АЭРОДИСК
мы делаем будущее

Практика применения метрокластера для системы хранения данных АЭРОДИСК

Дата: 26.08.2024
Версия: 2.0

Оглавление

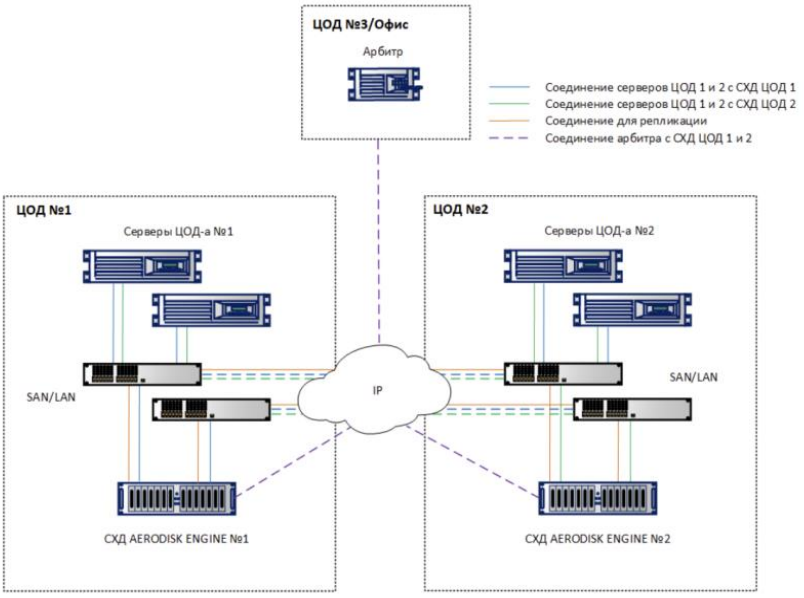
Метрокластер: основные понятия

Две аппаратные системы хранения данных (СХД) АЭРОДИСК, объединенные в метрокластер и разнесенные на расстояние до нескольких десятков километров, позволяют обеспечить бесперебойной работу и защиту от катастрофы на одной из площадок размещения. В метрокластере две копии данных физически находятся в двух разных, отдаленных на значительное расстояние, местах, и их синхронизация обеспечивается по сети Ethernet (синхронизация по сети Fibre Channel не поддерживается). Можно использовать две похожие, но не обязательно однотипные, СХД АЭРОДИСК. Решение метрокластера базируется на использовании синхронной репликации между двумя СХД АЭРОДИСК на уровне отдельных блочных устройств (логических томов), репликационные связи которых можно добавлять и удалять в онлайн-режиме.

Размещенный на третьей площадке арбитр управляет автоматическим переключением между двумя площадками без вмешательства системного администратора и предотвращает ситуацию «разделения кластера» (Split-Brain). Арбитр представляет из себя виртуальную машину, которая может работать на любом популярном гипервизоре: vAIR, ESXi, Hyper-V. В случае аварии на одной площадке остаётся полная копия данных на второй площадке и с помощью арбитра обеспечивается автоматическое переключение работы конечных приложений на оставшуюся работоспособной СХД.

Архитектура решения

Архитектура решения метрокластера представлена на картинке ниже.



Количество одновременно реплицируемых пар логических томов явно неограниченно и может составлять несколько десятков пар. Направления репликации могут отличатся между логическими томами: часть томов реплицируется с первой СХД на вторую СХД, другая часть реплицируется со второй СХД на первую, что позволяет распределить нагрузку конечных приложений между обеими аппаратными СХД. Реплицируемая пара логических томов на двух аппаратных СХД АЭРОДИСК должна размещаться на сравнимых по параметрам производительности пулах хранения с учетом используемых типов носителей, их количества и уровня RAID-защиты.

При работе в режиме метрокластера серверы обработки подключаются к СХД *только по протоколу iSCSI, и каждый аппаратный сервер на обеих площадках должен иметь сетевой доступ к обеим системам хранения.*

Планирование метрокластера

Для метрокластера необходимы три территориально разнесенные площадки: на двух размещаются две аппаратные СХД и сервера обработки, на третьей работает арбитр.

Физическое расстояние между двумя аппаратными СХД явно не ограничивается и может составлять десятки километров. Во избежание деградации производительности СХД и замедления работы конечных приложений необходимо обеспечить прохождение сетевых пакетов между двумя СХД с низкими сетевыми задержками в канале передачи. Рекомендуемые к выполнению требования по сетевому взаимодействию двух площадок размещения СХД:

- оптоволоконная среда передачи с временем приема-передачи (RTT) не более 4 мс;
- сетевые коммутаторы с пропускной способность портов не менее 10 Гбит/с.

Расстояние между СХД и арбитром на третьей площадке также может составлять десятки километров. Для работы арбитра необходимо обеспечить прохождение сетевых пакетов между каждой из СХД и арбитром с максимальной сетевой задержкой не более 100 мс.

Во избежание критичной деградации в работе конечных приложений необходимо закладывать минимально допустимый уровень производительности в условиях временной работы на одной аппаратной СХД в случае «падения» одной из площадок размещения.

Топология сети Ethernet

Сетевые соединения между площадками размещения двух аппаратных СХД и серверов обработки должны выполняться через коммутаторы Ethernet, пропускная способность

которых является достаточной для продуктивной работы конечных приложений и синхронизации обрабатываемых данных (передаче всех изменений, операций записи).

Для работы метрокластера необходимо определить несколько логических подсетей для различных типов передаваемых трафиков:

- управляющая подсеть – для управления двумя аппаратными СХД и работы арбитра;
- подсеть хранения данных – растянутая подсеть размещения аппаратных СХД и серверов обработки, по которым сервера обработки получают доступ к ресурсам СХД по протоколу iSCSI. Из этой подсети назначаются VIP-адреса типа «метрокластер»;
- подсеть для репликации – подсеть на двух площадках размещения аппаратных СХД и серверов обработки для репликации, по которой будут синхронизироваться данные между двумя аппаратными СХД. Из этой подсети назначаются VIP-адреса типа «репликация» при создании репликационных связей.

Разделение различных типов передаваемых трафиков по разным подсетям позволит оптимизировать обработку трафика. Особенно важно отделить трафик ввода-вывода серверов обработки и трафик синхронизации данных между двумя аппаратными СХД при большой загрузке канала.

Настройка арбитра

Арбитр представляет собой виртуальную машину на основе Альт Линукс 10.1. Арбитр используется для организации функционала метрокластера между двумя аппаратными СХД. Установка осуществляется сервисной поддержкой АЭРОДИСК из образа виртуальной машины. После установки нужно поменять IP-адрес арбитра на выделенный заказчиком и указать IP-параметры для подключения к подсети управления. Никаких других дополнительных настроек не требуется, арбитру необходимо лишь обеспечить сетевой доступ ко всем управляющим интерфейсам на всех четырёх контроллерах аппаратных СХД по протоколам ICMP и ssh.

Настройка репликационных связей реплицируемых логических томов

На обеих системах хранения для каждого логического тома, который будет реплицироваться на другую аппаратную СХД, создаётся уникальный виртуальный IP-адрес (VIP) типа «репликация», который используется для создания репликационной связи между парой логических томов на двух аппаратных СХД.

Для каждой реплицируемой пары логических томов создаётся репликационная связь с типом связи «синхронная» и максимальным количество узлов «2».

Создание репликационной связи выполняется в интерфейсе управления одной из СХД, которая будет выполнять роль «локального узла» для создаваемой репликационной связи, при этом другая СХД будет выполнять роль «удаленного узла». Репликационная связь на локальной СХД будет создана с ролью «Primary», а репликационная связь на удаленной СХД будет создана с ролью «Secondary». Репликационная связь создаётся на «удаленном узле» автоматически, и после создания репликационной связи запускается процесс синхронизации данных с «локального узла» на «удаленный узел». При создании репликационных связей в текущей реализации доступный объем тома уменьшается на объем метаданных. Необходимо, чтобы тома были свободны от данных перед созданием репликационных связей.

Настройка метрокластера

Перед созданием связи метрокластера необходимо убедиться, что все логические тома, находящиеся в репликационной связи, завершили первоначальную синхронизацию (процесс синхронизации отображается в интерфейсе управления). Далее будем указывать действия для настройки, производимые отдельно на СХД1 и СХД2 в составе метрокластера (СХД1 – система хранения данных, на которой создавались репликационные связи). Связи находятся в синхронизированном состоянии.

СХД1:

1. Необходимо создать IP-ресурс на каждую пару реплицируемых логических томов (VIP-адрес метрокластера). В интерфейсе СХД1 на любом из контроллеров выбрать: IP ресурс –> Создать ресурс –> Тип “Метрокластер» -> Выбрать Ethernet-интерфейсы, принадлежащие сети доступа.
2. Подключить СХД1 к метрокластеру. Выполняется в пункте меню Репликация -> Удаленная репликация -> по правой кнопке мыши выбрать «Подключить к метрокластеру» -> Выбираем IP-адрес метрокластера (VIP со статусом «метрокластер» для группы, в которой находятся логические тома).
3. Создать «iSCSI Таргеты» для VIP «Метрокластер» и указать маппинг томов для инициаторов серверов доступа.
4. Инициализировать метрокластер, указав IP-адрес арбитра и IP-адреса обоих контроллеров другой СХД. Удаленная репликация -> Метрокластер -> Сконфигурировать. После этого необходимо перезапустить связи и сервисы в репликации. Для этого в интерфейсе управления есть опция «Всё перезапустить».

Необходимо выполнить операцию «Всё перезапустить» в интерфейсе управления СХД на двух контроллерах.

СХД2:

1. Необходимо создать IP-ресурс VIP-адрес метрокластера на каждую пару реплицируемых логических томов. В интерфейсе СХД2 на любом из контроллеров выбрать: IP ресурс –> Создать ресурс –> Тип “Метрокластер» -> Выбрать Ethernet-интерфейсы, принадлежащие подсети доступа. IP-адреса должны совпадать с адресами доступа, ранее созданными на СХД1.
2. Сменить роли в репликации с «Secondary» на «Primary» для всех репликационных связей в СХД2: выполняется в пункте меню «Репликация -> Удаленная репликация -> по правой кнопке мыши выбрать репликацию со статусом «Secondary» и выбрать «Сделать первичным».
3. Подключить СХД2 к метрокластеру. Выполняется в пункте меню «Репликация -> Удаленная репликация -> по правой кнопке мыши выбрать «Подключить к метрокластеру» -> Выбираем IP-адрес метрокластера (VIP со статусом «метрокластер» для группы, в которой находятся логические тома).
4. Создать «iSCSI Таргеты» для VIP «Метрокластер» и указать маппинг томов для инициаторов серверов доступа. Маппинг для томов в репликации должен быть идентичен маппингу СХД1.
5. Инициализировать метрокластер, указав IP-адрес арбитра и IP-адреса обоих контроллеров другой СХД. Удаленная репликация -> Метрокластер -> Сконфигурировать. После этого необходимо перезапустить связи и сервисы в репликации. Для этого в интерфейсе управления есть опция «Всё перезапустить». Необходимо выполнить операцию «Всё перезапустить» в интерфейсе управления СХД на двух контроллерах.

В один момент времени каждый из VIP-адресов метрокластера активен только на одном контроллере из четырёх, входящих в состав метрокластера. Названия таргетов и групп на обеих СХД при настройке iSCSI подключения должны быть идентичны, чтобы при смене ролей в репликации сервера обработки подключались без изменений. В случае «перехода» с отказавшей системы хранения на другую, у серверов обработки сохраняется доступ к данным по тем же самым VIP-адресам метрокластера.

Необходимо проверять, что сгенерированный уникальный NAA идентификатор блочного устройства одинаков на обеих СХД для всех пар логических томов в метрокластере.

Сценарии отказов

Логика работы метрокластера решает задачу предотвращения ситуации «разделения кластера» (Split Brain), которая приводит к запуску обеих аппаратных СХД с ролью «основного узла» для всех пар реплицированных логических томов и обеспечивает автоматическую обработку возникающих отказов в работе всего окружения метрокластера.

В таблице ниже представлены варианты отказов и действия в метрокластере для обеспечения продолжения работы серверов обработки в автоматическом режиме с помощью арбитра.

Вариант отказа	Что с трафиком серверов обработки на площадке А?	Что с трафиком серверов обработки на площадке Б?
Отказ контроллера владельца на СХД с ролью «локального узла»	Кратковременная остановка, серверы обработки продолжат работать с оставшимся в работе контроллером на той же СХД с ролью «локального узла»	Кратковременная остановка, серверы обработки продолжат работать с оставшимся в работе контроллером на той же СХД с ролью «локального» узла»
Отказ контроллера на СХД с ролью «локального узла», не владеющего дисковой группой	Штатная работа	Штатная работа
Отказ контроллера владельца на СХД с ролью «удаленного узла»	Штатная работа	Штатная работа
Отказ контроллера на СХД с ролью «удаленного узла», не владеющего дисковой группой	Штатная работа	Штатная работа
Отказ репликационной связи между двумя СХД	Штатная работа	Штатная работа
Отказ связи между арбитром и СХД одной из площадок	Штатная работа	Штатная работа
Отказ СХД одной из площадок целиком	Кратковременная остановка, работа серверов обработки переводится на СХД второй площадки	Кратковременная остановка, работа серверов обработки переводится на СХД второй площадки

Далее описаны более сложный сценарий отказов, который также обрабатывается метрокластером.

Сценарий сложного отказа: со стороны одной СХД потеряно сетевое соединение с арбитром и с другой СХД

Событием-сигналом об отказе аппаратной СХД является отсутствие ICMP пакета (пинга) с обоих контроллеров аппаратной СХД в течение 5 секунд. При потере арбитром сетевой связи с одной из СХД (далее СХД1) арбитр отправляет запрос на другую СХД (далее СХД2), чтобы убедиться, что недоступная СХД1 действительно не работает. Для этого выполняются следующие действия:

1. СХД2 проверяет доступность управляющих интерфейсов СХД1;
   - если управляющие интерфейсы СХД1 доступны с СХД2, то никаких действий в метрокластере не выполняется, в журнале событий фиксируется ошибка сетевого взаимодействия арбитра с СХД1;
   - если управляющие интерфейсы СХД1 недоступны с СХД2, то проверяется статус репликационных связей между СХД:

   - если статус репликационных связей «Вне связи» на СХД2, то репликационные связи с ролью «Secondary» будут переключены в роль «Primary» и соответствующие VIP-адреса метрокластера будут подняты на СХД2 для обслуживания запросов серверов обработки;

   - если статус репликационных связей «Синхронизирован», то СХД1 оказывается изолированной, все репликационные связи на СХД1 с ролью «Primary» меняются на «Secondary». В этом случае для восстановления работы требуется «ручное» вмешательство и устранение возникших проблем сетевой недоступности СХД1.

Используемые сетевые протоколы и порты

Для обеспечения сетевого взаимодействия между участниками метрокластера необходимо убедиться, что на межсетевых экранах и других средствах сетевой безопасности разрешены следующие сетевые взаимодействия между СХД1, СХД2, Арбитром, Серверами доступа. Стрелками -> указано направление взаимодействия.

Подсеть управления:

СХД1 <-> Арбитр: (HTTP/HTTPS) 80/443 tcp

СХД1 <-> Арбитр: (SSH) 22 tcp

СХД1 <-> Арбитр: ICMP

СХД2 <-> Арбитр: (HTTP/HTTPS) 80/443 tcp

СХД2 <-> Арбитр: (SSH) 22 tcp

СХД2 <-> Арбитр: ICMP

СХД2 <-> СХД1: SSH port 22 tcp

СХД2 <-> СХД1: ICMP

Подсеть репликации:

СХД1 -> СХД2: port 7000 tcp

СХД2 -> СХД1: port 7000 tcp

Подсеть доступа iSCSI:

Серверы доступа <-> СХД1: tcp any -> 3260

Серверы доступа <-> СХД2: tcp any -> 3260

Влияние на производительность

Синхронная репликация снижает максимальную производительность системы хранения данных и увеличивает задержки на некоторых операциях. Приведем основные наблюдения:

1. Для операций записи увеличивается latency более чем на 3 мс.
2. Общая производительность при штатной работе уменьшается более чем на 60% (репликационные связи в статусе “Синхронизировано”).
3. Функция динамического управления полосой пропускания для репликации данных при начальной синхронизации или передача изменений после простоя Secondary томов несущественно влияет на производительность блочного доступа (около 10%). Однако характер нагрузки влияет на скорость восстановления репликации или скорость начальной синхронизации.
4. При проектировании систем для работы в режиме репликации/метрокластер необходимо учитывать увеличенную среднюю задержку на запись и меньшую производительность, выдаваемую дисковыми группами в такой конфигурации.

