Опыт эксплуатации распределенного кластерного хранилища vSAN в ИПА РАН

© В. А. Яковлев, И. А. Безруков, А. И. Сальников

ИПА РАН, Санкт-Петербург, Россия

Реферат

Технология виртуальных машин на базе VMware используется в ИПА РАН с 2009 г. Виртуальные машины применялись как для вторичной обработки данных РСДБ-наблюдений, так и в качестве серверов буферизации в режиме e-РСДБ. По мере развития информационной сети ИПА РАН многие служебные сервисы, такие как почтовый сервер или веб-сервер, были перемещены с физических серверов на виртуальные машины. Возникла необходимость обеспечить бесперебойную работу этих сервисов или минимизировать время простоя в случае отказа оборудования или проведения профилактических работ.

С 2014 г. в ИПА РАН использовались два сервера виртуальных машин на двух географически разнесенных площадках. Однако единой точкой отказа оставалось сетевое файловое хранилище. В 2020 г. было принято решение внедрить технологию распределенного кластерного хранилища vSAN для решения этой проблемы.

В статье представлено описание технологии миграции инфраструктуры виртуальных машин на новое хранилище, проанализированы преимущества и недостатки этой технологии, сделаны выводы и описан практический опыт после одного года использования vSAN.

Ключевые слова: виртуальные машины, отказоустойчивость, масштабирование, обработка данных.

Контакты для связи: Безруков Илья Алексеевич (bezrukov@iaaras.ru).

Для цитирования: Яковлев В. А., Безруков И. А., Сальников А. И. Опыт эксплуатации распределенного кластерного хранилища vSAN в ИПА РАН // Труды ИПА РАН. 2021. Вып. 59. С. 26–29.
https://doi.org/10.32876/ApplAstron.59.26-29

Operational Experience of VMware vSAN Cluster Storage Technology at IAA RAS

© V. A. Yakovlev, I. A. Bezrukov, A. I. Salnikov

Institute of Applied Astronomy of the of the Russian Academy of Sciences, Saint Petersburg, Russia

Abstract

Virtual machines (hosted by VMware products) have been used at IAA RAS since 2009. Virtual machines were used both for processing of VLBI observation data and as buffering servers in the e-VLBI mode. With the development of the IAA RAS information network, many service services, such as a mail server or a web server, were moved from physical servers to virtual machines. So, there was a need arose to ensure uninterrupted operation of these services and to minimize downtime in the event of equipment failure or preventive maintenance.

Since 2014 IAA RAS has used two virtual machine servers at two independent sites. However, network file storage remained the single point of failure. In 2020 a decision was made to implement vSAN distributed cluster storage technology to address this issue.

The article describes the technology for migrating virtual machine infrastructure to new storage, the advantages and disadvantages of this technology, conclusions and practical experience after one year of using vSAN.

Keywords: virtual machines, fault tolerance, scaling, data processing.

Contacts: Il'ya A. Bezrukov (bezrukov@iaaras.ru).

For citation: Yakovlev V. A., Bezrukov I. A., Salnikov A. I. Operational experience of VMware vSAN cluster storage technology at IAA RAS // Transactions of IAA RAS. 2021. Vol. 59. P. 26–29.
https://doi.org/10.32876/ApplAstron.59.26-29

Введение. Классические конфигурации инфраструктуры VMware vSphere

Инфраструктура VMware vSphere состоит из хостов (серверов с операционной системой ESXi, запускающих виртуальные машины) и хранилищ (устройств, обеспечивающих хранение файлов дисков виртуальных машин). В качестве хранилищ могут выступать локальные дисковые устройства, подключенные непосредственно к хостам, и объединенные в массивы RAID или сетевые хранилища (NAS) (Стэмпковский, Гаязов, 2009). В общем случае последние представляют собой выделенный сервер с локальным хранилищем, которое предоставляется хостам VMware с помощью сетевых файловых систем, например NFS.

Конфигурация с локальными хранилищами (Direct Attached Storage, DAS) (рис. 1) является простейшей. Виртуальная машина (ВМ) исполняется на том же хосте, где хранится ее диск. Такая конфигурация не обеспечивает никакой защиты для ВМ. В случае отказа любого узла схемы — хоста или хранилища — ВМ, относящиеся к этому узлу, останавливаются до восстановления работоспособности узла. Время простоя зависит от причины отказа. В случае выхода из строя дискового контроллера или нескольких дисковых устройств одновременно возможна потеря данных. Однако при такой конфигурации доступна возможность горячей миграции ВМ в случае запланированных технических работ при помощи технологии vMotion: ВМ вместе со своим диском перемещается по сети на соседний хост без остановки работы. Следует особо отметить, что подобная миграция возможна лишь при полной работоспособности всех хостов ESXi и недоступна при авариях.

Организация выделенного сетевого хранилища (рис. 2) повышает отказоустойчивость. Диски ВМ располагаются на выделенном хранилище, а ВМ исполняются на одном из хостов. При такой конфигурации точкой отказа является хранилище. В случае отказа одного из хостов ВМ, которые исполнялись на нем, будут перезапущены на втором хосте. Время простоя этих ВМ составляет порядка нескольких минут, в зависимости от скорости загрузки операционной системы ВМ. Однако выход из строя хранилища или потеря сетевой связности с ним приведет к остановке всех без

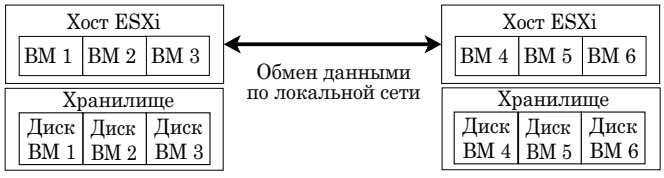

Рис. 1. Два хоста ESXi с локальными хранилищами (DAS). Возможность горячей или холодной миграции при отказе хоста отсутствует

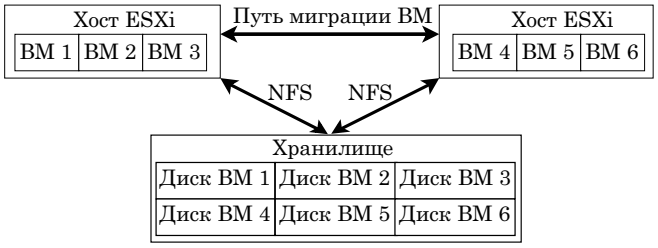

Рис. 2. Два хоста ESXi не имеют локальных хранилищ. Их единственная функция — исполнять ВМ. Файлы дисков ВМ хранятся на отдельном сервере (NAS), предоставляющем данные по протоколу NFS. В случае отказа одного из хостов ВМ перезапускаются на втором хосте

исключения ВМ, диски которых находились на этом хранилище. Вероятность такого отказа можно снизить, используя для организации хранилища конфигурации RAID с высокой избыточностью (RAID6 или RAIDZ-2), а также агрегацию сетевых каналов (например, с помощью LACP).

В ИПА РАН до 2013 г. использовалась схема «хост + локальное хранилище» (Сальников и др., 2012), так как количество используемых ВМ было относительно мало и объем используемых ими данных был незначителен. Также отсутствовала техническая возможность организовать выделенный сервер с большим объемом дискового пространства для создания хранилища.

В 2013 г. между площадками на наб. Кутузова, 10 и ул. Ждановской, 8 был организован выделенный канал связи с пропускной способностью 10 Гбит/с. Этой скорости достаточно для работы по схеме «хосты + сетевое хранилище». Сервер хранилища был размещен на второй площадке (Ждановская, 8). Опыт эксплуатации показал, что на этой площадке гораздо реже происходят перебои с электропитанием. Там же располагался один из хостов ESXi. Второй хост ESXi находился на первой площадке (наб. Кутузова, 10).

Развертывание схемы с двумя хостами и хранилищем позволило организовать миграцию ВМ с одного хоста на другой в случае плановых профилактических работ на какой-либо из площадок, а также снизить время простоя в случае потери связи с одним из хостов ESXi.

По мере эксплуатации возникла потребность в замене дисков хранилища, находившегося под серьезной нагрузкой в режиме 24/7 несколько лет. Дисковый контроллер и другие компоненты сервера хранилища сильно превысили свои гарантийные сроки работы. Было произведено несколько плановых остановок хранилища для замены вышедших из строя компонентов, что сопровождалось многочасовым простоем всех ВМ ИПА РАН. Стало очевидно, что необходим поиск схемы, которая исключит единую точку отказа инфраструктуры и позволит продолжать работу даже при полном отключении одной из площадок института.

Облачное хранилище vSAN

Облачное хранилище vSAN (VMware vSAN) предполагает организацию кластера, состоящего из структурно одинаковых узлов, выполняющих функции как хоста, так и хранилища. На рис. 3 представлена простейшая схема такого кластера, состоящего из двух узлов и хоста-свидетеля. Каждый узел содержит реплику дисков всех ВМ в кластере и исполняет некоторые из ВМ. Хост-свидетель отслеживает доступность и работоспособность хостов, а также синхронизацию реплик дисков ВМ между собой.

Конфигурация кластера vSAN, состоящая из двух хостов ВМ и хоста-свидетеля, называется «растянутым кластером». Все три узла кластера размещены в географически удаленных дата-центрах и объединены между собой скоростными каналами связи.

В ИПА РАН узлы кластера размещаются следующим образом: один хост ВМ — на первой площадке, один хост ВМ — на второй площадке, и хост-свидетель — в обсерватории «Светлое».

Миграция существующих виртуальных машин происходит без остановки их работы с помощью технологии vMotion. Виртуальные машины

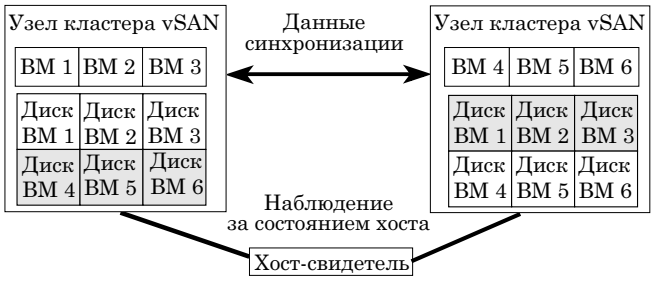

Рис. 3. Схема кластера vSAN. Выход из строя любого узла кластера не приводит к его отказу за счет избыточности. Каждый диск ВМ имеет копию (выделены серым), которая синхронизируется с оригинальным диском. Хост-свидетель в зависимости от состояния узлов определяет, какой экземпляр диска считать активным

перемещались вручную одна за другой, общее время миграции составило около двух суток.

Такая схема позволяет поддерживать работоспособность всех ВМ даже в случае полной потери связи с одной из площадок. При отказе одного из хостов ВМ хост-свидетель отмечает его как временно недоступный, и ВМ, работавшие на нем, перезапускаются на втором хосте. Так как диски ВМ полностью реплицируются, необходимости переносить данные с хоста на хост нет. При отказе хоста-свидетеля кластер продолжает работу в обычном режиме, однако рекомендуется запустить новый временный хост-свидетель для поддержания отказоустойчивости. Эта процедура занимает несколько минут.

Преимущества vSAN. High Availability и Fault Tolerance

Результаты сравнения возможностей и отказоустойчивости различных схем инфраструктуры представлены в таблице. По сравнению с классическими схемами организации инфраструктуры vSphere, vSAN устраняет единую точку отказа — теперь допустима полная потеря работоспособности целого дата-центра. В качестве дополнительных преимуществ можно отметить следующие:

1. High Availability (HA) — технология, позволяющая автоматически перезапускать ВМ в случае отказа хоста, на котором она исполнялась. Время простоя ВМ составляет несколько минут в зависимости от скорости загрузки операционной системы внутри этой ВМ. Эта технология также доступна и в конфигурации с выделенным сетевым хранилищем.

2. Fault Tolerance (FT) — технология, позволяющая мгновенно переключать исполнение ВМ с одного хоста на другой при отказе основного хоста. На втором хосте создается теневая копия работающей ВМ, синхронизирующая свое состояние с исходной ВМ. При отказе основного хоста хост-свидетель отмечает теневую копию как главную, в результате работа ВМ не прерывается.

Таблица

Сравнение возможностей и отказоустойчивости различных схем инфраструктуры vSphere

Схема vSphere	Единые точки отказа	Время восстановления работоспособности	vMotion	HA	FT
Хост + DAS	Хост, хранилище	Несколько часов; возможна потеря данных	Да	Нет	Нет
Хосты + NAS	Хранилище	Несколько минут при отказе хоста; несколько часов при отказе хранилища	Да	Да	Нет
vSAN	отсутствуют	Несколько минут при отказе хоста (затронуты только ВМ, исполнявшиеся на нем); 0 при использовании FT	Да	Да	Да

Недостатки vSAN

В схеме vSphere с использованием vSAN можно выделить следующие недостатки:

1. Используется только часть доступного дискового пространства. Каждый узел хранит полную реплику всех дисков ВМ, в результате полезное используемое пространство составляет лишь $1/N$ от общего объема, где N — количество хостов. В случае двух хостов это 50 % объема, в случае трех — уже 33 % и т. д.

2. Для обеспечения доступности ВМ в случае выхода из строя одного дата-центра необходимо построить распределенную сетевую инфраструктуру: все локальные сети, используемые в одном дата-центре, должны быть также доступны и в другом. В случае географически разнесенных площадок эта задача нетривиальна и требует настройки динамической маршрутизации или логического объединения маршрутизаторов двух дата-центров.

3. Аппаратное обеспечение узлов кластера, исполняющих ВМ, должно быть приблизительно идентичным — в частности требуется полная совместимость центральных процессоров; желательны сопоставимые объемы оперативной памяти и дискового пространства; требуется доступ к сети с пропускной способностью до 10 Гбит/с для синхронизации данных между хостами.

4. Требуется обязательное использование дисков SSD (3–4 на каждом хосте) для хранения кэша.

5. vSAN является проприетарной технологией, и каждый новый хост требует приобретения лицензии. Таким образом, стоимость добавления каждого хоста в кластер составляет несколько тысяч долларов США.

Результаты первого года эксплуатации

В течение года эксплуатации vSAN в ИПА РАН произошло несколько сбоев, потребовавших миграции виртуальных машин. Сбои были связаны либо с масштабным отключением электропитания на площадке (более 2 часов), либо с отключением площадки от сети. Во всех случаях технологии HA и FT обеспечили перенос виртуальной инфраструктуры на другую площадку, время простоя служб, использующих технологию HA, составило не более 15 мин. Потерь данных не произошло. Однако в некоторых случаях возникала проблема сетевой связности — не все подсети, используемые виртуальными машинами, присутствуют на обеих площадках. В результате некоторые виртуальные машины оказывались отключены от сети после миграции, хоть и продолжали работу. Решение этой проблемы запланировано в будущем при помощи технологий CARP или VRRP.

Выводы:

1. Технология vSAN хорошо себя зарекомендовала как средство повышения отказоустойчивости комплекса виртуальных машин.

2. Внедрение технологии vSAN обеспечило возможность проведения работ по обслуживанию и модернизации серверов виртуализации без остановки виртуальных машин.

3. Планируется дальнейшее использование технологии vSAN в информационной сети ИПА РАН.

Литература

Стэмпковский В. Г., Гаязов И. С. Использование технологии виртуальных машин в центре обработки и анализа данных ИПА РАН // Труды ИПА РАН. 2009. Вып. 20. С. 314–318.

Сальников А. И., Яковлев В. А., Безруков И. А. Применение технологии виртуальных машин в режиме e-РСДБ на радиоинтерферометрическом комплексе «Квазар-КВО» // Приборы и техника эксперимента. 2012. № 6. С. 30–34.

VMware vSAN Documentation [Электронный ресурс] URL: https://docs.vmware.com/en/VMware-vSAN/index.html (дата обращения: 03.08.2021).