



## Машина больших данных Скала^р МБД.КХ

Программно-аппаратный комплекс для быстрых аналитических витрин с реляционным доступом на базе технологии ClickHouse (Arenadata QuickMarts)

### Технический обзор



## ОГЛАВЛЕНИЕ

1. Введение .....	3
2. Основы аналитических СУБД с колоночным хранением .....	4
3. Отличительные черты .....	6
4. Подтвержденная безопасность .....	7
5. Особенности сценариев применения .....	9
6. Способы работы с данными .....	10
7. Технологические особенности решения .....	12
8. Состав решения .....	16
9. Специфичные черты Машины Скала^р МБД.КХ .....	31
10. Гарантированное качество .....	33
11. Реакция на возможные отказы .....	34
12. Вариативность решения .....	36
13. Требования к размещению решения .....	37
14. Техническая поддержка .....	38
15. Лицензирование ПО Машины больших данных .....	40
О компании .....	41

### 1. ВВЕДЕНИЕ

**Машина больших данных Скала^р МБД.КХ** — это программно-аппаратный комплекс быстрых аналитических витрин с реляционным доступом под управлением СУБД Apenadata QuickMarts (ADQM), созданной на основе ClickHouse.

**Скала^р МБД.КХ** — готовое решение, обеспечивающее высокую производительность и отказоустойчивость, позволяет снизить затраты за счёт проработанной интеграции аппаратного и программного обеспечения, оптимизации алгоритмов для используемых технологий, широкого применения методов обеспечения надёжности, комплексности решения, специальных моделей лицензирования.

**Скала^р МБД.КХ** более чем в 100 раз быстрее реляционных СУБД общего назначения и 2—20 раз быстрее классических аналитических реляционных СУБД с колоночным хранением на нагрузках витринного типа.

**Скала^р МБД.КХ** комплексное решение, включающее узлы сегментов баз данных, систему резервного копирования, высокоскоростную сетевую среду, 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 надёжных комплектующих, специализированной версии СУБД (ADQM) с резервированием критических компонентов, использованием устойчивых сетевых протоколов.

**Скала^р МБД.КХ** содержит все необходимые элементы для функционирования высоконагруженной СУБД ADQM. Подключение к внешним сетям осуществляется с помощью стандартного интерфейса Ethernet.

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

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

**Машина больших данных Скала^р МБД.КХ** впервые была представлена в 2021 году как продукт в линейке Скала^р МБД8.АДКМ.

Решение внедрено в крупных корпоративных и государственных организациях, инсталляционная база составляет более 40 узлов.

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

## 2. ОСНОВЫ АНАЛИТИЧЕСКИХ СУБД С КОЛОНОЧНЫМ ХРАНЕНИЕМ

Реляционные СУБД общего назначения очень хорошо справляются с задачами обработки и изменения единичных записей в транзакционных системах, однако, при выполнении аналитических запросов такая СУБД требует значительных аппаратных ресурсов и их производительность, как правило, недостаточна. Строчное хранение, используемое в реляционных системах общего назначения, обеспечивает эффективную обработку транзакций ценой крайне медленной работы сложных аналитических запросов, в тех случаях, когда необходимо агрегировать информацию из миллиардов строк и всего лишь нескольких.

При создании отчётов в строковых СУБД приходится анализировать множество связанной информации, это приводит к необходимости считывания всех строк, необходимых для выполнения запроса, при этом в записях используется лишь некоторые из полей, а остальные данные просто являются сопутствующей нагрузкой, что загружает ресурсы системы. На скорость построения выборки из такого количества записей не влияет даже оптимизация, настроенные индексы и ключевые поля. Агрегация данных происходит уже на последующем этапе.

Поэтому для эффективной обработки аналитических запросов применяется колоночное хранение: данные физически хранятся не по строкам, а по столбцам, что позволяет эффективно сжимать данные (поскольку данные в столбцах одного типа) и производить доступ только к запрошенным столбцам, в результате объёмы данных, пропускаемые через подсистему ввода-вывода, сокращаются на порядки. При этом вставка единичных записей при такой организации хранения намного более ресурсозатратна.

Для аналитических задач характерны следующие свойства:

- большинство запросов поступает на чтение данных
- данные добавляются и обновляются достаточно большими пачками более 1000 строк, а не по одной, или не добавляются и не обновляются
- при чтении используется значительное количество строк данных, и небольшое количество столбцов.

Таким образом, неэффективность коротких вставок в базу данных со колоночной организации хранения в аналитических системах не является критическим недостатком, а эффект, получаемый при доступе к данным — существенный.

На Рис. 1 представлено сравнение колоночного хранения данных (по столбцам) и строкового хранения данных. Во втором случае объем считываемых данных существенно больше.



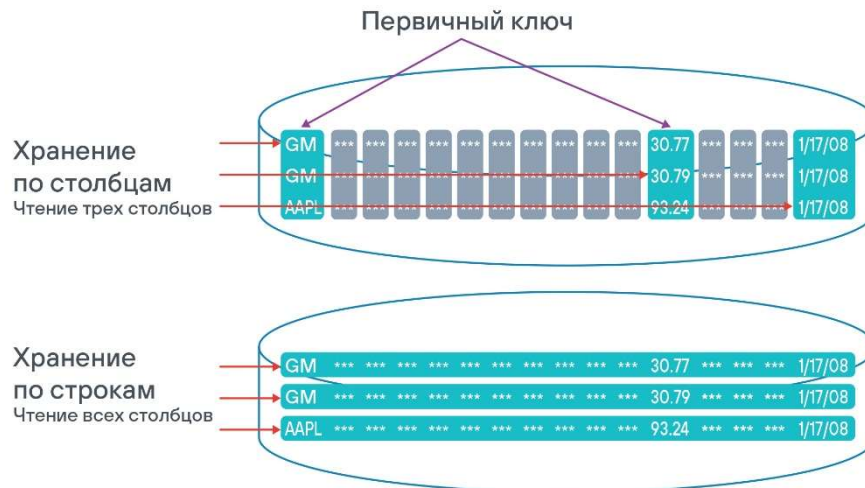


Рис. 1. Схема сравнения

Операции выборки при колоночном и строчном хранении СУБД ADQM созданы таким образом, чтобы можно было быстро строить аналитические отчёты по данным, не прошедшим предварительную агрегацию. Такой подход позволяет намного эффективнее, в сравнении с другими системами, решать задачи подготовки аналитической информации, поэтому его активно используют в мониторинге и бизнес-аналитике, а также для анализа данных телеметрии.

Ниже приведены архитектурные особенности ADQM:

- хранение данных в столбцах позволяет считывать информацию только из нужных колонок, а также обеспечивать сжатие однотипной информации;
- поддерживаются приближённые вычисления для частичных выборок данных, что позволяет снизить число обращений к подсистеме хранения и повышает скорость обработки запросов;
- физическая сортировка данных по первичному ключу позволяет быстро получать конкретные значения или диапазоны;
- векторные вычисления по фрагментам данных столбцов позволяют снизить издержки на диспетчеризацию и повысить эффективность использования процессоров;
- применение параллельных операций как в пределах одного сервера на несколько процессорных ядер, так и в рамках распределённых вычислений на кластере за счёт механизма сегментирования позволяет существенно повысить производительность системы;
- линейная масштабируемость позволяет построить кластер со многими десятками узлов;
- отказоустойчивость реализована за счёт репликации сегментов.

ADQM поддерживает множество клиентских программ для подключения: консольный клиент, HTTP API, компоненты Python, PHP, Node.js, Perl, Ruby, R и многие другие. Также для ADQM применяются ODBC, JDBC и Golang драйверы.

### 3. ОТЛИЧИТЕЛЬНЫЕ ЧЕРТЫ

Важное преимущество **Машины больших данных Скала^р МБД.КХ** по сравнению с аналогичными решениями — это **линейная масштабируемость**. Поставляемая Машина способна линейно масштабироваться на петабайты данных. Применение решения cross-datacenter позволяет даже «растянуть» решение на несколько удалённых друг от друга центров обработки данных.

Система построена так, что сервис **будет всегда доступен** за счет применения кластеризации. С помощью штатной настройки можно получить требуемый коэффициент репликации и распределения.

**Машина больших данных Скала^р МБД.КХ обеспечивает сжатие данных в десятки, сотни и даже тысячи раз** за счет колоночного хранения и большого количества оптимизаций.

Поддерживаемый **диалект SQL** имеет разнообразные дополнительные функции для приблизительных вычислений, для обработки URL и для работы с дополнительными типами данных.

**Машина больших данных Скала^р МБД.КХ** имеет следующие особенности ввода-вывода:

- **Сокращенное считывание.** Для выполнения аналитического запроса требуется прочитать небольшое количество столбцов таблицы, в этом случае система сканирует только необходимые столбцы.
- **Блочное сжатие.** Так как данные читаются блоками, то их проще сжимать. Данные, лежащие по столбцам, также лучше сжимаются. За счёт этого дополнительно уменьшается объём ввода-вывода.

В **Машине больших данных Скала^р МБД.КХ** используется **векторный движок**. Операции осуществляются не над отдельными значениями, а над наборами значений с использованием векторных процессорных инструкций (AVX). За счёт этого издержки на диспетчеризацию становятся пренебрежимо малыми.

Так же одной из особенностей вычислений **Машины больших данных Скала^р МБД.КХ** является **кодогенерация** — осуществляется динамическая компиляция запросов в нативный код, что обеспечивает лучшую производительность в сравнении с интерпретируемым выполнением.

## 4. ПОДТВЕРЖДЕННАЯ БЕЗОПАСНОСТЬ

Машина больших данных Скала^р МБД.КХ 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

**Машина больших данных Скала^р МБД.КХ** используется сертифицированная СУБД Arenadata QuickMarts (Сертификат ФСТЭК 4682 от 08.06.2023, действует до 08.06.2028), которая:

### 1. Может применяться для защиты информации:

- В значимых объектах критической информационной инфраструктуры 3 категории.
- В государственных информационных системах 3 класса защищенности
- В автоматизированных системах управления производственными и технологическими процессами 3 класса защищенности
- В информационных системах персональных данных при необходимости обеспечения 3 и 4 уровня защищенности персональных данных

### 2. Соответствует требованию следующих нормативных документов:

- «Требования по безопасности информации, устанавливающие уровни доверия к средствам технической защиты информации и средствам обеспечения безопасности информационных технологий» (ФСТЭК России, 2020, приказ № 76) - по 6 уровню доверия



## 5. ОСОБЕННОСТИ СЦЕНАРИЕВ ПРИМЕНЕНИЯ

Машина больших данных Скала^р МБД.КХ эффективно применяется в сценариях, где подавляющее большинство запросов выполняются на чтение. Максимальная эффективность достигается при соблюдении следующих условий:

### Оптимальные сценарии чтения данных

- при чтении извлекается достаточно большое количество строк из БД и небольшое подмножество столбцов
- запросы осуществляются к таблицам с большим количеством столбцов и огромным числом строк, и при этом используются соединения только с небольшими таблицами (словарями, «справочниками»)
- запросы поступают относительно редко (не более сотни в секунду на сервер)
- при выполнении простых запросов допустимы задержки порядка 50 мс
- результат выполнения запроса существенно меньше исходных данных;
- результат выполнения запроса помещается в оперативную память одного узла
- требуется высокая пропускная способность при обработке одного запроса (до миллиардов строк в секунду на один сервер)

### Особенности хранения и обработки данных

В ADQM не поддерживаются механизмы изоляции транзакций. В качестве значений в столбцах должны быть записаны числа или относительно небольшие строки (пример — 60 байтов для URL).

### Сферы применения

Благодаря своим возможностям, **Машина больших данных Скала^р МБД.КХ** на основе СУБД ADQM может использоваться в самых разных сферах, для которых характерны задачи аналитики над накапливаемыми данными журнального характера — платёжных систем, систем диспетчеризации, для веб-проектов, мобильных приложений.

Сверхвысокая скорость на аналитических запросах к огромным ненормализованным таблицам — это одно из главных преимуществ решения, которое в режиме реального времени генерирует данные для аналитических отчётов, что позволяет с успехом использовать её для нужд государственных органов, телекоммуникаций, ритейла, в области электронной коммерции, торгов в реальном времени, интернета вещей.

Отслеживание бизнес-показателей предоставляет широкие возможности для бизнес-аналитики, анализа поведения пользователей в различных системах, таких как интернет-магазины или онлайн-игры. Возможность создания витрин данных без предварительной разработки моделей хранилищ данных и сложных ETL-процессов позволяет безопасно предоставлять пользователям только нужный контент.

## 6. СПОСОБЫ РАБОТЫ С ДАННЫМИ

В **Машине больших данных Скала<sup>А</sup>р МБД.КХ** используется набор движков для работы с данными. Движок — это код, встроенный в таблицу, который определяет или обеспечивает:

- как и где хранятся данные, куда их записывать и откуда считать
- какие запросы поддерживаются и каким образом
- конкурентный доступ к данным
- использование индексов, если они есть
- возможность многопоточного выполнения запроса
- параметры репликации данных

В **Машине больших данных Скала<sup>А</sup>р МБД.КХ** используются движки, показанные в таблице ниже.

*Движки Скала<sup>А</sup>р МБД.КХ*

Семейство	Описание	Движок
MERGETREE	Основное семейство для колоночного хранения, обеспечивают относительно быструю вставку данных без синхронизации с последующим слиянием в фоновом режиме.	MergeTree
		ReplacingMergeTree
		SummingMergeTree
		AggregatingMergeTree
		CollapsingMergeTree
		VersionedCollapsingMergeTree
		GraphiteMergeTree
LOG	Движки разработаны для сценариев, когда необходимо быстро записывать много таблиц с небольшим объёмом данных (менее 1 миллиона строк), а затем читать их целиком.	TinyLog
		StripeLog
		Log
Движки для интеграции	Конфигурирование интеграционных движков для интеграции с внешними системами осуществляется с помощью запросов.  С точки зрения пользователя, настроенная интеграция выглядит как обычная таблица, но запросы к	Kafka
		MySQL
		ODBC
		JDBC

Семейство	Описание	Движок
	<p>ней передаются через прокси во внешнюю систему. Этот прозрачный запрос является одним из ключевых преимуществ этого подхода по сравнению с альтернативными методами интеграции, такими как внешние словари или табличные функции, которые требуют использования пользовательских методов запроса при каждом использовании.</p>	S3
		EmbeddedRocksDB
		RabbitMQ
		PostgreSQL
<p>Специальные движки</p>	<p>Несгруппированные в другие семейства движки таблиц, уникальные по назначению.</p>	Distributed
		MaterializedView
		Dictionary
		Merge
		File
		Null
		Set
		Join
		URL
		View
		Memory
		Buffer

## 7. ТЕХНОЛОГИЧЕСКИЕ ОСОБЕННОСТИ РЕШЕНИЯ

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

### Технологические принципы

- Дублирование критичных компонентов
- Применение высокопроизводительных компонентов
- Горизонтальное масштабирование вычислительных ресурсов
- Сохранение работоспособности при отказе отдельных элементов системы (в отдельных случаях — со снижением производительности)

### Технические решения

- Блочная-модульная архитектура
- 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

Архитектура **Машины больших данных Скала^р МБД.КХ** базируется на следующих принципах:

### Программный RAID

- Производительнее аппаратного RAID-контроллера
- Не подвержен аппаратным сбоям
- Минимальное использование RAM (требуется менее 4 GB RAM)
- Минимальная снижение производительности в режиме восстановления RAID

### Спрогнозированная нагрузка

- Распараллеливание нагрузки достигается с помощью сегментирования
- Производительность можно выбирать встраиванием согласованного с задачей движка

### Выделенный интерконнект

- Высокоскоростная сеть интерконнекта по схеме «Spine-Leaf» ускоряет взаимодействие между узлами, что отражено на (Рис. 2)
- Параллельная обработка запросов на узлах приводит к суммированию мощностей всех узлов
- Создание параллельной синхронной копии не влияет на выполнение задания
- Все узлы взаимодействуют между собой с одинаковой скоростью

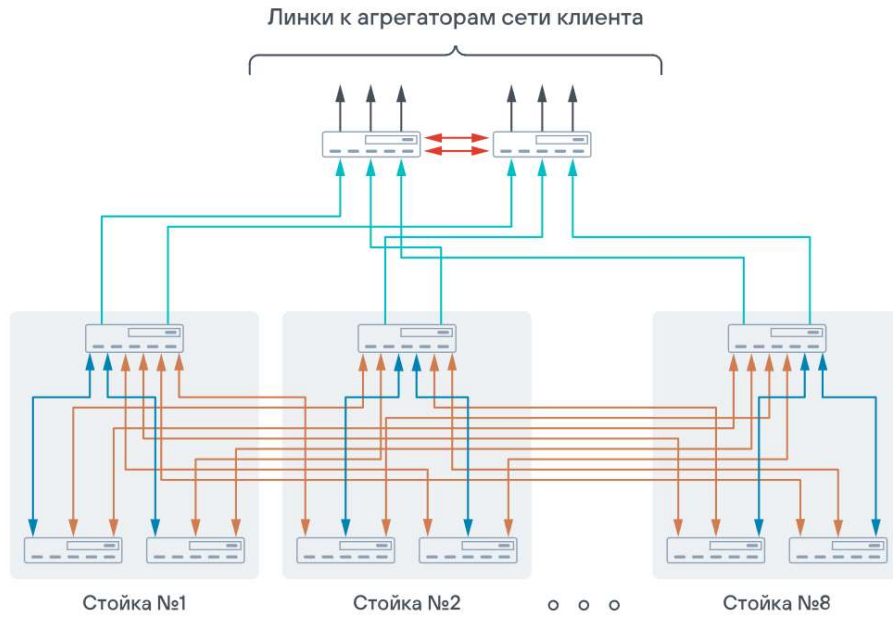


Рис. 2. Схема внутренних соединений Leaf-Spine с увеличением скорости при горизонтальном масштабировании

### Отказоустойчивость

В **Машина больших данных Скала^р МБД.КХ** осуществляется синхронное копирование сегментов, что позволяет обеспечивать стабильную доступность данных (Рис. 3).

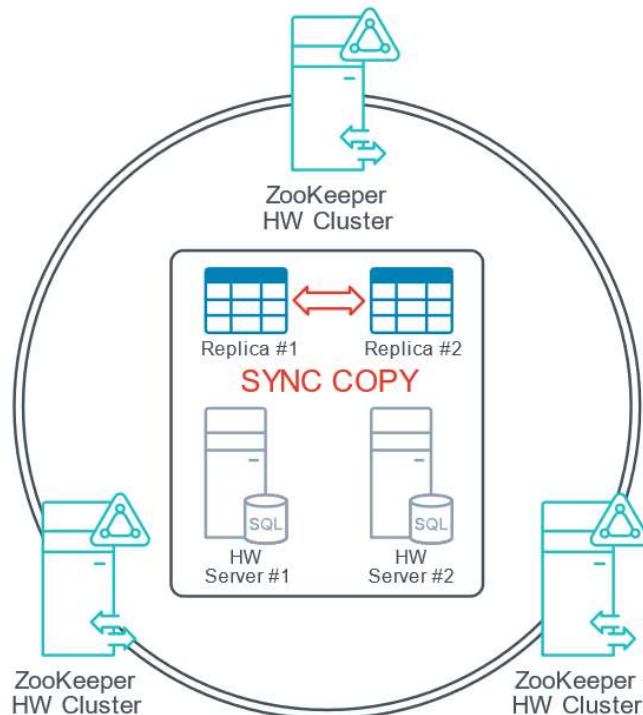


Рис. 3. Схема синхронного копирования БД

Синхронная репликация доступна для следующих таблиц семейства MergeTree:

- ReplicatedMergeTree
- ReplicatedSummingMergeTree



- ReplicatedReplacingMergeTree
- ReplicatedAggregatingMergeTree
- ReplicatedCollapsingMergeTree
- ReplicatedVersionedCollapsingMergeTree
- ReplicatedGraphiteMergeTree

Метаинформация о репликах хранится в трехточечном кластере Модуля Управления Машины.

Репликация работает на уровне отдельных таблиц, а не всей базы данных. На каждом узле могут быть расположены одновременно реплицируемые и нереплицируемые таблицы. Репликация не зависит от сегментирования, не привязана к именам таблиц и основана на запросах INSERT и ALTER.

### Повышенная надежность

Асинхронная репликация позволяет повысить надежность системы и осуществлять распараллеливание запросов (Рис. 4).

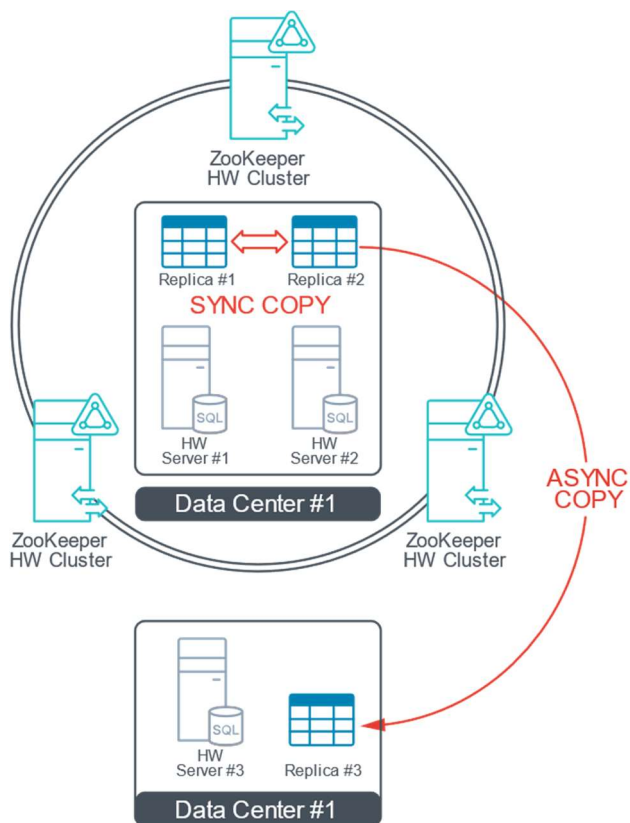


Рис. 4. Схема асинхронного копирования БД

Машина больших данных Скала^р МБД.КХ использует распределенный кластер, что позволяет увеличить надежность за счет децентрализации и отсутствия единой точки отказа.

Асинхронная многосторонняя репликация обеспечивает реплицирование данных в фоновом режиме. СУБД поддерживает полную идентичность данных на разных репликах, автоматически восстанавливая их после сбоев.

В Машине используется кворумный режим записи данных. Запись считается успешной только после того, как информация записана на несколько узлов — обеспечен кворум.

Отставание асинхронной реплики определяется шириной канала связи и задержками.

Асинхронная реплика может быть использована для создания резервных копий, чтобы не останавливать основной синхронный контур.

### Расширение объёма данных — сегментирование

Основные преимущества сегментирования:

- Снимает ограничение ресурсов одного узла, увеличивая объем базы на десятки и сотни узлов
- Позволяет распараллеливать выполнение запросов, увеличивая скорость в десятки и сотни раз
- Позволяет с помощью реплик строить защиту на любом уровне — таблицы, узла, кластера, стойки, ряда стоек, зала, центра обработки данных

На схеме (Рис. 5) приведен вариант сегментирования и кольцевой репликации, при котором можно допустить выход из строя целой серверной стойки без потери данных.

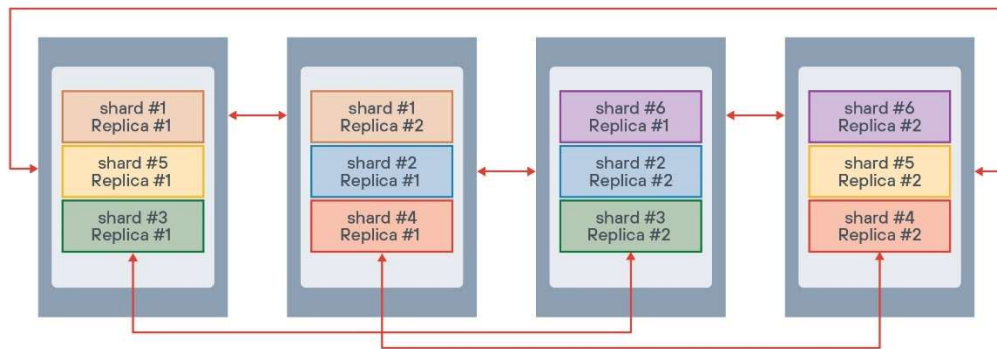


Рис. 5. Вариант сегментирования и кольцевой репликации

### Сокращенное считывание

Для выполнения аналитического запроса, требуется прочитать небольшое количество столбцов таблицы, что обеспечивается колоночным хранением, используемым в ADQM.

### Блочное сжатие

Так как данные читаются блоками, то их проще сжимать. Данные, лежащие по столбцам, также лучше сжимаются, за счет этого дополнительно уменьшается объем ввода-вывода.

## 8. СОСТАВ РЕШЕНИЯ

Ниже приведены термины, используемые для комплектации **Машины больших данных Скала^р МБД.КХ**.

**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 модулей, выполняющих единую функцию в одной или нескольких стойках.

**Модуль** — это единица поставки Машин Скала^р, 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 Машин Скала^р, объединенных в один серверный шкаф.

Формирование решения основано на принципе разделения на блоки и модули. 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.

Решение **Скала^р МБД.КХ** состоит из следующих блоков:

Блок вычисления и хранения

Блок коммутации и агрегации

Блок управления и распределения

Блок мониторинга и регистрации

Блок резервного копирования



**Блок вычисления и хранения**

Блок вычисления и хранения состоит из типовых Модулей вычисления и хранения на базе аппаратных узлов и программной платформы ADQM.

Образуемый узлами модулей кластер ADQM не имеет мастер-узла и единой точки входа. Запросы могут отправляться к любому из узлов в кластере. Архитектура кластера представлена на Рис. 6.

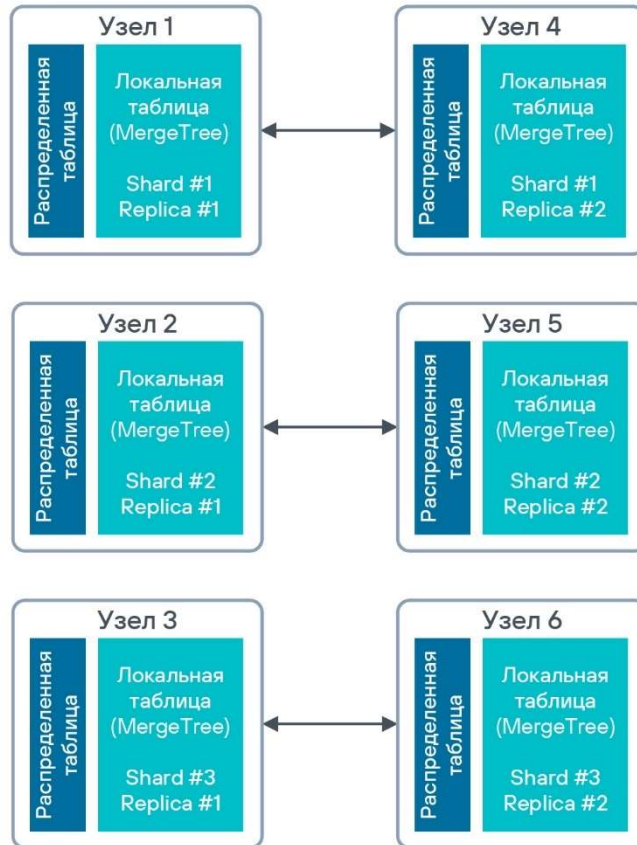


Рис. 6. Архитектура кластера

Локальные таблицы (Local table) отвечают за хранение и репликацию данных.

Распределенные таблицы (Distributed table) не хранят данные, они позволяют сделать запрос на несколько локальных таблиц, распределенных на хостах, объединённых в виртуальный кластер (remote\_servers).

Сегменты (shards) — части данных, распределенные по серверам для распараллеливания запросов. Реплики (replicas) позволяют обеспечить отказоустойчивость на уровне узлов (hosts).

Модуль вычисления и хранения состоит из двух или трёх узлов и соединен с Модулем сетевого взаимодействия и сетевым узлом управления (Рис. 7).



Рис. 7. Модуль вычисления и хранения

### Каждый отдельный узел вычисления и хранения:

- содержит выделенные накопители SAS SSD для загрузки ОС
- использует локальные SAS SSD 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 512 GB RAM в узле 1-го типа и 384 GB RAM в узле 2-го типа

### Применяемое программное обеспечение:

- ОС: Linux CentOS
- специальная версия платформы ADQM
- управление резервным копированием: специализированные программные модули платформы ADQM
- управление кластером средствами ADQM



**Блок коммутации и агрегации**

Обеспечивает передачу данных между элементами **Машиной больших данных Скала^р МБД.КХ** (интерконнект) и информационный обмен с внешними сетями, а также низкоуровневое управление узлами Машины. Соответственно эффективное сетевое взаимодействие является важным фактором для быстрого и надежного функционирования кластера.

Схема сетевого взаимодействия представлена на Рис. 8.

**Реализованные подсети:**

- Internal VLAN — сеть интерконнекта, для внутреннего взаимодействия между узлами БД, сеть резервного копирования, сеть кластерного взаимодействия на схеме ниже обозначена стрелками (оранжевого цвета)
- External VLAN — сеть для подключения к сервисам БД внешних пользователей и прикладных систем, подключение к серверу управления на схеме ниже обозначена стрелками (черного цвета)
- Сети мониторинга и управления - обмен служебными данными, данными для мониторинга и сеть управления узлами Машины на схеме ниже обозначена стрелками (синего цвета)
- Стрелками зеленого цвета обозначены соединения коммутаторов между стойками с агрегированным трафиком

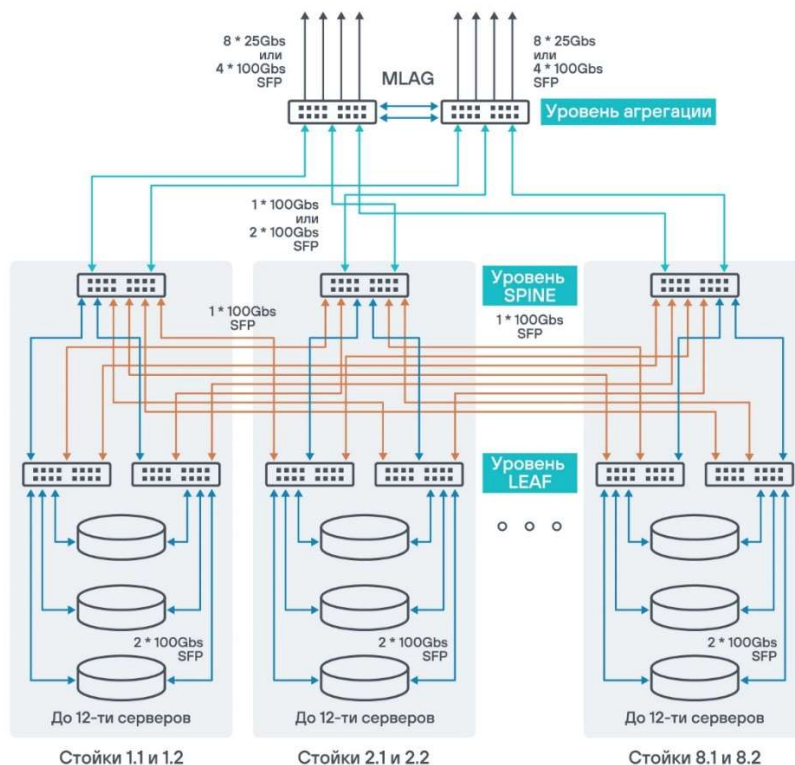


Рис. 8. Общая схема сети интерконнекта

**Сети мониторинга и управления Машины:**

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

На рисунке Рис. 9 представлена архитектура сети управления.

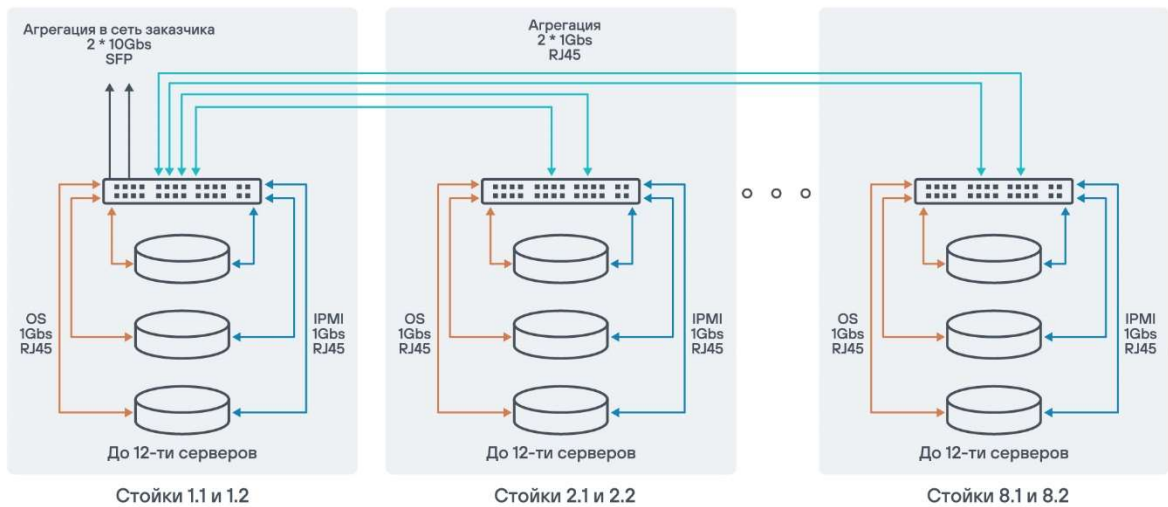


Рис. 9. Общая схема сетей мониторинга и управления

**Сетевой узел управления**

Сетевой узел управления состоит из одного коммутатора для организации сетей мониторинга, управления и служебного обмена (Рис. 10).

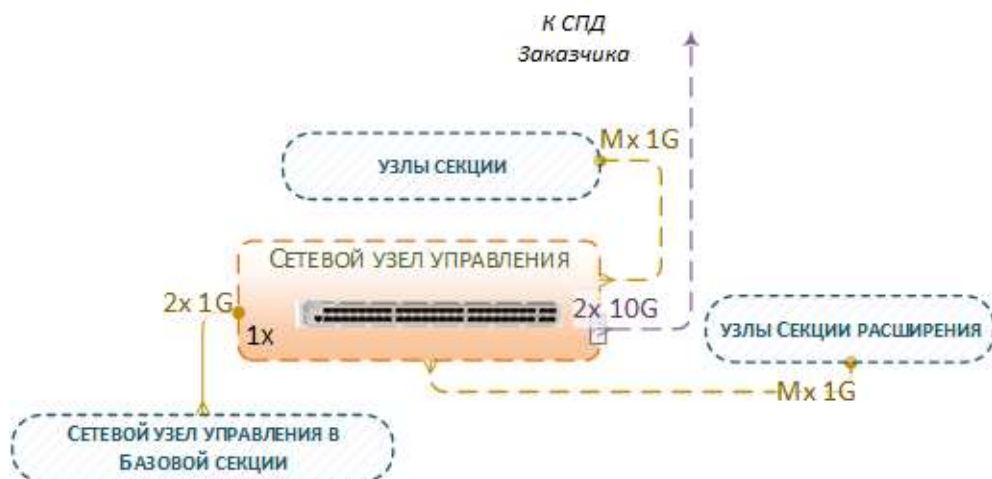


Рис. 10. Сетевой узел управления

**Модуль агрегации**

Модуль агрегации — пара коммутаторов, собранных по технологии MLAG для агрегации коммутаторов интерконнекта.

Модуль агрегации устанавливается в Базовый Модуль в тех случаях, когда уже нет доступных портов в базовом комплекте узлов сетевого взаимодействия.

Модуль состоит из узлов сетевого взаимодействия (Рис. 11).



Рис. 11. Модуль агрегации

## Блок управления и распределения

Основное предназначение блока — управление синхронизацией реплик БД и поддержание отказоустойчивого кластера средствами Zookeeper. Он состоит из трехузлового кластера показанного на Рис. 12.



Рис. 12. Модуль управления и распределения

### Каждый узел управления и распределения:

- 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 10 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 128 GB RAM

## Блок мониторинга и регистрации

Блок мониторинга и регистрации состоит из двух или трех узлов вычисления и хранения — высокопроизводительных серверов одного из двух типов на выбор. Он обеспечивает управление Машиной на всех этапах жизненного цикла программных и аппаратных компонентов.

### Основные функции блока:

- мониторинг и визуализация работы сети и оборудования, входящего в состав Машины
- мониторинг и визуализация функционирования платформы ADQM, связи реплик платформы и компонентов физической инфраструктуры (Рис. 13)
- 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 ADQM для автоматизированной установки

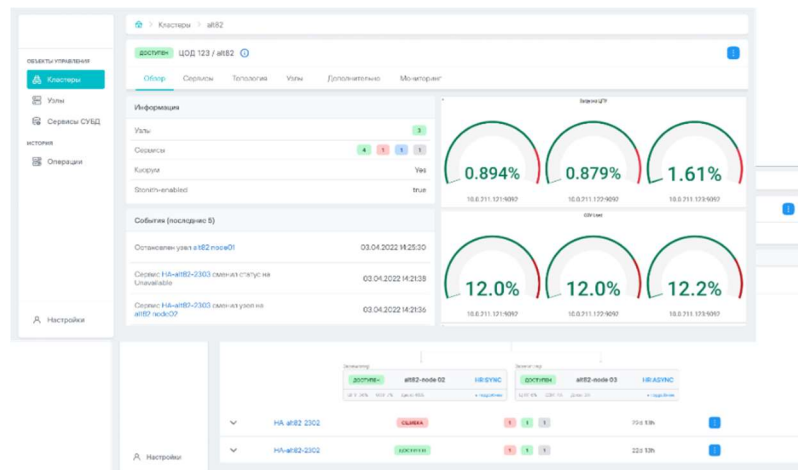


Рис. 13. Пример экрана подсистемы мониторинга МБД.КХ

Модуль мониторинга и регистрации состоит из двух высокопроизводительных узлов, объединенных в зеркальный кластер (Рис. 14).





Рис. 14. Модуль мониторинга и регистрации

**Каждый узел мониторинга и регистрации представлен специализированным серверным узлом:**

- 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 256 GB RAM

**Применяемое программное обеспечение:**

- ОС: Альт Линукс, сервер с виртуализацией Базис vCore
- мониторинг и управление: ПО Скала^р «Визион»
- управление жизненным циклом: ПО Скала^р «Геном»

## Блок резервного копирования

Основное предназначение блока — запись хранение резервных копий базы данных, а также их восстановление. Блок собирается из Модулей резервного копирования. Каждый такой модуль состоит из одного узла резервного копирования (Рис. 15).



Рис. 15. Модуль резервного копирования

### Каждый отдельный узел резервного копирования:

- содержит выделенные накопители SAS SSD для загрузки ОС — обеспечение отказоустойчивости
- применяется аппаратный RAID
- оснащен 14 HDD для хранения данных
- 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

### Применяемое программное обеспечение

Для создания резервных копии используются встроенные средства резервного копирования.

Резервное копирование рассчитывается для объема от 4-х копий максимальной базы.

## Базовый Модуль

Базовый Модуль — основа любого решения. Может содержать все виды модулей и узлов Машины (Рис. 16) и способен доукомплектовываться отдельно стоящими секциями (стойками) расширения. Состоит из узлов сетевого взаимодействия, мониторинга и регистрации и узла сетевого управления.

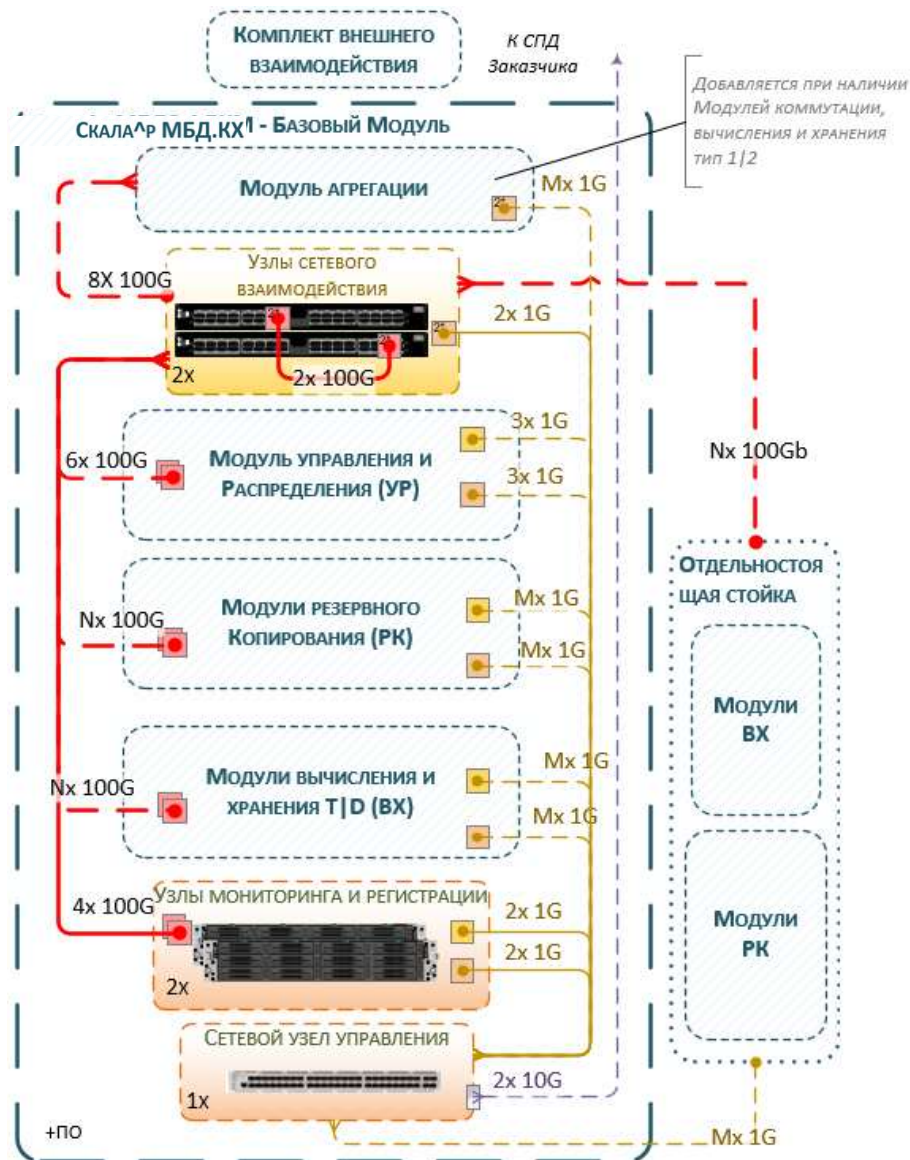


Рис. 16. Базовый Модуль

## Модуль коммутации, вычисления и хранения

Модуль комплектуется в отдельную стойку и применяется для горизонтального масштабирования Машины. Содержит узлы сетевого взаимодействия и управления, наполняется следующими видами модулей (Рис. 17):

- Модули вычисления и хранения
- Модули резервного копирования

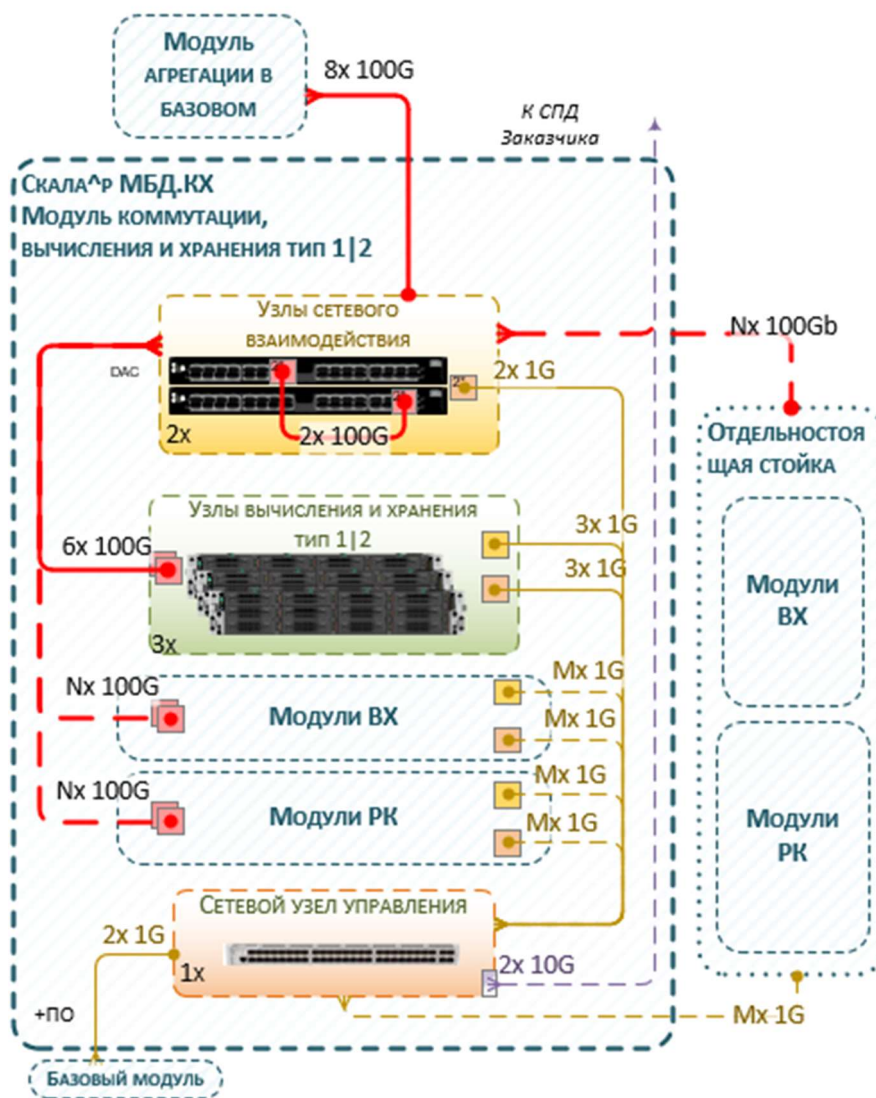


Рис. 17. Модуль коммутации, вычисления и хранения

### Секция расширения (отдельно стоящая стойка)

Секция расширения (Рис. 18) предназначена для размещения дополнительных модулей в дополнение к Базовому Модулю и/или Модулю коммутации, вычисления и хранения и содержит следующие виды модулей:

- Модули вычисления и хранения
- Модули резервного копирования

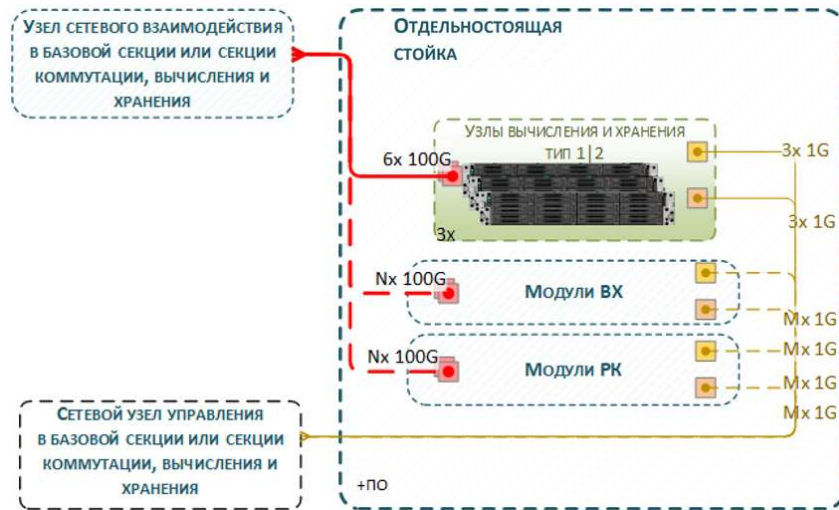


Рис. 18. Секция расширения

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 любому основному модулю можно добавить одну или две дополнительных секции (стойки).

### Схема размещения блоков и их модулей в стойках по этапам роста масштаба Машины

На диаграмме ниже (Рис. 19) представлен пример размещения базовых блоков и вариантов расширения при двух этапах горизонтального масштабирования.

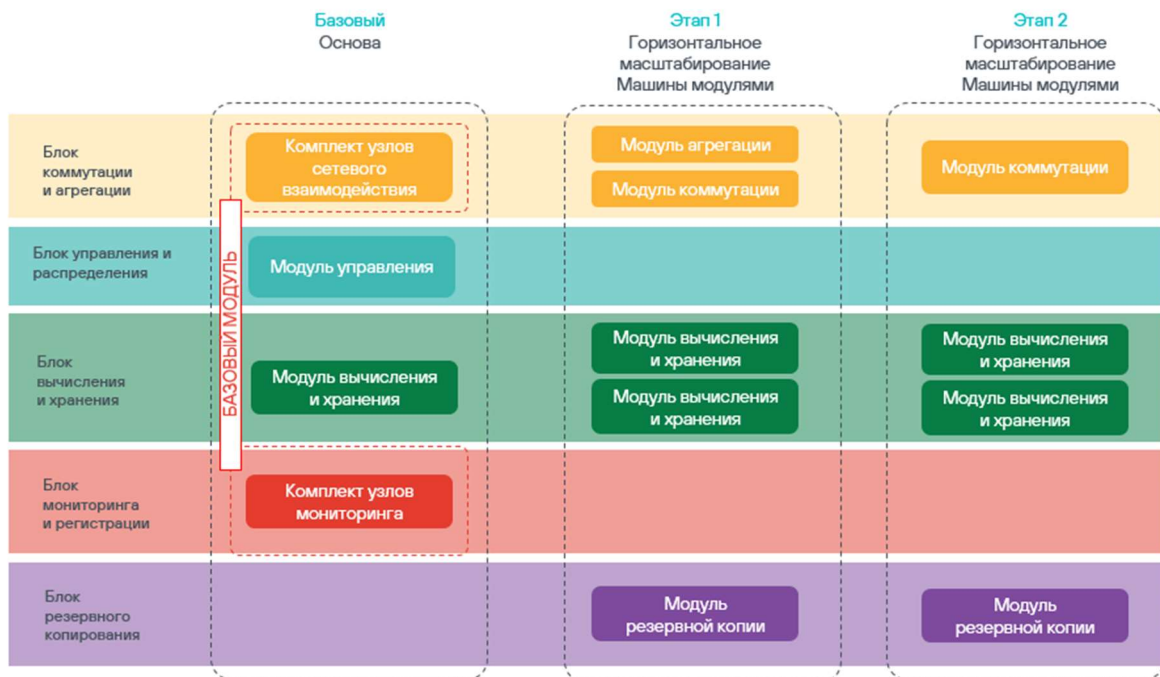


Рис. 19. Пример размещения базовых блоков и вариантов расширения



Этапы масштабирования приведены ниже:

- Первый этап — это установка модулей в Базовый Модуль. Как правило, это возможно, пока позволяет предельная мощность энергопотребления первой стойки Базового Модуля
- Далее устанавливается дополнительная отдельно стоящая стойка, в которой размещаются Модули вычисления и хранения, а также резервного копирования (при необходимости). Такое расширение возможно в одной или двух таких дополнительных стойках, и до достижения ограничения по питанию стоек или до заполнения количества портов в узлах внутреннего взаимодействия первой стойки — Базового Модуля. Важно, что в этих дополнительных стойках нет сетевого оборудования
- Установка Модуля коммутации, вычисления и хранения выполняется при заполнении портов внутреннего взаимодействия в Базовом Модуле. Стойка этого модуля включает в себя набор узлов сетевого взаимодействия и требует установки дополнительного модуля агрегации в Базовый Модуль (первая стойка на схеме)

Пример такого размещения в стойках представлен на диаграмме ниже (Рис. 20).

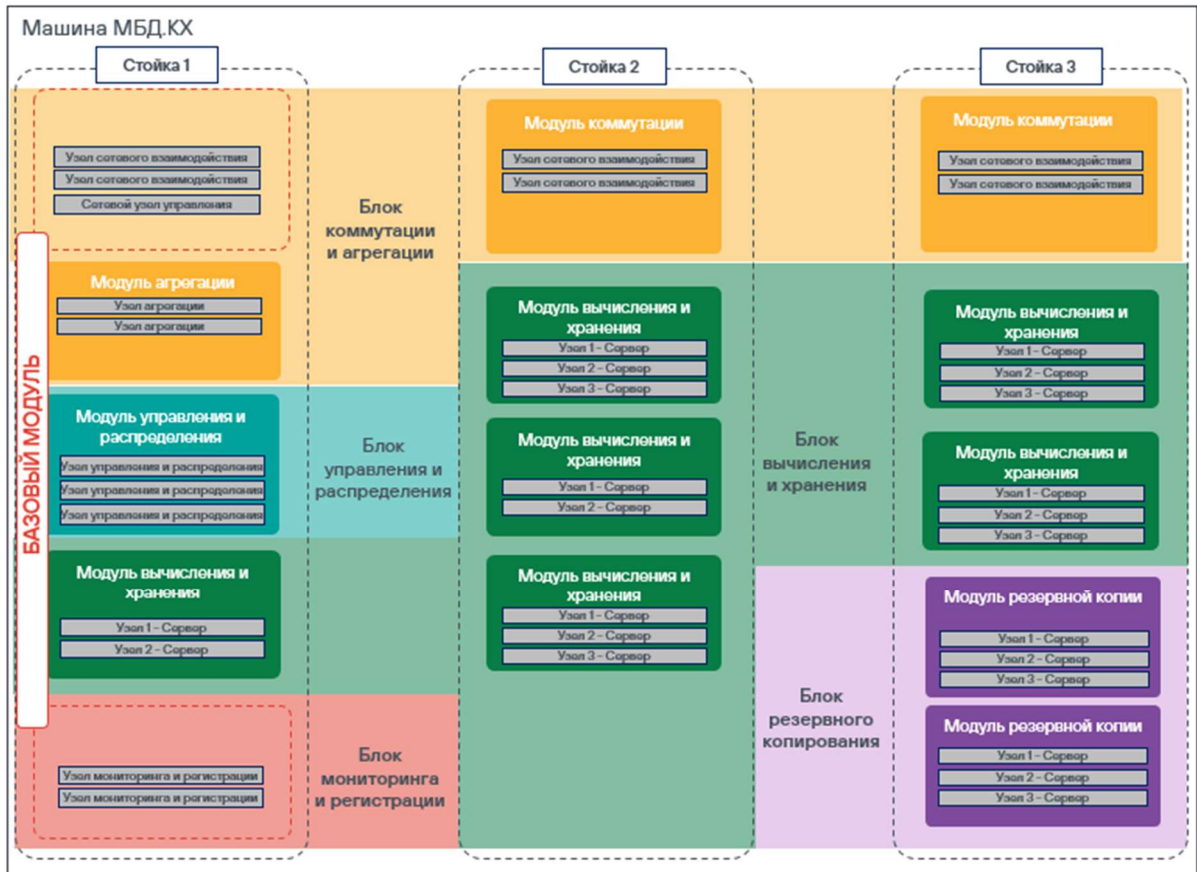


Рис. 20. Пример размещения базовых блоков, модулей и узлов в серверных стойках



## Система управления жизненным циклом Скала^р Геном

В **Маchine больших данных Скала^р МБД.КХ** применяется специализированное программное обеспечение Скала^р.Геном (Рис. 21).

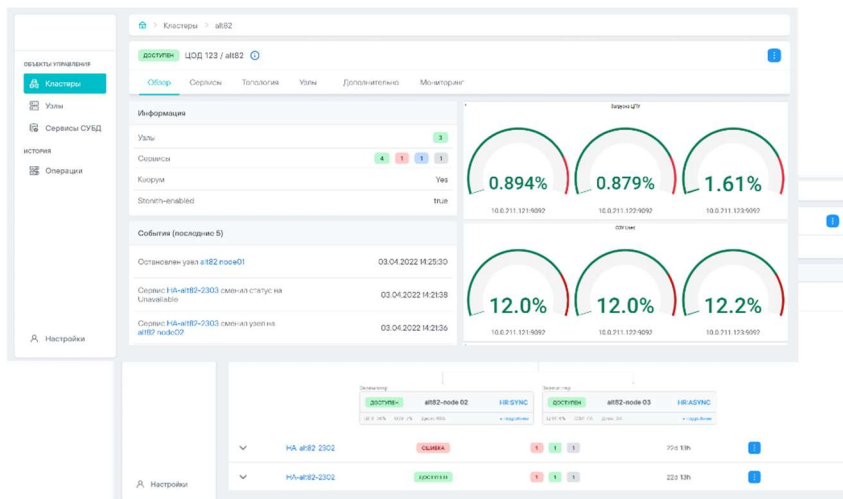


Рис. 21. Пример интерфейса Скала^р Геном

Данный программный продукт обеспечивает:

- контроль развертывания компонентов Машины
- ведение электронного паспорта Машины
- отслеживание состояния узлов
- отслеживание конфигурации программно-аппаратного состава Машины
- снижение влияния человеческого фактора — сокращение рисков, связанных с ошибками эксплуатирующего персонала

## 9. СПЕЦИФИЧНЫЕ ЧЕРТЫ МАШИНЫ СКАЛА^Р МБД.КХ

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

### *Приоритет обеспечения сохранности данных перед повышенной доступностью*

Такой подход обеспечивает гарантию сохранности данных при любых единичных отказах оборудования и быстрое восстановление из резервных копий в случае фатальных сбоев. В ряде сценариев возможны одновременные отказы разных компонентов без потери данных.

### *Реализация вычислительного блока на аппаратном решении вместо использования виртуальной среды*

Это позволяет достичь максимума производительности на данном оборудовании (нет потерь на работу среды виртуализации, прочие сведены к минимуму) и повысить надёжность (нет дополнительного программного слоя).

### *Использование локальных дисков вместо сетевой системы хранения для снижения затрат на передачу данных*

Так как СУБД ADQM изначально создана для работы с накопителями прямого подключения, ей не требуются сети для доступа к данным на внешних массивах, что обеспечивает линейную масштабируемость системы и отсутствие порога по суммарной пропускной способности подсистемы ввода-вывода. Отсутствие дополнительного сложного элемента в виде сетей хранения данных также способствует повышению надёжности решения, а снижение стоимости решения обеспечивается благодаря отсутствию расходов на внешнюю систему хранения в целом, затраты идут только на SSD.

### *Применение стандартного высоконадёжного и производительного оборудования в качестве платформы для размещения компонентов решения взамен уникальных аппаратных разработок*

Обеспечение стабильного уровня производительности достигается за счет использования компонентов, проверенных временем и лабораторией Скала^р. Отсутствие уникальных элементов также повышает надёжность решения.

*Использование программных RAID отечественных производителей вместо отдельных аппаратных RAID*

Это позволяет достичь более высокой производительности, высокой гибкости в настройках (в зависимости от требований), снижение зависимости от производителей оборудования и гарантирует оптимальность алгоритмов RAID.

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

Это позволяет сохранить ранее сделанные инвестиции в системы управления ИТ-инфраструктурой и дает возможность построения сквозных систем управления, в которые интегрируются Машины линейки Скала^р МБД.

## 10. ГАРАНТИРОВАННОЕ КАЧЕСТВО

Качественные показатели **Машины больших данных Скала^р МБД.КХ** обеспечиваются её соответствием проверенному стандартному варианту, формированием из блоков и модулей, производством работ высококвалифицированными специалистами и процедурами приемки качества.

### Производство (сборка Машины и предустановка ПО)

При производстве используются только высококачественные комплектующие, а сборка продукции осуществляется строго в соответствии с технологическими картами. Первичное развертывание ПО осуществляется в автоматическом режиме, дополнительные настройки ПО осуществляются в соответствии с утверждённой пошаговой инструкцией. Сформированная Машина тщательно тестируется, таким образом отклонения от типового решения **Скала^р МБД.КХ** исключены.

### Передача в эксплуатацию

**Машина больших данных Скала^р МБД.КХ** полностью сформирована, протестирована, готова к размещению в сети заказчика и установке прикладного ПО заказчика. В комплекте с Машиной передаётся Паспорт изделия, эксплуатационная документация, Сертификат на техническую поддержку.

## 11. РЕАКЦИЯ НА ВОЗМОЖНЫЕ ОТКАЗЫ

### Отказы, связанные со стандартными элементами Скала^р МБД.КХ

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

- узлов (дублирование процессоров, источников питания и др.)
- подсистемы ввода-вывода (программный RAID)
- сети интерконнекта (дублирование сетевых интерфейсов и самих коммутаторов)
- системы резервного копирования

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 небольшое снижение производительности. После устранения неисправности исходная производительность **Скала^р МБД.КХ** также восстанавливается.

### Отказы, связанные с узлами кластера баз данных

#### Аппаратные сбои

Архитектура программного обеспечения, лежащего в основе **Скала^р МБД.КХ**, позволяет построить отказоустойчивый многоузловой кластер, в том числе катастрофоустойчивый кластер.

Отказоустойчивость обеспечивается за счет настройки репликации таблиц между узлами, а также избыточности экземпляров сервиса ZooKeeper, который отвечает за хранение метаданных, необходимых для репликации. При этом отказоустойчивым считается кластер, состоящий как минимум из трех узлов ZooKeeper и двух узлов сегментов ADQM.

Отказ любых двух узлов с разными ролями, либо отказ всех узлов с ролью ADQM, кроме одного, не влияет на работоспособность кластера и позволяет ему обрабатывать запросы как на чтение, так и на запись. При этом общая производительность системы снижается.

В случае, если ZooKeeper недоступен, реплицируемые таблицы остаются доступными только для чтения, а запросы на запись приведут к выдаче исключения. При восстановлении доступности сервиса данные будут автоматически синхронизированы, если это возможно, а в противном случае неизвестные участки данных будут перемещены в подкаталог detached. Если блоки данных внутри файлов оказались повреждены, запрос SELECT приведет к исключению, после чего будет предпринята попытка их проверки и восстановления.

Если после сбоя локальный набор данных критически отличается от ожидаемого, сработает защитный механизм, и потребуются ручное (полуавтоматическое) восстановление.

Если данные на одной из реплик полностью утеряны, доступен вариант ручного восстановления.

#### Программные сбои и человеческий фактор

Кроме обеспечения отказоустойчивости, необходимо организовать систему резервного копирования, исходя из ресурсов и потребностей.

Одним из вариантов решения данной задачи является дублирование данных в Hadoop или S3, например, путем подписки дополнительных подписчиков на темы Kafka.

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.КХ** могут быть предоставлены по запросу.

## 12. ВАРИАТИВНОСТЬ РЕШЕНИЯ

### Приоритет производительности

Область применения:

- дополнение к аналитической части систем класса Data Warehouse
- хранение информации оперативного доступа
- системы с множественными аналитическими запросами

Варианты решения:

- увеличенный объём оперативной памяти
- повышение базовой частоты работы процессоров
- высокопроизводительные SSD
- RAID 10

### Приоритет объёма хранения

Область применения:

- база аналитической информации предприятия
- данные с датчиков, устройств интернета вещей
- журналы действий пользователей
- данные исторического анализа

Вариант решения:

- стандартные параметры вычислительного модуля
- SSD повышенного объёма

### Специальный тюнинг

Вариант решения:

- может использоваться в комплексе с любым из вариантов
- требуется участие разработчиков прикладных систем
- достигается адаптацией настроек и конфигурации оборудования под структуру данных заказчика, типы и периодичность запросов и т.п.



### 13. ТРЕБОВАНИЯ К РАЗМЕЩЕНИЮ РЕШЕНИЯ

Решение Машины представляет собой серверный монтажный шкаф 19", высота 42U, с дальнейшей возможностью модульной расширяемости до 14 стоек.

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

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

Расчётная потребляемая мощность шкафа составляет от 6 до 11 кВт.

В месте установки должны быть предусмотрены соответствующие мощности по отводу тепла.

Для подключения к локальной сети заказчика необходим резервированный канал до 4x100 Gigabit Ethernet или до 8x10/25 Gigabit Ethernet. Требуемые трансиверы определяются на этапе формирования спецификации Машины.

При развёртывании решения на нём будут выполнены 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 настройки прикладных программных систем.

## 14. ТЕХНИЧЕСКАЯ ПОДДЕРЖКА

**Машина больших данных Скала^р МБД.КХ** поставляется с обязательной 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 предоставляется непосредственно производителем или сертифицированным партнёром. 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 больших данных Скала^р МБД.КХ**.

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

- Пакет услуг по технической поддержке на первый год включен в поставку
- Заказчик может выбирать пакет в базовом режиме 9×5, или в расширенном режиме 24×7 (опция для критической функциональности)
- Срок начально приобретаемой технической поддержки может быть увеличен до 3 и 5 лет, также доступна пролонгация поддержки
- Возможно включение в состав стандартных пакетов дополнительных опций и услуг

Состав типовых пакетов услуг по технической поддержке представлен в таблице ниже.

*Пакеты услуг по технической поддержке Скала^р МБД.КХ*

Услуга	Пакет «9×5»	Пакет «24×7»
Режим «Обслуживание комплекса Скала^р МБД.КХ в режиме 9х5» (в рабочее время по рабочим дням)	+	+
Режим «Обслуживание комплекса Скала^р МБД.КХ в режиме 24×7» (круглосуточно)	—	+
Предоставление доступа к системе регистрации запросов/инцидентов Service Desk	+	+
Предоставление доступа к базе знаний по продуктам Скала^р	+	+
Предоставление обновлений лицензионного ПО Скала^р	+	+
Диагностика, анализ и устранение проблем в работе комплекса Скала^р МБД.КХ, включая: <ul style="list-style-type: none"> <li>■ устранение аппаратных неисправностей;</li> </ul>	+	+

Услуга	Пакет «9×5»	Пакет «24×7»
<ul style="list-style-type: none"> <li>■ техническое сопровождение ПО.</li> </ul>		
Консультации по работе комплекса Скала^р МБД.КХ	+	+
«Защита конфиденциальной информации» (неисправные носители информации не возвращаются Заказчиком)	Опция	Опция
Замена и ремонт оборудования по месту установки	+	+
Доставка оборудования на замену за счет производителя	+	+
Расширенные опции обслуживания	—	+
Времена реагирования и отклика, не более:		
Время регистрации обращений	30 минут, рабочие часы (9×5)	30 минут, круглосуточно (24×7)
Подключение специалиста к решению инцидентов критичного и высокого уровней	В течение 1 рабочего часа (9×5)	В течение 1 часа (24×7)

Примечание к срокам ремонта оборудования: комплекс **Скала^р МБД.КХ** архитектурно является устойчивым к выходу из строя отдельных компонентов и даже узлов, поэтому нет необходимости в обеспечении дорогостоящего сервиса срочного восстановления оборудования в течение суток и менее. В комплексе предусмотрено, как минимум, двойное резервирование основных компонентов, позволяющее сохранять данные и работоспособность даже при выходе из строя нескольких дисков и/или серверов.

## 15. ЛИЦЕНЗИРОВАНИЕ ПО МАШИНЫ БОЛЬШИХ ДАННЫХ

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

### Лицензирование ПО Машины больших данных Скала^р МБД.КХ

Программное обеспечение СУБД Arenadata QuickMarts (ADQM) лицензируется согласно объёму ресурсов в Модуле вычисления и хранения, при этом на каждый модуль выдается единая лицензия.

Программное обеспечение **Скала^р Визион**, **Скала^р Геном** поставляется исключительно в составе **Машины больших данных Скала^р МБД.КХ**, и лицензируется по количеству модулей в ней.

### Варианты лицензирования

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

- фиксированная — приобретается бессрочная (постоянная) лицензия
- временная — приобретается лицензия на период времени

### Политика обновления ПО

Обновления функционального и системного ПО предоставляются по минорным и мажорным версиям, выпускаемым партнером АренаДата и иными, в течение действия технической поддержки бесплатно.

Также команда Скала^р активно занимается развитием собственных программных продуктов и утилит для **Машины больших данных Скала^р МБД.КХ** и предоставляет обновления по мере их появления.

### 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](http://www.skala-r.ru).