Отказоустойчивость системы VK WorkSpace

Документация для системных инженеров

Портал с документацией: https://biz.mail.ru/docs/on-premises/

Оглавление

Назначение документа

В документе описаны технологии достижения отказоустойчивости в кластерных инсталляциях VK WorkSpace. Документ будет полезен системным инженерам и архитекторам.

Введение

В документе используются следующие понятия:

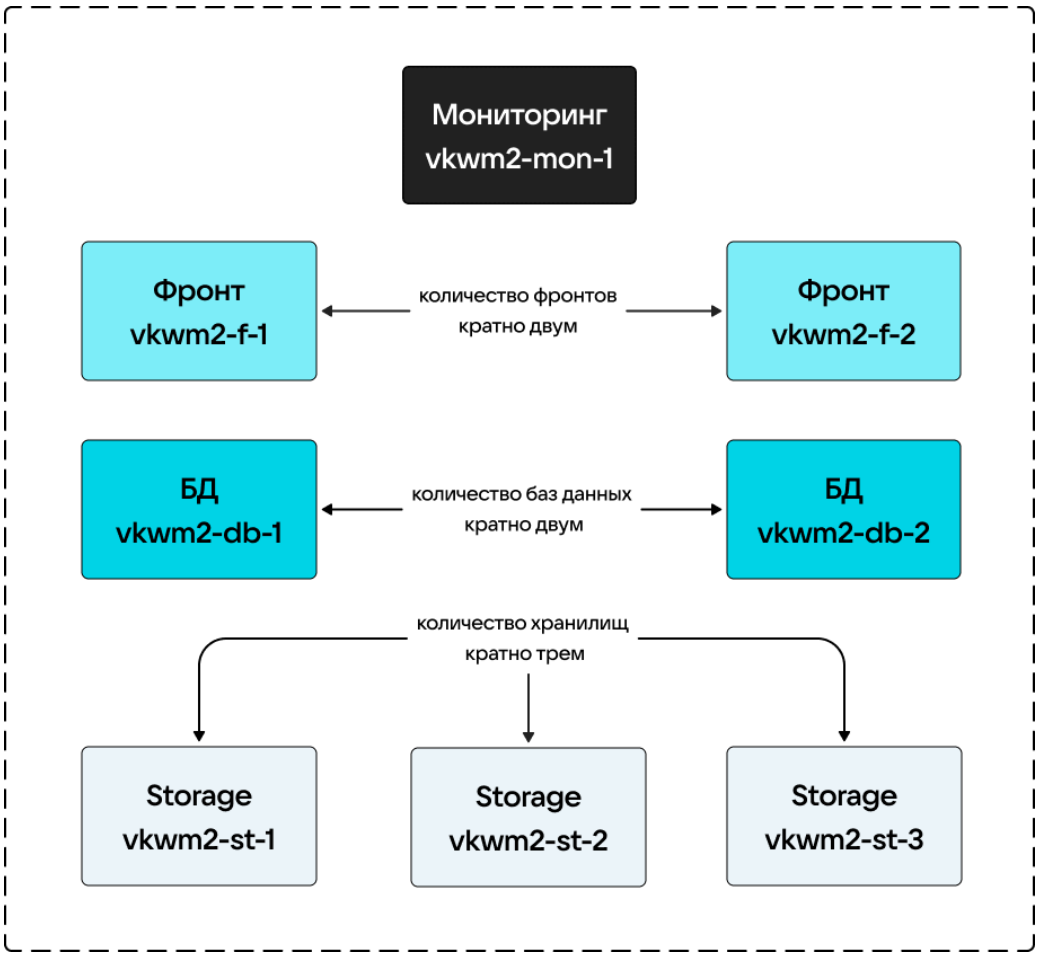
Отказоустойчивость — способность системы переносить без потерь выход из строя одного или нескольких компонентов.

Минимальная отказоустройчивая конфигурация состоит из:

- **двух** ВМ (виртуальных машин) или физических серверов под фронты,
- **двух** ВМ или физических серверов под базы данных,
- **трех** ВМ или физических серверов под хранилища,
- **одной** ВМ или физических серверов под установщик + мониторинг.

Таким образом, в отказоустойчивом кластере должно быть минимум **8 ВМ**.

Схема кластера для 8 ВМ выглядит следующим образом:



Важно

Количество ВМ может быть увеличено кратно, согласно схеме выше.

Единица отказоустойчивости — непосредственно сервер, на котором установлены компоненты VK WorkSpace, поэтому компоненты должны быть разнесены по трем выделенным площадкам. Например, ВМ могут находиться на разных полках/стойках/шкафах, но в одном ЦОД. Также может использоваться несколько ЦОД. Размещение на трех георасределенных дата-центрах обеспечивает инсталляции катастрофоустройчивость.

Фронт — единица отказоустойчивости, которая обслуживает не только сервисы, ответственные за отображение информации, но и бизнес-логику, почтовый транспорт и API.

Требования к сотрудникам

Для успешного построения отказоустойчивой инфраструктуры от персонала требуется:

- Знания Linux на уровне системного администратора
- Базовые знания по организации сетей
- Базовые знания почтовых протоколов

Условия для создания отказоустойчивой конфигурации

Перед установкой отказоустойчивой конфигурации Почты нужно обеспечить следующее:

- Три выделенные площадки размещения
- Минимум 8 ВМ с указанными выше ролями
- Сетевая связанность всех компонентов системы обеспечена каналами не менее чем 10 Гб/с
- Входящие каналы связи до фронтов — от 2Мб/с на пользователя

Обеспечение отказоустойчивости на уровне фронтов

Для создания отказоустойчивой конфигурации на уровне фронтов необходимы минимум две единицы отказоустойчивости, размещенные на разных выделенных площадках.

В базовой конфигурации настраивается балансировка через DNS.

> **Важно**
>
> Для оптимального георасределения потока входящих соединений с учетом загрузки и доступности сервисов рекомендуется использовать внешние средства балансировки.

Реализована проверка работоспособности сервисов VK WorkSpace, которая производится раз в минуту с помощью Envoy. В случае если какой-либо компонент не отвечает, запрос перенаправляется на одноименный.

Обеспечение отказоустойчивости на уровне баз данных

Отказоустойчивость на уровне БД достигается с помощью асинхронной master-slave репликации. Выбор лидера обеспечивается за счет алгоритма Raft.

Базы данных используются в двух экземплярах — им не нужен кворум из трех, однако количество БД можно увеличить кратно двум, тем самым увеличив количество копий данных.

Основные используемые базы данных:

- PostgreSQL

- MySQL
- Tarantool

> **Информация**
>
> В интерфейсе установщика есть возможность запросить статусы БД с помощью кнопки **Опросить все Overlord'ы**

PostgreSQL

Реплицируемые базы данных, оркестрируются Patroni. В случае выхода из строя мастера Patroni выберет нового.

Голосование происходит через etcd: какая из реплик первой сделает запись в etcd, та станет мастером. Это работает и в случае наличия одной реплики — она автоматически становится мастером.

Split brain исключен, так как бывший мастер переводится в режим read-only.

MySQL

Реплицируемые базы данных, управляются с помощью Consul + Orchestrator. Для MySQL orchestrator выбирает нового мастера (голосование проводится через Consul).

Tarantool

Реплицируемые базы данных. Управляются с помощью внутреннего сервиса Overlord (голосование проводится через etcd).

Обеспечение отказоустойчивости на уровне хранилищ

На уровне хранилищ отказоустойчивость достигается за счет разделения каждого из них на дисковые пары.

> **Информация**
>
> Под дисковой парой подразумеваются связанные разделы дисков, которые размещены **на двух разных ВМ-хранилищах**.

Минимальная отказоустойчивая конфигурация: 3 машины, на каждой из которых по 2 дисковых раздела — всего 6 разделов.

При такой конфигурации:

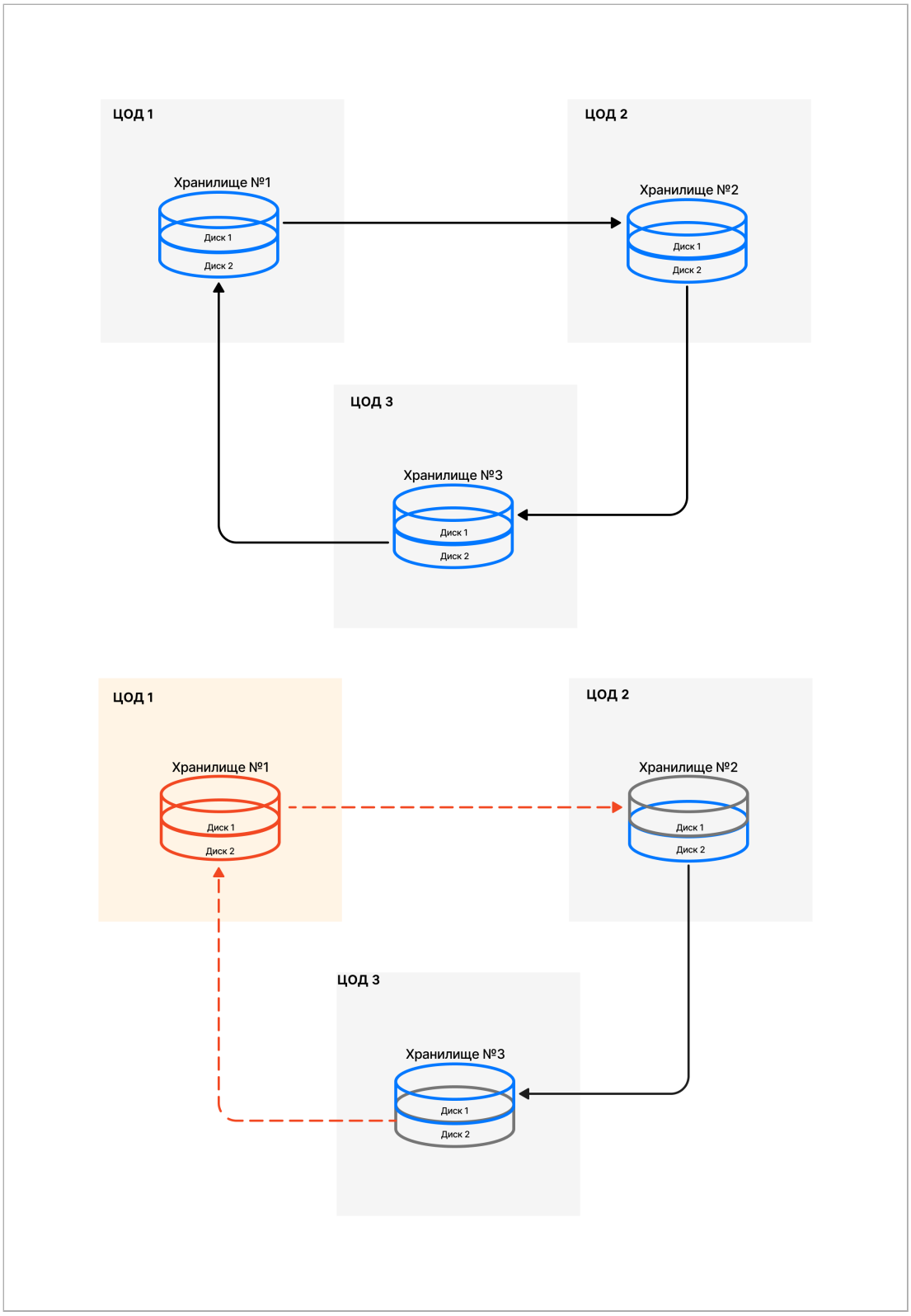
- Всегда есть пара на запись.
- Остальные пары доступны для чтения.

Таким образом:

- Хранилище №1 разделено на 2 части
- Хранилище №2 разделено на 2 части
- Хранилище №3 разделено на 2 части

Всего 6 разделов хранилищ (3 дисковые пары): 2 на одном сервере, 2 — на втором, еще 2 — на третьем.

При сборке хранилищ дисковые пары собираются из дисков, размещенных на разных выделенных площадках. Объединение происходит по принципу: 1-2, 2-3, 3-1.



Примечание

Стрелки на изображении показывают, какие диски объединены в пару. На нижней части изображения демонтрируется ситуация, когда одно из хранилищ вышло из строя.

При выходе из строя одного из хранилищ данные не будут утеряны, так как остается еще по экземпляру данных, которые были в read-only.

Автор: Груздев Никита

18 декабря 2024 г.