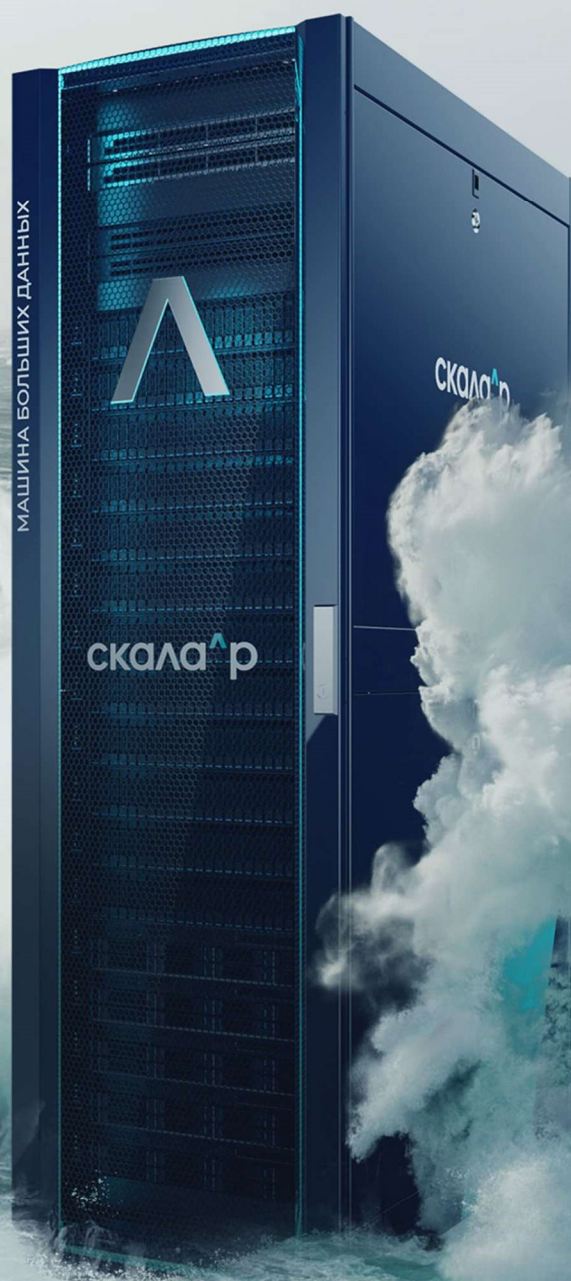




Машина больших данных Скала^р МБД.Т

Программно-аппаратный комплекс для развертывания высокопроизводительных программных систем на основе резидентной СУБД Tarantool

Технический обзор



ОГЛАВЛЕНИЕ

1. Введение	3
2. Основы резидентных СУБД	4
3. Варианты применения резидентных вычислений	5
4. Отличительные черты Скала^р МБД.Т	7
5. Подтвержденная безопасность	9
6. Принципы проектирования	11
7. Состав решения	13
8. Специфичные черты	28
9. Гарантированное качество	30
10. Реакция на возможные отказы	31
11. Примеры комплектов поставки решения	32
12. Лицензирование решения	33
13. Вариативность решения	34
14. Требования к размещению решения	35
15. Примеры работающих решений	36
О компании	37

1. ВВЕДЕНИЕ

Машина больших данных Скала^р МБД.Т — это программно-аппаратный комплекс для обработки и хранения данных, специально предназначенный для развертывания высокопроизводительных программных систем на основе резидентной СУБД Tarantool с использованием программного обеспечения Picodata.

Tarantool — не только СУБД, но и сервер приложений, работающий в парадигме резидентных вычислений. Это быстрая, надёжная и хорошо масштабируемая основа для построения распределённых систем. Программное обеспечение Picodata — существенно доработанная и улучшенная версия Tarantool, обеспечивающая гарантии строгой согласованности данных и среду выполнения для безопасных и ресурсоэффективных приложений на языке Rust.

Скала^р МБД.Т повышает производительность и отказоустойчивость, снижает затраты за счёт проработанной интеграции аппаратного и программного обеспечения, использования передовых технологий, широкого применения методов обеспечения надёжности и специальных моделей лицензирования.

Скала^р МБД.Т — решение для размещения баз данных объёмом от 1 Тбайт до 30 Тбайт, в зависимости от выбранного комплекта оборудования и приоритета производительности или отказоустойчивости при формировании кластера.

Скала^р МБД.Т — комплексное, горизонтально масштабируемое решение, включающее в себя модули для проведения вычислений и хранения данных, систему резервного копирования, сверхскоростную сетевую среду, систему интеллектуального управления.

Высокая производительность решения достигается за счёт использования уникальной по своим характеристикам резидентной платформы, применением оптимальных комплектующих, в том числе — современных NVMe-накопителей, а также стогигабитных сетей.

Отказоустойчивость обеспечивается применением надёжных комплектующих, специализированной версии распределённой платформы Picodata, оптимизацией структуры программного кластера, резервированием критических компонентов, использованием устойчивых сетевых протоколов.

Скала^р МБД.Т содержит все необходимые элементы для функционирования высокопроизводительной распределённой базы данных на платформе Picodata. Подключение к внешним сетям осуществляется с помощью высокоскоростных интерфейсов Ethernet.

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

Машина Скала^р МБД.Т была разработана в качестве ответа на потребности в сокращении времени отклика при кратном увеличении объёмов хранения, обеспечении высокого уровня отказоустойчивости и производительности.

Проведённые нами исследования позволили сформировать единый программно-аппаратный комплекс, имеющий проработанную архитектуру, собранный из согласованных и устойчиво взаимодействующих компонентов, с применением современных технологий формирования резидентных баз данных.

Программно-аппаратные комплексы **Скала^р МБД.Т** включены в Единый реестр российской радиоэлектронной продукции и работают на ПО, включённом в реестр Минцифры РФ.

2. ОСНОВЫ РЕЗИДЕНТНЫХ СУБД

Традиционные СУБД оптимизированы под хранение данных на энергонезависимых блочных носителях — жёстких дисках и твердотельных накопителях. И хотя они разработаны с учетом требований обработки транзакций, операции вставки, обновления, частичной выборки в этих СУБД остаются весьма медленными из-за необходимости частой и дублирующей синхронизации с устройством постоянного хранения.

Заметного роста производительности можно достичь, разместив традиционную реляционную базу данных непосредственно в оперативной памяти. И ряд коммерческих СУБД, в том числе Oracle Database, IBM DB2, Microsoft SQL Server, допускают такие варианты применения.

Ещё больший эффект достигается за счёт отказа от использования реляционных структур и снижения требований к согласованности данных: резидентные NoSQL-СУБД классов «ключ — значение» и «семейство столбцов» позволяют достичь высокой производительности благодаря горизонтальному масштабированию в таких условиях. Нереляционные агрегатные модели данных дают возможность разделять базы данных на сегменты по формальному признаку, например, значению хэш-функции от ключа, и сегменты, будучи в нереляционных условиях фактически независимыми, могут управляться отдельными экземплярами СУБД, располагаемыми на различных аппаратных узлах. Снижение требований к согласованности позволяет сохранять уровень производительности на уровне близком к тому, как если бы база данных не была распределена по сети и не содержала реплик данных, полагаясь на «согласованность в конечном счёте», когда рано или поздно для отдельно взятой записи система придёт в согласованное состояние.

Использование распределённой схемы позволяет достичь не только горизонтальной масштабируемости и заметного роста производительности при добавлении новых узлов, но и увеличить устойчивость системы путем наличия нескольких копий данных. Как следствие — потеря отдельного узла и даже группы узлов — не приводит к потере данных, хотя общая производительность при этом может уменьшиться.

Резидентные технологии ускоряют доступ к данным и их обработку. Так, традиционная СУБД выдерживает нагрузку до пары десятков тысяч запросов в секунду, резидентная СУБД на аналогичном оборудовании способна обрабатывать **сотни тысяч запросов** за то же время.

СУБД Tarantool, изначально разработанная как резидентная NoSQL-СУБД класса «ключ — значение», успешно применяется в качестве распределённых кэшей, для сервисов метаданных, организации брокеров сообщений. Применить её в качестве основной СУБД для систем, требующих строгой согласованности, позволяют доработки, выполненные в Picodata — благодаря им система прочно занимает место в классе, обозначаемом как «NewSQL», соединяя свойства горизонтальной масштабируемости, отказоустойчивости и сверхвысокой производительности из мира NoSQL и обеспечивая максимально строгие транзакционные требования, которые ранее удовлетворялись только в классических реляционных СУБД.

Обеспечение строгой согласованности, реализация Raft-консенсуса, поддержка языков запросов SQL и GraphQL, а также сервер приложений на языках Rust, встроенный в Picodata и обеспечивающий прямой доступ ко резидентной базе, позволяют работать с ней как с полноценным транзакционным решением корпоративного уровня, обладающим при этом свойствами горизонтальной масштабируемости и сверхвысокой производительностью, унаследованным от распределённых резидентных NoSQL-систем.

3. ВАРИАНТЫ ПРИМЕНЕНИЯ РЕЗИДЕНТНЫХ ВЫЧИСЛЕНИЙ

Резидентные вычисления наиболее востребованы в задачах, требующих быстрой реакции и минимального времени отклика.

Цель, применения резидентных вычислений в современных программных системах — это скорость

Обозначим несколько типовых сценариев применения резидентных СУБД в современных условиях.

1. Повышение производительности унаследованных систем

Часто традиционные и унаследованные системы лежат в основе ключевых бизнес-процессов организации. При этом они технически достигли предела производительности, а также не способны поддерживать современные потребности, например, работу онлайн-сервисов.

Проблема может быть устранена путем создания промежуточного слоя между работающими решениями и цифровыми сервисами с применением резидентных вычислений. Этот слой будет собирать, хранить и обрабатывать наиболее востребованные данные, технологически соответствовать новым сервисам и значительно ускорять их работу.

2. Оперативный доступ к основным данным

В крупных организациях используется, как правило, большое число разнообразных программных решений. При этом разнообразная информация о поставщиках, клиентах, продуктах, услугах, нормативно-справочной информации оказывается распределенной по всем системам, вынужденно дублируется, сложно актуализируется, не является полной.

Решением может быть ведение единых справочников основных данных, работа с «золотыми записями» о клиентах, поставщиках как стандарта для компании. Применение резидентных вычислений для этих целей вполне оправдано, поскольку позволит аккумулировать разрозненные данные из множества источников в режиме реального времени и отобразить их в едином формате без потери производительности и без задержек. В сравнении с традиционными MDM-системами, резидентные основные данные доступны онлайн, и могут обслуживаться без каких-либо регламентных пакетных процессов, требующих остановки системы на время их выполнения.

3. Маркетинг в режиме реального времени

Одним из наиболее актуальных способов удержания клиента и повышения объемов продаж является умение своевременно сделать ему персонализированное предложение. Так, онлайн-торговля содержит множество полезных для этого данных, включая возможность отслеживания фактически всех действий клиента как в прошлом, так и в текущий момент — что он искал, какие товары он отобрал, на какие не стал обращать внимание... Однако максимальная выгода будет достигнута только в случае очень быстрой реакции продавца, пока клиент не завершил свои действия и не отключился или изменил предпочтения. Порой речь идет о считанных секундах для коммуникации персонифицированного предложения.

Реальный опыт и примеры ряда компаний подтверждают возможность решения всего комплекса аналитических задач — фиксации действия, анализа контекста и реакции, формирования целевого предложения в крайне сжатые сроки — на базе технологий резидентных вычислений. Конечно, такая задача заметно сложнее уже приведенных сценариев, но она крайне актуальна, и достигает цели, позволяя повысить выручку компаний и лояльность клиентов.

4. Кэш

Еще одним аспектом, существенно влияющим на эффективность онлайн-продаж, является отсутствие задержки при загрузке информации. Ряд исследований показывают, что до четверти покупателей отказываются от дальнейших действий, сталкиваясь с медлительностью систем, а общая конверсия продаж теряет при этом до 7%.

Ускорение обработки запросов может быть достигнуто путем создания кэширующего слоя — вспомогательного хранилища, содержащего всю необходимую для конкретного сервиса информацию в оперативном доступе. Реализация витрины данных с применением резидентной СУБД позволит полностью устранить замедление при пользовательских запросах, одновременно обеспечив и регулярную актуализацию кэша.

5. Хранение данных

В последнее время, за счет продолжающегося развития технологий, удалось избавиться резидентные СУБД от целого ряда ранее характерных недостатков. В частности, проблема возможной потери данных в случае хранения их непосредственно в оперативной памяти была преодолена за счет развития схем репликации и применения алгоритмов иерархического хранения данных на устройствах разного типа: оперативной памяти, энергонезависимой байтоадресуемой памяти, высокоскоростных SSD, высокоёмких SSD, HDD, архивных систем хранения.

Современные резидентные базы данных стали полноценной альтернативой традиционным, не уступая последним по надежности, и при этом — заметно выигрывая в производительности.

4. ОТЛИЧИТЕЛЬНЫЕ ЧЕРТЫ СКАЛА^Р МБД.Т

1. Резидентная обработка сверхбольших объёмов «горячих» данных и надёжное хранение

- Надёжный источник данных с поддержкой долговечности (в том числе гарантированного восстановления данных после любых сбоев), репликации и сегментирования на платформе Picodata
- Производительность превышает 100 000 RPS (запросов в секунду) на одном ядре CPU
- Стабильная работа при большом проценте запросов на запись с одновременным чтением данных
- Объём базы данных от 1 Тбайт до 30 Тбайт, в зависимости от масштаба и выбранного уровня резервирования узлов. Названные объёмы касаются фактического объёма хранения, применение алгоритмов сжатия информации позволяет размещать в этом объёмекратно большие базы данных

2. Высокая производительность

- Сбалансированный комплект оборудования
- Архитектурная оптимизация производительности
- Высокоскоростные локальные накопители для размещения образов памяти и временного хранения журналов предзаписи (WAL)
- Оптимизированный по производительности программный RAID
- Специальные настройки программного обеспечения
- Продуманные алгоритмы резервного копирования и восстановления
- Горизонтальное масштабирование ресурсов Машины

3. Отказоустойчивость на всех уровнях

- Отказоустойчивая платформа
- Надёжные комплектующие
- Резервирование значимых компонентов на аппаратном уровне
- Оперативная восстанавливаемость при сбоях

4. Приоритет сохранности данных

- Поддержка ACID-транзакционности (atomicity, consistency, isolation, durability — атомарности, согласованности, изоляции, устойчивости)
- Синхронная и асинхронная репликация локально и на удаленных узлах
- Единый механизм упреждающей записи в журнал (WAL, Write Ahead Log)
- Разделение доступа на основе ACL-модели (Access Control List)
- Непрерывное формирование полных резервных копий данных и журналов

5. Обеспечение качества при развёртывании

- Оптимальность настроек проверена тестами
- Автоматизированное развёртывание исключает человеческие ошибки

- Стандартизация развёртывания гарантирует соответствие решения заявленным характеристикам

6. Непрерывный контроль состояния

- Мониторинг работоспособности платформы и оборудования
- Преднастроенные пороговые значения критичных параметров наблюдения
- Различные каналы информирования об отклонениях

7. Гибкие возможности администрирования

- Предусмотрены дополнительные решения для управления
- Сохранены все стандартные механизмы управления платформой Picodata
- Проработаны рекомендации по выполнению отдельных мероприятий

8. Обеспечение эксплуатации

- Централизованная поддержка решения единым сервисным центром
- Единая ответственность за весь комплекс
- Оперативный выпуск исправлений и рекомендаций
- Паспорт Машины в комплекте
- Обучение персонала заказчика

9. Экономическая эффективность

- Специальные условия лицензирования платформы Picodata
- Сокращённые сроки ввода в эксплуатацию
- Только обоснованно необходимые компоненты
- Минимальные затраты на комплексную поддержку решения

5. ПОДТВЕРЖДЕННАЯ БЕЗОПАСНОСТЬ

Машина больших данных Скала^р МБД.Т 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

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.САВЗ.Б2.13» (ФСТЭК России, 2012), «Профиль защиты средств антивирусной защиты типа В второго класса защиты. ИТ.САВЗ.В2.ПЗ» (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

6. ПРИНЦИПЫ ПРОЕКТИРОВАНИЯ

Машина больших данных Скала^р МБД.Т — это программно-аппаратный комплекс, предназначенный для развертывания платформы Picodata. Ориентация на конкретную реализацию определяет набор принципов, заложенных при разработке архитектуры решения.

Более ярко данные принципы будут видны при сравнении с традиционными решениями по хранению и обработке данных.

Традиционные СУБД

Схема развёртывания традиционной реляционной СУБД для крупномасштабных задач обычно включает узлы вычисления, системы хранения, а также узлы резервного копирования. Для обеспечения надежности решение может быть сформировано в виде кластера. При этом фактическая работа ведется с одним вычислительным узлом — остальные используются для резервирования. При росте объемов хранения сложность отработки транзакций растет нелинейно, постепенно приводя к исчерпанию вычислительных ресурсов и заметному увеличению времени отклика.

Одним из путей повышения производительности традиционных систем стало применение специализированных программно-аппаратных комплексов — Машин баз данных, спроектированных под применение конкретных СУБД (в линейке продуктов Скала^р присутствуют такие решения).

Оптимизация комплекта оборудования, тонкие настройки ПО СУБД и программно-управляемых систем хранения дают определенный скачок производительности, и в ряде ситуаций его вполне хватает для решения стоящих задач. Однако если объемы хранения продолжают активно расти или требуется кратно повысить производительность — такого рода подход в перспективе себя не оправдывает.

Основными ограничениями при повышении объемов хранимых данных являются единственный доступный вариант масштабирования — вертикальный, а также необходимость размещения данных на энергонезависимых устройствах хранения. Указанные ограничения быстро приводят к достижению предела производительности, непреодолимым даже более при переходе на аппаратные компоненты категории Hi-End.

Архитектурные особенности платформы Picodata

- Архитектура кластера предполагает систему отдельных экземпляров (instance) — программных узлов, входящих в состав кластера
- Каждый узел выполняет свою роль, хранения данных, сервера приложения, или служебную роль координатора кластера
- Все экземпляры работают с единой схемой данных и кодом приложения
- Каждый процесс базы данных выполняется на одном процессорном ядре и хранит используемый набор данных в оперативной памяти
- Любой отдельный экземпляр является частью набора реплик (replicaset)
- Набор реплик может состоять из одного или нескольких экземпляров — дубликатов одного и того же набора данных
- Внутри набора реплик всегда есть активный экземпляр и, если реплик больше одной, то некоторое число резервных экземпляров, обеспечивающих

отказоустойчивость системы в случае выхода из строя или недоступности активного экземпляра

- Число реплик определяется коэффициентом репликации, заданным в глобальных настройках

Машина Скала^р МБД.Т создана для платформы Picodata

Целью разработки **Машины больших данных Скала^р МБД.Т** было создание полного тиражируемого комплекта аппаратного и программного обеспечения, адаптированного под In-Memory резидентную СУБД Picodata, с учетом ее архитектурных особенностей и возможностей расширения.

Связующей основой решения является внутренняя сеть интерконнекта, построенная с применением стогигабитного Ethernet. Примененные технологические решения позволяют строить Машины, содержащие в совокупности от 3 до 148 высокопроизводительных узлов вычисления и хранения, и до 16 узлов резервного копирования данных.

Преднастроенные алгоритмы резервного копирования совместно с применением производительных NVMe-накопителей, используемых для формирования моментальных образов базы, позволяют формировать резервные копии без отрицательного влияния на производительность системы в целом.

Все программные компоненты Скала^р МБД.Т подобраны оптимальным образом

Основными программными элементами решения Скала^р МБД.Т являются операционная система, ПО распределенного сервера приложений со встроенной распределенной резидентной СУБД Picodata, ПО мониторинга и администрирования, ПО резервного копирования.

В рамках Скала^р МБД.Т обеспечена оптимизация, тонкая настройка и доработка перечисленных компонентов для обеспечения их большей производительности и функционального соответствия потребностям решения в целом.

В ходе развития решения Скала^р МБД.Т были оптимизированы **настройки ядра операционной системы** узлов вычисления и хранения под применение платформы Picodata.

Оптимизация функционирования платформы Picodata и ее отказоустойчивость достигается путём изменения настраиваемых параметров только после развертывания и формирования кластера. По этой причине соответствующие мероприятия не могут быть завершены в полном объеме до выполнения настройки платформы заказчиком. Поставщик готов оказать соответствующие услуги дополнительно к поставляемому решению по индивидуальному запросу.

Существенные улучшения производительности были достигнуты за счёт доработки и **совершенствования ПО управления RAID-массивами**, используемыми для хранения образов данных и журналов предзаписи. Дополнительно это привело к полному отказу от аппаратной реализации RAID-массивов.

Постоянно ведётся деятельность по **развитию систем мониторинга и управления**, относящихся к узлу управления Скала^р МБД.Т.

7. СОСТАВ РЕШЕНИЯ

Решение Скала^р МБД.Т состоит из следующих блоков:

Блок вычисления и хранения

Блок коммутации и агрегации

Блок мониторинга и регистрации

Блок резервного копирования



Для обеспечения отказоустойчивости и высокой производительности при проектировании программно-аппаратного комплекса были заложены технологические принципы и применён ряд технических решений, описанных ниже.

К технологическим принципам относятся:

- Дублирование критичных компонентов
- Применение высокопроизводительных компонентов
- Горизонтальное масштабирование вычислительных ресурсов
- Сохранение работоспособности при отказе отдельных элементов системы (в отдельных случаях — со снижением производительности)
- Специальное ПО управления и мониторинга
- Глубокая адаптация компонентов для совместной работы в составе продукта
- Многоуровневое тестирование комплекса и его узлов и компонентов при производстве, для исключения отказов

Модульная структура решения

Формирование решения основано **на принципе модульности**. Каждый из блоков комплектуется из набора стандартных модулей. Этим обеспечивается универсальный подход, более высокий уровень технологичности и надежности эксплуатации. Модули, в свою очередь, формируются из одного или нескольких узлов для выполнения определенных задач в соответствии с архитектурой комплекса.

Блок коммутации и агрегации

Эффективное сетевое взаимодействие является основой кластера на платформе Picodata

Основные функции сети:

- Передача данных между элементами Скала^р МБД.Т (интерконнект).
- Обеспечение информационного обмена с внешними сетями.

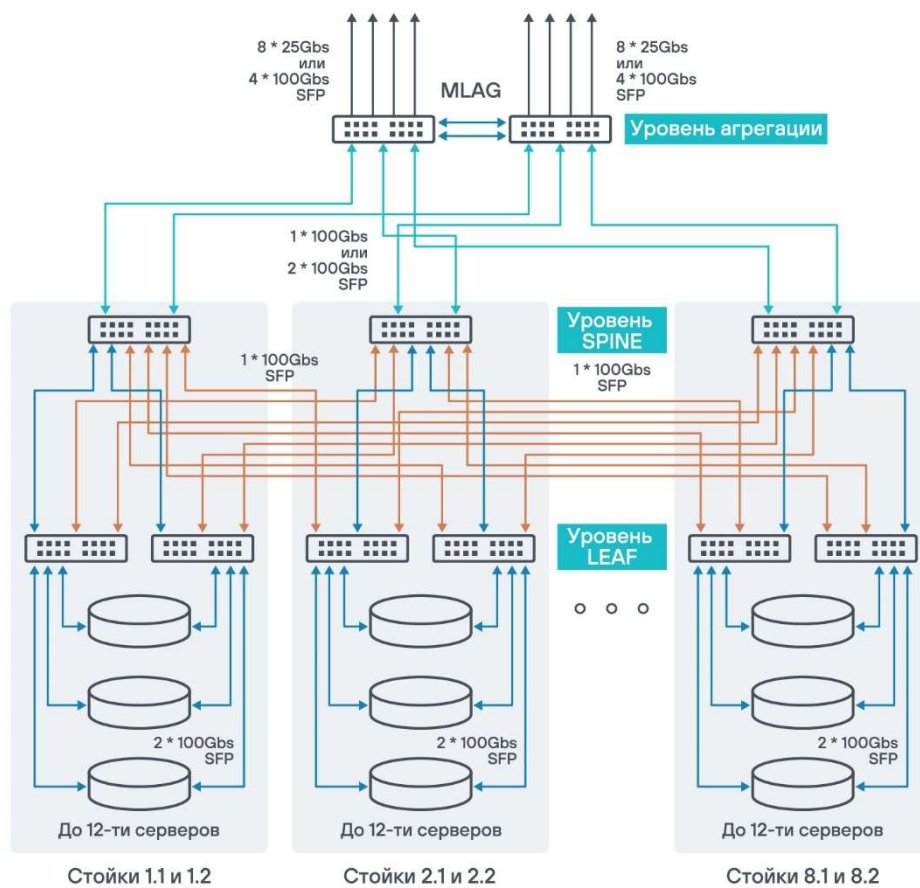


Рис. 1. Общая схема сети интерконнекта

Реализованные подсети:

- External VLAN — сеть для подключения внешних пользователей и прикладных систем, подключение к серверу управления
- Internal VLAN — сеть для внутреннего взаимодействия между узлами Машины, сеть резервного копирования, сеть кластерного взаимодействия

Сети мониторинга и управления — основа обеспечения бесперебойного функционирования решения Скала^р МБД.Т

Основные функции сети:

- Обмен служебными данными, данными для мониторинга и управления.

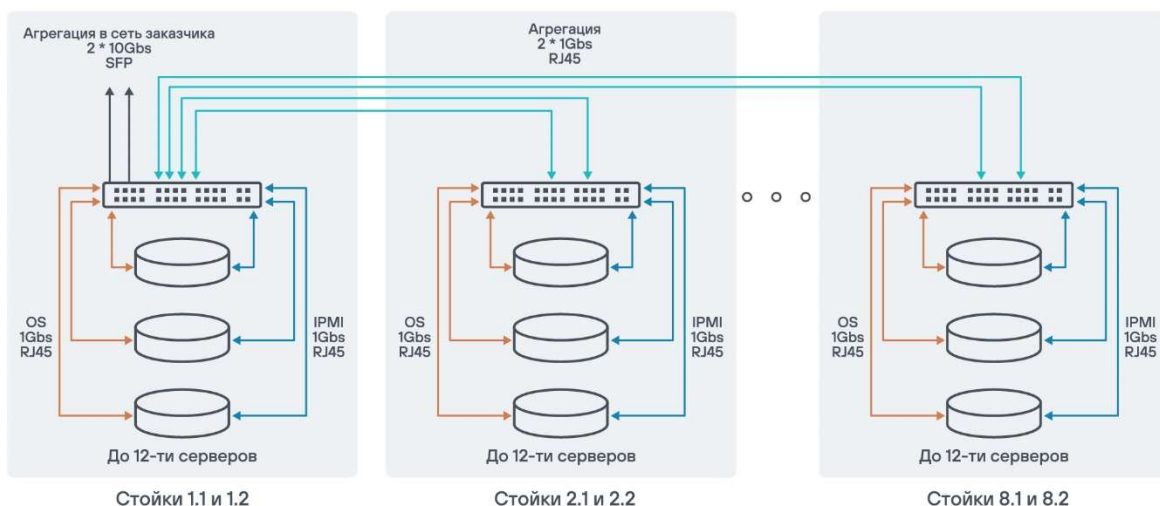


Рис. 2. Общая сетей мониторинга и управления

Реализованные подсети:

- PXE (OS) VLAN — сеть для развёртывания операционной системы по PXE, платформы МБД, мониторинга
- Ring VLAN — резервная сеть кластерного взаимодействия, доступ к IPMI
- IPMI VLAN — сеть управления оборудованием через интерфейсы удалённого управления

Модули и состав оборудования блока коммутации и агрегации

Модуль агрегации

Модуль агрегации применяется в случаях, когда в коммутационном оборудовании Базового модуля исчерпан лимит свободных коммутационных портов.

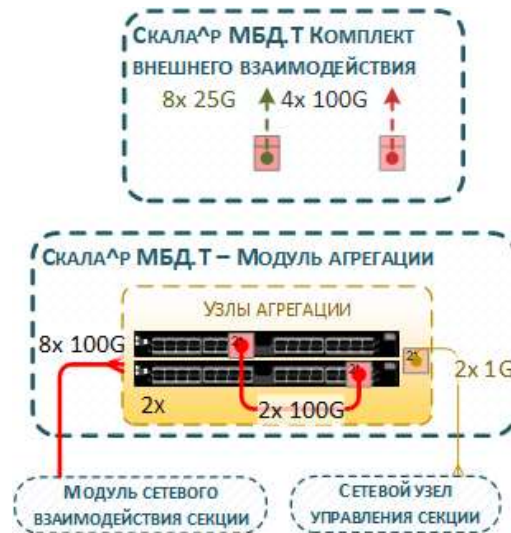


Рис. 3. Модуль агрегации

Модуль агрегации — виртуальный коммутатор по технологии MLAG из двух аппаратных коммутаторов для интерконнекта.

Для обеспечения внешних соединений устанавливается 2 узла взаимодействия с сетями 100 GbE, 25 GbE, 10 GbE, соединенных по технологии MLAG.

Модуль сетевого взаимодействия

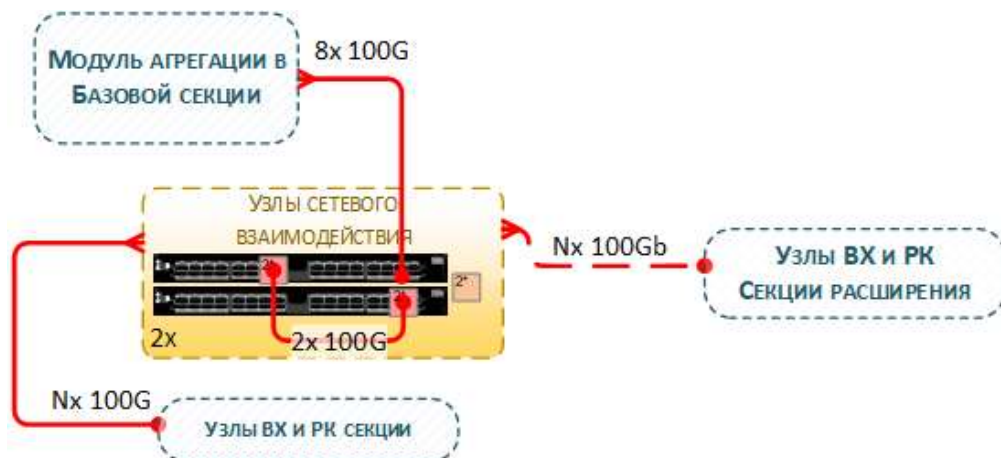


Рис. 4. Модуль сетевого взаимодействия

Модуль сетевого взаимодействия — виртуальный коммутатор по технологии MLAG из двух аппаратных коммутаторов для сетевого доступа.

Модуль сетевого управления



Рис. 5. Модуль сетевого управления

Модуль сетевого управления состоит из одного узла сетевого управления — коммутатора для организации сетей мониторинга, управления и служебного обмена.

Блок вычисления и хранения

Программно-определяемый кластер состоит из типовых модулей вычисления и хранения на базе аппаратных узлов и программной платформы Picodata.

Архитектура кластера Picodata предполагает систему отдельных программных экземпляров — узлов, представленных серверами в составе кластера. Все экземпляры работают с единой схемой данных и кодом приложения. Каждый процесс базы данных выполняется на одном процессорном ядре и хранит используемый набор данных в оперативной памяти. Любой отдельный экземпляр является частью набора реплик. Набор реплик состоит из одного или нескольких дубликатов одного и того же набора данных. Внутри набора реплик всегда есть активный экземпляр и, если реплик больше одной, то некоторое число резервных экземпляров, обеспечивающих отказоустойчивость системы в случае выхода из строя или недоступности активного экземпляра. Число реплик определяется коэффициентом репликации, заданным в глобальных настройках.

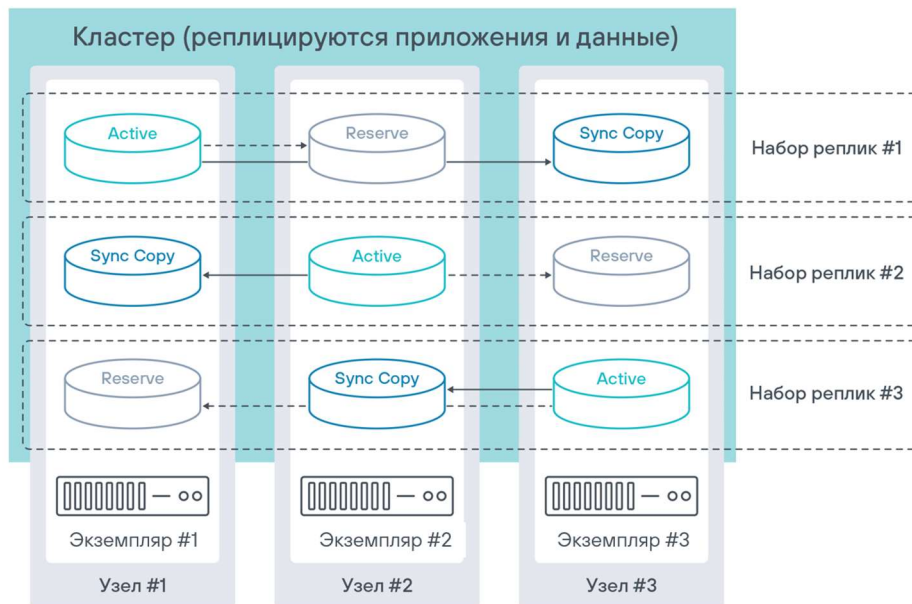


Рис. 6. Общая схема кластера

Хранение данных организуется следующим образом: внутри каждого набора реплик существует «корзина» (*bucket*) — виртуализированная неделимая единица хранения, обеспечивающая локальность данных (например, хранение нескольких связанных с клиентом записей на одном физическом узле сети). Сама по себе корзина не имеет ограничений по емкости и может содержать любой объем данных.

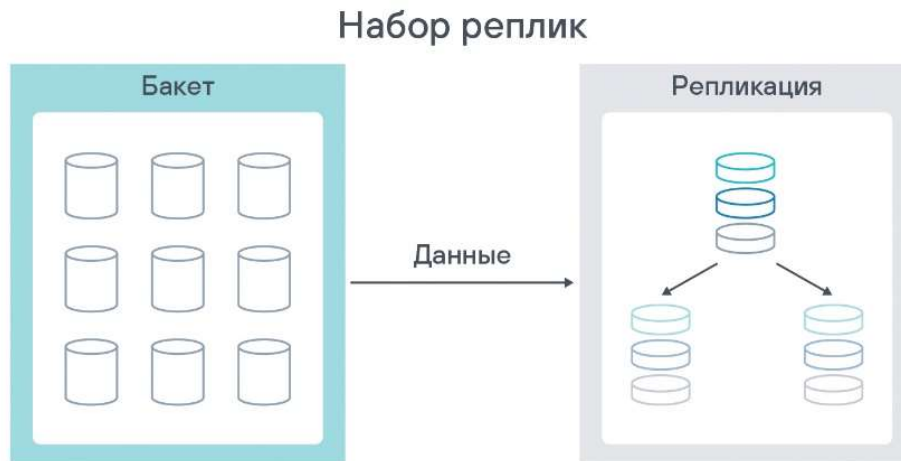


Рис. 7. Хранение данных

Горизонтальное масштабирование позволяет распределить корзины по разным сегментам, оптимизируя производительность кластера путем добавления новых реплицированных экземпляров. Чем больше наборов реплик входит в состав кластера, тем меньше нагрузка на каждый из них. Корзина хранится физически на одном наборе реплик и является промежуточным звеном между данными и устройством хранения. В каждом наборе реплик может быть много корзин (или не быть ни одного). Внутри корзины данные задублированы по всем экземплярам в рамках набора реплик в соответствии с коэффициентом репликации. Количество корзин может быть задано при первоначальной настройке кластера. По умолчанию кластер Picodata использует 3000 корзин.

- Отказоустойчивость обеспечивается наличием нескольких реплик внутри набора реплик, что обеспечивают его отказоустойчивость. Дополнительно для повышения надежности каждый экземпляр кластера внутри набора реплик находится на разных физических серверах.
- Для обеспечения отказоустойчивости применяется сегментирование — это распределение корзин между различными наборами реплик. Для определения, в каком сегменте располагается запись, используется хэш-функция. Каждый набор реплик является сегментом, и чем больше наборов реплик имеется в кластере, тем гибче система с точки зрения масштабируемости. При добавлении новых экземпляров в кластер или формировании новых наборов реплик система автоматически равномерно распределит корзины с учетом новой конфигурации.

Каждый сегмент содержит реплики двух типов — мастер и синхронная реплика. Реплика, исполняющая роль мастера, определяется платформой динамически, в зависимости от доступности и настроек приоритета.

- Обеспечение отказоустойчивости: в случае отказа мастера его функция исполняется любой синхронной репликой
- Повышение производительности за счет разделения данных по сегментам и распараллеливания их обработки
- Повышение производительности за счет доступности для чтения сразу нескольких реплик



Рис. 8. Модуль вычисления и хранения

Каждый отдельный узел вычисления и хранения:

- Обеспечивает высокую производительность за счет применения резидентной базы данных
- Использует NVMe-накопители для обеспечения высокой производительности при формировании образов данных из памяти и ведения журнала предзаписи
- Содержит выделенные накопители SAS SSD для загрузки операционной системы
- Использует локальные накопители для размещения образов данных и журналов (RAID 10), что обеспечивает повышение производительности (нет необходимости дополнительного внешнего обмена с системой хранения)
- Имеет дублированные интерфейсы данных (стандарт IEEE 802.3ad LACP) — повышение производительности, отказоустойчивость (в случае отказа одного из интерфейсов возможно снижение производительности)
- Оснащен двухпортовыми сетевыми картами 100 Gigabit Ethernet для интерконнекта в рамках Машины
- Содержит два блока питания в режиме резервирования по схеме (1 + 1)
- Имеет два процессора Xeon не ниже 2-го поколения
- Использует 1536 GB RAM

Применяемое программное обеспечение:

- ОС: Linux CentOS 7.9
- Специальная версия платформы Picodata
- Управление резервным копированием: специализированные программные модули платформы Picodata
- Управление кластером средствами Picodata

Блок резервного копирования

Основное предназначение — хранение резервных копий базы данных. Дополнительно обеспечивается хранение настроек и метаданных, а также формирование пространства для задач ETL.



Рис. 9. Модуль резервного копирования

Модуль резервного копирования состоит из одного узла резервного копирования.

Каждый отдельный узел резервного копирования:

- Содержит выделенные накопители SAS SSD для загрузки ОС — обеспечение отказоустойчивости
- Имеет интерфейсы данных дублированы (стандарт IEEE 802.3ad LACP) — повышение производительности, отказоустойчивость (в случае отказа одного из интерфейсов возможно снижение производительности).
- Оснащен двухпортовыми сетевыми картами 100 Gigabit Ethernet для интерконнекта в рамках Машины
- Содержит два блока питания в режиме резервирования по схеме (1 + 1)
- Имеет два процессора Xeon не ниже 2-го поколения
- Использует 384 GB RAM
- Оснащен 14-ю дисками для хранения данных

Применяемое программное обеспечение:

- Программное обеспечение для создание программных RAID-массивов: Raidix.
- Специализированные программные модули Picodata для сохранения текущих копий данных и постоянной синхронизации обновлений копий.

Блок мониторинга и регистрации

Модуль мониторинга и регистрации состоит из двух или трех узлов вычисления и хранения, представленных высокопроизводительными серверами одного из двух типов.

Обеспечивает управление на всех этапах жизненного цикла программных и аппаратных компонентов.

Основной функционал реализуется с помощью ПО Скала^р:

- Мониторинг и визуализация работы сети и оборудования, входящего в состав Машины
- Мониторинг и визуализация функционирования платформы Picodata, связи реплик платформы и компонентов физической инфраструктуры
- Накопление данных о функционировании Машины для автоматизированной и/или ручной оптимизации настроек аппаратной и программной платформы
- Автоматизированное реагирование на неблагоприятные события и отклонения параметров функционирования Машины
- Репозиторий пакетов для ОС и платформы Picodata для автоматизированной установки

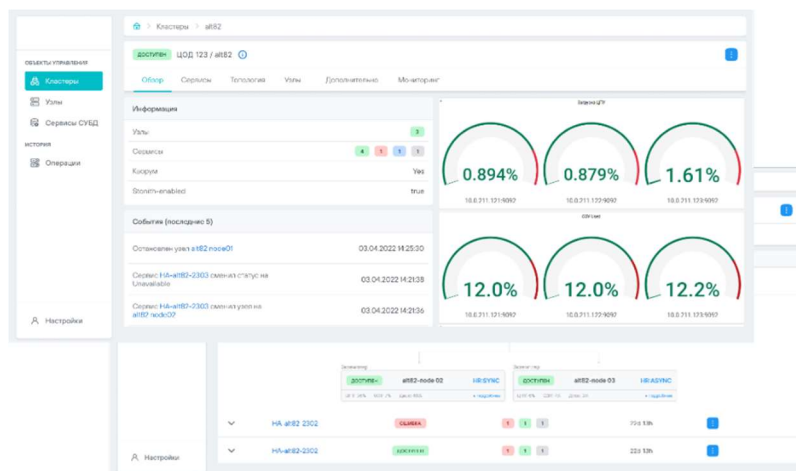


Рис. 10. Пример экрана подсистемы мониторинга Скала^р МБД.Т

Модуль мониторинга и регистрации состоит из двух высокопроизводительных узлов мониторинга и регистрации, объединенных в зеркальный кластер.



Рис. 11. Модуль мониторинга и регистрации

Каждый узел мониторинга и регистрации — специализированный сервер:

- Использует SSD для обеспечения высокой производительности при хранении служебных данных
- Содержит выделенные SSD для загрузки ОС
- Оснащен двухпортовыми сетевыми картами 100 Gigabit Ethernet для интерконнекта в рамках Машины
- Имеет 2 порта 1 Gigabit Ethernet для сетей управления и IPMI
- Содержит два блока питания в режиме резервирования по схеме (1 + 1)
- Имеет два процессора Xeon не ниже 2-го поколения.
- Использует 384 GB RAM

Применяемое программное обеспечение:

- ОС: Альт Линукс, сервер с виртуализацией Базис vCore
- Мониторинг и управление: разработка ПО Скала^р Визион
- Управление жизненным циклом: ПО Скала^р Геном

Компоновка решения Скала^р МБД.Т

Решение **Скала^р МБД.Т** состоит из нескольких отдельно стоящих секций (стоек), каждая из которых содержит определенные наборы модулей.

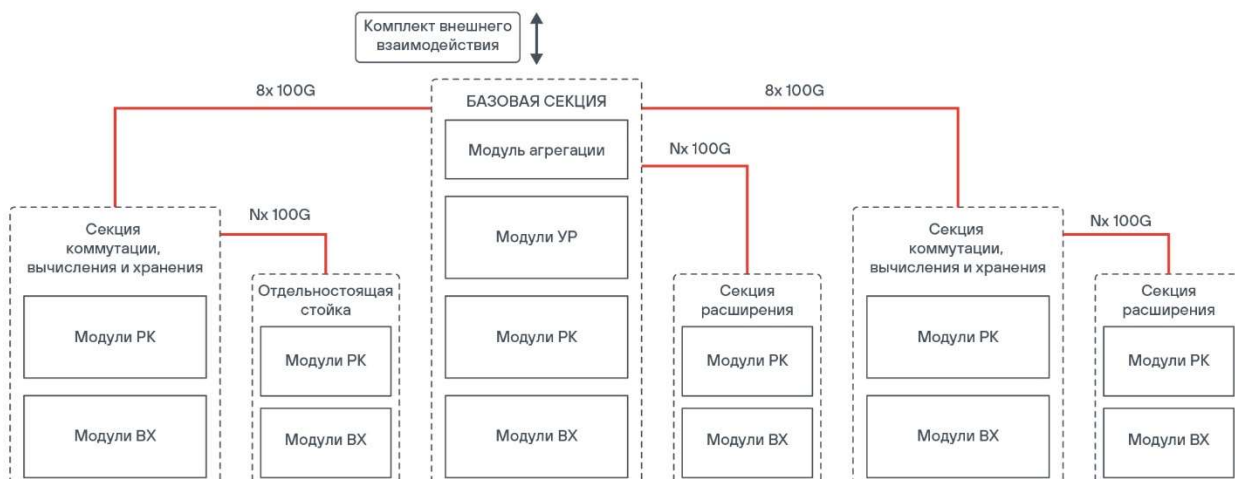


Рис. 12. Секции решения Скала^р МБД.Т

Секции решения Скала^р МБД.Т бывают трех типов:

- Базовая секция — основа решения, содержит все виды модулей, включая модуль агрегации
- Секция коммутации, вычисления и хранения — является типовой секцией для горизонтального масштабирования расширения; в отличие от базовой не имеет модуля агрегации
- Секция расширения (дополнительная стойка) — используется совместно с базовой секцией или секцией коммутации; в отличие от них не имеет сетевых узлов (установлены только узлы вычисления и хранения и узлы резервного копирования)

Базовая секция

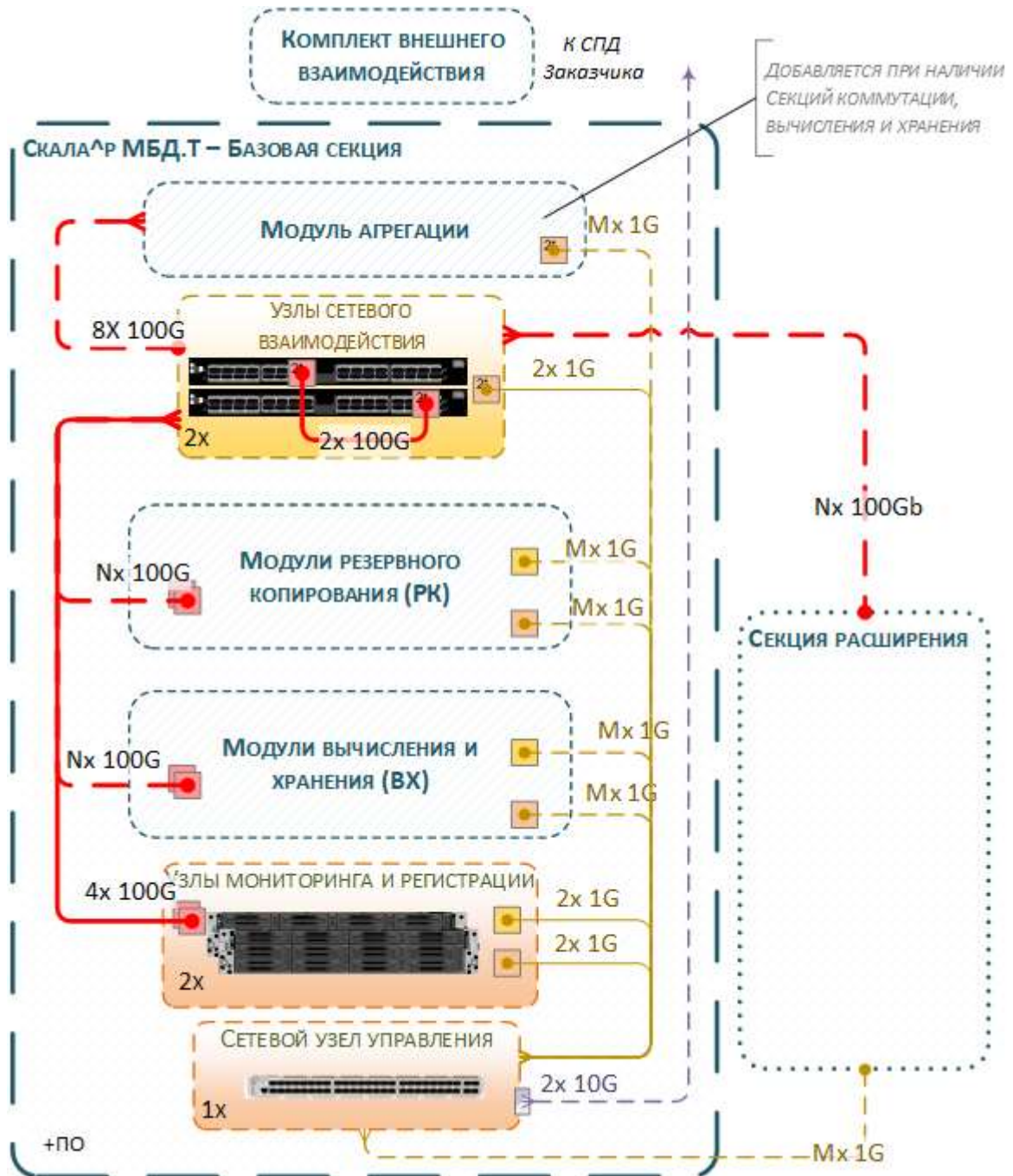


Рис. 13. Базовая секция

Базовая секция — основа любого решения содержит все виды модулей Машины и может доукомплектовываться отдельно стоящей секцией расширения.

Секция коммутации, вычисления и хранения

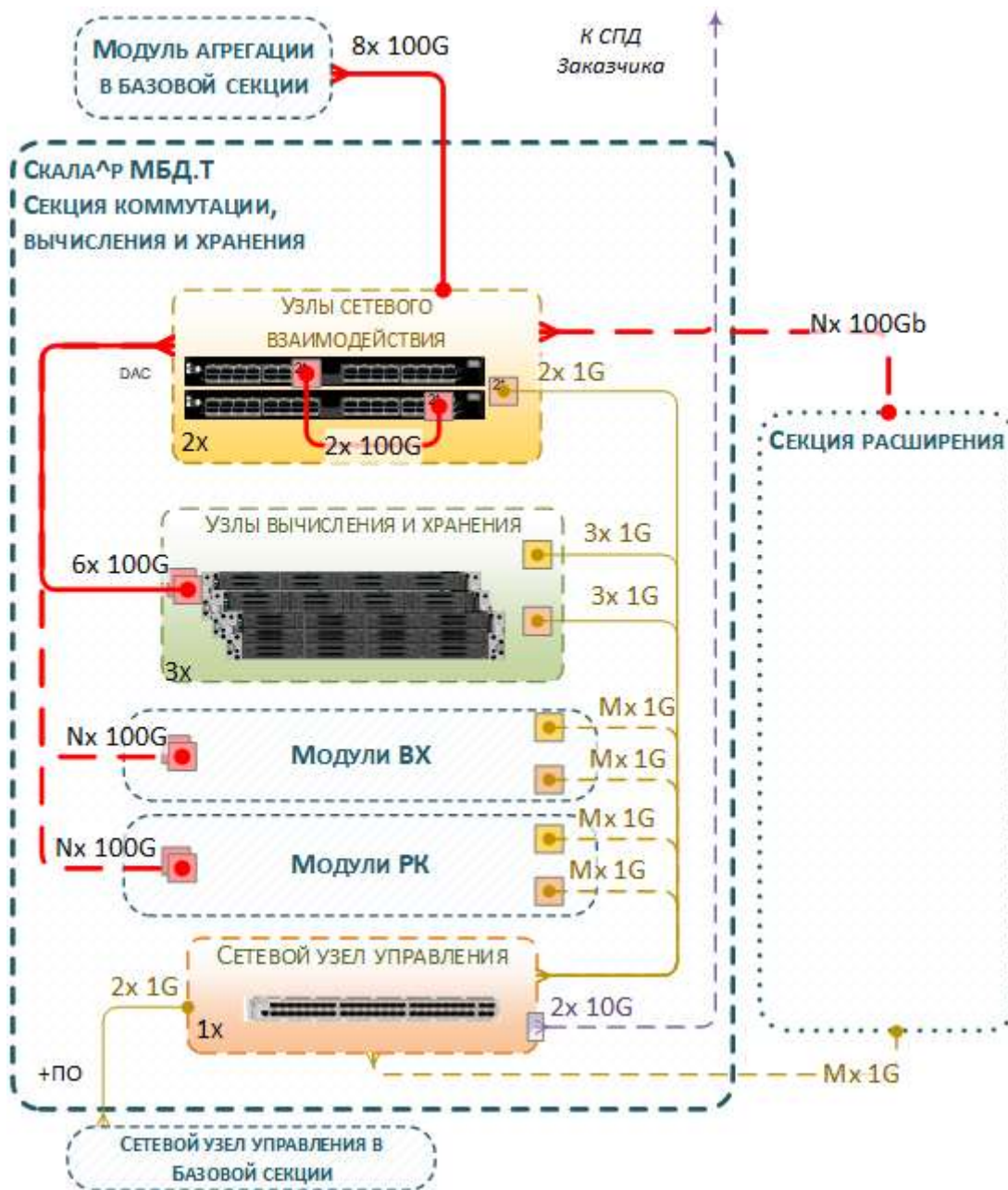


Рис. 14. Секция коммутации, вычисления и хранения

Секция применяется для горизонтального масштабирования Машины и может содержать следующие виды Модулей решения:

- Узлы сетевого взаимодействия
- Модули вычисления и хранения
- Модули резервного копирования

Также может доукомплектовываться отдельно стоящей секцией расширения.

Секция расширения (дополнительная стойка)

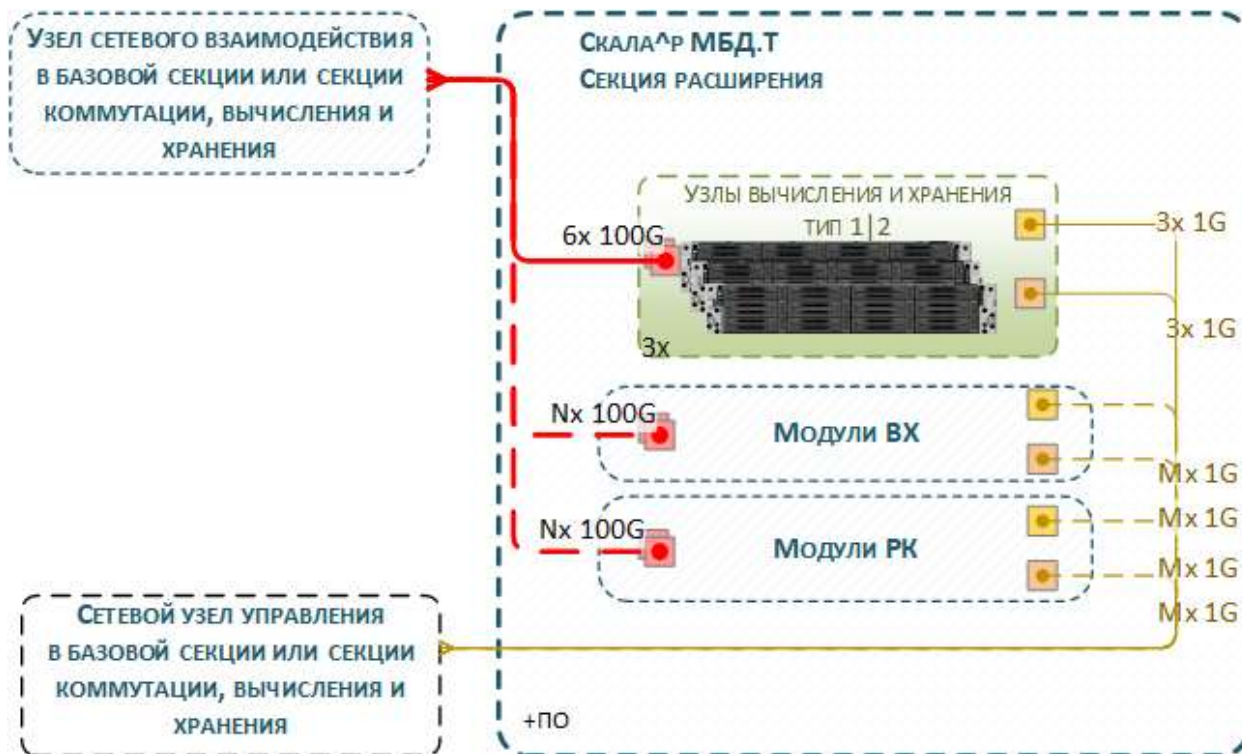


Рис. 15. Секция расширения

Секция предназначена для размещения дополнительных блоков в дополнение к базовой секции и/или секции коммутации, вычисления и хранения, содержит следующие виды блоков решения:

- Блоки вычисления и хранения
- Блок резервного копирования

Применение дополнительных стоек обуславливается требованиями и ограничениями инженерной инфраструктуры заказчика, в том числе — по допустимой потребляемой мощности и допустимому тепловыделению на отдельный серверный монтажный шкаф (стойку).

К любой основной секции можно добавить одну или две дополнительных стойки.

8. СПЕЦИФИЧНЫЕ ЧЕРТЫ

Проектирование и реализация решения **Скала^р МБД.Т** осуществлялись с учётом ряда выбранных приоритетов, оказывающих непосредственное влияние на функциональные и эксплуатационные показатели. Наиболее значимые из них следующие:

Размещение данных в оперативной памяти и минимизация ввода-вывода

Эффект:

- Максимально возможная производительность подсистемы хранения
- Повышение производительности (данные хранятся в той же памяти, что и исполняемые программные процедуры)
- Повышение надёжности работы решения (исключено влияние отказов дисков на базу данных)

Программно-определяемый приоритет повышенной устойчивости решения или повышенной производительности

Приоритет повышенной устойчивости:

- Гарантия сохранности данных при любых отказах
- В случае сбоя система продолжает работать без снижения производительности, простоя или потери данных (возможна кратковременная задержка на период назначения платформой другой реплики в качестве мастера вместо вышедшей из строя — несколько микросекунд)

Приоритет производительности:

- Снижение времени отклика за счет большего распараллеливания

Выбор аппаратного решения для реализации подсистемы вычисления и хранения

Эффект:

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

Применение стандартного высоконадёжного и производительного оборудования в качестве платформы для размещения компонентов решения

Эффект:

- Обеспечение стабильного уровня производительности (компоненты проверены временем)
- Повышение надёжности решения (нет уникальных элементов)
- Снижение стоимости сопровождения (доступность элементов при выходе из строя)

Применение программных RAID отечественного производства

Эффект:

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

Выбор ПО с открытым кодом и отечественных разработок

Эффект:

- Повышение производительности за счёт доработки ПО (силами Скала^р и партнёров)
- Повышение надёжности решения (снижение рисков недоступности поддержки)
- Снижение зависимости от импортных производителей ПО

9. ГАРАНТИРОВАННОЕ КАЧЕСТВО

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

Производство (комплектование и развёртывание ПО)

- При производстве используются высококачественные комплектующие
- Сборка продукции осуществляется строго в соответствии с утверждённым планом размещения компонентов
- Первичное развёртывание ПО осуществляется в автоматическом режиме
- Дополнительные настройки ПО осуществляются в соответствии с утверждённой пошаговой инструкцией
- Осуществляется тестирование сформированной Машины
- Отклонения от типового решения Скала^р МБД.Т исключены

Передача в эксплуатацию

- Скала^р МБД.Т полностью сформирована, протестирована, готова к размещению в сети заказчика и развёртыванию решения на базе платформы Picodata
- В комплекте со Скала^р МБД.Т передаётся паспорт решения, эксплуатационная документация, сертификат на поддержку
- Проводится обучение специалистов заказчика работе со Скала^р МБД.Т (опция по запросу)

Поддержка

- Скала^р МБД.Т 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.
- 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.Т.

Сопровождение

Возможна реализация дополнительных требований по модернизации или развитию Скала^р МБД.Т (по запросу), в том числе:

- Аппаратная модернизация Машины
- Тестирование приложений, производительности приложений или иное другое запрошенное тестирование

Работы выполняются с участием архитекторов и инженеров, непосредственно участвовавших в разработке **Скала^р МБД.Т**.

10. РЕАКЦИЯ НА ВОЗМОЖНЫЕ ОТКАЗЫ

Отказы, связанные со стандартными элементами Скала^р МБД.Т

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

- серверов (дублирование процессоров, источников питания и др.)
- дисковых подсистем (программный RAID)
- сетей интерконнекта и доступа (полное дублирование)
- сетей управления (резервный вариант - использование сети интерконнекта)

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

Отказы, связанные с узлами кластера баз данных

Для обеспечения бесперебойности доступа и сохранности данных в решении реализован многоузловой кластер, состоящий из набора сегментов (наборов реплик). В рамках каждого сегмента один из узлов выполняет функцию мастер-сервера БД с возможностью чтения и записи данных, остальные — функцию реплики с возможностью чтения.

В случае отказа любого узла кластера, исполняющего роль мастера, платформа в автоматическом режиме назначает на эту роль один из узлов с репликой, изменяет общую структуру кластера (уменьшив на одну реплику соответствующий сегмент) и продолжает обработку запросов в обновленной конфигурации. В случае отказа узла с репликой — переключение не требуется, выполняется только актуализация конфигурации кластера. Для пользователей такие отказы остаются незаметными, за исключением потенциального повышения времени отклика в периоды пиковых нагрузок.

После завершения обслуживания или устранения причины отказа и восстановления узла размещенные в сегменте данные будут скопированы на него средствами платформы и произведено переконфигурирование кластера (добавление узла в сегмент).

В случае полного отказа всех узлов сегмента платформа произведет переконфигурирование кластера путем выделения под данные этого сегмента узлов из других сегментов или путем включения данных сегмента в другой сегмент. При этом последующее восстановление данных сегмента осуществляется из актуальной резервной копии.

Конкретный алгоритм реагирования кластера определяется заказчиком в ходе формирования решения на платформе и может быть скорректирован в произвольный момент времени.

Детальный порядок обеспечения отказоустойчивости кластера и рекомендации по действиям при его администрировании в той или иной конкретной ситуации с конкретным экземпляром **Машины больших данных Скала^р МБД.Т** могут быть предоставлены по запросу.

11. ПРИМЕРЫ КОМПЛЕКТОВ ПОСТАВКИ РЕШЕНИЯ



**Скала^р
МБД.Т К-1**



**Скала^р
МБД.Т К-2**



**Скала^р
МБД.Т К-3**

Рис. 16. Варианты поставки МБД.Т (примеры)

Параметры	Модель	К-1	К-2	К-3
Количество секций (стоек)		2	16	1
Общее энергопотребление		до 15 кВт	до 120 кВт	до 15 кВт
Энергопотребление на стойку		до 7,5 кВт	до 7,5 кВт	до 15 кВт
Количество серверов - узлов вычисления и хранения		16	128	16
Количество In-Memory платформ		1	1	1
Общий объем БД, Тбайт (при минимум двух репликах на один шард)		до 6	до 30	до 6
Объем хранения системы резервного копирования (СРК), Тбайт		до 180	до 1440	до 180

12. ЛИЦЕНЗИРОВАНИЕ РЕШЕНИЯ

Все наименования ПО лицензируются по модулям Машины и их количеству в Машине.

Лицензирование ПО Скала^р МБД.Т

Программное обеспечение платформы Picodata лицензируется согласно объёму ресурсов, при этом на каждый модуль выдается единая лицензия.

Программное обеспечение **Скала^р Визион** и **Скала^р Геном** поставляется исключительно в составе Машин **Скала^р МБД.Т**, и лицензируется количеством модулей в ней.

Варианты лицензирования

Лицензирование ПО комплекса **Скала^р МБД.Т** имеет две редакции:

- Фиксированная — приобретается бессрочная (постоянная) лицензия
- Временная — приобретается лицензия на период времени

Политика обновления ПО

Команда **Скала^р** активно занимается развитием собственных программных продуктов **Скала^р МБД.Т**. Направления развития формируются на основе анализа мирового опыта использования систем подобного класса и пожеланий заказчиков и партнеров. Новые функции реализуются в форме мажорных и минорных релизов: мажорные релизы выпускаются раз в квартал, минорные релизы выпускаются при необходимости более быстрого введения в эксплуатацию небольших улучшений в системе.

13. ВАРИАТИВНОСТЬ РЕШЕНИЯ

Варианты компоновки решения при фиксированном объеме хранимых данных.

Приоритет удельного энергопотребления

Вариант решения:

- Увеличенное количество секций
- Низкая степень заполнения стоек
- Использование менее производительных процессоров

Приоритет занимаемой площади

Вариант решения:

- Максимальное заполнение секций оборудованием
- Сокращение количества секций

Приоритет производительности

Вариант решения:

- Увеличенное количество секций
- Увеличенного количество узлов вычисления и хранения (высокая степень заполнения стоек)
- Использование высокопроизводительных процессоров

Пространственно-разнесенный отказоустойчивый кластер (как опция)

Вариант решения:

- Размещение отдельных секций расширения на удалении до 500 метров от базовой секции решения
- Настройки платформы осуществляются с учетом размещения реплики на разных секциях (локальных или удаленных)

14. ТРЕБОВАНИЯ К РАЗМЕЩЕНИЮ РЕШЕНИЯ

Решение поставляется в виде одного или нескольких серверных монтажных шкафов (стоек) 19", высота 42U.

Наполнение стоек оборудованием и совокупный вес зависит от выбранного варианта решения и может составлять от 300 до 800 кг.

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

Потребляемая мощность стойки может составлять от 6 до 15 кВт.

Должны быть предусмотрены соответствующие мощности по отводу тепла.

Подключение к локальной сети осуществляется в соответствии с требованиями заказчика, возможны варианты 4×100 Gigabit Ethernet, 8×10/25 Gigabit Ethernet или иные.

При развёртывании решения на нём 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 путём проведения настройки прикладных программных систем.

15. ПРИМЕРЫ РАБОТАЮЩИХ РЕШЕНИЙ

Государственная организация федерального уровня

Внедрено решение как один из компонентов комплекса Машин Скала^р в составе крупного проекта. Машина внедрена в крупных организациях, инсталляционная база составляет более 30 модулей.

Реализована инфраструктура платформы подписания документов в двух географически разнесенных ЦОД. Общий объем данных для обработки в памяти БД составляет около 1 Тбайт. Архитектура решения предполагает дублирование данных с репликацией. Система имеет минимальные задержки ответа до 75 наносекунд, что позволяет обеспечить возможность визуализации данных на географической карте и в сетях, изолированных от Интернет в реальном времени.

Надёжность, производительность решения Скала^р МБД.Т подтверждается проведёнными тестами, практическим использованием решений в течение ряда лет.

Дополнительная информация по решению Скала^р МБД.Т 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 развёртывание 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.