Машина баз данных Скала^р МБД.П

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

Технический обзор

версия 2.3 от 08.04.2024

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

ОГЛАВЛЕНИЕ

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.

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

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

1. ВВЕДЕНИЕ

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

Машина баз данных Скала^р МБД.П повышает производительность и отказоустойчивость, снижает затраты за счет проработанной интеграции аппаратного и программного обеспечения, оптимизации алгоритмов для используемых технологий, широкого применения методов обеспечения надежности, комплексности решения, специальных моделей лицензирования.

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

Машина баз данных Скала^р МБД.П — комплексное решение, включающее в себя масштабируемые модули для проведения вычислений и хранения данных, системы резервного копирования, а также сверхскоростную сетевую среду и систему интеллектуального управления.

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

Отказоустойчивость обеспечивается применением надежных комплектующих, специализированной версии СУБД (Postgres Pro Enterprise), кластеризацией, физическим резервированием критических компонентов, использованием устойчивых сетевых протоколов.

Машина баз данных Скала^р МБД.П содержит все необходимые элементы для функционирования высоконагруженной СУБД Postgres. Подключение к внешним сетям 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 Баз Данных 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 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 Единый реестр российской радиоэлектронной продукции и работают на ПО, включенном в реестр российского программного обеспечения. **Машина баз данных Скала^р МБД.П** также находится в реестре российского программного обеспечения.

2. ОТЛИЧИТЕЛЬНЫЕ ЧЕРТЫ

1. Надежное хранение и высокопроизводительная обработка больших объемов данных

- Объем баз данных до 40 ТБ при высоких нагрузках и с минимальным временем на резервное копирование и восстановление
- Объем баз данных до 160 ТБ при средних и низких нагрузках, и для систем с большим окном времени для резервного копирования и восстановления
- Производительность более 65 000 TPS по тестам pgbench
- Формирование катастрофоустойчивых решений

2. Высокая производительность

- Сбалансированный комплект оборудования
- Архитектурная оптимизация производительности
- Оптимизированная локальная система хранения
- Специальные настройки программного обеспечения
- Улучшенная функциональность ведения журналов предзаписи
- 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 данных программным 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 собственная система мониторинга Скала^р Визион
- Мониторинг работоспособности СУБД и оборудования
- Преднастроенные пороговые значения критичных параметров
- Различные каналы информирования об отклонениях

7. Улучшенные возможности администрирования

- Встроенная собственная система автоматизации администрирования и обслуживания Машины Скала^р Геном
- Автоматизированные действия по выполнению сложных операций с кластером и узлами
- Расширяемая библиотека сценариев для управления Машиной
- Сохранены все стандартные механизмы управления Postgres

8. Обеспечение эксплуатации

- Централизованная поддержка решения
- Единая ответственность 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 обоснованно 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

3. 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:

1. 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

2. 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:

1. 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

2. 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

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 27.12.2025):

   - «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.Б2.13» (ФСТЭК России, 2012), «Профиль защиты средств антивирусной защиты тина В второго класса защиты. ИТ.САВ3.В2.П3» (ФСТЭК России, 2012), «Профиль защиты средств антивирусной защиты типа Г второго класса защиты

2. Сертифицированное средство доверенной загрузки ПК «Соболь» версия 4:

   - Подтверждает соответствие требованиям руководящих документов к средствам доверенной загрузки, а также 2 уровню доверия средств технической защиты безопасности и обеспечения безопасности информационных технологий и возможность использования в ИСПДн до УЗ1 включительно, в ГИС до 1-го класса защищенности включительно и в ЗОКИИ до 1 категории включительно

4. 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 присутствует в Едином реестре российской радиоэлектронной продукции Минпромторга РФ (РЭП МПТ) согласно Постановлению Правительства РФ № 878.

Машина баз данных Скала^р МБД.П может 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 представлены ниже (Таблица 1).

Таблица 1. Товарные позиции Машин и модулей

Код по ОКПД 2	Примечание
26.20.14.160. Программно-аппаратные комплексы, созданные на серверах или устройствах, содержащие в своем составе один или более вычислительных узлов	Для Машин
26.20.14. Машины вычислительные электронные цифровые, поставляемые в виде систем для автоматической обработки данных	Для модулей

Наличие **Машины баз данных Скала^р МБД.П** на сайте государственной информационной системы промышленности показано ниже (Рисунок 1, фрагмент страницы https://gisp.gov.ru/goods/#/product/3738080).

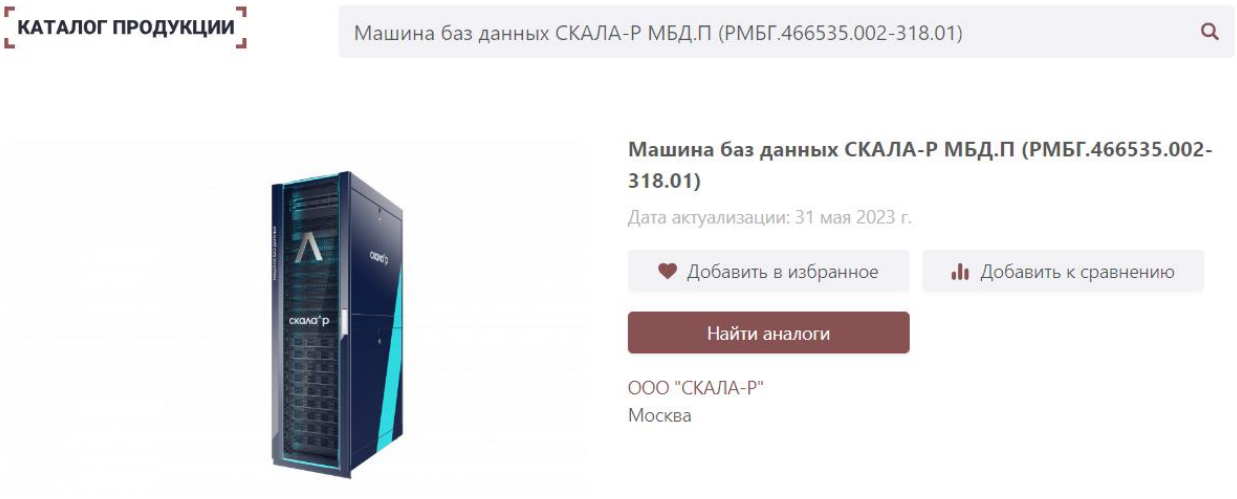


Рисунок 1. Машина баз данных Скала^р МБД.П на сайте государственной информационной системы промышленности (ГИСП)

Подробная информация о Машинах и модулях **Машины баз данных Скала^р МБД.П**, включенных в Единый реестр российской радиоэлектронной продукции, представлена в таблицах ниже (Таблица 2 и Таблица 3).

Машины и модули различаются исполнением (код .0X в номере конструкторской документации). Предлагается на момент написания данного документа 6 исполнений, различающихся серверной платформой (материнской платой производства РФ).

Таблица 2. Подробная информация о Машинах Скала^р МБД.П, включенных в РЭП МПТ

Наименование Машины (разработан согласно Техническим условиям РМБГ.466535.002ТУ)	**Код изделия по ОКПД2**
Машина баз данных Скала^р МБД.П (РМБГ.466535.002-318.0x)	26.20.14.160

Таблица 3. Подробная 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.466535.003-10.0x)	26.20.14
СКАЛА-Р Модуль баз данных (РМБГ.466535.003-60.0x)	26.20.14
СКАЛА-Р Модуль резервного копирования (РМБГ.466535.003-40.0x)	26.20.14
СКАЛА-Р Модуль расширения СРК (РМБГ.466535.003-68.0x)	26.20.14

5. 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 массиву хранения данных. Узел используется для размещения программного обеспечения СУБД, сами данные хранятся в массиве и по мере необходимости передаются по сети. Используются 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. В **Машине баз данных Скала^р МБД.П** используются исключительно SSD диски, при этом журнал WAL (Write Ahead Log) ведется на выделенных сверхбыстрых NVMe SSD дисках.

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, таким образом, ведущий узел может нести только транзакционную нагрузку, освобождаясь от ресурсоемких выборок, которые могут быть выполнены в синхронном режиме с синхронной реплики или в асинхронном режиме с асинхронной.

Проработанность всех программных компонентов

Основными программными элементами **Машины баз данных Скала^р МБД.П** являются ПО СУБД Postgres, ПО мониторинга и администрирования, ПО управления кластером СУБД, ПО резервного копирования и ряд других.

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 действительно высокую производительность решения, в то же время

был выявлен ряд направлений, в отношении которых были возможны дальнейшие улучшения.

Поскольку аппаратные возможности в **Машине баз данных Скала^р МБД.П** используются практически полностью, дальнейшие улучшения возможны только за счет развития интеллектуальных составляющих и в первую очередь используемого программного обеспечения.

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 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.

Настройки параметров потоковой репликации и архивирования журналов предзаписи (WAL) позволяют добиться существенного повышения производительности решения, а также заметно сократить время на восстановление нормального функционирования кластера.

Дополнительно при настройке системы управления кластером решается ряд проблем, в том числе связанных с обеспечением корректного прохождения трафика между узлами кластера, связи узлов кластера с узлом управления, с разрешением имен и других.

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, конкретного комплекта оборудования. В указанных условиях ряд настроек осуществляется непосредственно при развертывании решения.

Существенные улучшения производительности были достигнуты за счет доработки и **совершенствования ПО управления RAID-массивами**, используемыми непосредственно для хранения данных. Дополнительно это привело к возможности полного отказа от аппаратной реализации RAID-массивов.

Следующими ключевыми компонентами, кардинально улучшенными в **Машине баз данных Скала^р МБД.П**, явились **алгоритмы и настройки ПО управления резервным копированием.** Резервные копии важны не только для случаев фатальных аварий, но и при незначительных сбоях, а также при проведении регламентных работ на узлах. Время их формирования и время восстановления из них узла кластера являются

значимыми, а часто — критичными параметрами. В результате проведенных доработок и улучшения применяемых алгоритмов удалось добиться существенного сокращения этих показателей.

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 «одного окна», а не от нескольких разных поставщиков, как бывает при самостоятельном подборе, развертывании и настройке компонентов в случае традиционного подхода.

6. СОСТАВ РЕШЕНИЯ

Ниже приведены 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.

Блок — группа однотипных модулей или узлов, выполняющих единую функцию в одной или нескольких стойках.

Модуль — это структурный элемент **Машины баз данных Скала^р МБД.П**, выполняющих 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 функциональных блоков модульной архитектуры **Машины баз данных Скала^р МБД.П**, объединенных в один серверный шкаф.

Формирование решения основано на принципе разделения на блоки и модули.

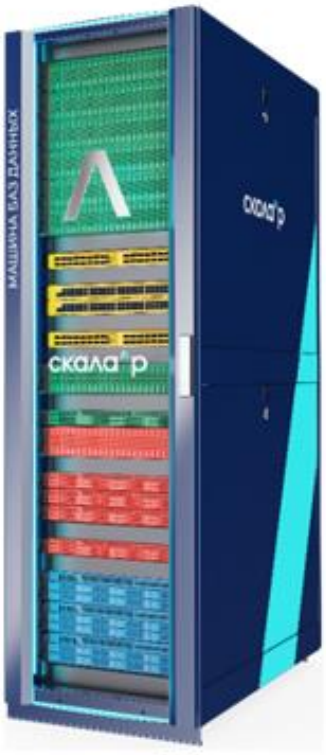
Машина баз данных Скала^р МБД.П состоит из следующих блоков:

Блок вычисления и хранения

Блок резервного копирования

Блок мониторинга и регистрации

Блок коммутации и агрегации



Блок вычисления и хранения объединяет модули баз данных. Блок мониторинга и регистрации содержит узел управления. Блок коммутации и агрегации объединяет сетевые узлы Машины. Блок резервного копирования содержит узлы резервного копирования и его расширения или модули хранения.

Машина баз данных Скала^р МБД.П основана на принципе модульности. Машина 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 на разных серверных платформах. 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 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, описанных ниже.

К технологическим принципам проектирования относятся*:

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 помечены символом P, если они ориентированы на обеспечение производительности (Performance).*

Блок вычисления и хранения содержит кластер из трех узлов, в котором находятся базы данных и журналы WAL.

Взаимодействие между узлами кластера, а также между блоками **Машины баз данных Скала^р МБД.П** осуществляется с помощью блока коммутации и агрегации, который обеспечивает внутренний интерконнект на высокой скорости. Блок коммутации и агрегации имеет выделенную сеть для управления и мониторинга, а также блок имеет возможность подключения к внешним сетям.

Блок мониторинга и регистрации содержит узел управления, 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).

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

Схема внутренней коммутации **Машины баз данных Скала^р МБД.П** представлена ниже (Рисунок 2).

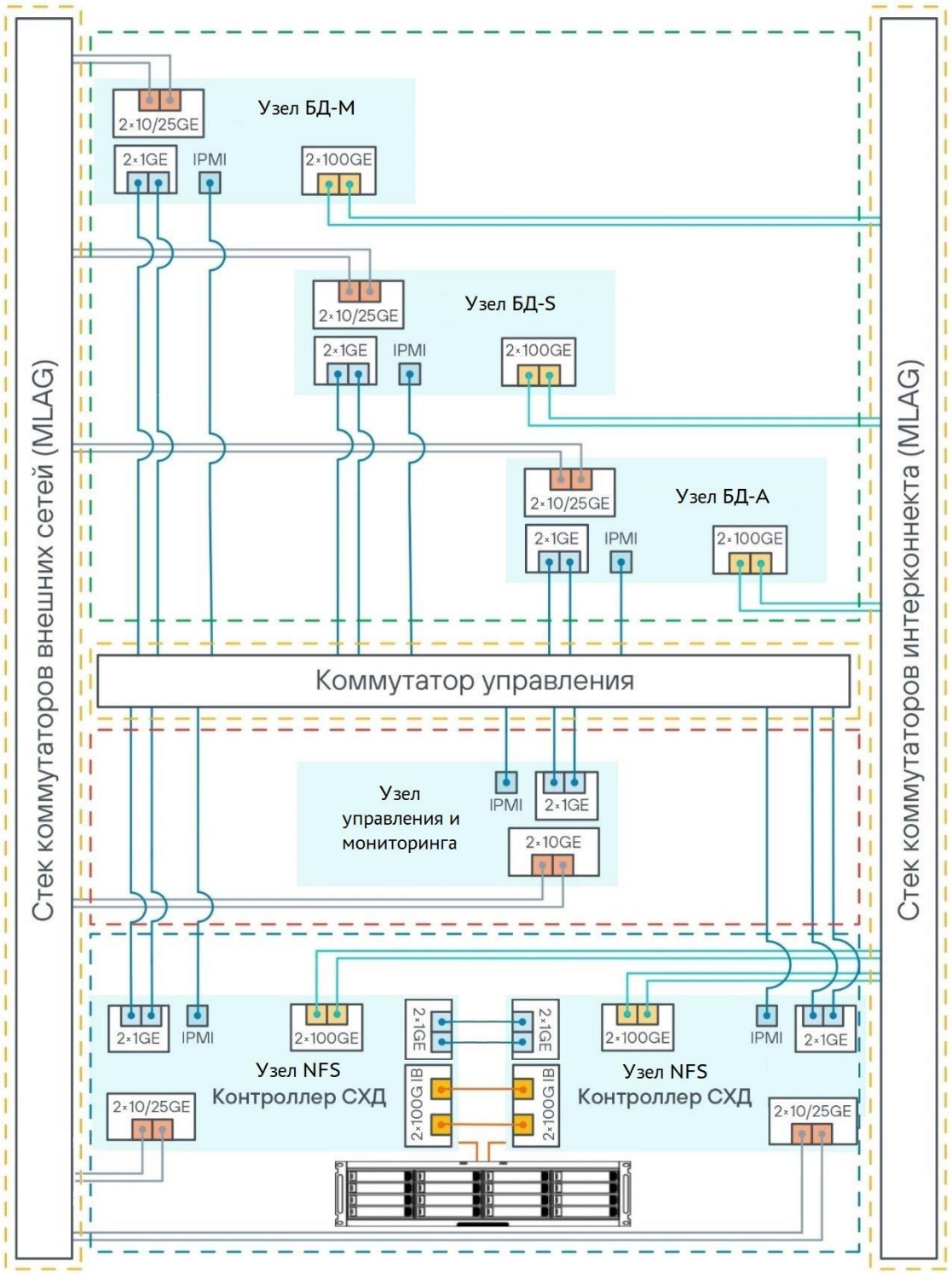


Рисунок 2. Схема внутренней коммутации Скала^р МБД.П

Блок вычисления и хранения

Блок вычисления и хранения состоит одного или нескольких Модулей 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 ниже (Рисунок 3).

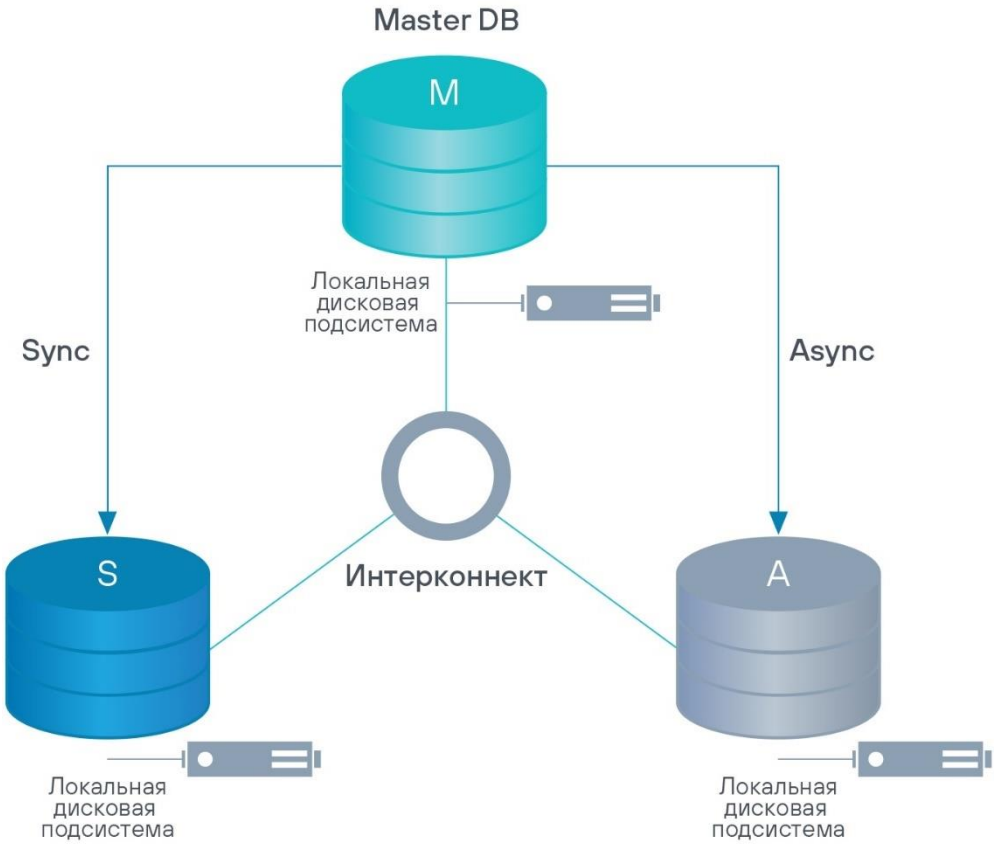


Рисунок 3. Реализация трехузлового кластера

Каждый отдельный узел БД:

P — использование SAS SSD для томов данных и NVMe SSD для 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 ниже (Рисунок 4).

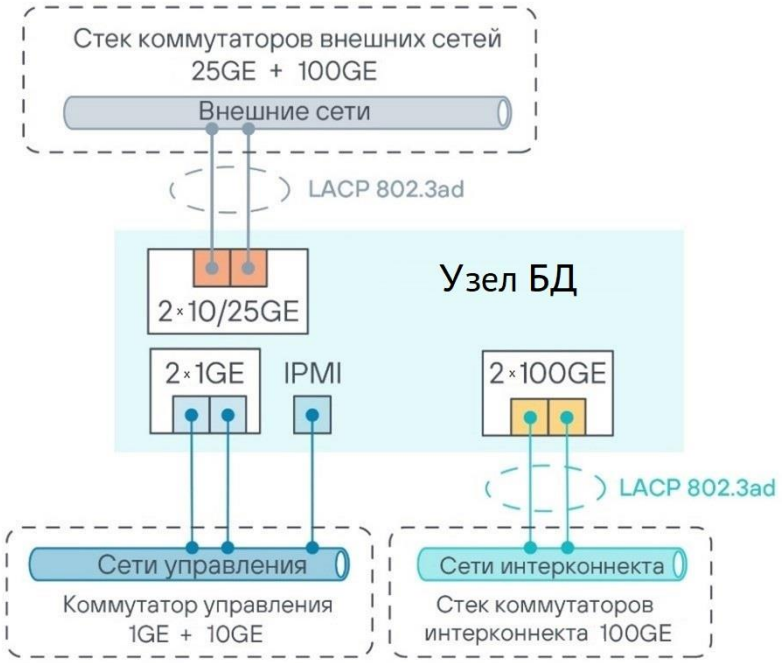


Рисунок 4. Схема сетевого взаимодействия вычислительного узла

Применяемое программное обеспечение:

T — ОС: Альт 8 СП, Альт 10 СП, Astra Linux (как варианты)

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 — программное обеспечение для управления локальными дисками: Raidix Era, в случае использования программно-определяемого способа создания RAID-массивов на узлах Модуля БД (предпочтительный и более производительный по сравнению аппаратными RAID-контроллерами способ).

Блок резервного копирования

Блок резервного копирования предназначен для хранения архивных файлов транзакционного журнала WAL, создания полных и инкрементальных резервных копий баз данных кластеров, восстановления баз данных, проверки целостности данных и резервных копий, управления 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.

Подключение узлов резервного копирования к узлам баз данных возможно различными протоколами (при установке карт Fibre Channel – также по протоколу FC), но в **Машинах баз данных Скала^р МБД.П** используется исключительно протокол NFS по высокоскоростным соединениям Ethernet.

Реализация сервиса управления дисками СРК представлена ниже (Рисунок 5).

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.

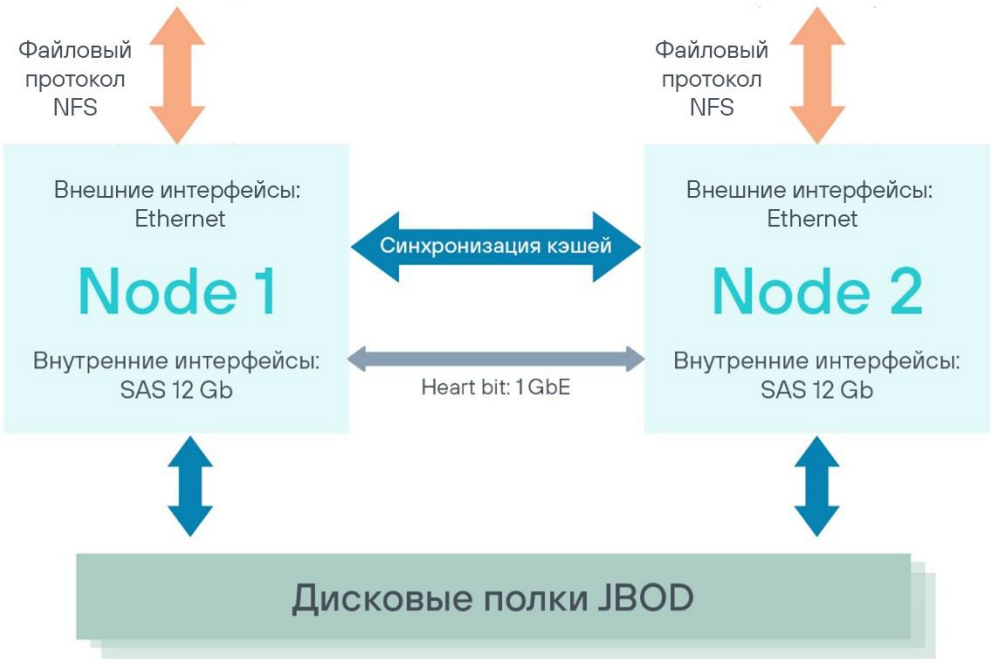


Рисунок 5. Реализация сервиса управления дисками СРК

Каждый отдельный контроллер:

P — использование общей дисковой полки с 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 блоке 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 Скала^р SDS 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, который используется для служебного трафика кластера (Heartbeat).

Схема сетевого взаимодействия контроллеров СХД представлена ниже (Рисунок *6*).

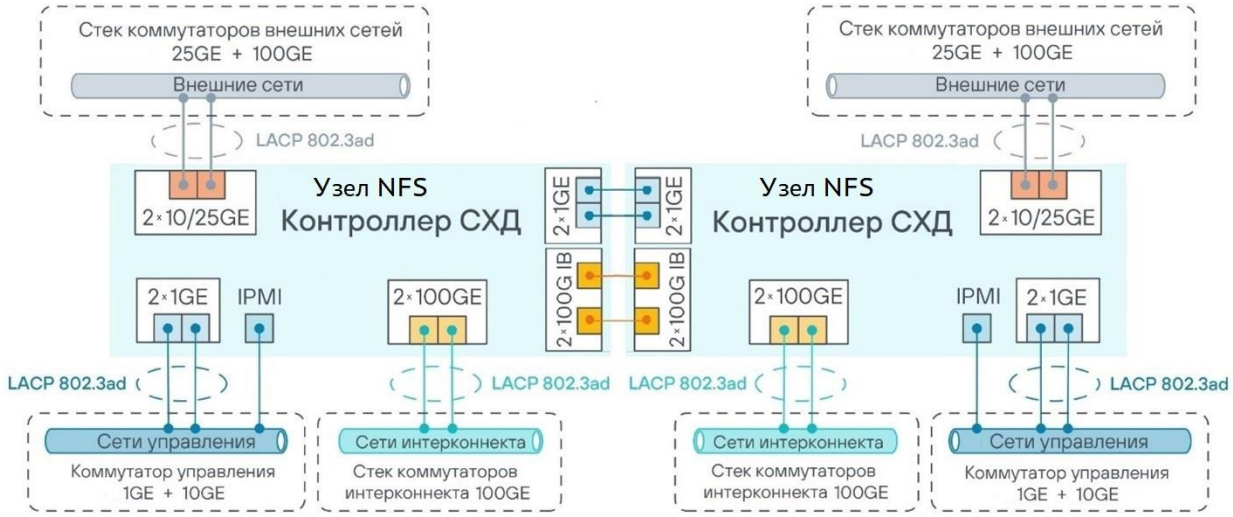


Рисунок 6. Схема сетевого взаимодействия контроллеров СХД

Применяемое программное обеспечение:

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

Блок мониторинга и регистрации

Блок мониторинга и регистрации реализован на одном физическом узле, на который 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 репозитарий необходимых пакетов для операционных систем, оборудования и СУБД Postgres Pro.

Для установки операционной системы используются два локальных диска, собранных в RAID 1. В качестве операционной системы узла мониторинга используется Альт Сервер 8.4 SP. Для хранения данных мониторинга и репозитария 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, обеспечивающих 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, обеспечивается через установленного агента 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 ниже (Рисунок *7*).

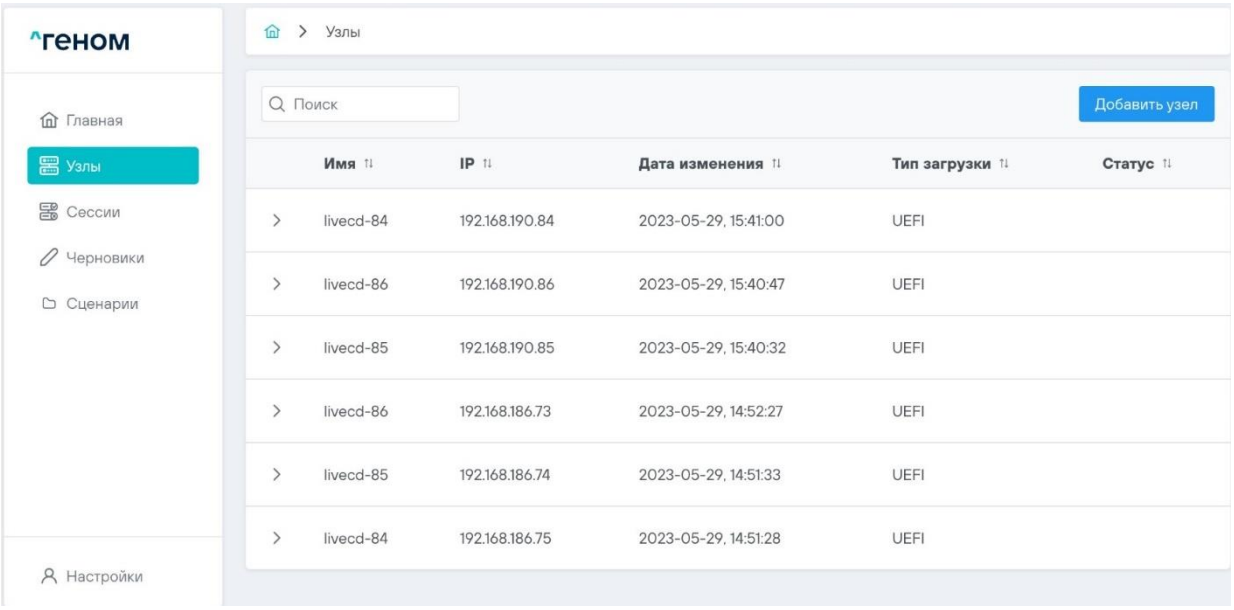


Рисунок 7. Пример интерфейса системы управления эксплуатацией Геном

Пример интерфейса системы управления кластером представлен ниже (Рисунок *8*).

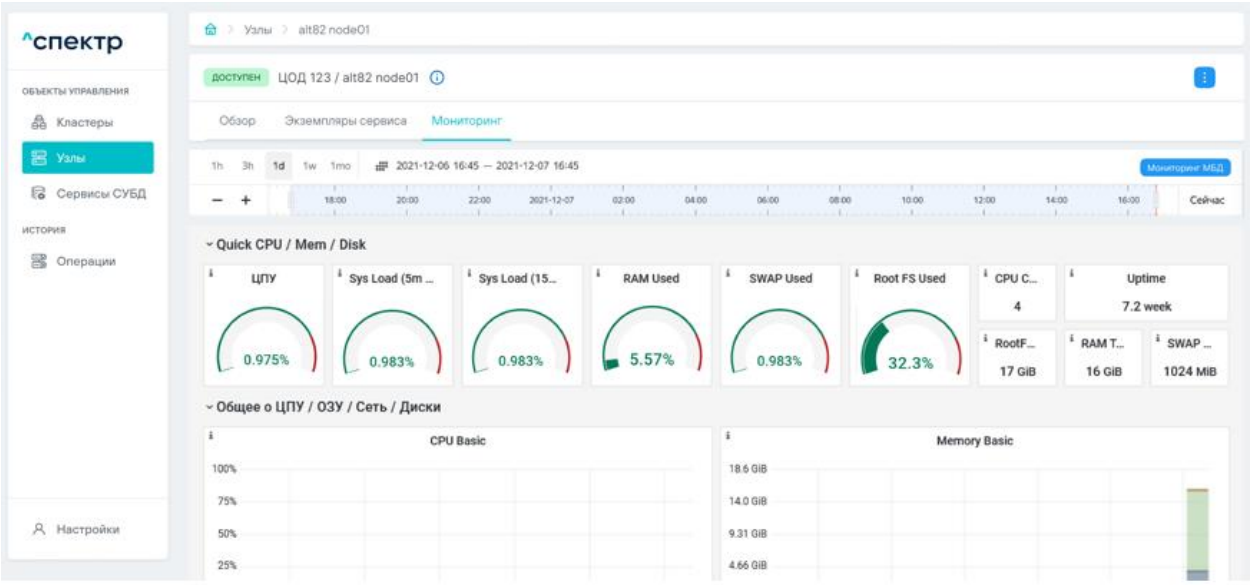


Рисунок 8. Пример интерфейса системы мониторинга кластера Спектр

Пример интерфейса системы мониторинга представлен ниже (Рисунок *9*).

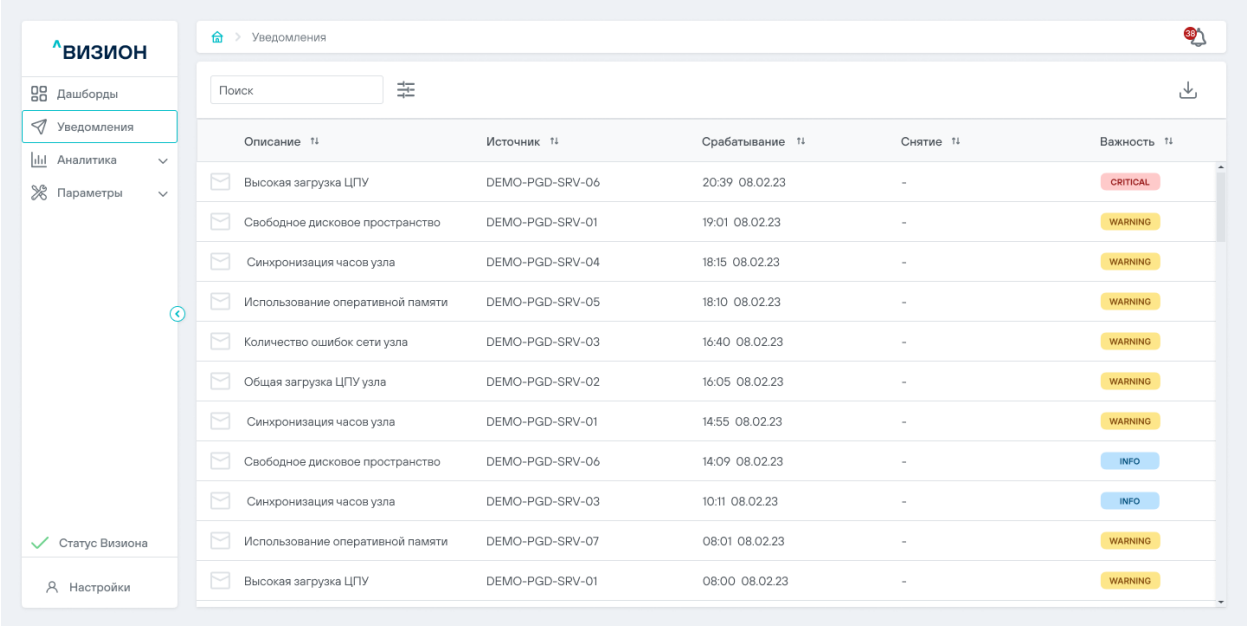


Рисунок 9. Пример экрана системы мониторинга Визион

В блоке мониторинга и регистрации 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
- конфигурационную:
  - 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 оборудования:

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

Для приведенных параметров объектов мониторинга блок 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 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 сервисам.

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 ниже (Рисунок 10)

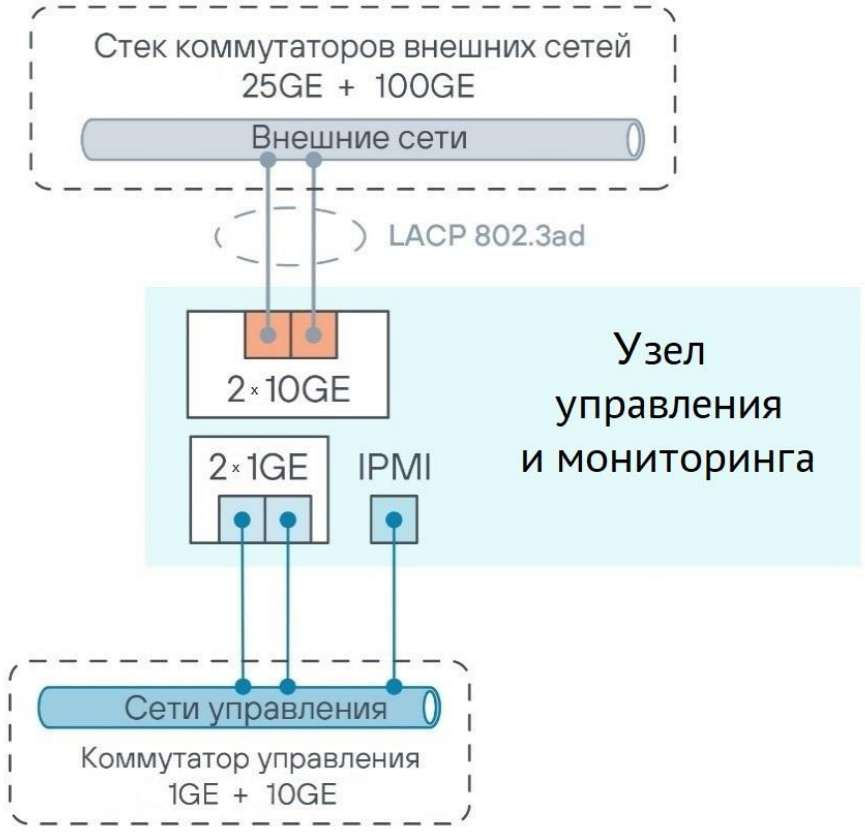


Рисунок 10. Схема сетевого взаимодействия узла управления

Применяемое программное обеспечение:

T — ОС: Альт Линукс, Альт 10 СП, Astra Linux (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.

Блок коммутации и агрегации

Блок коммутации и агрегации 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 ниже (Рисунок 11).

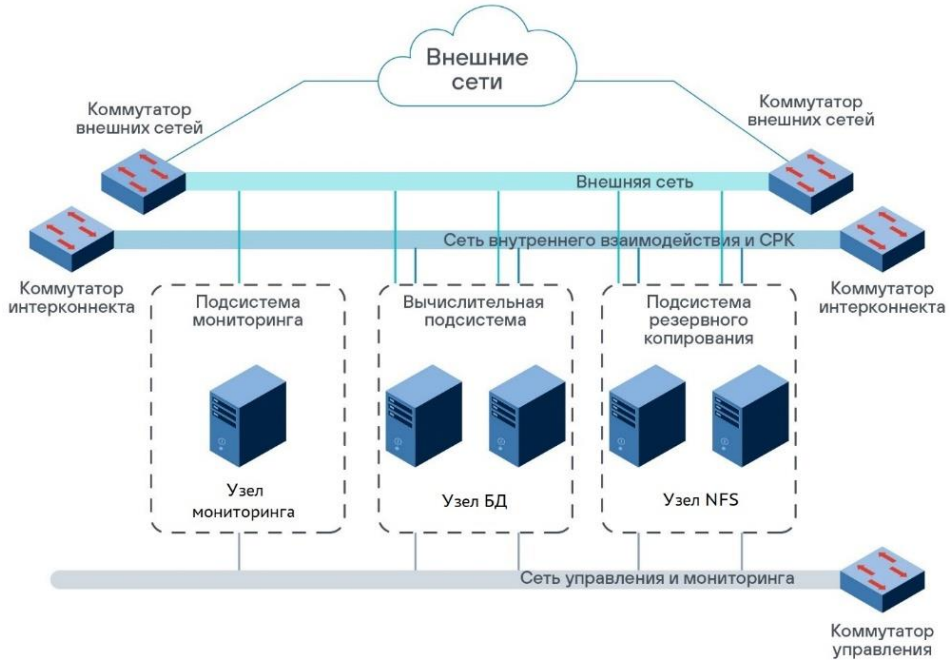


Рисунок 11. Общая схема сетевого взаимодействия

Состав оборудования блока:

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.

7. 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 блока 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 дисковой подсистемы (нет использования сетей для доступа к данным)
- Повышение производительности (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 диски 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 отдельных аппаратных RAID в пользу программных RAID отечественных производителей

Эффект

- Обеспечение более высокой производительности
- Высокая гибкость в настройках (в зависимости от требований)
- Уверенность в реализации оптимальных алгоритмов
- Снижение зависимости от производителей оборудования

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, в которых **Машина баз данных Скала^р МБД.П** — лишь один элемент

8. 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 и развертывание ПО)

- 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**
- Отклонения от типового решения **Машины баз данных Скала^р МБД.П** исключены

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**.

9. 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.)
- дисковой подсистемы (RAID)
- внешних сетей и интерконнекта (полное дублирование)
- системы резервного копирования (дублирование контроллеров)

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.

10. ТИПОВЫЕ КОМПЛЕКТЫ РЕШЕНИЯ

Типовые комплекты поставки **Машины баз данных Скала^р МБД.П** приведены ниже (Рисунок 12).

***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**

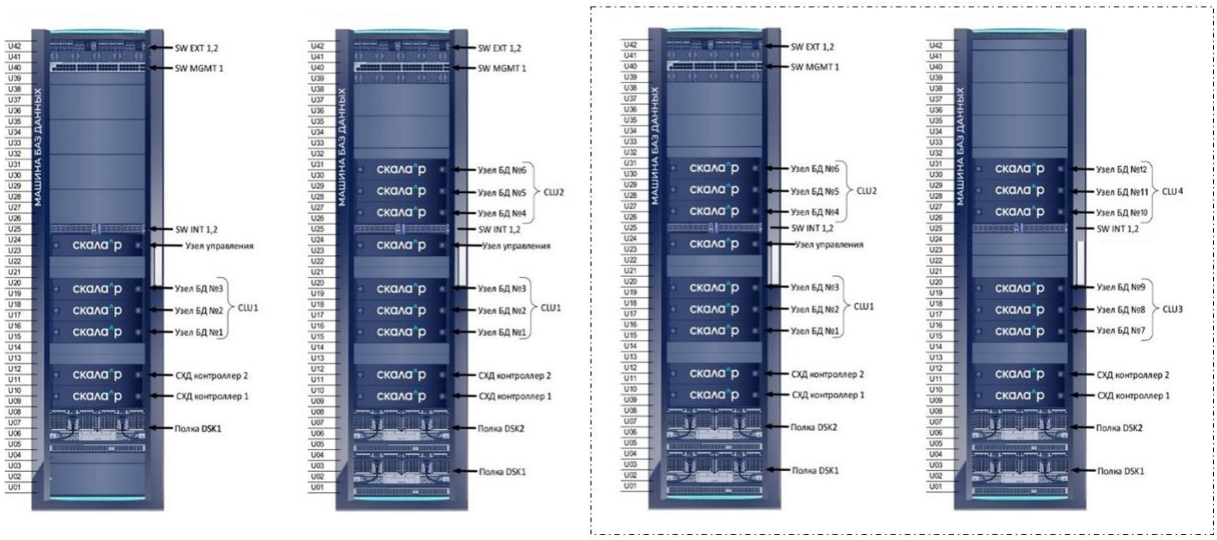


Рисунок 12. Типовые комплекты поставки Машины баз данных Скала^р МБД

Таблица 1. Параметры поставляемых моделей (производительный профиль, до 160 ТБ)

Параметры / Модель	М-1	М-2	М-3	М-4
Количество Модулей баз данных	1	2	3	4
Количество узлов БД	3	2x3	3x3	4x3
Количество экземпляров СУБД	до 3	до 6	до 9	до 12
Общий полезный объем всех БД, ТБ	до 160	до 2x160	до 3x160	до 4x160
Полезный объем хранения системы резервного копирования (БД+WAL), на одну полку, ТБ	190-260	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.

11. ВАРИАТИВНОСТЬ РЕШЕНИЯ

Малый или тестовый ландшафт

Вариант решения:

- достаточно памяти RAM (384 ГБ)
- высокопроизводительные накопители SSD (8 × 1,92 TB + 2x NVMe для WAL)
- программный RAID
- RAID 10

Высокопроизводительный продуктивный ландшафт OLTP

Вариант решения:

- больше памяти RAM (768 ГБ)
- твердотельные накопители повышенного объема (16x 1,92 TB + 4x NVME)
- программный RAID
- RAID 50 или RAID 10 в зависимости от приоритета на объем или высокую интенсивность

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

12. 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.

13. ПРИМЕРЫ РАБОТАЮЩИХ РЕШЕНИЙ

Пример: георасределенный кластер

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

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

- 1 аппаратный кластер (3 узла)
- 1 экземпляр БД
- СУБД Postgres Pro Enterprise Certified
- Объем БД ~ 20 ТБ

Показатели производительности системы резервного копирования

- Вечерняя пятничная полная резервная копия ~11 ТБ (сжатие) за 5 часов (режим archive) + 500 ГБ архивных журналов
- Утренняя инкрементальная резервная копия в понедельник ~200–400 ГБ за 30–50 минут (режим page) + 10–20 ГБ архивных журналов
- Инкрементальная резервная копия каждые 2 часа в рабочие дни ~10–30 ГБ за ~3–5 минут (режим page) + 5–15 ГБ архивных журналов

Специфика

- Повышение производительности за счет использования синхронной реплики
- Оптимизация чтения путем доработки прикладной системы

Реализация георасределенного кластера

- Вторая, взаимодействующая **«Машина баз данных Скала^р МБД #2»** (полная копия основной)
- Удаленность порядка 500 км
- Канал 10 Гбит/с
- Асинхронная реплика основного кластера
- Каскадная репликация

Пример реализации геокластера приведен ниже (Рисунок 13).

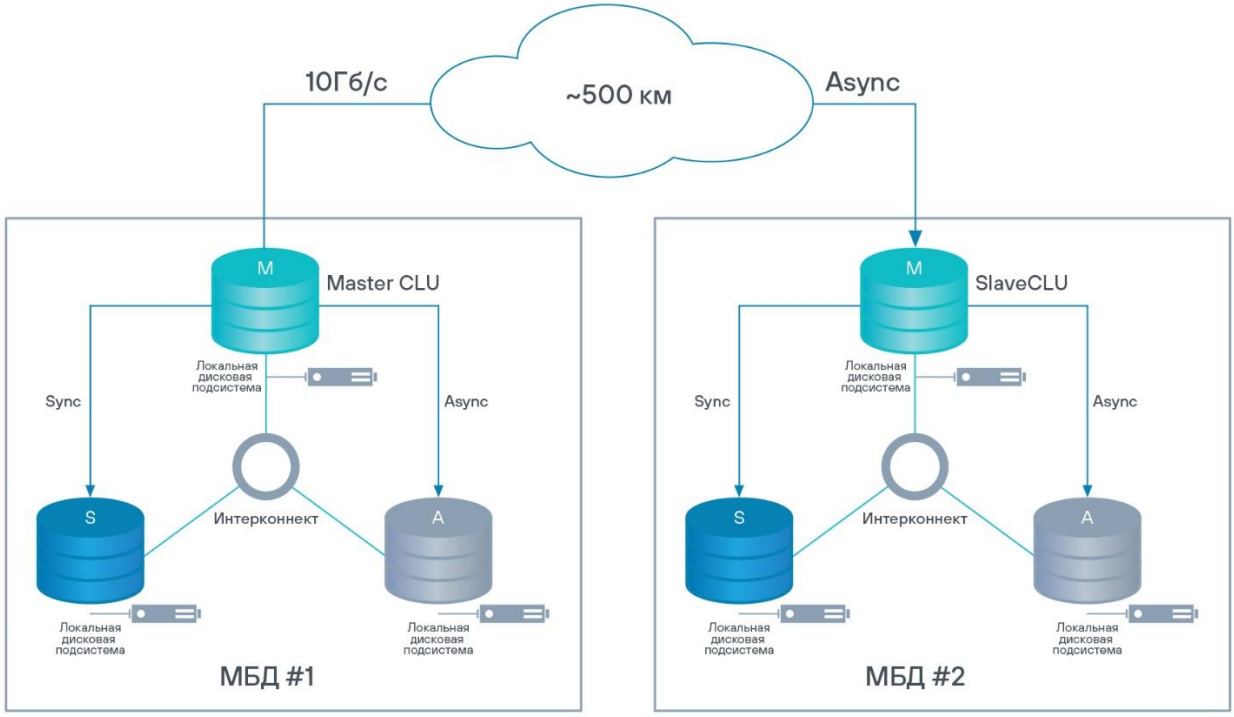


Рисунок 13. Пример реализации геокластера

Пример: гибридный тип нагрузки на МБД (HTAP)

Решение — **Машина баз данных Скала^р МБД.П:**

- 1 аппаратный кластер (3 узла)
- 1 экземпляр БД
- СУБД Postgres Pro Enterprise Certified
- объем БД — ~22,5 ТБ

Пример локального кластера для HTAP-нагрузки приведен ниже (Рисунок 14).

Показатели производительности системы резервного копирования

- Полная резервная копия раз в неделю ~14 ТБ (сжатие) за 5,5 часов (режим archive) + 6,5 ГБ архивных журналов
- Инкрементальная резервная копия ежедневно ~10–60 ГБ за ~1–8 мин (режим page) + 100 МБ архивных журналов

Специфика

- Повышение производительности за счет объединения в пул внешних соединений
- Повышение производительности за счет использования реплики

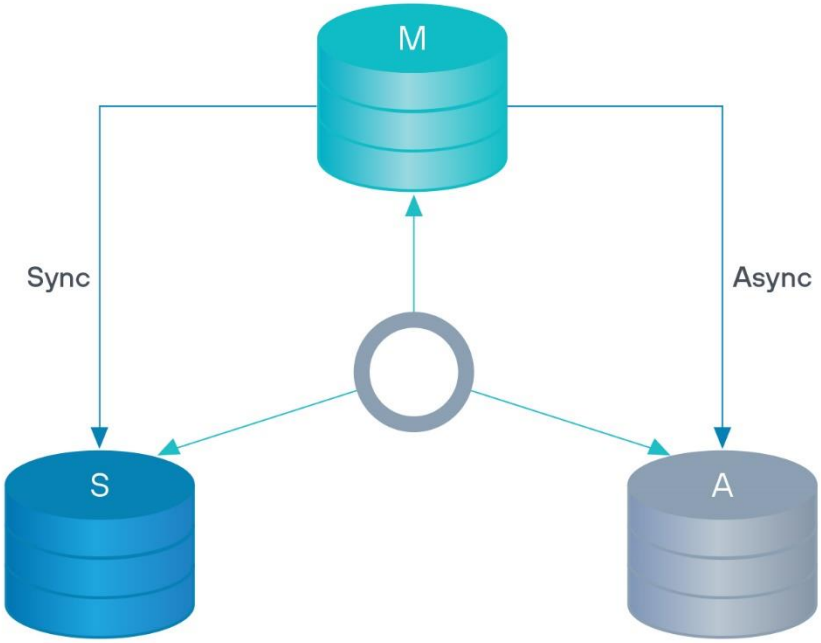


Рисунок 14. Пример локального кластера для HTAP-нагрузки

Пример: транзакционный тип нагрузки

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

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

«Шахматное» размещение БД по узлам кластера приведено ниже (Рисунок 15).

Показатели производительности системы резервного копирования

- Полная резервная копия БД № 1 ~ 9,5 ТБ (сжатие) за 4,5 часа (режим archive) + 11 ГБ архивных журналов
- Инкрементальная резервная копия БД № 1 ежедневно ~30–100 ГБ за ~5–25 мин. (режим page) + 500–1000 МБ архивных журналов
- Полная резервная копия БД № 2 ~10 ТБ (сжатие) за 4 часа 20 минут (режим archive) + 50 ГБ архивных журналов (проводиться одновременно с резервным копированием БД № 1)
- Инкрементальная резервная копия БД № 2 ежедневно ~60–200 ГБ за ~10–40 мин. (режим page) + 2–10 ГБ архивных журналов

Специфика

- Повышение производительности за счет объединения в пул внешних соединений
- «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 15. «Шахматное» размещение БД по узлам кластера

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
- 512 ГБ оперативной памяти
- Контроллер LSI 9361-16i
- Диски SSD

Параметр pgbench sf = 75 000 приводит к созданию БД объемом ~1 ТБ; функция объединения в пулы pgbench не использовалась, для каждого пользователя использовалось выделенное подключение к СУБД. Тесты проводились на ОС Альт 8СП для сертифицированной версии Postgres Pro Enterprise.

Результаты сравнительного тестирования приведены ниже (Рисунок 16 и Рисунок 17).

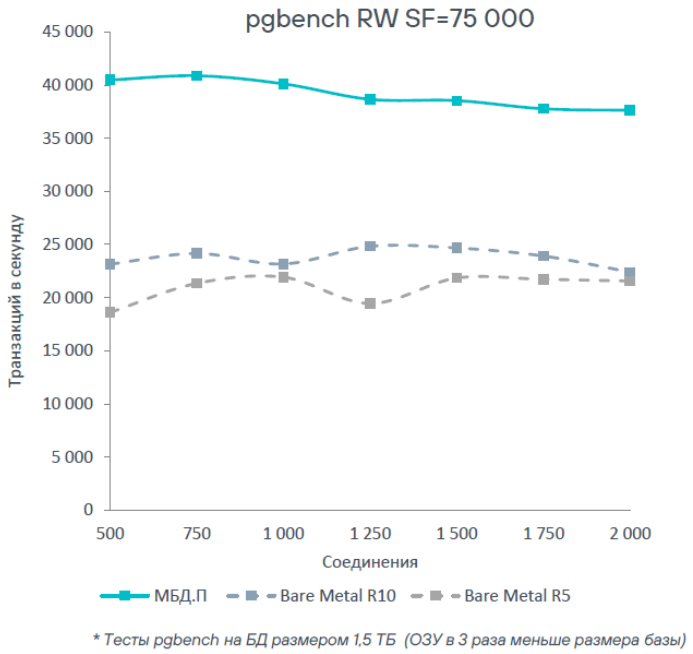


Рисунок 16. Производительность Машины баз данных Скала^р МБД.П по результатам теста pgbench

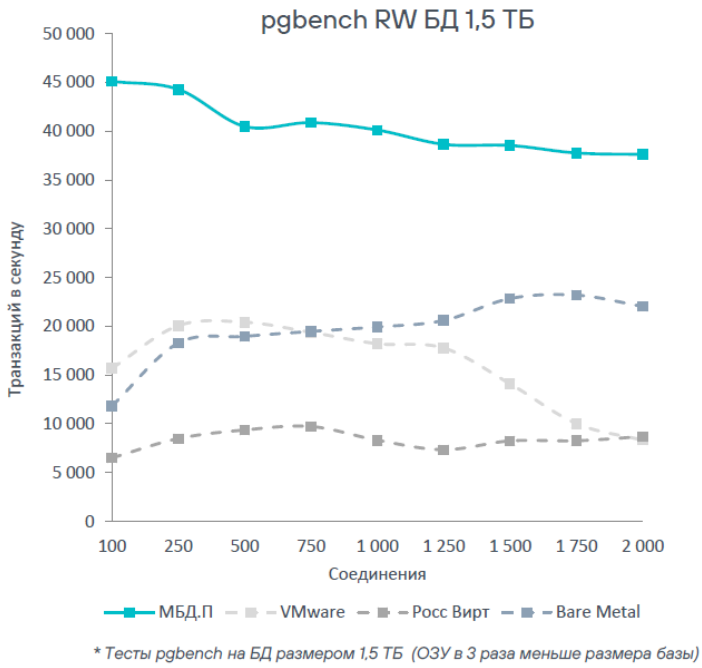


Рисунок 17. Скорость обработки запросов Машины баз данных Скала^р МБД.П по результатам теста pgbench

Далее приведены диаграммы, которые представляют результаты тестирования обновленной версии **Машины баз данных Скала^р МБД.П** (Рисунок 18 и Рисунок 19).

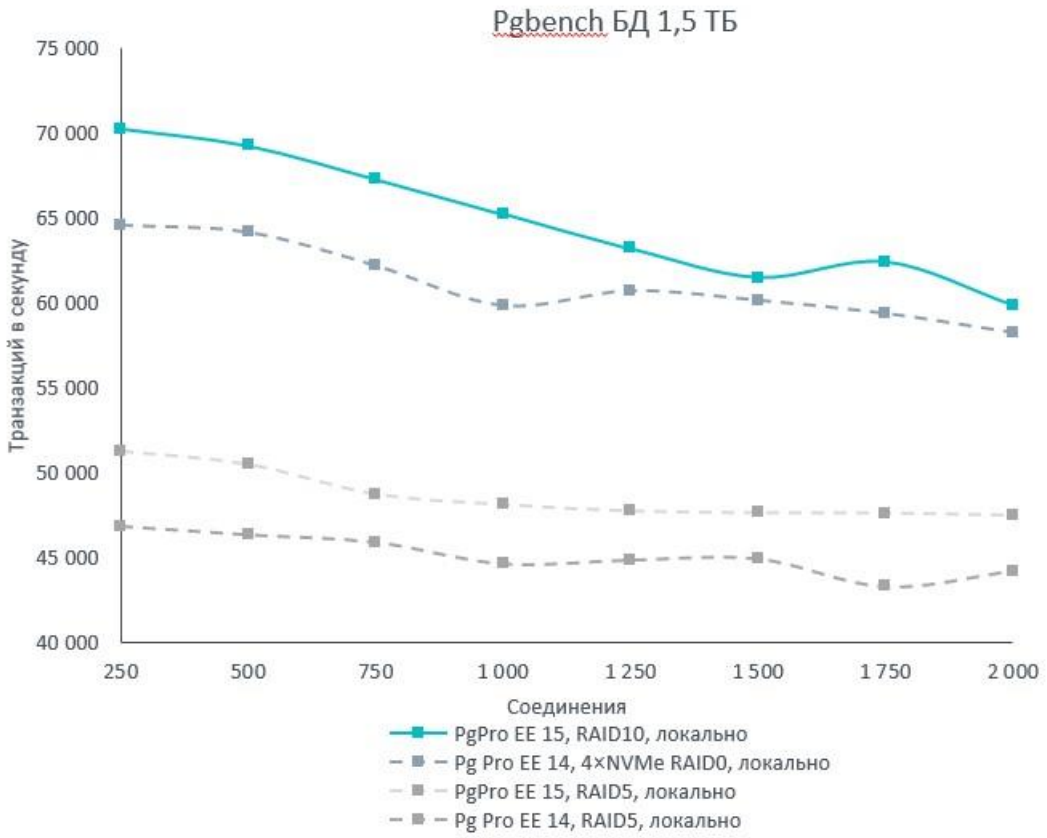


Рисунок 18. Производительность Машины баз данных Скала^р МБД.П при разных версиях СУБД Postgres Pro Enterprise

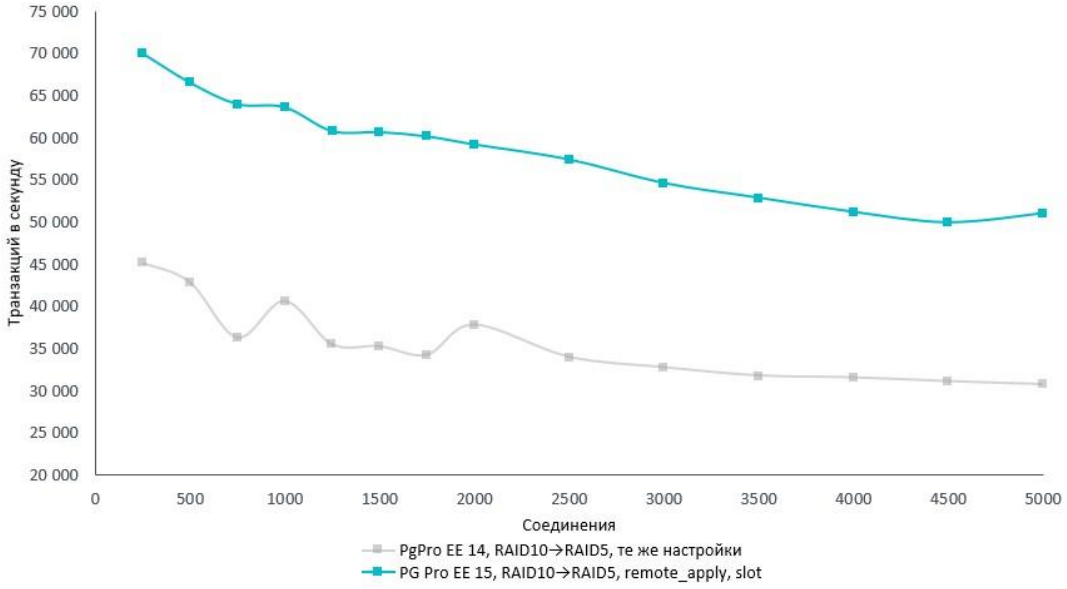


Рисунок 19. Производительность Машины баз данных Скала^р МБД.П при разных версиях СУБД Postgres Pro Enterprise при увеличенном количестве подключений

14. 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 ниже (Рисунок 20).

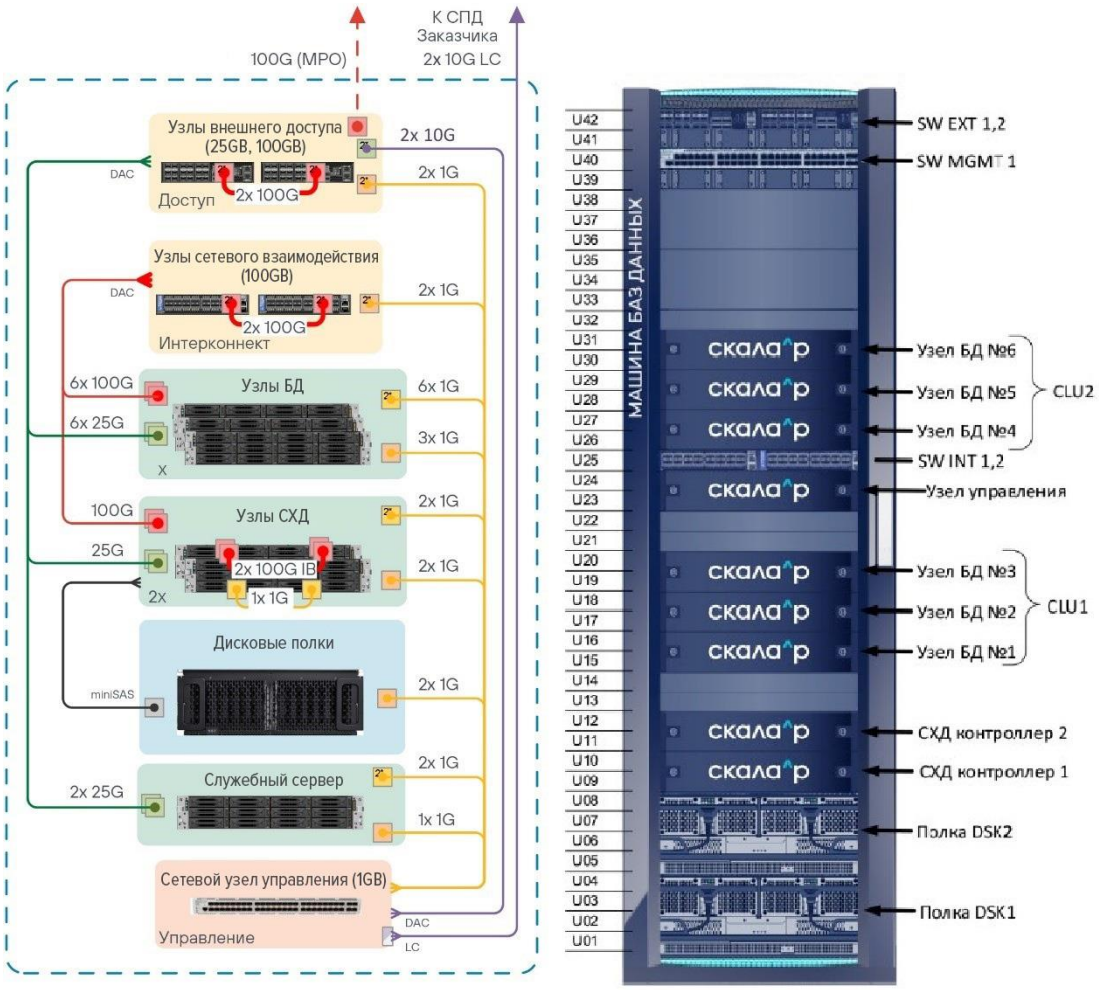


Рисунок 20. Схема взаимодействия для расчета надежности Машины баз данных Скала^р МБД.П

Основные понятия и ограничения расчетной модели

Под работоспособностью решения понимается состояние, в котором решение **выполняет свои функции в полном объеме.**

Вероятность безотказной работы P(t) — вероятность того, что система будет работоспособна в течение заданного времени работы при заданных условиях эксплуатации.

Коэффициент готовности Кг — вероятность того, что система окажется в работоспособном состоянии в произвольный момент времени.

Принято, что как вся система, так и любой ее элемент могут находиться **только в одном из двух возможных состояний** — работоспособном или неработоспособном — и отказы элементов независимы друг от друга, при этом состояние системы (работоспособное или неработоспособное) определяется состоянием элементов и их сочетанием.

Для расчета показателей надежности была **сформирована структурно-логическая схема** решения, учитывающая влияние каждого элемента решения на вероятность его безотказной работы. Далее задача расчета показателей надежности была сведена к задаче расчета соответствующих показателей для каждого из модулей модели, соответствующих блокам, входящих в состав решения.

Исходные данные по блокам решения были сформированы на основании информации, предоставленной производителями. В ряде случаев соответствующие значения были получены расчетным путем на основании более глубокой детализации соответствующих блоков и анализа данных по составляющим эти блоки элементам (например, в отношении серверного оборудования и дисковых массивов).

Дополнительно был проведен анализ накопленных за более чем 5 лет эксплуатации данных по возникшим сбоям и отказам эксплуатируемых решений. Это позволило оценить структуру и ключевые источники возникавших проблем. Проведенный анализ позволил сделать вывод о высоком уровне стабильности функционирования программного обеспечения.

Тем не менее поскольку программная часть решения в существенной степени зависит от действий администраторов заказчика, а события, возникающие в ней, могут интерпретироваться заказчиком по-разному, итоговые значения параметров надежности **Машины баз данных Скала^р МБД.П** представили в виде зависимости от уровней надежности ПО. Полученные результаты приведены ниже (Рисунок 21, Рисунок 22, Рисунок 23).

Рисунок 21. График зависимости вероятности отказа решения от коэффициента готовности комплекта программного обеспечения

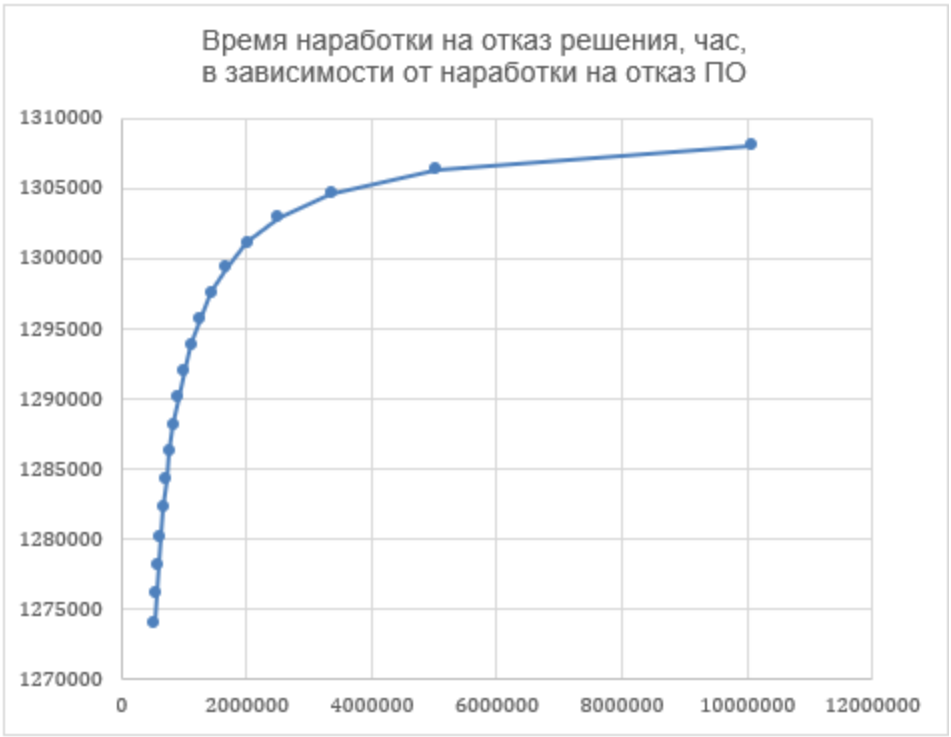


Рисунок 22. График зависимости времени наработки на отказ решения от времени наработки на отказ комплекта программного обеспечения

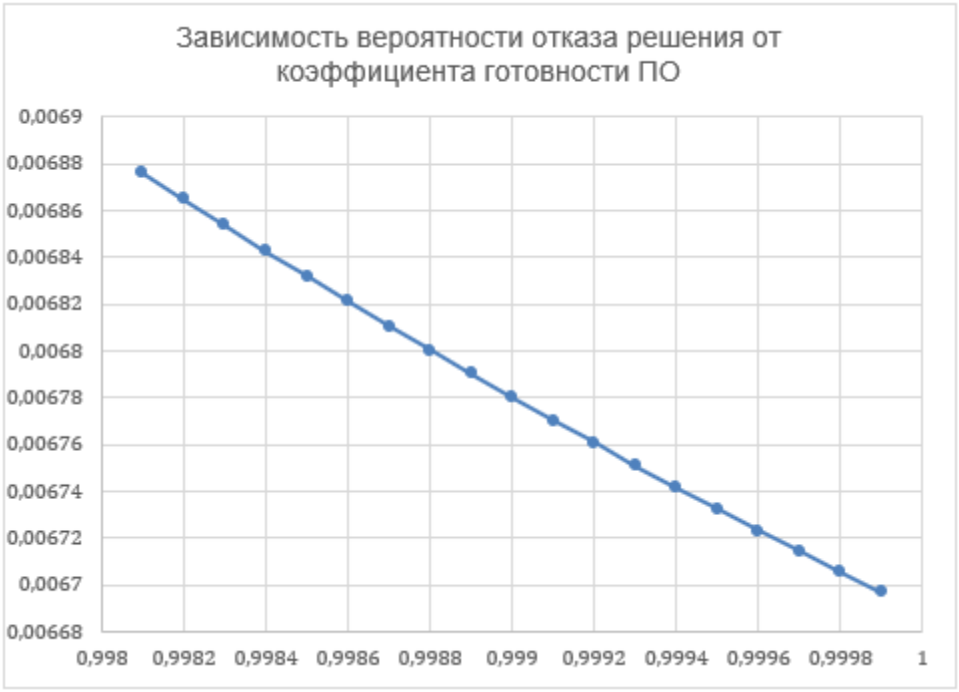


Рисунок 23. График зависимости вероятности отказа решения от коэффициента готовности программного обеспечения

С учетом принятых допущений и ограничений модели, приведенные результаты могут быть использованы для определения значений показателей надежности решения, в том числе при выполнении проектных работ.

Разработанный математический аппарат, накопленные расчетные и статистические данные о функционировании работающих систем позволяют выполнить аналогичные расчеты для иных вариантов конфигурации **Машины баз данных Скала^р МБД.П**.

ЗАКЛЮЧЕНИЕ

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

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

- Доступность
- 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 Минпромторга (в том числе как ПАК) и реестры ПАК Минцифры

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

- Базовый модуль (Блок коммутации и агрегации/Блок мониторинга и регистрации)
- Модуль баз данных (Блок вычисления и хранения)
- Модуль резервного копирования (Блок 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

Компания Скала^р — разработчик и производитель модульной платформы для высоконагруженных корпоративных и государственных информационных систем.

Машины Скала^р являются серийно выпускаемыми преднастроенными комплексами и позволяют осуществлять быстрое развертывание и ввод в эксплуатацию.

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.