Steve Garraway
Jon Tate

VMware Multipathing with the SAN Volume Controller and the Causes of SCSI-2 Reservation Conflicts

Introduction

In order to provide a reliable connection between VMware ESX hosts and their disk storage, multiple paths through the SAN are required. This is accomplished by using multiple host bus adapters (HBAs) on the host, multiple ports on the disk subsystem, and redundant routing through the switched SAN fabric. ESX provides *multipathing* support to understand and exploit the resulting multiple paths between host and storage.

To maintain integrity during updates by way of multiple paths, locking mechanisms are required to control access. When performing *metadata* updates (such as creating a new virtual machine), VMware ESX will utilize SCSI-2 reserve/release commands for LUN-level locking. This has the potential to reduce I/O performance if reservation conflicts occur.

This IBM® Redpapers™ publication will explain the variations in multipathing support for different types of storage controller, and explain where reservation based locking is used.

Storage controller types

Storage subsystems normally have dual controllers to provide redundancy against controller failure. These controllers can either be active/active or active/passive.

Active/Active

With this type of subsystem, all LUNs are accessible by either controller simultaneously with equal performance.

Active/Passive

With active/passive, LUNs are only accessible through one (active) controller, with the other (passive) controller taking over if the active controller fails.

To avoid wasting the capabilities of the passive controller, some disk subsystems split the LUNs so that controller A is active for half the LUNs, while controller B is active for the remaining LUNs. The two controllers then act as passive backups for the opposing set of LUNs.

SAN Volume Controller

While the SAN Volume Controller (SVC) assigns VDisks to a preferred node within an I/O group, there is only a small performance penalty in accessing the VDisk by way of the non-preferred node. As such, the SVC can be viewed as an active/active disk subsystem.

> **Tip:** As of Version 4.3.1.5, there is no performance penalty in accessing the VDisk by way of either node.

Multipathing and failover

Multipathing and failover are two related concepts, which together provide improved availability through a fault tolerant I/O design.

Purpose of multipathing

VMware supports *multipathing* in order to maintain the connection between the host and its storage in the event of a failure of an HBA, SAN switch, SAN cable or storage port. This requires at least two instances of each of these components, as shown in Figure 1.

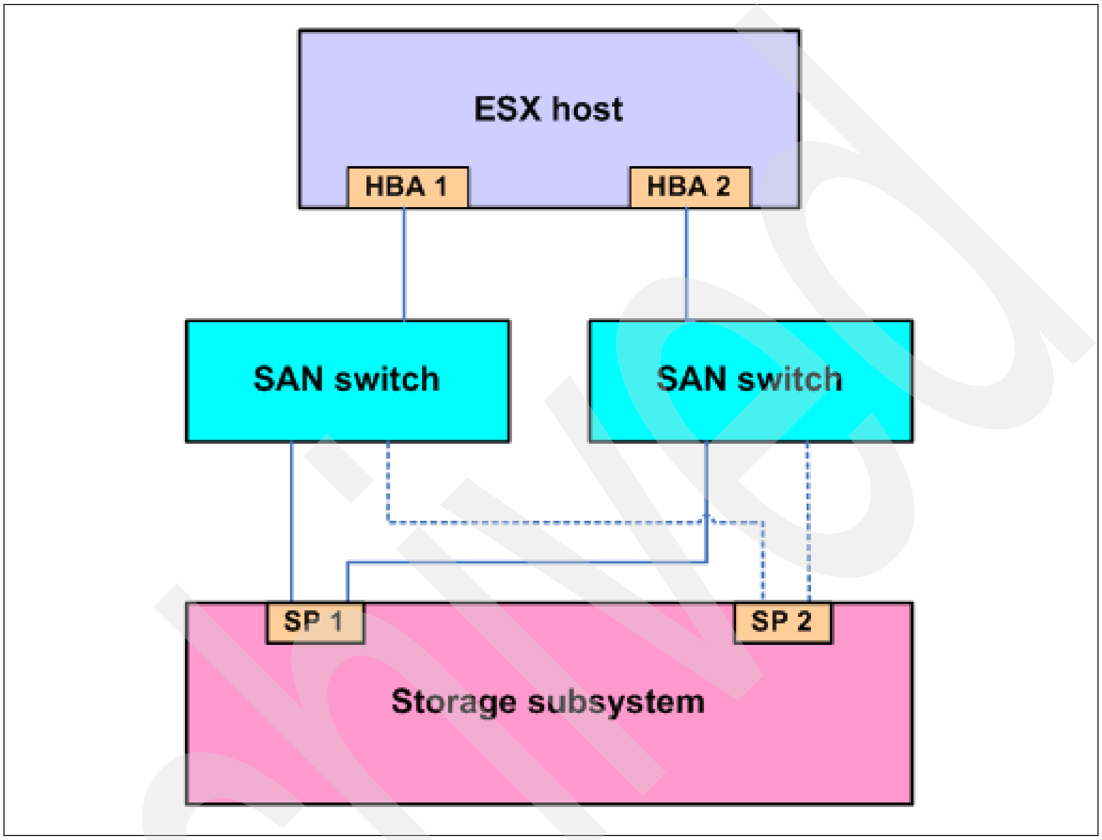

Figure 1 Single host with multiple storage paths

Without multipathing support, this configuration would cause a disk LUN to appear to the host as four different LUNs, and there would be no data integrity.

Failover

The term *failover* refers to a redundant component taking over when the active component fails. An example of HBA failover would be HBA 2 in Figure 1 taking over I/O from HBA 1 if it failed.

Pauses during failover

Depending on the type of failover, and the configuration of the hardware involved, failover processing can cause a significant delay in I/O processing while the SAN fabric stabilizes or the disk subsystem recovers. VM operating systems should be configured to handle a delay of up to 60 seconds, which in the case of Windows® may require adjusting the SCSI TimeOutValue, as described in the section “Setting Operating System Timeout” of Chapter 6 of the *Fibre Channel SAN Configuration Guide*, found at:

http://www.vmware.com/pdf/vi3_35/esx_3/r35u2/vi3_35_25_u2_san_cfg.pdf

Policies

VMware ESX prior to vSphere 4 supported two multipathing policies, namely *Most Recently Used (MRU)* and *Fixed*. vSphere 4 added support for the *Round-Robin Load Balancing* option.

The ESX server selects the appropriate policy based on the make and model of disk subsystem detected. If the detected disk is not recognized, it is treated as active/active. Refer to the **Storage/SAN** tab on the following Web page for support information:

http://www.vmware.com/resources/compatibility/search.php

Note: All VMware native multipathing policies only use one path at a time between the host and a given LUN.

MRU policy

ESX selects the path most recently used and only switches to another path if the active path fails. It then continues using the new active path until it fails or an administrator takes action to cause ESX to switch back to the original path.

This policy is required for active/passive storage subsystems.

Prior to vSphere 4, ESX defaulted to the MRU policy for SVC. As of vSphere 4, ESX will default to the Fixed policy.

Tip: If you wish, you can override the pathing policy at the individual LUN level in ESX to use Fixed pathing to the preferred node for the VDisk.

Potential performance impact

It is quite possible when using MRU pathing that all I/O will default to the same host HBA and the same disk subsystem target port for all LUNs in the subsystem (normally this is the first discovered HBA and target). This is obviously not optimal for performance.

It is possible to change the designated preferred path for each LUN to distribute I/O load across HBAs and target disk ports. To learn how to view the current path status for a LUN, and change it (if necessary), refer to the section “Multipathing” in Chapter 6 of the *Fibre Channel SAN Configuration Guide,* found at:

http://www.vmware.com/pdf/vi3_35/esx_3/r35u2/vi3_35_25_u2_san_cfg.pdf

Important: With ESX V3.5, there must be no I/O going to the selected storage volume when this operation is done or it will not work.

Fixed policy

ESX uses the designated preferred path unless it fails, at which time it will switch to an alternate path. If the preferred path recovers, ESX will switch back to it.

This policy is recommended for active/active storage subsystems, and as of vSphere 4 it is the default for SVC. For any versions prior to vSphere 4, use the default option as well.

Note: We do not recommend using the Fixed policy for active/passive storage.

Potential performance impact

It is also quite possible when using Fixed pathing that all I/O will default to the same host HBA and the same disk subsystem target port for all LUNs in the subsystem. Again, this is obviously not optimal for performance.

As with the MRU policy described in "MRU policy" on page 4, it is possible to change the designated preferred path for each LUN to distribute I/O load across HBAs and target disk ports using the same solution.

Managing paths in this way is not ideal for a large number of LUNs and paths, as it requires a great deal of manual work.

Round robin policy

ESX will automatically rotate path allocation to individual LUNs across all available paths in order to spread I/O load across all paths with the goal of optimizing performance. If the host in Figure 1 on page 3 had eight LUNs, the paths would be allocated in a similar way to that shown in Table 1.

Table 1 Round robin path allocation example

HBA	Storage port	LUNs
1	1	1 and 5
1	2	2 and 6
2	1	3 and 7
2	2	4 and 8

If a path fails, ESX will switch to another working path. There is no need for fail back with this policy.

It is also possible to configure ESX to switch paths after a certain amount of I/O has happened. See the following VMware technical note for full details of this task:

http://www.vmware.com/resources/techresources/1016

Consideration: Round robin is not supported prior to vSphere 4 for production use.

Potential performance gain

Round robin pathing should greatly alleviate the potential performance issues with the other two pathing policies, as it should cause all HBAs and target ports to be used for disk access. It also has the benefit that it simply has to be enabled for each LUN, without the need to manually choose the initial paths.

For details about how to select this policy for a LUN, refer to the sections "Path Scanning and Claiming" and "Path Management and Manual, or Static, Load Balancing" in Chapter 6 of the *Fibre Channel SAN Configuration Guide*, found at:

http://www.vmware.com/pdf/vsphere4/r40/vsp_40_san_cfg.pdf

Although this algorithm will distribute I/O for multiple LUNs across all available paths, it does not perform true load balancing (where I/Os to the same LUN would be routed down the best performing path available each time).

vSphere 4 and the Pluggable Storage Architecture

vSphere 4 introduces a new layer in the VMkernel called the *Pluggable Storage Architecture* (PSA), which is responsible for managing storage paths. It also provides an open framework supporting multipathing plug-ins to provide enhanced multipathing. The default plug-in is VMware's Native Multipathing Plugin (NMP), which assigns the appropriate path algorithm for all storage in the VMware Hardware Compatibility List (HCL). Use the Web page shown in "Policies" on page 4 to determine the default policy for different storage subsystems.

The details of handling path failover for a storage array are controlled by the Storage Array Type Plugin (SATP), while the choice of which available path to use for I/O is handled by the Path Selection Plugin (PSP).

The SATP includes the following features:

- Array specific actions for failover
- Path health monitoring
- Reporting path state changes to the NMP

The following PSPs are available by default:

- MRU
- Fixed
- Round robin

In addition, storage vendors can supply their own SATP and PSP modules, which can enhance the multipathing support for their disk arrays. One enhancement could be the implementation of true load balancing at the individual LUN level.

Metadata updates

Since VMFS is a shared file system, it is necessary to use locks to serialize disk updates. For normal I/O, this is handled at the file level, and therefore we avoid conflicts at the LUN level. Metadata updates require the use of a SCSI-2 reserve command to temporarily lock the entire LUN. The reserve is only held long enough to obtain a metadata lock, and is then released.

The following is a list of common metadata actions:

- Creating a VMFS datastore
- Expanding a VMFS datastore onto additional extents
- Powering on or off a virtual machine
- Acquiring a lock on a file
- Creating or deleting a file
- Creating a template
- Deploying a virtual machine from a template
- Creating a new virtual machine
- Migrating a virtual machine with VMotion
- Taking Snapshots
- Growing a file, for example, a Snapshot™ file or a thin provisioned virtual disk

Note: The impact of such metadata operations applies to all disk subsystems, not just the SVC.

SCSI-2 reservations

A SCSI-2 reserve can only be held by one host on a single path[1], which prevents other paths from the same or another host from accessing the reserved LUN. This guarantees the holder of the reserve exclusive access to the LUN to perform updates with data integrity.

Reserve retries

If a host attempts to access a reserved LUN, it reports a SCSI reservation conflict. It is normal to see these conflicts, but if the reservation lasts too long, then I/O will start failing. As such, reserves should be held for as short a time as possible, and then released to reduce their impact on other host I/O.

When a host experiences a reservation conflict, it will retry the I/O up to 80 times, with a randomized interval between each retry. If the number of retries is exhausted, the I/O will fail. For each retry, the VMkernel log shows the following message, detailing the number of retries remaining:

```
Sync CR at <number>
```

Recent ESX releases have introduced new algorithms to reduce the number of locking activities, and therefore reduce the occurrence of SCSI reserve conflicts. However, these improvements can be offset by using larger LUNs containing more VMs, as more VMs are likely to result in more metadata operations being performed against the LUN.

Note: We recommend that operations such as those listed above be performed at less busy times, and that only one such operation be performed at a time.

Recommendations for avoiding reservation conflicts

If conflicts are causing poor performance, then the following actions should be considered:

- Serialize metadata operations so that they are not attempted concurrently.
- Perform metadata operations outside of peak hours.
- Try performing metadata operations from a different ESX host.
- If possible, reduce the number of snapshots, as these generate a great deal of SCSI-2 reserves.

If none of these actions help, or are not possible, then change your LUN configuration to allocate more LUNs of a smaller size in order to reduce the number of VMs per LUN, and thus the potential for conflict. Spreading the workload over more LUNs may also improve performance by increasing the number of I/O queues used.

Summary

While VMFS will scale well to many VMs on a single VDisk for normal I/O, metadata actions that cause SCSI-2 reserves against the VDisk have the potential to cause significant performance problems if they occur frequently enough. Whenever possible, these operations should be scheduled so that they occur outside peak hours and do not conflict with each other.

[1] A path is the connection between a single initiator (host HBA port) and a single target (storage subsystem port).

A policy for sizing LUNs used as VMFS extents should ideally be based on the peak I/O rate and the expected rate of metadata operations, rather than simply on the size of the VMs and their data. This should help avoid overloading the LUNs with small but very active VMs.

If it is not possible to predict the workload in advance, then, for medium and large clusters, start with LUNs of about 500 GB, monitor the performance, and adjust as required.

Finally, try to limit metadata changes to off-peak times.

Related material

The publications listed in this section are considered particularly suitable for a more detailed discussion of the topics covered in this paper. Note that some of the documents referenced here may be available in softcopy only.

- *IBM SVC and Brocade Disaster Recovery Solutions for VMware*, REDP-4626
- *VMware Proof of Practice and Performance Guidelines on the SVC*, REDP-4601

The team who wrote this IBM Redpapers publication

This paper was produced by a team of specialists from around the world working at the International Technical Support Organization, San Jose Center.

Steve Garraway is an IT specialist with the Storage and SAN Services team at the IBM Hursley Laboratory in the United Kingdom. He has 22 years of experience in storage management, mainly with z/VM® and IBM Tivoli® Storage Manager, and has worked for IBM for 23 years. His areas of expertise include IBM System z® disk and tape storage, McDATA and Brocade SANs, and the SAN Volume Controller. He has previously co-authored four IBM Redbooks® publications on enterprise disk and SAN. Steve is a Brocade Certified Fabric Professional.

Jon Tate is a Project Manager for IBM System Storage™ SAN Solutions at the International Technical Support Organization, San Jose Center. Before joining the ITSO in 1999, he worked in the IBM Technical Support Center, providing Level 2 and 3 support for IBM storage products. Jon has 24 years of experience in storage software and management, services, and support, and is both an IBM Certified IT Specialist and an IBM SAN Certified Specialist. He is also the UK Chairman of the Storage Networking Industry Association.

Thanks to:

Andrew Martin
Bill Passingham
Thomas Vogel
Steven G White
IBM

Notices

This information was developed for products and services offered in the U.S.A.

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service.

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not give you any license to these patents. You can send license inquiries, in writing, to:
IBM Director of Licensing, IBM Corporation, North Castle Drive, Armonk, NY 10504-1785 U.S.A.

The following paragraph does not apply to the United Kingdom or any other country where such provisions are inconsistent with local law: INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some states do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this IBM product and use of those Web sites is at your own risk.

IBM may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation to you.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to the names and addresses used by an actual business enterprise is entirely coincidental.

COPYRIGHT LICENSE:

This information contains sample application programs in source language, which illustrate programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to IBM, for the purposes of developing, using, marketing or distributing application programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. IBM, therefore, cannot guarantee or imply reliability, serviceability, or function of these programs.

This document REDP-4627-00 was created or updated on December 18, 2009.

Send us your comments in one of the following ways:

- Use the online **Contact us** review Redbooks form found at:
  ibm.com/redbooks
- Send your comments in an email to:
  redbooks@us.ibm.com
- Mail your comments to:
  IBM Corporation, International Technical Support Organization
  Dept. HYTD Mail Station P099
  2455 South Road
  Poughkeepsie, NY 12601-5400 U.S.A.

Trademarks

IBM, the IBM logo, and ibm.com are trademarks or registered trademarks of International Business Machines Corporation in the United States, other countries, or both. These and other IBM trademarked terms are marked on their first occurrence in this information with the appropriate symbol (® or ™), indicating US registered or common law trademarks owned by IBM at the time this information was published. Such trademarks may also be registered or common law trademarks in other countries. A current list of IBM trademarks is available on the Web at http://www.ibm.com/legal/copytrade.shtml

The following terms are trademarks of the International Business Machines Corporation in the United States, other countries, or both:

IBM®
Redbooks®
Redpaper™
Redpapers™
Redbooks (logo) ®
System Storage™
System z®
Tivoli®
z/VM®

The following terms are trademarks of other companies:

Snapshot, and the NetApp logo are trademarks or registered trademarks of NetApp, Inc. in the U.S. and other countries.

VMotion, VMware, the VMware "boxes" logo and design are registered trademarks or trademarks of VMware, Inc. in the United States and/or other jurisdictions.

Windows, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

Other company, product, or service names may be trademarks or service marks of others.