Tarantool Clusters Federation

Руководство пользователя

Версия 1.0.0 от 23.08.23

Термины и сокращения

В настоящем документе используются сокращения, перечисленные в таблице ниже.

Сокращение	**Расшифровка**
БД	База данных
ВМ	Виртуальная машина
ОС	Операционная система
ПО	Программное обеспечение
СУБД	Система управления базами данных
API	Application Programming Interface - Аппаратно-программный интерфейс. Описание способов (набор классов, процедур, функций, структур или констант), которыми одна компьютерная программа может взаимодействовать с другой программой
JavaScript	Мультипарадигменный язык программирования. Поддерживает объектно-ориентированный, императивный и функциональный стили
JSON	JavaScript Object Notation – текстовый формат обмена данными, основанный на JavaScript
Узел, Инстанс, Нода	Единичный экземпляр (процесс) запущенного ПО
Балансировщик	Программное или аппаратное решение, используемое для распределения (балансировки) всех запросов приложения по кластеру с целью оптимизации использования вычислительных ресурсов
etcd	Устойчивая к сбоям key-value распределенная база данных и работает поверх консенсус-протокола RAFT
ЦОД	Центр обработки данных

1. Введение

1.1. Область применения

Tarantool Clusters Federation (далее по тексту TCF) представляет собой инструмент позволяющий организовать катастрофоустойчивую конфигурацию из двух и более кластеров Tarantool Enterprise Edition, находящихся в разных дата-центрах. Данная конфигурация обязательна при построении Mission Critical систем хранения данных, для которых необходимо выполнение следующих требований:

- гарантия постоянной доступности данных
- обеспечение отсутствия потерь данных
- минимизация времени отклика для Клиента

1.2. Краткое описание возможностей

TCF используется для реализации механизма переключения кластеров и синхронизации данных в геораспределенных кластерах и поддерживает следующие сценарии работы:

- Ручное переключение активного кластера
- Автоматическое переключение активного кластера

TCF позволяет минимизировать последствия от возникновения нештатных ситуаций при работе с изолированными кластерами и обеспечить бесперебойный доступ к данным в условиях построения Mission Critical систем хранения данных. К нештатным ситуациям могут быть отнесены:

- ошибки при проведении регламентных работ на узлах распределенной системы (изменение схемы хранения данных)
- чрезвычайные ситуации, приводящие к полной недоступности одного из узлов распределенной системы (выход из строя ЦОД или нарушение сетевой связанности)
- ошибки при разработке прикладного ПО / SDK, приводящие к полной недоступности кластера во всех ЦОДах

1.3. Требования к квалификации пользователя

К пользователям TCF предъявляются следующие требования:

- Прохождение обучения от компании ООО «ВК ЦИФРОВЫЕ ТЕХНОЛОГИИ» по использованию компонентов, разработанных вендором;
- владение персональным компьютером на уровне уверенного пользователя;
- знание функциональных возможностей Системы и особенностей работы с ними;
- знать принципы построения систем управления базами данных;
- обладать навыками разработки сервисов взаимодействия;
- иметь навыки работы с серверным оборудованием;
- иметь расширенные знания в области поддержки пользователей;
- знать основы администрирования ОС, серверов приложений и серверов баз данных;
- обладать навыками использования терминала;
- обладать навыками использования Tarantool Cartridge;
- знать основы работы вычислительной техники и программного обеспечения в локальных сетях, а также настроек системной политики прав пользователей в операционных системах семейства Linux.

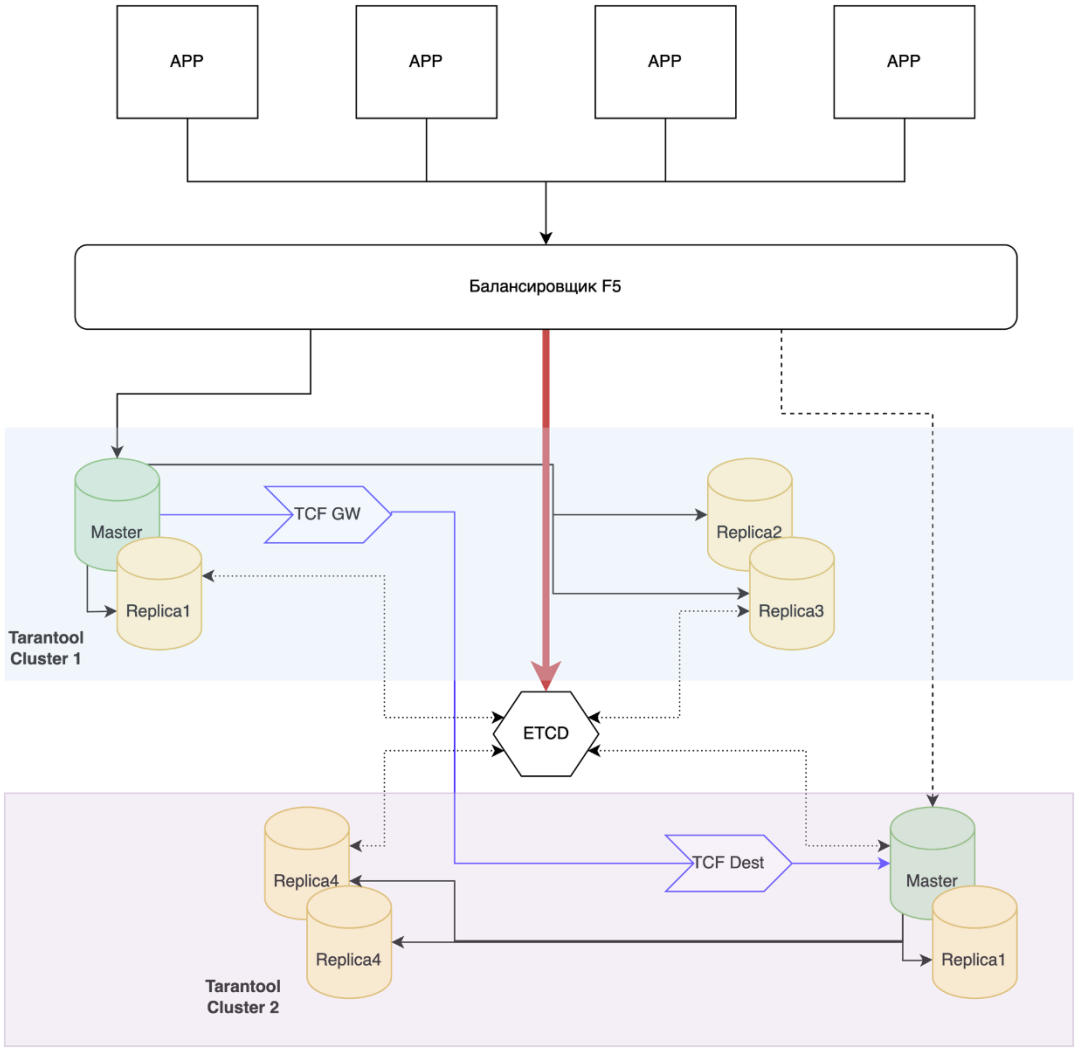
1.4. Перечень эксплуатационной документации

Перечень эксплуатационной документации, с которой необходимо ознакомиться:

- Инструкция по установке экземпляра программного обеспечения TCF;
- Описание функциональных характеристик TCF;
- Руководство по установке TCF;
- Описание процессов поддержания жизненного цикла TCF;
- Документация к Tarantool на сайте tarantool.io.

2. Принцип работы решения

2.1. Выбор активного кластера



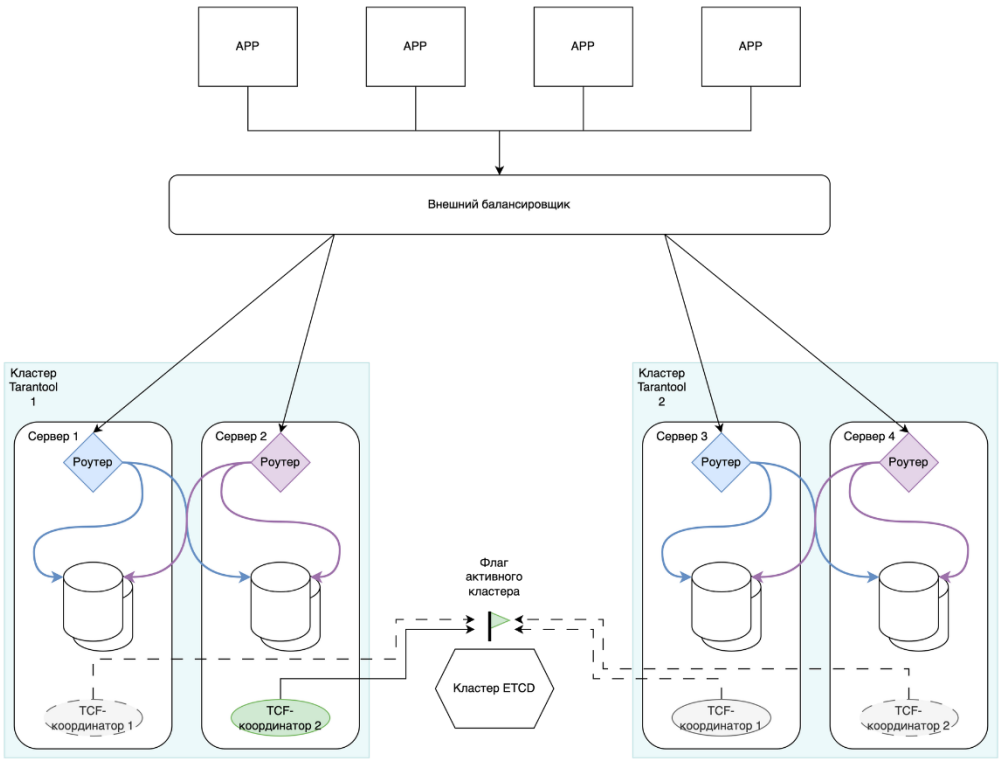
Настроено и работает два идентичных кластера Tarantool Enterprise, ТС1 и ТС2. Кластер ТС1 является ведущим. С ним работает на чтение и запись приложение. Трафик между приложением и кластером балансируется через внешний балансировщик. Кластер ТС2 является ведомым кластером. При штатной работе решения трафик от приложения к кластеру ТС2 не направляется.

На балансировщике созданы 2 пула адресов, соответствующие списку роутеров каждого кластера. Балансировщик периодически опрашивает ключ etcd `/stand_in/<cluster_name>/is_active` для каждого кластера, и проверяет ответ. Если статус возвращаемого HTTP-запроса равен 200, то балансировщик направляет трафик на пул адресов, соответствующих этому кластеру. Если статус возвращаемого HTTP-запроса равен 400, то трафик на пул адресов, соответствующих кластеру, не направляется. Одновременно может быть активным только один кластер. То есть в любой момент времени для одного кластера возвращаемый из ectd статус запроса будет равен 200, а для другого 400.

При смене состояний кластеров их статусы в etcd меняются на противоположенные.

○ 2.2. Автоматическое переключение кластера

За автоматическое переключение отвечает отдельная роль в каждом кластере "TCF-координатор". Экземпляров Tarantool с данной ролью должно быть запущено в каждом кластере два и более (для отказоустойчивости).



В каждый момент времени активным является только один из координаторов суммарно в двух кластерах. Активный координатор может изменять флаг `target_status` своего кластера, назначая его активным.

TCF-координаторы (все, не только активный) использует модуль `membership` для мониторинга состояния своих. При смене активного координатора, новый лидер сможет быстрее включится в работу и назначить активный кластер.

Если TCF обнаруживает, то его активный кластер не соответствует критериям "здоровья" (недоступны все роутеры, недоступен сегмент целиком), то он складывает с себя полномочия, переставая обновлять флаги выбора активного координатора. Это даст возможность захватить флаг координаторам из второго кластера. Другие координаторы сбойного кластера, зная его проблемное состояние, выходят из гонки по захвату флага.

Возможна ситуация, при которой координаторы обоих кластеров считают свои кластеры сбойными. В этом случае оба кластера будут находится в пассивном режиме. Причинами такого состояния может быть потеря именно координатором связи с кластерами (потеря связности сети) либо более общей проблемой в кластерах или сети. И принимать решение о автоматическом переключении опасно. Такая ситуация потребует вмешательства оператора.

○ 2.3. Ручное переключение кластера

В TCF реализована возможность ручного переключения активных кластеров, либо с помощью изменения значения флага `/stand_in/toggle` , либо с помощью POST запроса на /tcf/toggle на любой ноде.

Пример переключения активного кластера через etcd

$$curl\ -i\ http://127.0.0.1:4001/v2/keys/stand\_in/toggle\ -XPUT\ -d\ value=true$$

Пример получения информации об активном кластере

$$\$\ curl\ http://127.0.0.1:4001/v2/keys/stand\_in/active$$

$$\{"action":"get","node":\{"key":"/stand\_in/active","value":"cluster\_a"\ldots.\}\}$$

Пример получения статуса с кластера являющегося активным в данный момент времени

$$\$\ curl\ -i\ http://127.0.0.1:4001/v2/keys/stand\_in/cluster\_a/is\_active$$

$$HTTP/1.1\ 200\ OK$$

$$\{"action":"get","node":\{"key":"/stand\_in/cluster\_a/is\_active","value":"true"\ldots\}\}$$

Пример получения статуса с кластера не являющегося активным в данный момент времени

$$curl\ -i\ http://127.0.0.1:4001/v2/keys/stand\_in/cluster\_b/is\_active$$

$$HTTP/1.1\ 404\ Not\ Found$$

$$\{"errorCode":100,"message":"Key\ not\ found","cause":"/stand\_in/cluster\_b/is\_active"\ldots\}$$

○ 2.4. ETCD ключи

Ключ	Описание
`/stand_in/active`	Имя текущего активного кластера
`/stand_in/<cluster_name>/is_active`	Флаг, установленный на активном кластере
`/stand_in/<cluster_name>/target_status`	Целевой статус кластера (возможно еще не достигнутый)
`/stand_in/<cluster_name>/status`	Текущий статус кластера
`/stand_in/<cluster_name>/replicasets/<replicaset_uuid>/replicas`	JSON со статусами каждого инстанса данного репликасета
`/stand_in/<cluster_name>/replicasets/<replicaset_uuid>/status`	Текущий статус данного репликасета
`/stand_in/<cluster_name>/replicasets/<replicaset_uuid>/flag`	Флаг для синхронизации доступа к `./replicas` и `./status`

3. Аварийные ситуации

В случае возникновения аварийной ситуации пользователю необходимо обратиться к администратору системы и выполнить следующие действия:

- Передать информацию о том, что было сделано, прежде чем проблема появилась;
- Сформулировать и описать, в чем именно заключается проблема.