

VMware® VMFS Volume Management

VMware Infrastructure 3

The VMware Virtual Machine File System (VMFS) is a powerful automated file system that simplifies storage management for virtual machines. VMFS also serves as a cluster volume manager to ease the management of shared storage resources within a cluster of VMware ESX hosts. As a volume manager, VMFS serves as the interface for virtualization of several storage options that a storage area network provides. These options include supporting snapshots that might be made by the storage subsystems, changing disk serial numbers, and presenting a given disk to two separate servers as different LUN numbers. VMFS can handle these different cases with a few settings that enable VMFS volumes to be assigned new volume signatures. These settings enable copies of VMFS volumes to be mounted, ignored, or given an updated signature so they are recognized as new VMFS volumes.

The behavior of these logical volume manager (LVM) settings is not always intuitive. You make the settings for individual ESX hosts. To make the settings correctly, you must understand the reason a particular volume is being detected as a copy and how changing the LVM advanced settings affects the VMFS volumes. When you understand the types of changes in the underlying storage configuration that cause ESX to detect a mismatch between what is calculated for the VMFS volume metadata and the metadata stored in the VMFS volume header, you can make a more logical choice of which setting to select. This paper addresses those options and outlines best practices in setting these advanced logical volume settings to give the desired outcome.

If you use third-party storage snapshot and replication capabilities in conjunction with ESX and VMFS, you must understand what VMFS volume manager settings to use and under what conditions you might want to leverage the use of the third-party copy and replication technology. In addition, you need to understand what effects using these VMFS logical volume advanced settings might have on the future status of the datastore that resides on the VMFS volume.

This guide explains the mechanics of VMFS volume header metadata and how that metadata is related to using the volume signatures and resignature options available in VMware Infrastructure 3. It explains why you might need the ability to change the default settings. It also explains some additional reasons that identifying a given volume can be a complex task.

This guide covers the following topics:

- [“VMFS Volume Creation and Identification”](#) on page 2
- [“VMFS Volume Manager Management Mechanics”](#) on page 2
- [“VMFS Volume Signature Settings”](#) on page 3
- [“Causes of a VMFS Volume Metadata Mismatch”](#) on page 4
- [“Handling Array Snapshots with VMFS”](#) on page 4
- [“Handling Disk Signature Changes”](#) on page 5
- [“Handling Array-based Replication with VMFS”](#) on page 5
- [“Handling Disk Signature Changes”](#) on page 6

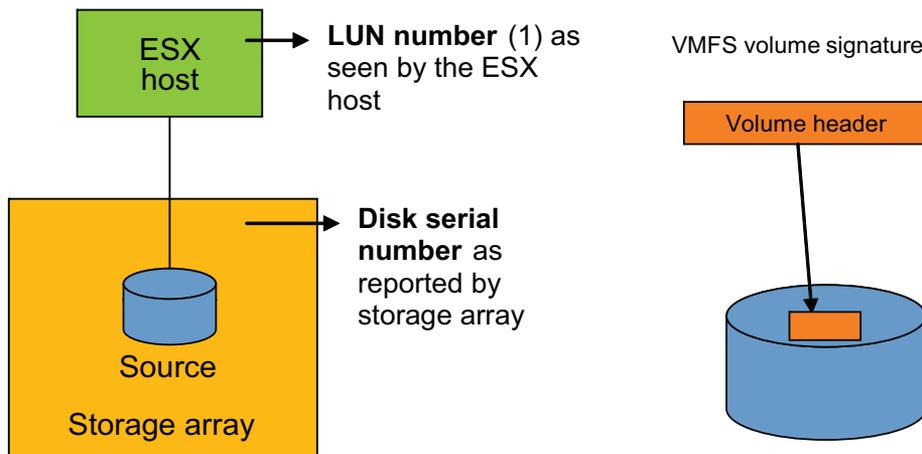
- “Best Practices Summary” on page 7
- “Conclusion” on page 7
- “Appendix: VMFS Volume Manager Terminology” on page 7

VMFS Volume Creation and Identification

As a volume manager, VMFS needs to manage a pool of storage resources (block devices) that are used to create datastores. VMFS is both a volume manager and a file system that is laid down on the aggregate of storage resources made available to either a single ESX host or a cluster of ESX hosts. When you define a new VMFS volume on a block device, VMFS calculates a volume signature and writes it in the header of the VMFS volume. This volume signature is a combination of the LUN number, as seen by that ESX host, and the disk serial number as the storage array reports it to the ESX host.

VMFS volume signature = LUN number + disk serial number

Figure 1. VMFS Volume Signature—Calculated and Stored



VMFS also uses disk signatures to keep track of block devices (LUNs) that are members of VMFS volumes it manages. These signatures are tracked by the VMFS instance on the ESX hosts that access the storage resources. With one single ESX host that has one VMFS volume containing a single extent, there is not much confusion. However, once you adds more variables, the complexity grows.

The variables can include:

- Multiple ESX hosts accessing the same storage resources (LUNs)
- Multiple copies of the LUNs (third-party storage snapshots)
- Changes to settings in the storage array that changes the LUN
- Presentation of the same LUN to two hosts as different LUN numbers

VMFS Volume Manager Management Mechanics

After a VMFS volume is created, other ESX hosts can discover it if they are in the same cluster as shown in the VI Client and if they share access to a common pool of storage. The VMFS volume manager on another ESX host can scan the storage resources and determine which of the LUNs it finds belong to which VMFS volumes. It is important for the VMFS instance to identify which volumes are which to avoid corruption of data and to prevent the improper use of copies instead of the original. When VMFS rescans the storage resources available to it, it checks to see if the VMFS metadata stored in the volume header matches the metadata it calculates. In short it compares what it finds from calculating the volume signature with what was written to disk (Figure 1) If there is a mismatch, the volume might not be the original copy that was there at the time it was first created.

VMFS Volume Signature Settings

To address the complexity in the way the VMFS volume manager sees a set of shared storage resources, VMFS provides settings you can use to eliminate possible confusion.

A few LVM options affect the way VMFS handles mismatches between observed and stored VMFS disk signatures. You can view and modify these settings from the command line or in vCenter under the advanced LVM settings for a particular ESX host.

The `LVM.DisallowSnapshotLUN` option controls whether the VMFS volume manager checks for snapshots. You can set it to either 1 or 0. The default is 1. Changing the value to 0 enables the VMFS volume manager to mount LUNs even when the LUN number and disk signature do not match. Setting this value to 0 turns off the check for snapshots and could allow the ESX host to find duplicate copies of the same VMFS volume and use them as if they were the same volume. This could lead to corruption of data on the VMFS volume, thus you should not use this setting without considering the full impact of the change. You should change this value to 0 only if you are certain only one copy of the volume is presented to the ESX host.

The `LVM.EnableResignature` option controls whether a discovered disk in a VMFS volume should be assigned a new VMFS disk signature. You can set it to 0 or 1. The default is 0. If you change the value to 1, a discovered disk in a VMFS volume is assigned a new VMFS disk signature and the newly identified VMFS volume is mounted by the ESX host as a new datastore. All virtual machines on that newly discovered and mounted datastore must then be unregistered and registered again in the vCenter Management Server. All paths in `/vmfs/volumes/<old UUID>` are changed to `/vmfs/volumes/<new UUID>` and the datastore name for the VMFS volume with a new signature is changed from `<label>` to `snap-<xxxxxx>-<label>` where `<xxxxxx>` is a system-generated number for each resignature done for that VMFS volume.

Changing `LVM.EnableResignature` from 0 to 1 overrides the setting of `LVM.DisallowSnapshotLUN` when it is set to 0. Specifically, if you give a new signature to a VMFS volume that is detected as not being the original, and if at the same time you allow snapshots to be mounted, VMFS gives a new signature to the VMFS volume instead of allowing the duplicate to be mounted without changing the signature.

Table 1. Summary of LVM Advanced Settings and the Resulting Actions

Disallow Snapshot	Resignature	Action
0	0	Mount VMFS volume without new signature
0	1	Give new signature to VMFS volume and mount
1	0	Do not mount VMFS volume
1	1	Give new signature to VMFS volume and mount

A new setting, introduced in ESX 3.5, is `SCSI.CompareLUNNumber`. It controls whether VMFS considers the LUN number when it compares the LUN to the value found in the VMFS volume metadata. You can set it to either 1 or 0. The default is 1. The default setting specifies that the VMFS does consider the LUN number when it compares the LUN to what is found in the VMFS volume metadata as it determines whether there is a mismatch. You can change this setting in ESX Server Advanced Settings (on the Configuration tab) in VI Client. Changing the value to 0 disables consideration of the LUN number when the metadata is compared to what is returned from the discovery of a given LUN by an ESX host. When the setting is 0, only the disk serial number is compared to determine if a mismatch exists.

Another feature, introduced in ESX Server 3.1.x, is use of NAA (Network Address Authority) identifiers for storage arrays that export this standard. In that release and subsequent releases, if the array supports per-LUN NAA UUIDs, the LUN numbers exported to the ESX host are not compared to detect a possible snapshot.

Causes of a VMFS Volume Metadata Mismatch

Many storage area networks (iSCSI and Fibre Channel) can provide a rich set of features and variables that affect the way a given LUN is presented, copied, or replicated. As a result, you must determine whether a copy that the ESX host detects as a snapshot is really a copy of the original LUN or whether it appears to be a snapshot because some other variable changed, thus changing the way the LUN is presented to the ESX host. More specifically, you need to understand what change caused the mismatch in order to avoid further issues with access to the shared storage resource.

In a number of cases, the ESX host sees a mismatch between the volume signature it calculates and the volume signature written in the volume header. The possible causes of this mismatch include the following:

- The disk is a snapshot copy of a device served by the storage array.
- The disk serial numbers (SCSI IDs) have been changed (by changing array settings).
- The LUN number is not the same as the LUN number on the ESX host where the VMFS volume was originally created. A mismatch is detected in this case only when:
 - The arrays do not use NAA and are connected to a host running an ESX version earlier than 3.5.
 - The SCSI.`CompareLUNNumber` option is set to 1 (the default) for arrays not using NAA that are connected to a host running ESX 3.5 or a later version).

If your objective is to have a snapshot presented to a second ESX host and mounted for use there, and you never want to allow it to be seen by the original ESX host, you can allow snapshots to be mounted by turning off the `LVM.DisallowSnapshotLUN` option (change the setting from 1 to 0).

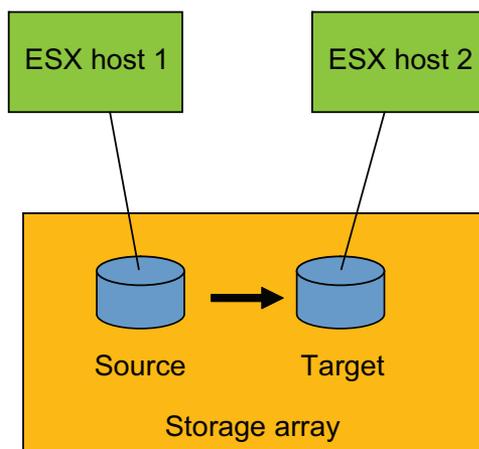
You might also want to use this option if you are replicating an entire VMFS volume to a secondary site.

Handling Array Snapshots with VMFS

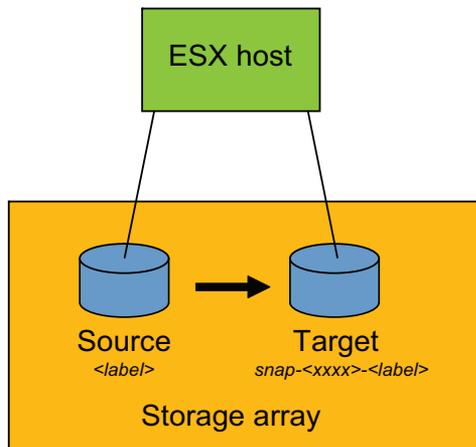
When a storage array takes a snapshot, the array assigns a new disk serial number (disk UUID) to the snapshot copy. That change causes VMFS to recognize that the VMFS disk ID written to the disk header does not match what the VMFS instance calculates when it scans the storage resources.

If the snapshot copy of a LUN, created by the storage array, is not visible to the original ESX host and can be seen only by a second server (Figure 2), you can use the `LVM.DisallowSnapshotLUN` setting to enable the second ESX host to mount the VMFS volume. For best results, ESX host 1 should be managed by a different vCenter Management Server than ESX host 2 to avoid confusing VMFS disk IDs.

Figure 2. Snapshot Copy Not Visible to Original ESX Host



However, if the ESX host connected to the source version of that disk (LUN) can also access the snapshot copy of the disk, as shown in Figure 3, you must give the snapshot copy a new volume signature. You can then mount both source and copy VMFS volumes on the same ESX host.

Figure 3. Snapshot Copy Visible to Original ESX Host

Handling Disk Signature Changes

Some arrays do not present the same disk serial number for a particular disk when it is in a shared storage configuration. They present one serial number to the first host and a different serial number to the second host. In addition, certain changes to disk array settings can cause the array to assign new disk serial numbers. Rebuilding a RAID group on some arrays can cause this to occur, as well. When VMFS encounters different serial numbers in any of these circumstances, it treats the existing VMFS volume devices as snapshots because the calculated disk signature does not match the value stored in the VMFS volume header.

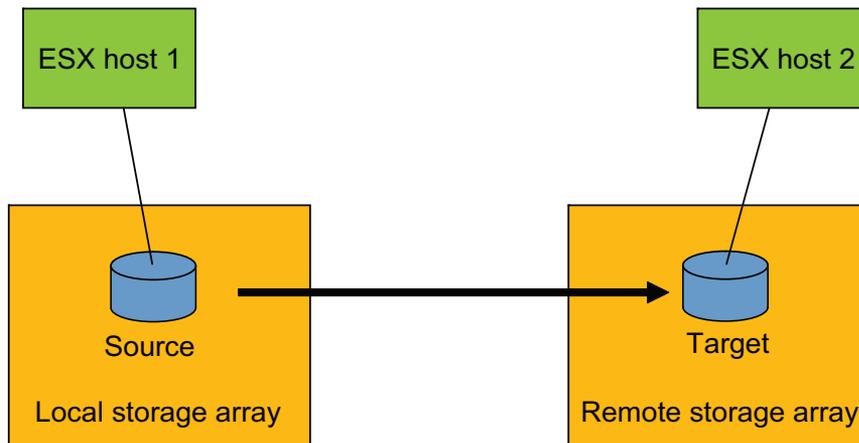
In such cases, you can allow the ESX host to mount the volume that VMFS on that host sees as a snapshot, or you can give the volume a new signature. The disadvantage of giving the volume a new signature is that you may need to use the vCenter management interface to unregister and reregister all the virtual machines on that volume. The reregistration is necessary because the VMFS volume is considered a new datastore and thus the virtual machines are considered new virtual machines. However, in some instances, giving the volume a new signature is the only option as a result of the way VMFS detects the identity of the members of a VMFS volume.

Handling Array-based Replication with VMFS

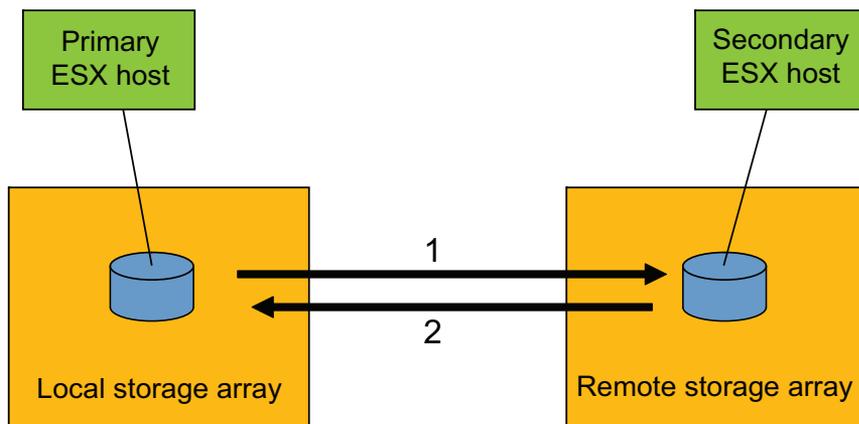
With array-based replication, you have two options, similar to those you have with local array-based snapshots:

- Allow the snapshot to be mounted.
- Change the VMFS volume signature before mounting the VMFS volumes at the remote site.

For most one-way replications schemes (Figure 4), you can use the option to allow snapshots to be mounted. However, if the ESX hosts at the remote site share an extended storage area network that enables an ESX host at the primary site to see the replicated VMFS volumes at the remote site, you should use the Enable Resignature option. You should take this approach because remote access is similar to case when a snapshot is seen by the server attached to the source volume (Figure 3), and incorrect configuration can lead to data corruption.

Figure 4. One-way Replication to a Remote Site

When using replication to a remote site that has the option of failing back to the primary site, you should not give a new signature to the volumes so that, in case of failback, virtual machines do not have to be reregistered in vCenter at the primary site. This approach assumes there is no risk of an ESX host at the primary site seeing both the remote copy and the local copy at the same time. That could lead to data corruption.

Figure 5. Replication to a Remote Site with Failback Option

NOTE If you are using raw device mappings (RDM), you must address replication of RDMs to a remote site because these device maps also depend on LUN number and disk serial number (SCSI ID). The VMFS volumes on the remote site need to have newly correlated RDMs mapped for the new LUN numbers and disk serial numbers for those replicated RDMs.

Handling Disk Signature Changes

It is considered best practice to present a shared LUN as the same LUN number to all ESX hosts in a cluster. If shared LUNs are not configured correctly, some ESX hosts see the VMFS volume as a snapshot because of the mismatch between what the ESX host calculates for the VMFS volume and what it reads from the volume header. The best solution is to have the storage administrator eliminate this inconsistency in the array configuration.

If you are using ESX 3.5 and arrays that support the NAA addressing scheme, you can resolve this issue by configuring the ESX host to use the NAA disk ID to compare disk identity instead of using the disk serial number and LUN number. By using NAA disk IDs, you can lessen the chances of a false mismatch.

Best Practices Summary

The following best practices can help you configure VMFS for best results in your environment:

- When starting out, use the default settings for both `LVM.DisallowSnapshotLUN` and `LVM.EnableResignature`.
- If you encounter issues that require addressing a mismatch in VMFS metadata and the observed VMFS disk signature values, make sure you understand the reason VMFS detects a mismatch for those volumes. When you understand the cause, make sure to change the proper settings to achieve the behavior you intend.
- If a snapshot created by your storage array is not addressable by the original ESX host (as in [Figure 2](#)), you can enable mounting of snapshot LUNs on the secondary ESX host so the snapshot can be mounted and available for use.
- If the snapshot is to be accessed by an ESX host that also accesses the source disk, you must give a new volume signature to the snapshot.
- If the disk array has provided separate disk serial numbers for the same disk to separate ESX hosts or changed the disk serial number after the VMFS volume was created, you can either allow snapshots or give the volume a new signature. If you give the volume a new signature, you must reregister each of the virtual machines on that datastore with vCenter.

NOTE If you use a resignature option, you should disable it after you have changed the signature of the desired volumes. Otherwise, you might get additional, unexpected changes later.

Conclusion

VMFS is a powerful interface for managing shared storage resources for virtualization environments. It has a cluster volume manager capability that enables virtualization to leverage array-based snapshots and replication functions in a manner that provides maximum flexibility. VMFS compliments storage-based features to maximize the benefits of enterprise environments. It offers flexible settings to serve storage objects to a virtualization environment in a manner that contributes to increased flexibility of storage virtualization. When you understand them, the volume manager functions can turn a diverse set of storage resources into a very easy to manage and highly productive shared resource pool.

Appendix: VMFS Volume Manager Terminology

This appendix defines key terms used to describe components of a logical volume.

Datastore A formatted file system that is layered on top of either a VMFS volume with block-based storage (iSCSI and FC) or a mount point for NFS storage. It is a shared storage resource in which VMhomes, virtual disks, and VM objects are stored.

Extent See “[VMFS volume extent](#)” on page 8.

LUN A single block storage allocation presented to a server. This logical unit number is the unique identification a host has assigned to a given block device resource (disk) it finds when it scans the storage array network. The term disk is often used interchangeably with LUN. From the perspective of an ESX host, a LUN is a single unique raw storage block device or disk.

ESX LUN number The host-based number assigned to the storage unit as seen by the ESX host addressing the storage. In most cases this can be set by the array and should be consistent across a set of ESX hosts that share a common pool of storage resources.

Disk serial number A unique number assigned to a LUN by the storage array for identification by hosts. This number is also called a SCSI disk ID, SCSI ID, LUN ID, or LUN UUID.

NAA disk ID National Address Authority disk identification that is presented by most storage arrays.

VMFS volume extent VMFS volumes can be made up of one or more disks (LUNs). Each disk (LUN) is called a VMFS extent. In many cases there is a 1:1 ratio of LUNs to VMFS volumes. However, some VMFS volumes have many extents, and in some rare cases several VMFS volumes might exist on a single LUN.

VMFS volume manager The process that scans, sorts, and manages status, membership, and protection of underlying components that are presented to an upper layer as a single resource.

VMFS volume signature A unique identification assigned to a given storage resource and used by VMFS to identify the disk as unique. It is a combination of the LUN number and the disk serial number.

VMFS volume A collection of storage resources managed as a single shared resource. In most cases the VMFS volume contains a single LUN. In those cases the datastore and the VMFS volume are identical. However, in some cases the VMFS volume might span two or more LUNs and be composed of multiple extents.

If you have comments about this documentation, submit your feedback to: docfeedback@vmware.com

VMware, Inc. 3401 Hillview Ave., Palo Alto, CA 94304 www.vmware.com

Copyright © 2009 VMware, Inc. All rights reserved. This product is protected by U.S. and international copyright and intellectual property laws. VMware products are covered by one or more patents listed at <http://www.vmware.com/go/patents>. VMware, the VMware "boxes" logo and design, Virtual SMP, and VMotion are registered trademarks or trademarks of VMware, Inc. in the United States and/or other jurisdictions. All other marks and names mentioned herein may be trademarks of their respective companies.

Revision: 20090303 Item: IN-086-INF-01-01
