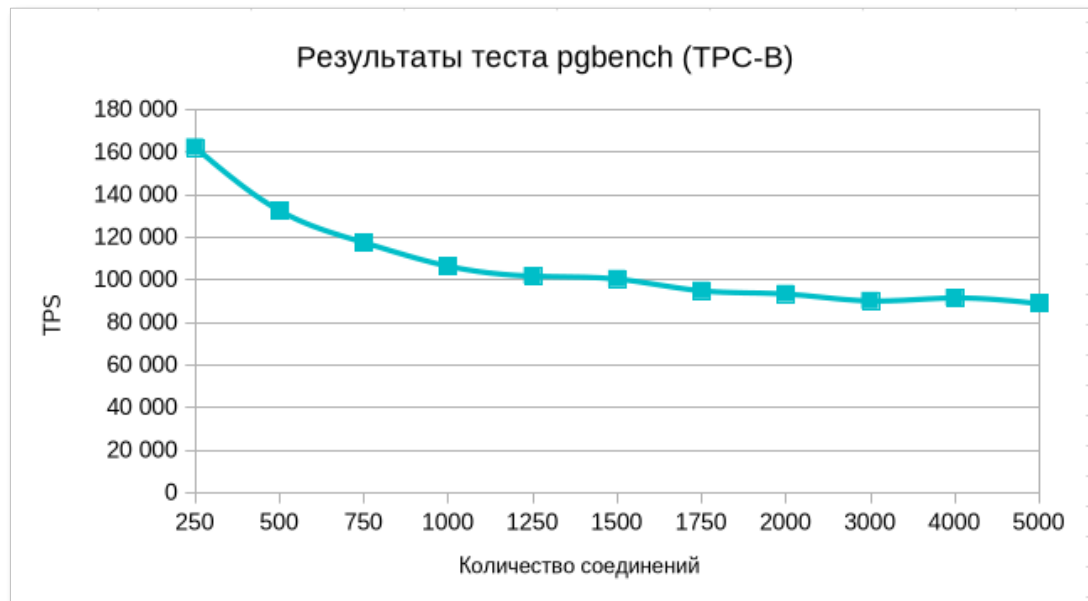


Рисунок 16. Производительность Машины баз данных Скала^р МБД.П по результатам теста pgbench

15 ТЕХНИЧЕСКАЯ ПОДДЕРЖКА

Поставка **Машины** осуществляется с предварительными сборкой, тестированием и настройкой оборудования согласно требованиям заказчика. Качественная поддержка **Машины** обеспечивается едиными стандартами гарантийного и постгарантийного технического обслуживания:

- пакет услуг по технической поддержке на первый год включен в поставку
- заказчик может выбирать пакет в базовом режиме 9x5 или в расширенном режиме 24x7 (опция для критической функциональности)
- срок начально приобретаемой годовой технической поддержки может быть увеличен до 3 и 5 лет, также доступна пролонгация поддержки по окончании сроков
- возможно включение в состав стандартных пакетов дополнительных опций и услуг

Состав типовых пакетов услуг по технической поддержке представлен в Таблице 4 ниже.

Таблица 4. Пакеты услуг по технической поддержке Скала^р

Услуга	Пакет «9x5»	Пакет «24x7»
Режим «Обслуживание комплекса Скала^р в режиме 9x5» (в рабочее время по рабочим дням)	+	–
Режим «Обслуживание комплекса Скала^р в режиме 24x7» (круглосуточно)	–	+
Предоставление доступа к системе регистрации запросов/инцидентов Service Desk	+	+
Предоставление доступа к базе знаний по продуктам Скала^р	+	+
Предоставление обновлений лицензионного ПО Скала^р	+	+
Диагностика, анализ и устранение проблем в работе комплекса Скала^р, включая: ▪ устранение аппаратных неисправностей ▪ техническое сопровождение ПО	+	+
Консультации по работе комплекса Скала^р	+	+
«Защита конфиденциальной информации» (неисправные носители информации не возвращаются заказчиком)	Опция	Опция
Замена и ремонт оборудования по месту установки	+	+
Доставка оборудования на замену за счет производителя	+	+

Услуга	Пакет «9x5»	Пакет «24x7»
Расширенные опции обслуживания	–	+
Времена реагирования и отклика, не более:		
Время регистрации обращений	30 минут, рабочие часы (9x5)	30 минут, круглосуточно (24x7)
Подключение специалиста к решению инцидентов критичного и высокого уровней	В течение 1 рабочего часа (9x5)	В течение 1 часа (24x7)

Примечание к срокам ремонта оборудования

Машина архитектурно является устойчивой к выходу из строя отдельных компонентов и даже узлов, поэтому нет необходимости в обеспечении дорогостоящего сервиса срочного восстановления оборудования в течение суток и менее. В **Машине** предусмотрено как минимум двойное резервирование основных компонентов, позволяющее сохранять данные и работоспособность даже при выходе из строя нескольких дисков и/или узлов вычисления (серверов).

Полное описание услуг поддержки доступно на сайте skala-r.ru.

16 ЗАКЛЮЧЕНИЕ

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

Основные черты **МБД.П**:

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

Машина и ее Модули произведены в РФ и внесены в реестры РЭП Минпромторга (в том числе как ПАК) и реестры ПАК Минцифры.

Структурно в **Скала^Ар МБД.П** входят:

- Базовый модуль
- Модуль баз данных
- Модуль резервного копирования

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

Надежность, производительность **МБД.П** подтверждается проведенными тестами, специальными расчетными данными, практическим использованием решений в течение ряда лет.

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

О КОМПАНИИ

Скала^р — модульная платформа для построения высоконагруженной ИТ-инфраструктуры (продукт Группы Rubytch).

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 Реестр промышленной продукции, произведенной на территории Российской Федерации, и в Единый реестр российских программ 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.

Дополнительная информация — на сайте www.skala-r.ru.