Катастрофоустойчивость 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 катастрофоустойчивости на трех георасределенных дата-центрах в кластерных инсталляциях VK WorkSpace. Документ будет полезен системным инженерам и архитекторам.

Дополнительная документация

Кластерная установка VK WorkSpace — подробная инструкция по кластерной установке VK WorkSpace на 8 машин.

Катастрофоустойчивость по схеме 2 ЦОД + witness — в документе рассматривается катастрофоустойчивая конфигурация из двух георасределенных ЦОД и дополнительного сервера, не находящегося в этих двух ЦОД.

Введение

Прежде чем перейти к методам достижения катастрофоустойчивости, необходимо пояснить некоторые понятия.

Катастрофоустойчивость — способность системы продолжать работу после значительных аварий: полное отключение электричества, потоп, землятресение и т.п. Катастрофоустойчивость достигается за счет размещения VK WorkSpace на **трех георасределенных ЦОД**.

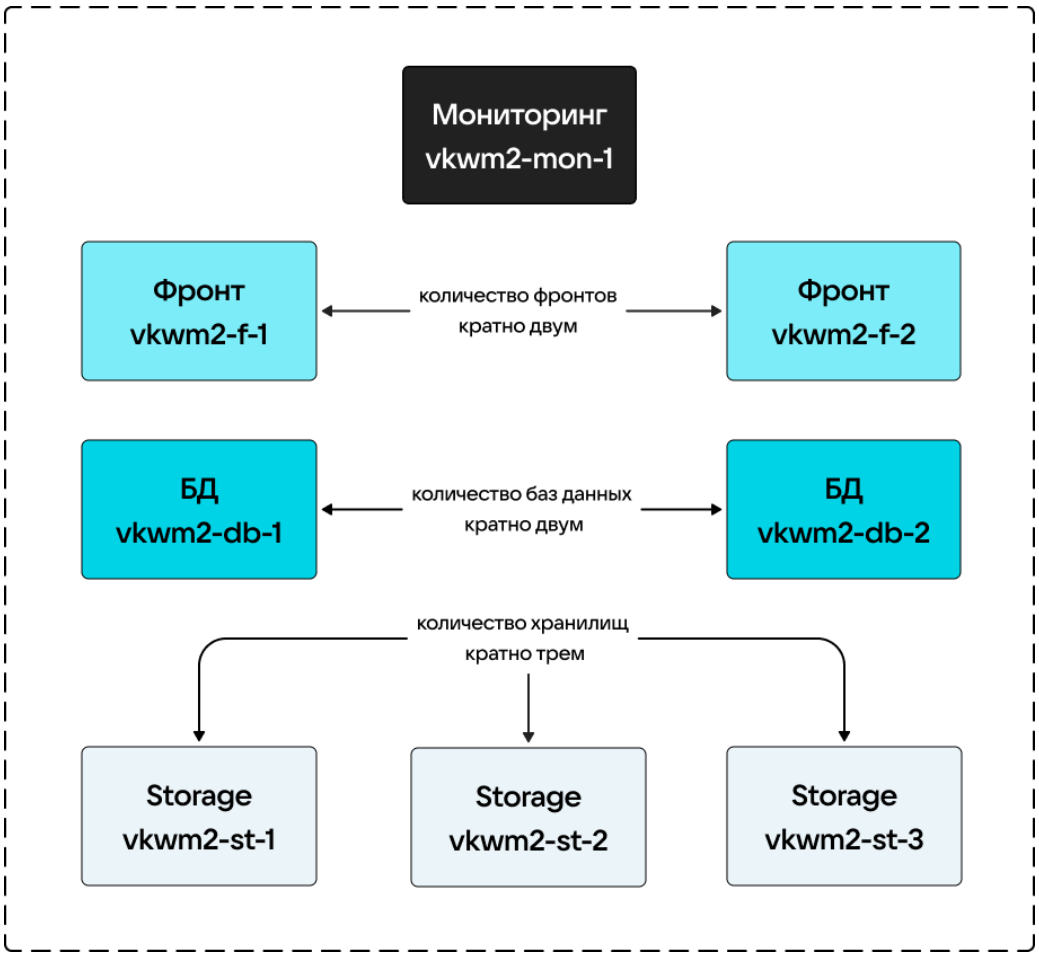
Катастрофоустойчивая конфигурация — конфигурация, в которой базы данных, фронты и хранилища разнесены минимум по трем георасределенным дата-центрам.

Минимальная катастрофоустойчивая 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 катастрофоустойчивом кластере должно быть минимум **8 ВМ**.

Схема кластера для 8 ВМ выглядит следующим образом:



Важно

Количество ВМ может быть увеличено кратно, согласно схеме выше.

Одноименные компоненты **не должны** быть размещены в одном ЦОД. **Запрещено** размещать в одном ЦОД два экземпляра баз данных, фронтов или хранилищ, иначе катастрофоустойчивости достигнуть будет невозможно.

Пример правильного распределения серверов по дата-центрам:

Компонент	ЦОД
Хранилище №1	1
Хранилище №2	2
Хранилище №3	3
БД №1	1
БД №2	2

Компонент	ЦОД
Фронт №1	1
Фронт №2	2
Установщик + мониторинг	3

На ЦОД №1 находится по экземпляру: хранилища, БД, фронта.

На ЦОД №2 находится по экземпляру: хранилища, БД, фронта.

На ЦОД №3 находится по экземпляру: хранилища, установщика + мониторинга.

Единица катастрофоустойчивости — непосредственно ЦОД, на котором установлены компоненты VK WorkSpace.

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

Требования к сотрудникам

Для успешного построения катастрофоустойчивой инфраструктуры от персонала требуется:

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

Условия для создания катастрофоустойчивой конфигурации

Перед установкой катастрофоустойчивой конфигурации VK WorkSpace нужно обеспечить следующее:

- Три и более георазпределенных ЦОД
- Минимум 8 ВМ или физических серверов с указанными выше ролями
- Сетевую связанность всех компонентов системы обеспечена каналами не менее чем 10 Гб/с
- Входящие каналы связи до фронтов — от 2Мб/с на пользователя

Обеспечение катастрофоустойчивости на уровне фронтов

Для создания катастрофоустойчивой конфигурации на уровне фронтов необходимы минимум 2 экземпляра фронтов, размещенные в разных ЦОД.

В базовой конфигурации настраивается балансировка через DNS.

> **Важно**
>
> Для оптимального геораспределения 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 катастрофоустойчивости на уровне баз данных

Катастрофоустойчивость 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 катастрофоустойчивости на уровне хранилищ

На уровне хранилищ катастрофоустойчивость достигается за счет разделения каждого из них на дисковые пары. Данные записываются сразу на два диска, за счет чего достигается избыточность x2 и обеспечивается катастрофоусточивость.

> **Информация**
>
> Под дисковой парой подразумеваются связанные разделы дисков, которые размещены в разных ЦОД.

Минимальная катастрофоустойчивая конфигурация: 3 машины (каждая из которых находится в разных геораспределенных дата-центрах), на каждой из которых по 2 дисковых раздела — всего 6 разделов.

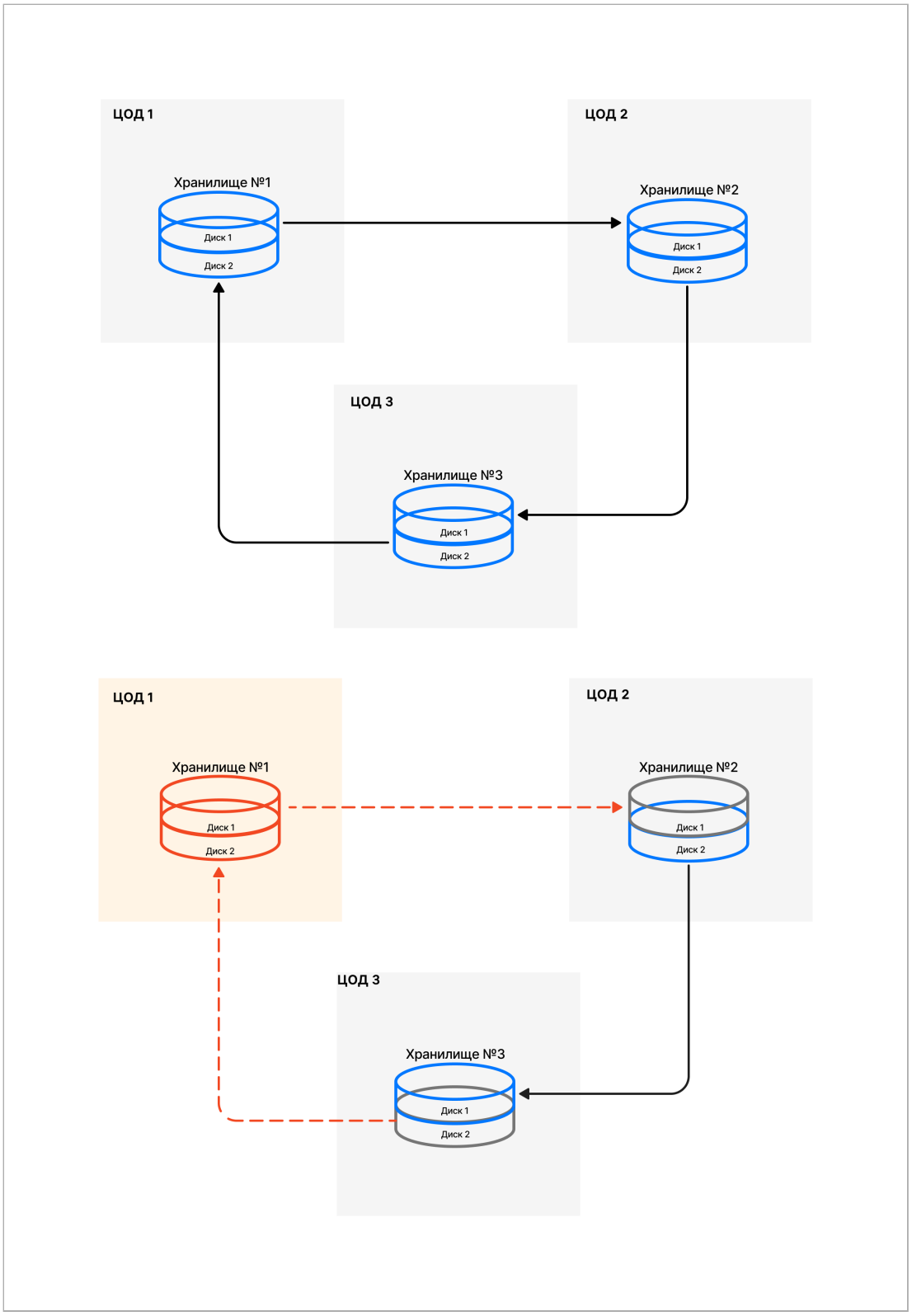
При такой конфигурации всегда есть пара на запись.

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, находящихся в разных ЦОД. Объединение происходит по принципу: 1-2, 2-3, 3-1.



Информация

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

При выходе из строя одного ЦОД данные не будут утеряны, так как останется еще по экземпляру данных. Половины утраченных пар, в свою очередь, переводятся в режим read-only.

Автор: Груздев Никита

18 декабря 2024 г.