скала^р

# Машина хранения данных Скала^р МХД.О

## Объектное хранилище S3

Технический обзор

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

# Оглавление

# 1. ВВЕДЕНИЕ

Машина хранения данных Скала^р МХД.О предназначена для создания горизонтально масштабируемого объектного хранилища, совместимого с Amazon S3. В зависимости от области применения и потребностей заказчика, Машины хранения данных могут поставляться в разных исполнениях.

S3 хранилище – это сервис хранения файлов с данными в форме объектов. От обычного хранения файлов хранение объектов в S3 хранилище отличается форматом хранения, наличием метаданных и уникальных идентификаторов объектов, которые дают возможность организации хранения миллионов файлов, что не доступно для большинства файловых хранилищ.

Основными преимуществами S3 хранилища являются:

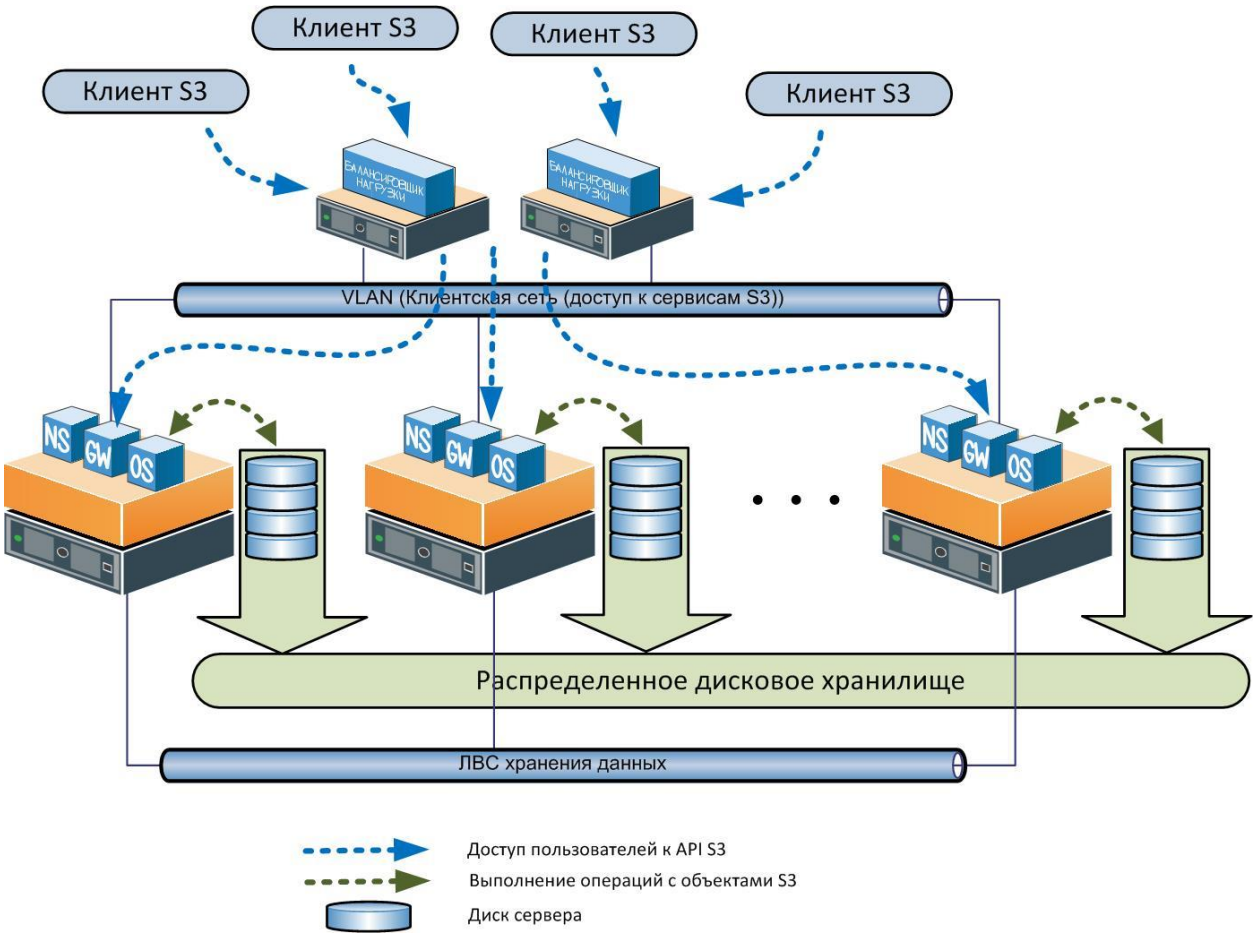
- Масштабируемость – возможность создания хранилищ неограниченных размеров;
- Хранение неограниченного количества объектов – одно из основных преимуществ, достигаемое благодаря тому, что адреса к объектам хранятся в виде ссылок, а не по именам;
- Сохранность данных за счет отсутствия единой точки отказа, исключающей потерю данных в результате единичных сбоев;
- Обеспечение катастрофоустойчивости за счет использования синхронной или асинхронной репликации;
- Обеспечение целостности данных на уровне хранилищ даже при использовании асинхронной репликации с использованием медленных каналов связи.

В настоящий момент Скала^р МХД.О является единственным продуктом российского производства, обеспечивающим реализацию объектного хранилища практически любого размера и совместимого с Amazon S3.

Начиная с 2014 года решения Скала^р планомерно развивались, превращаясь из решений, предназначенных для интернет-провайдеров, в полноценные системы корпоративного класса. На сегодняшний день решения Скала^р являются безусловными лидерами на российском рынке корпоративных средств виртуализации и хранения данных за счет своих функциональных возможностей и довольно простых процессов внедрения и эксплуатации. Производитель и партнеры на местах оказывают полную поддержку комплекса в части эксплуатации и развития.

# 2. АРХИТЕКТУРА СКАЛА^Р МХД.О

Скала^р МХД.О реализуется на серверах с архитектурой x86-64 с установленными накопителями, предназначенными для хранения данных. Архитектура Машины хранения данных представлена на рисунке ниже.




Скала^р МХД.О включает следующие основные подсистемы:

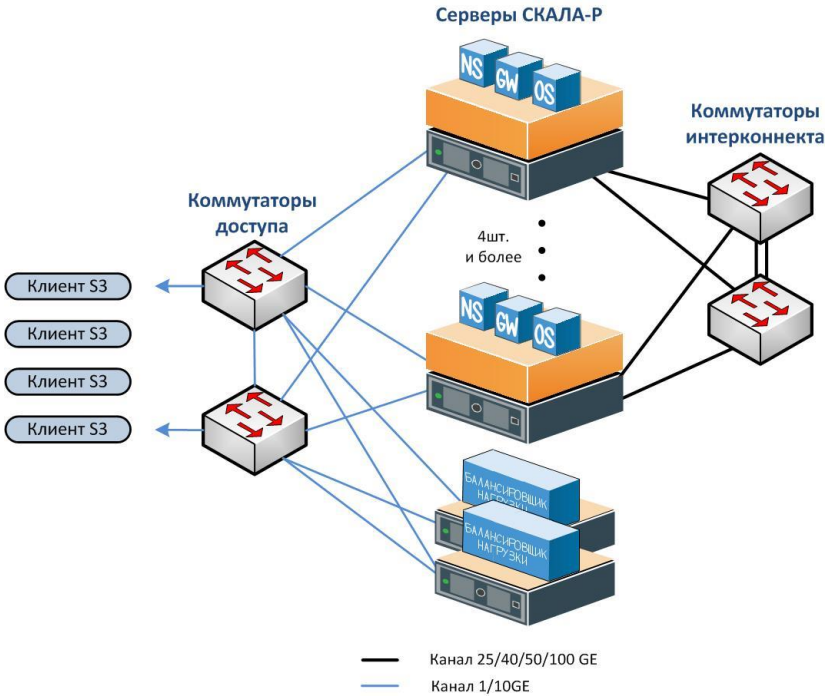
- Балансировщик нагрузки обеспечивает балансировку запросов между шлюзами S3, размещенными на разных физических серверах, что обеспечивает равномерную нагрузку на комплекс, а также исключает перенаправление запросов на неисправные серверы при авариях;
- Шлюз S3 (GW) является компонентом, реализующим S3 API для пользователей объектного хранилища. GW получает и обрабатывает запросы, сформированные с использованием протокола Amazon S3, выполняет аутентификацию пользователей S3 и проверку списков контроля доступа (ACL);
- Сервер объектов (OS) обеспечивает хранение данных объектов. Хранение объектов осуществляется на распределенном дисковом хранилище;

- Серверы имен (NS) обеспечивают хранение метаданных объектов, включающих имя объекта, его размер, список контроля доступа (ACL), расположение, владельца и др.;
- Распределенное дисковое хранилище представляет собой кластер с функционалом обеспечения высокой доступности служб и данных. Функционал высокой доступности задействуется и для обеспечения гарантированной доступности служб S3 (NS и OS) за счет их переноса на работающие серверы в случае сбоя сервера, на котором они работали.

Программные компоненты комплекса Скала^р МХД.О запускаются на серверах как службы, поэтому для работы сервиса S3 не нужны виртуальные среды, что упрощает эксплуатацию комплекса. Балансировщик нагрузки реализуется на двух выделенных серверах и работает в режиме отказоустойчивого кластера.

# 3. СТРУКТУРА СКАЛА^Р МХД.О

Структурная схема Скала^р МХД.О представлена на рисунке ниже.




В состав комплекса Скала^р МХД.О входит два типа серверов:

1) Серверы хранения – серверы, на которых функционируют сервисы хранилища S3 и осуществляется хранение данных;
2) Серверы балансировки нагрузки – обеспечивают работу балансировщика нагрузки в режиме отказоустойчивого кластера.

Серверы хранения подключены к сети хранения данных с использованием коммутаторов интерконнекта. В зависимости от требований к нагрузочной способности конкретного комплекса Скала^р МХД.О и требований заказчика могут использоваться коммутаторы с интерфейсами 25/40/50/100 Gigabit Ethernet. Коммутаторы интерконнекта работают в режиме взаимного дублирования с использованием технологии MLAG (Multi-Chassis Link Aggregation).

Серверы хранения и серверы балансировки нагрузки подключаются к коммутаторам доступа, которые обеспечивают доступ потребителей к сервису S3. Данные коммутаторы могут поставляться в составе комплекса Скала^р МХД.О или используется существующая емкость портов сети заказчика.

Количество серверов хранения и их конфигурация определяются требованиями к производительности S3 хранилища и объему хранимых объектов.

Хранение данных в комплексе Скала^р S3 реализуется с помощью установленных в серверы накопителей и программного обеспечения ПО Скала^р Управление — Программно-определяемое хранилище S3. ПО Скала^р Управление — Программно-определяемое хранилище S3 устанавливается на серверах и объединяет их дисковое

пространство в распределенный отказоустойчивый и масштабируемый программно-определяемый дисковый массив.

Распределенный дисковый массив, реализованный на базе ПО Скала^р Управление — Программно-определяемое хранилище S3, имеет следующие основные свойства:

- Высокая отказоустойчивость;
- Возможность работы в двух режимах: хранение 2 или более реплик на накопителях разных хостов комплекса Скала^р S3 или хранение блоков четности/избыточности (Erasure Code);
- Возможность комплектования любыми типами накопителей для получения характеристик, наиболее полно отвечающих требованиям;
- Емкость хранения до 8 Пбайт;
- Гибкие возможности по модернизации и обслуживанию.

## 3.1 Отказоустойчивость

Отказоустойчивость комплекса Скала^р МХД.О достигается за счет использования комплекса мер:

1) Применения отказоустойчивого распределенного дискового хранилища, не имеющего единой точки отказа;
2) Технологии высокой доступности для служб сервиса имен (NS) и сервиса хранения (OS). Стоит отметить, что сервис шлюза (GW) не хранит собственных данных и перезапускается автоматически при сбое платформы;
3) Наличия серверов балансировки нагрузки, организованных в кластер высокой доступности.

Отказоустойчивость распределенного дискового хранилища обеспечивается за счет хранения нескольких копий данных на разных физических серверах. Это позволяет сохранить данные при выходе из строя до двух серверов хранения комплекса Скала^р МХД.О.

Технология высокой доступности для служб сервисов S3 обеспечивает автоматический перезапуск служб, которые работали на сервере, вышедшем из строя, на рабочих серверах комплекса. Это позволяет даже при отказе нескольких сервера не снизить нагрузочной способности комплекса, например, при работе в режиме метрокластера возможна потеря половины всех серверов комплекса Скала^р МХД.О.

Серверы балансировки нагрузки организованы в кластер высокой доступности, который при отказе одного из серверов передает его функции на рабочий сервер. При работе серверы балансировки нагрузки отслеживают состояние серверов хранения, что исключает возможность перенаправления запроса пользователя на неработоспособный ресурс и отказ в предоставлении доступа к сервису.

## 3.2 Модернизация и обслуживание

Комплекс Скала^р МХД.О легко обслуживать и модернизировать. Основные возможности:

- Замена вышедшего из строя накопителя без остановки сервера и компонентов сервиса S3;
- Удаление нормально функционирующего накопителя (например, для замены на другой) без остановки сервера и компонентов сервиса S3;
- Установка дополнительного накопителя без остановки сервера и компонентов сервиса S3;
- Замена вышедшего из строя сервера хранения или сервера балансировщика нагрузки;
- Добавление дополнительного сервера хранения (для увеличения емкости хранения).

# 4. АППАРАТНАЯ ПЛАТФОРМА СКАЛА^Р МХД.О

Скала^р МХД.О реализуется на двухпроцессорных серверных платформах архитектуры x86-64, которые, в зависимости от требований к производительности и емкости хранилища, комплектуются дисками разных типов, включая ресурсные пулы (tiers) на SSD – SAS, SATA и HDD дисках (кроме варианта с метрокластером).

Не требуется установка RAID-контроллеров, платформа самостоятельно организует распределенное хранилище на имеющихся типах дисков.

Для ускорения работы HDD рекомендуется применять опцию кэширования за счет установки служебных SSD дисков небольшого объёма, типично 1 SSD кэш- накопитель на 5 дисков HDD. Сами кэш-диски не входят в лицензируемый объём полезного пространства (см. раздел 8).

# 5. СЦЕНАРИИ ИСПОЛЬЗОВАНИЯ СКАЛА^Р МХД.О

Благодаря своей универсальности, S3 хранилище на базе Скала^р МХД.О может использоваться для хранения файлов (объектов) самого разного типа, включая:

- Объекты, которые ранее хранились в СУБД;
- Статичные данные веб-сайтов;
- Разнообразные документы и их образы;
- Фотографии и видеозаписи;
- Большие данные (big data);
- Архивы и резервные копии.

То есть S3 хранилище на базе Скала^р МХД.О может использоваться практически в любом технологическом процессе, приложении или сервисе, где требуется хранение большого количества, файлов.

В связи с необходимостью перехода на технологии, свободные от санкционных рисков, S3 хранилища дают возможность повышения производительности реализуемых решений управления данными за счет выноса функции хранения больших объектов во внешнюю систему. Например, при использовании СУБД PostgreSQL хранение большого числа объектов в самой базе увеличивает ее размер и существенно снижает производительность. Хранение объектов за пределами СУБД PostgreSQL в S3 хранилище существенно уменьшает объем базы и повышает ее производительность, при этом доступ к объектам осуществляется по статичным ссылкам, хранимым в базе.

Существует три типовые схемы использования комплекса Скала^р МХД.О:

- Локальное хранилище S3;
- Распределенное хранилище S3 с асинхронной гео-репликацией;
- Метрокластер S3.

## 5.1 Локальное хранилище S3

Локальное хранилище S3 является стандартным вариантом использования комплекса. К основным достоинствам относятся:

- Возможность создания хранилищ объемом до 8 Пбайт;
- Возможность реализации высокопроизводительного комплекса со скоростью чтения/записи в десятки Гбайт/с;
- Возможность реализации архива максимальной емкости (8 Пбайт) всего в 3-х стандартных стойках размером 42U;
- Высокая устойчивость к единичным отказам оборудования, простота восстановления после сбоя.

## 5.2 Распределенное хранилище S3 с асинхронной гео-репликацией

Несколько комплексов Скала^р МХД.О могут быть объединены в единую систему, где есть головное и резервные хранилища S3. В обычном режиме работа ведется с головным хранилищем, и все изменения реплицируются в резервные хранилища.

К основным достоинствам такого решения относятся:

- Катастрофоустойчивость;
- Относительно невысокие требования к каналу связи между головным и резервными хранилищами.

В то же время есть и особенность, связанная с принципом «согласованности в конечном счете», реализуемом в объектных хранилищах. Согласованность в конечном счете означает, что если в головном хранилище изменить объект, то это изменение появится в ответах системы только после того, как оно будет реплицировано в резервное хранилище. До момента окончания репликации при запросе хранилищем будет выдаваться предыдущая версия объекта.

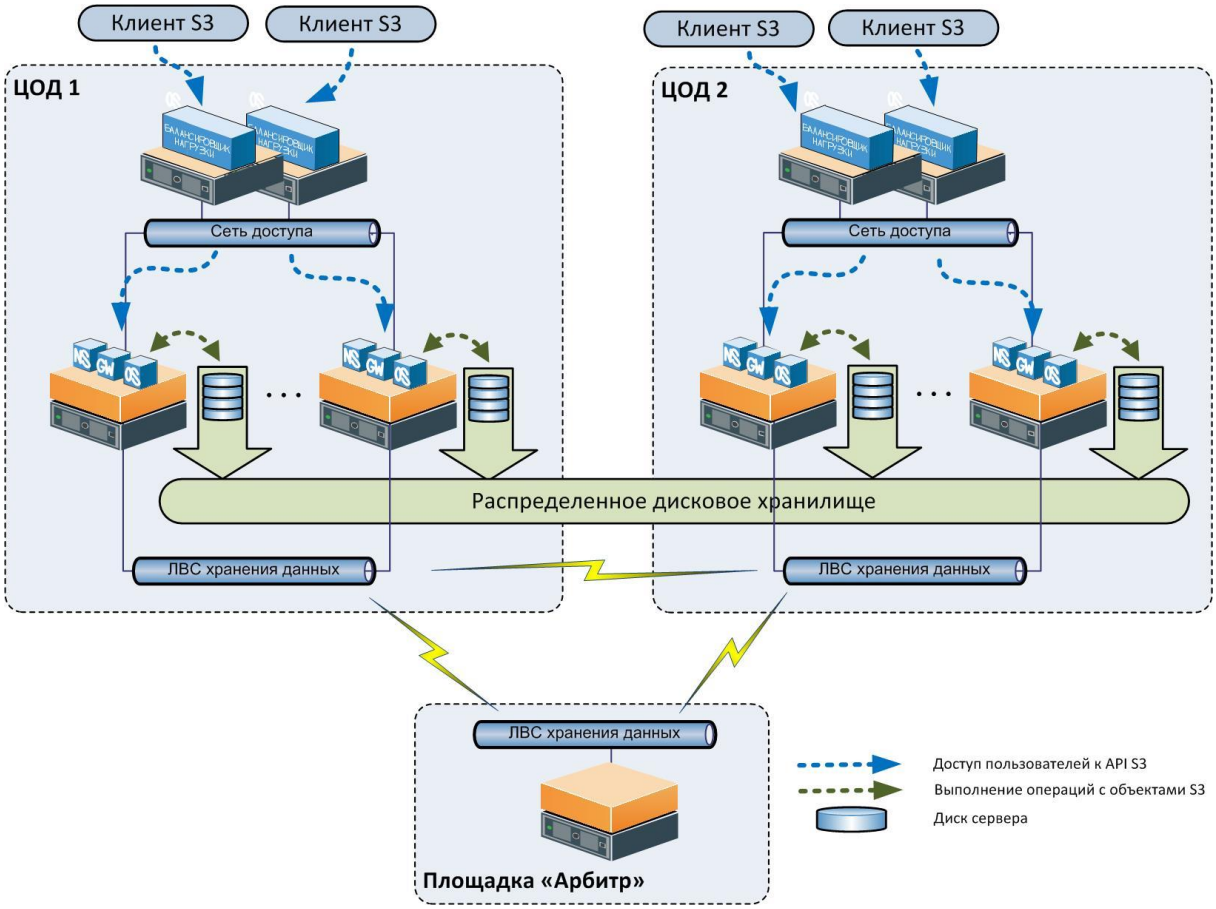
Указанный недостаток полностью исключен при реализации Скала^р МХД.О в режиме метрокластера (см. раздел 5.3).

## 5.3 Метрокластер S3

### 5.3.1 Архитектура

Метрокластер, как и обычный кластер, обеспечивает гарантированную доступность информационной системы при выходе из строя части оборудования. В отличие от обычного кластера, где все компоненты находятся в одном ЦОД, оборудование метрокластера разнесено на значительное расстояние, как правило, в пределах одного города. Это позволяет сохранить доступность информационных систем даже при полной утрате одного ЦОД. Задача метрокластера — сохранять доступность информационных систем автоматически, вне зависимости от внешних факторов и присутствия сотрудников службы эксплуатации на своих рабочих местах.

Метрокластер Скала^р МХД.О имеет классическую архитектуру, состоящую их двух центров обработки данных и площадки арбитра. На площадке арбитра размещается компонент, который автоматически выбирает площадку, где будет восстановлено функционирование всех информационных систем. Решение о выборе площадки принимается на основании данных о доступности оборудования, включая сетевое, и каналов связи.




Метрокластер Скала^р МХД.О имеет распределенное дисковое хранилище, обеспечивающее сохранность данных при выходе из строя одиночных экземпляров оборудования или полностью одного ЦОД.

### 5.3.2 Возможности

Распределенное дисковое хранилище реализуется на базе SSD-накопителей, использование HDD-накопителей не предусмотрено.

О возможностях системы говорят результаты тестирования, проведенного специалистами компании Скала^р. Краткая методика и результаты тестов описаны ниже.

Тестирование на производительность распределенного дискового массива проводилось со значениями:

- Тестовая конфигурация на SSD-накопителях Intel DC 3510;
- Канал между центрами обработки данных — 10 Гбит/с;
- Тип нагрузки: случайный смешанный доступ пакетами 8 КБ.

Тестирование было выполнено для разных уровней задержки между центрами обработки данных для имитации разного расстояния:

- 50 µsec ~ 15 км;
- 100 µsec ~ 30 км;
- 200 µsec ~ 60 км;
- 500 µsec ~ 150 км.

На графики также выведены показатели «3_2» — так обозначена производительность одиночного кластера (для сравнения).

Ниже приведены результаты тестов:

- скорость чтения (см. Рисунок 5.1);
- скорость записи (см. Рисунок 5.2).

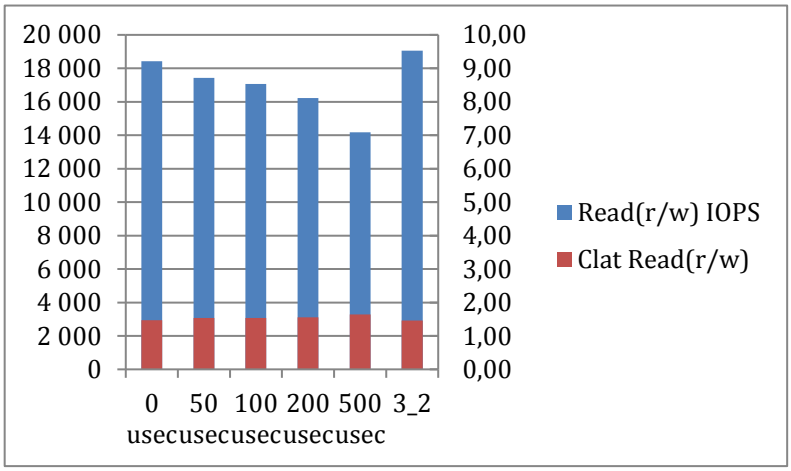



*Рисунок 5.1. Производительность дисковой подсистемы метрокластера на операциях чтения*

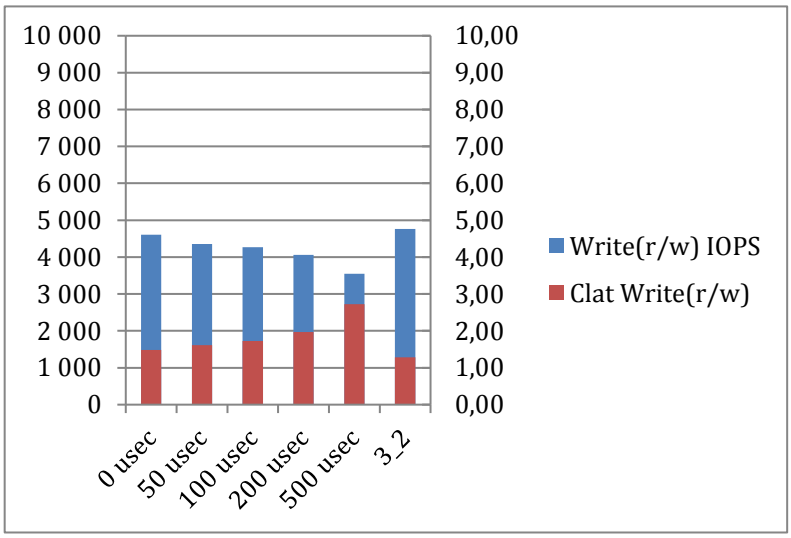



*Рисунок 5.2. Производительность дисковой подсистемы метрокластера на операциях записи*

# 6. ПЛАНИРОВАНИЕ ИНФРАСТРУКТУРЫ

Для интеграции Скала^р МХД.О в существующую ИТ-инфраструктуру необходимо выполнить следующие шаги:

- Выделить IP-адреса для сети распределенного дискового хранилища;
- Выделить IP-адреса для сети доступа комплекса Скала^р МХД.О;
- Выделить IP-адреса для сети управления;
- Предоставить доступ серверам Скала^р МХД.О к сервису времени NTP;
- Сформировать имя для сервиса S3;
- Приобрести сертификат или выдать самоподписанный сертификат для защиты трафика S3.

# 7. ТЕХНИЧЕСКАЯ ПОДДЕРЖКА

Поставка Скала^р МХД.О осуществляется с предварительным тестированием и настройкой оборудования согласно требованиям заказчика. Качественная поддержка Скала^р МХД.О обеспечивается едиными стандартами гарантийного и постгарантийного технического обслуживания:

- Пакет услуг по технической поддержке на первый год включен в поставку;
- Заказчик может выбирать пакет в базовом режиме 9x5, или в расширенном режиме 24x7 (опция для критической функциональности);
- Срок технической поддержки может быть увеличен до 5 лет на аппаратное обеспечение Скала^р МХД.О, и неограниченно - на входящее в состав комплекса программное обеспечение (при условии его обновления до актуальных версий);
- Возможно включение в состав стандартных пакетов дополнительных опций и услуг.

Разница между пакетами услуг по технической поддержке представлена ниже (Таблица 7.1).

*Таблица 7.1 Пакеты услуг по технической поддержке Скала^р МХД.О*

| Услуги | Пакет «9x5» | Пакет «24x7» |
|---|---|---|
| Режим «Обслуживание комплекса Скала^р МХД.О в режиме 9x5» (в рабочее время по рабочим дням) | + | — |
| Режим «Обслуживание комплекса Скала^р МХД.О в режиме 24x7» (круглосуточно) | — | + |
| Предоставление доступа к системе регистрации запросов/инцидентов Service Desk | + | + |
| Предоставление доступа к базе знаний по продуктам Скала^р | + | + |
| Предоставление обновлений лицензионного ПО Скала^р | + | + |
| Диагностика, анализ и устранение проблем в работе комплекса Скала^р МХД.О, включая:<br>• устранение неисправностей в аппаратной части;<br>• техническое сопровождение ПО. | + | + |
| Консультации по работе комплекса Скала^р МХД.О | + | + |

| Услуги | Пакет «9x5» | Пакет «24x7» |
|---|---|---|
| «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 (9x5) | 30 минут, кругло-суточно (24x7) |
| Подключение специалиста к решению инцидентов критичного и высокого уровней | 1 рабочего часа (9x5) | 1 часа (24x7) |

*Примечание к срокам ремонта оборудования:* комплекс Скала^р МХД.О 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.

# 8. ПОСТАВКА И ЛИЦЕНЗИРОВАНИЕ

Команда Скала^р активно занимается развитием программных продуктов Скала^р МХД.О. 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.

Информация об изменениях в версиях Скала^р МХД.О публикуется на сайте www.skala-r.ru в разделе «Новости».

## 8.1 Варианты лицензирования

Лицензирование ПО комплекса Скала^р МХД.О имеет две редакции:

- Корпоративная;
- Расширенная.

Сравнительная таблица с функциональными возможностями редакций продукта приведена в таблице ниже.

*Таблица 8.1 Варианты лицензирования Скала^р МХД.О*

| Возможности продукта | Корпоративная | Расширенная |
|---|---|---|
| Максимальное количество узлов в 1 комплексе | 16 | 128 |
| Режим высокой доступности | ● | ● |
| Модули сбора данных о системе | ● | ● |
| Интеллектуальные оповещения администраторам | ● | ● |
| Возможность реализации решения «Метрокластер» | | ● |

## 8.2 Лицензирование комплекса Скала^р МХД.О

В комплексе Скала^р МХД.О лицензируется только доступное пользователям дисковое пространство (только полезный объём, за вычетом системных затрат на схему репликации или применения блоков четности/избыточности). Метрикой является 1 Терабайт, лицензируется ближайшее целое число.

Возможно повышение уровня Редакции ПО при росте инсталляционной базы комплексов.

## О компании

Компания Скала^р — разработчик и производитель модульных платформ для высоконагруженных корпоративных и государственных информационных систем.

Вся продукция Скала^р включена в Единый реестр российской радиоэлектронной продукции.

Продукты Скала^р являются серийно выпускаемыми преднастроенными комплексами и позволяют осуществлять их быстрое развёртывание и ввод в эксплуатацию.

Модульный принцип обеспечивает интеграцию разнородных компонентов ИТ-инфраструктуры в единую платформу предприятий, корпораций и ведомств.

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