VMware Virtual SAN Data Management Operations

July 2014 Edition

TECHNICAL MARKETING DOCUMENTATION
VERSION1.1

vmware®

Table of Contents

Introduction

One of the most popular topics of discussion about Virtual SAN revolves around solution sizing and performance capabilities. For the most part, the majority of guidance around Virtual SAN designs has been focused on capacity sizing and performance characteristics of virtual machine workloads.

However, there are other aspects of sizing and design criteria for Virtual SAN, specifically those related to system-wide performance and availability during data management operations. The data management operations of Virtual SAN are focused around data resynchronization and rebalancing amongst all the all copies of data. The functions and impact of these operations should be part of all Virtual SAN design and sizing exercises for optimal results.

The design of data management operations is intrinsic to the value proposition of Virtual SAN. It is important to know the events that activate them and also understand the impact they introduce during normal operations. Inadequate size and design can have an impact on the overall performance expectation and availability capabilities of the solution.

Virtual SAN Data Management Operations

This document provides the description of the Virtual SAN data management operation, and highlights their functionalities with respect to recoverability and performance characteristics.

- **Resynchronization**
- **Rebalance**

Resynchronization Operation

The resynchronization operation in Virtual SAN is activated when a device failure occurs, or when changes are made to a VM storage policy. There are two different types of failure events that are recognized by Virtual SAN, and it is important to know and understand the behavior of these events and the impact the resynchronization operation introduce. Below is a description of the failure events and their characteristics.

Hardware Device Failure Events

Virtual SAN hardware failures are recognized in two different types, absent and degraded failures. The details and characteristics of each failure event are described below:

Absent Failure

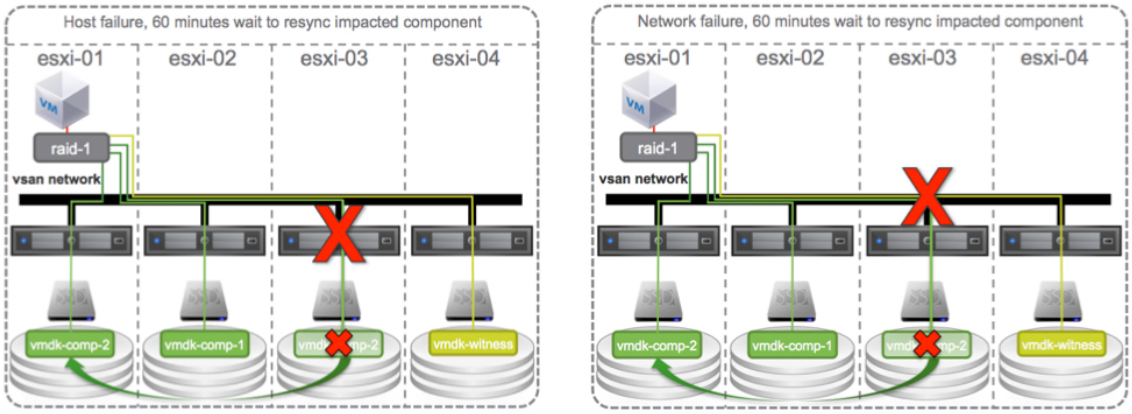
Absent failure events are recognized in Virtual SAN whenever I/O failures are detected with any of the following hardware devices:

- **Physical Network**
- **Network Interface Cards (NIC)**
- **Host Failures**

This type of a failure event activates the resynchronization operation 60 minutes from the detection and acknowledgement of the failure event. The logic behind this is that the types of failures listed above can often be transitory (e.g. network link down, host rebooting due to some other condition, etc.), and so the system waits for these components to come back online before committing to a resynchronization.

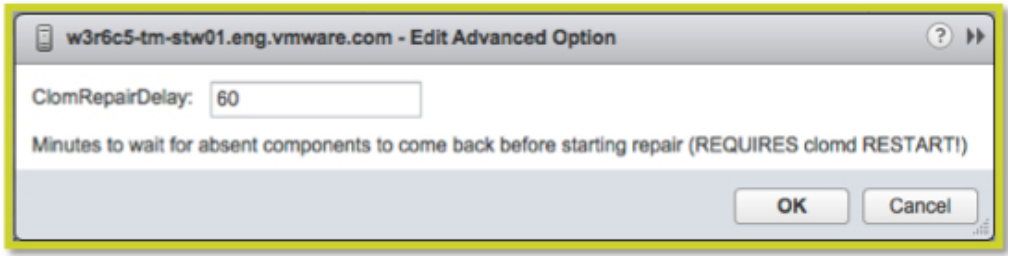
The default 60 minutes delay is an adjustable system parameter that can be increased or decreased.

Figure 1: Absent Failures



To modify the object and component resynchronization operation time for absent failure events, go to the advance setting of each host in the cluster and set a suitable time for your environment.

Figure 2: Absent Failure Advanced Setting



Note: KB article 2075456 provides instructions to modify the settings without having to restart the hosts.

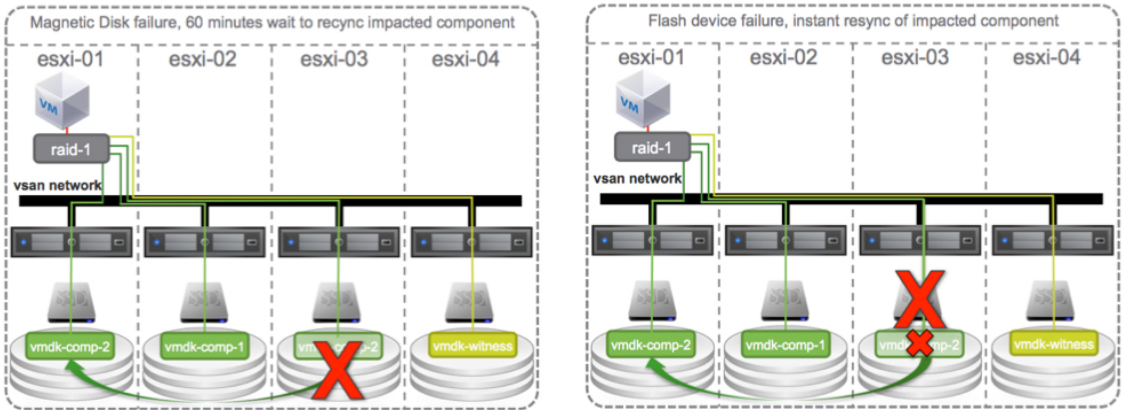
Degraded Failures

Degraded failure events are recognized in Virtual SAN whenever I/O failures are detected with any of the following hardware devices:

- **Magnetic Disks**
- **Flash Based Devices**
- **Storage Controllers**

Unlike absent failure events, degraded failure events immediately activate the resynchronization in the cluster operation, and is not configurable. The logic behind this is that when the devices listed above fail, it is not likely to be temporary; hence there is no need to wait for them to come back online.

Figure 3: Degraded Failures



The resynchronization operation is done for objects that are no longer in compliance with their policies due to a failure. The operation creates new replicas of the data that existed on the failed device or component, using the remaining replica (which exists elsewhere in the cluster, either on other hosts or on other magnetic disks of the same host) as the source.

The replicas of individual objects are not all created in the same place; rather, the replicas are distributed around the rest of the cluster wherever there is spare capacity. Thus, the entire cluster is used as a "hot spare".

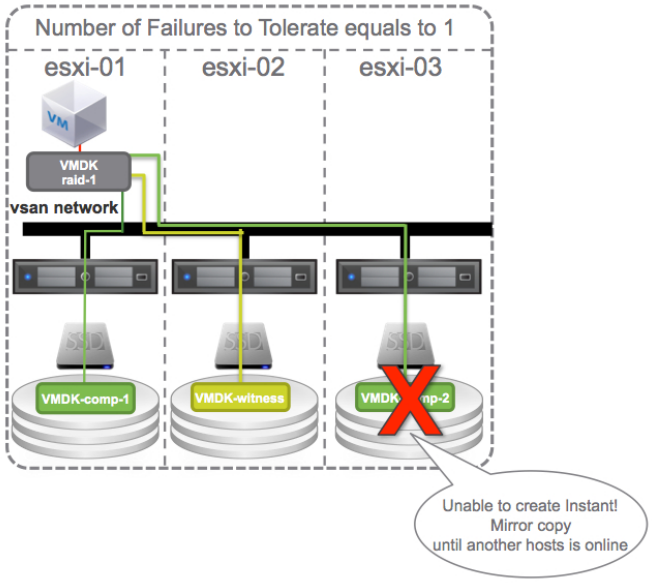
Regardless of when is activated, it contends with virtual machines and the resource available in the cluster. The operation could potentially have a detrimental impact to the overall capabilities of the Virtual SAN cluster if not sized and designed correctly. From a performance perspective, the resynchronization operation essentially limits the amount of IOPS available to virtual machines in the cluster because of the operations being performed to recover the affected objects and components.

From a data availability perspective, whenever the resynchronization operation is unable to be completed due to an inadequate number of hosts in a cluster, data accessibility could be at risk depending on the number of concurrent failures the cluster can support.

For example, in a scenario with a three node cluster is configured with the default availability policy setting of FTT=1. When a host failure occurs, the remaining two hosts are excluded from the resynchronization

operation because the remaining nodes are already hosting objects and components for the affected virtual machines.

Figure 4: Host Excluded From Resynchronization Operation



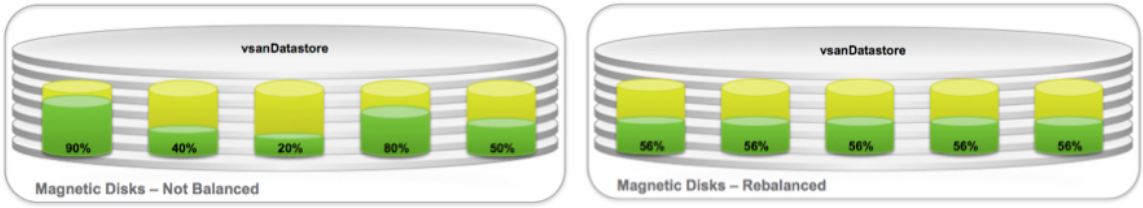
In this scenario, the resynchronization operation waits until the failed node is brought back online, or a new one is added to the cluster in order to resume the resynchronization operation and restore the compliance for data availability.

Rebalance Operation

The rebalance operation in Virtual SAN is designed to proactively re-distribute data throughout the cluster in order to maintain a balanced consumption and distribution of storage capacity.

By default, the rebalance operation is automatically activated whenever the storage capacity of the magnetic disks reaches 80% of utilization. The purpose of the operation is to distribute the data evenly throughout the cluster until the storage utilization is below the system's defined threshold of 80%.

Figure 5: Rebalance Operation



Operations such as maintenance mode and events such as hardware failures are responsible for activating

the rebalance operation. The vSphere host maintenance mode feature was modified in order to interoperate with Virtual SAN and adhere to its data management functionalities.

Two out of the three maintenance mode Virtual SAN data migration options activate the rebalancing operation:

- **Ensure accessibility** – partial data migration and distribution throughout the cluster.
- **Full data migration** – full data migration and distribution throughout the cluster.

Hardware failure events are also capable of activating the rebalance operation whenever the magnetic disks capacity utilization exceeds 80% due to resynchronization operation.

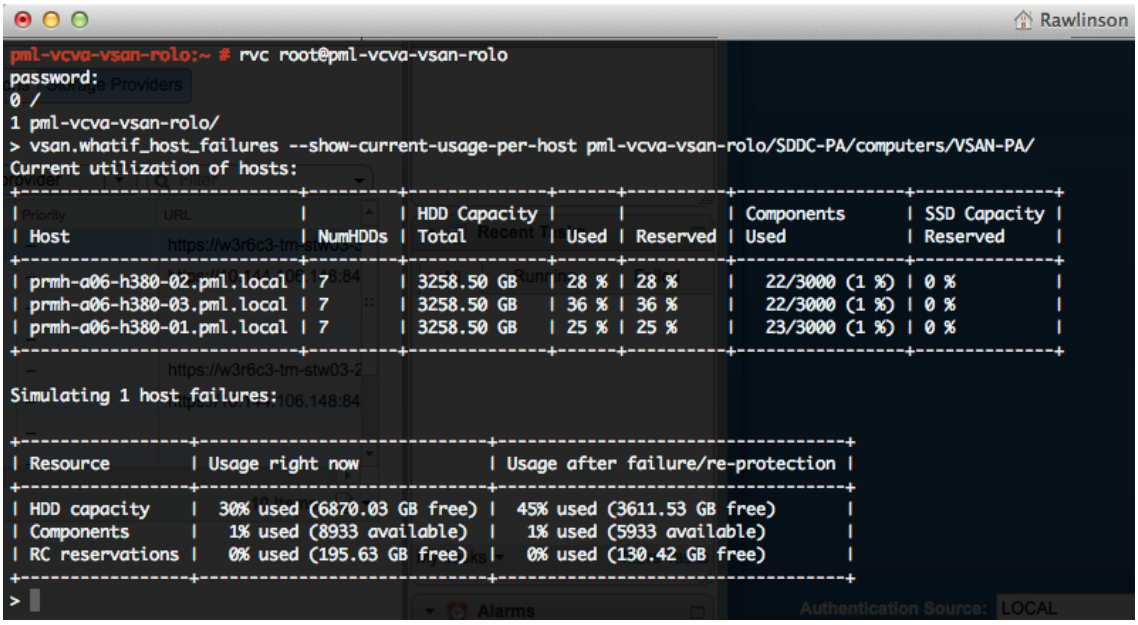
Adding new storage resources to the cluster does not automatically activate the rebalance operation. Newly added resources are recognized as additional capacity, and the capacity is only utilized by the system when new virtual machine are provisioned and resynchronization operations are initiated.

It is possible to force the system to perform a rebalance operation by manually placing a host into maintenance mode and choosing one of the options to migrate data throughout the cluster. This action will essentially force the utilization of the new available devices and storage capacity.

The RVC CLI tool can be used to identify and understand the storage resource capacity and consumption values as well as perform "what if" failure scenario calculations.

The "**vsan.whatif_host_failures**" command identifies the current storage resource utilization per host, and also performs simulation of host failures and their impact to the cluster.

Figure 6: RVC - vsan.whatif_host_failures



Recommendations

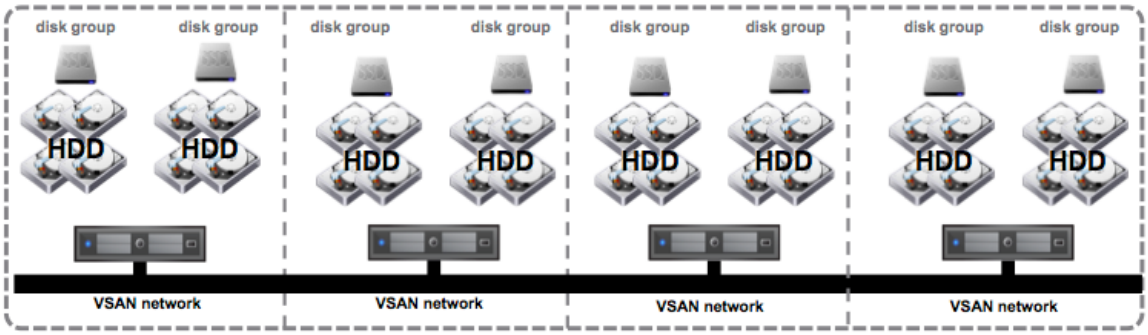
For better all around performance, availability, and recoverability results, VMware recommends the use of a 10GbE network as well as storage controllers with high queue depth supportability.

The use of a 10GbE network over 1GbE diminishes the impact introduced by the rebalance and resynchronization operations by enhancing the speed and time in which these operations are completed.

Also, the explicit use of storage controllers that provide support for queue depth of 256 or higher are recommended. The choice of storage controllers is crucial to sustain performance during resynchronization operations.

Queue depth of 256 or higher prevents the controller's IO queue from being overloaded, thus limiting IO operation time outs, high latency, and unresponsive virtual machines.

Figure 7: Wide Cluster and Disk Groups Design



Lastly, for faster resynchronization completion times, consider the deployment of wider cluster (larger number of nodes) configurations with multiple disk groups per host. Resynchronization operations will be performed in parallel onto all available hosts and disk group

Acknowledgments

Would like to thank Ankur Pai, Manager of VMware R&D Virtual SAN Team, Charu Chaubal, group manager of the Storage and Availability Technical Marketing team for reviewing this paper.

About the Author

Rawlinson Rivera is a Senior Architect in the Cloud Infrastructure Technical Marketing Group at VMware focused on Software-Defined Storage technologies primarily responsible for Virtual SAN, Virtual Volumes, and OpenStack. As a previous Architect in VMware's Cloud Infrastructure & Management Professional Services Organization, Rawlinson specialized on vSphere and Cloud enterprise architectures for VMware's fortune 100, 500 customers.

Rawlinson is amongst the few VMware Certified Design Experts (VCDX#86) in the world, and author of multiple books based on VMware and other technologies.

Follow Rawlinson’s blogs:

- http://blogs.vmware.com/vsphere/storage
- http://www.punchingclouds.com

Follow Rawlinson on Twitter:

- @PunchingClouds