



Recommendations for VMware Multipathing with the Hitachi Adaptable Modular Storage 2000 Family

October 2008



Executive Summary

New Hitachi storage system availability and performance features along with VMware fixed multipathing policies allow organizations to maximize the effectiveness of applications deployed on Hitachi Adaptable Modular Storage 2000 family systems with VMware virtual infrastructure environments. This document describes configuration alternatives and explains the benefits of using fixed rather than most recently used (MRU) multipathing policy. The document does not address use of the experimental round robin load-balancing feature of VMware virtual infrastructure environments.


With the introduction of automatic load balancing to the 2000 family, LUNs are optimally balanced between dual controllers to spread the I/O workload across resources. In addition, the active-active design of the 2000 family makes all LUNs accessible regardless of the physical port or the server from which the access is requested. This document provides specific recommendations to ensure effective deployments of VMware virtual infrastructure environments, particularly in server farm environments.

For best results use Acrobat Reader 8.0.



Contents

Hitachi Adaptable Modular Storage 2000 Family Features	1
Access from Multiple Servers.....	3
Server Farm Environments.....	3
Modular Storage Traditionally Treated as Active-Passive by ESX Server.....	3
Virtual Infrastructure Environment Multipathing.....	4
Occasional Revert to MRU on Reboot	4
Multipathing During Normal Operation.....	5
Multipathing During Failure or Reduced Service Conditions	6
Conclusion.....	7



Recommendations for VMware Multipathing with Hitachi Adaptable Modular Storage 2000 Family

Availability, balanced performance and ease of management are becoming more important as an increasing number of business critical applications are deployed on VMware virtual infrastructure environments. This document describes configuration options that assure balanced performance and enhanced availability. It also explains why Hitachi Data Systems recommends use of fixed multipathing policy rather than most recently used (MRU) policy for Hitachi Adaptable Modular Storage 2000 family.

Hitachi Adaptable Modular Storage 2000 Family Features

The 2000 family provides a reliable, flexible, scalable and cost-effective modular storage system for Microsoft® SQL Server. The 2000 family is ideal for more demanding application requirements and delivers enterprise-class performance, capacity and functionality at a midrange price.

The 2000 family is the only midrange storage product with symmetric active-active controllers that provide integrated, automated hardware-based front-to-back-end I/O load balancing. This ensures I/O traffic to back-end disk devices is dynamically managed, balanced and shared equally across both controllers, even if the I/O load to specific logical units (LUs) is skewed. Storage administrators are no longer required to manually define specific affinities between LUs and controllers, simplifying overall administration. In addition, this new controller design is fully integrated with standard host-based multipathing, thereby eliminating mandatory requirements to implement proprietary multipathing software.


No other midrange storage product that scales beyond 100TB has a serial attached SCSI (SAS) drive interface. The new point-to-point back-end design virtually eliminates I/O transfer delays and contention associated with Fibre Channel arbitration and provides significantly higher bandwidth and I/O concurrency. It also isolates any component failures that might occur on back-end I/O paths.

Flexibility

- Choice of Fibre Channel and iSCSI server interfaces or both
- Resilient performance using LUs that can be configured to span multiple drive trays and back-end paths
- Choice of high-performance SAS and low-cost SATA disk drives
- Lowered costs using SAS or SATA drives that can be intermixed in the same tray
- Support for all major open systems operating systems, host bus adapters (HBAs) and switch models from major vendors

Scalability

- Ability to add capacity, connectivity and performance as needed
- Concurrent support of large heterogeneous open systems environments using up to 2048 virtual ports with host storage domains and 4096 LUs
- Ability to scale capacity to 472TB

- 
- Ability to scale performance to more than 900K IOPS
 - Seamless expansion due to data-in-place upgrades from Adaptable Modular Storage 2100 to Adaptable Modular Storage 2300 and to Adaptable Modular Storage 2500
 - Large-scale disaster recovery and data migration using integration with Hitachi Universal Storage Platform™ V and Hitachi Universal Storage Platform VM
 - Complete lifecycle management solutions within tiered storage environments

Availability

- Outstanding performance and nondisruptive operations using Hitachi Dynamic Load Balancing Controller
- No single point of failure
- Hot swappable major components
- Dual battery backup for cache
- Nondisruptive microcode updates
- Flexible drive sparing with no copy back required after a RAID rebuild
- Host multipathing capability
- In-system SQL Server and Microsoft Exchange backup and snapshot support through Microsoft Windows Volume Shadow Copy Service (VSS)
- Remote site replication
- RAID-5, RAID-1, RAID-1+0 and RAID-0 (SAS drives) support
- RAID-6 dual parity support for enhanced reliability when using large SATA and SAS drives
- Hi-Track® Monitor support

Performance

- No performance bottlenecks in highly utilized controllers due to Hitachi Dynamic Load Balancing Controller
- Point-to-point SAS backplane with a total bandwidth of 96 gigabits per second (Gb/sec) and no overhead from loop arbitration
- Full duplex 3Gb/sec SAS drive interface that can simultaneously send and receive commands or data on the same link
- Up to 32 concurrent I/O paths provide up to 9600 megabytes per second (MB/sec) of total system bandwidth
- 4Gb/sec host Fibre Channel connections
- Cache partitioning and cache residency to optimize or isolate unique application workloads

Simplicity

- Simplified RAID group placement using SAS backplane architecture
- Highly intuitive management software that includes easy-to-use configuration and management utilities
- Command line interface (CLI) and command control interface (CCI) that match GUI functionality
- Seamless integration with Hitachi storage systems, managed with a single set of tools using Hitachi Storage Command Suite software

- Consistency among most Hitachi software products whether run on Hitachi modular storage systems or the Hitachi Universal Storage Platform family

Security

- Role-based access to Adaptable Modular Storage management systems
- Ability to track all system changes with audit logging
- Ability to apply system-based “write once, read many” (WORM) data access protection to logical volumes to provide regulatory compliant protection
- Encrypted communications between management software and storage system using SSL and TSL
- Internet Protocol version 6 (IPv6) and Internet Protocol Security (IPsec) compliant maintenance ports

Access from Multiple Servers

In a VMware virtual infrastructure environment, it is desirable to have servers access LUNs through preferred paths to achieve balanced performance and to intentionally distribute the I/O load across several paths. Dynamic load balancing and active-active access from any port to any LUN is automatically enabled on 2000 family systems. This eliminates the need to design for LUN ownership or data sharing modes and allows I/O through any path for common ESX Server farm access including path health checks (or future host load balancing).

Server Farm Environments

VMware virtual infrastructure environments allow virtual machines to transfer between physical hosts using a process called VMotion. This requires that multiple physical hosts (nodes) all have access to the same LUNs containing virtual machine data. In general, server farms include multiple independent paths from each node to the storage holding their LUNs, typically through two separate SANs to ports on different controllers on the storage. This design increases resilience of the solution to failures along the data path, including the loss of an HBA port, a Fibre Channel cable failure or a switch outage. During a path outage, whether transient or sustained, I/O from the VMware virtual infrastructure environment is automatically rerouted to alternate paths on the storage system.

Modular Storage Traditionally Treated as Active-Passive by ESX Server

Early versions of ESX Server treated Hitachi modular storage systems as traditional active-passive designs and selected MRU as the default multipathing policy. On a traditional active-passive system, when a disk ownership changes due to errors in the primary path, any hosts still using the non-owner controller ports receive a Unit Not Ready or illegal command response. This triggers failover for all the hosts using the affected ports, and results in all hosts using ports on the new owner controller.

However, as long as the failure occurs elsewhere in the path and not on the storage ports themselves, controller ownership is no longer a consideration. 2000 family systems continue to accept access through all ports without issuing a Unit Attention response, and no failover is triggered. Treating Hitachi modular storage as traditional active-passive can result in sub-optimal configurations for ESX Server farms.

Virtual Infrastructure Environment Multipathing

VMware virtual infrastructure environments support two types of multipathing policies: MRU and fixed. In both cases, path failover order is determined by hardware discovery order. Generally paths are discovered on one HBA, then additional paths are discovered on the other.

MRU multipathing policy causes the server to continue to use a currently active path until a failure is detected, at which time it switches to the next path in discovery order. If the next path responds positively to a Test Unit Ready inquiry, this path becomes the new active MRU path. If the Test Unit Ready inquiry fails, a Start Unit command is used to try to bring the LUN online, followed by a second Test Unit Ready attempt before trying the next path in the list.

With Hitachi modular storage, all paths to a LUN respond positively to a Test Unit Ready inquiry. Upon encountering a fault in a path, the VMware virtual infrastructure environment fails over that path to the next available path in device discovery order.

In Hitachi Adaptable Modular Storage 2000 family storage systems — both in single VMware virtual infrastructure environment installations and farm environments — access to LUNs through paths from other nodes or through other ports is supported transparently. Consequently, using MRU policy can result in nodes continuing to use temporary paths with no mechanism to fail back to the original path when the fault is corrected. The 2000 family controllers automatically balance the load between storage LUNs correctly. However, over time the continued use of temporary MRU paths might lead to a port imbalance with I/O being directed down an artificially limited number of paths even when all paths are functioning correctly.

Fixed multipathing policy uses the configured path to a LUN in preference to any other. In the event of a path failure, VMware virtual infrastructure environment fails to the next available path. However, VMware virtual infrastructure environment continues to test the failed path and when the preferred path recovers, VMware virtual infrastructure environment again routes all activity through the preferred path for that LUN. Fixed multipathing allows a configuration to dynamically repair itself. Access reverts to the preferred path when the path fault is corrected and port balance is maintained.

Occasional Revert to MRU on Reboot

Prior to the release of VMware Infrastructure 3, the reboot of a physical host might cause paths to revert to MRU operation under some circumstances. This results in a host continuing to use non-preferred paths to some or all of its storage.

With Virtual Infrastructure 3, the multipathing policy is associated with the LUN and is stored in `/etc/vmware/esx.conf`.

VMware Infrastructure 2.x and 3.x include tools to report whether a fixed or MRU multipathing policy is selected for each path. These can be added to a startup script to report any reversion to MRU operation, as shown in the following examples.

For VMware Infrastructure 2.x

```
if vmkmultipath -q |grep -q mru
then
    echo Host `hostname` has `vmkmultipath -q |grep mru` paths in MRU mode.
    >tempfile
    mail vmware_admin@somehost.somewhere <tempfile
    rm tempfile
fi
```

For VMware Infrastructure 3.x

```
if esxcfg-mpath -l |grep -q "Most Recently Used"
then
    echo Host `hostname` has `esxcfg-mpath -l |grep "Most Recently Used"` paths in
    MRU mode. >tempfile
    mail vmware_admin@somehost.somewhere <tempfile
    rm tempfile
fi
```

Multipathing During Normal Operation

During normal operation, assigning preferred paths allows the load to be distributed across multiple paths across the SAN by configuring the preferred paths appropriately for balance, as shown in Table 1.

Table 1. Preferred Paths

LUN	Preferred Path
00	HBA 0—port 0A
01	HBA 0—port 0B
02	HBA 1—port 1A
03	HBA 1—port 1B
04	HBA 0—port 0A
...	...

This layout allows greater throughput by spreading the workload across many paths. Where performance is critical, the LUNs can also be distributed across multiple RAID groups of 15k rpm Fibre Channel disks.

However, when using MRU multipathing policy, balanced path assignments are difficult to maintain. In fact, on a restarted VMware virtual infrastructure environment using MRU multipathing policy, all LUNs are accessed through the first discovered path, typically HBA 0 - target 0, while the other paths are unused. When failures occur, access is gradually distributed to other paths, but in a less than optimal arrangement.

Multipathing During Failure or Reduced Service Conditions

Fixed and MRU multipathing policies have different operational effects during path failure conditions.

HBA or Fibre Channel Path Failure

Following a single path failure, the next available path in device discovery order is used. No means exists to control selection of a specific alternate path. Within a server farm, the majority of nodes continue to use their preferred paths and only the server with a failure is redirected to an alternate path.

When the failed path is recovered, the use of fixed multipathing policy causes the server to fail back to the preferred path. When using MRU policy, fail back does not happen: the server that experiences a failure continues to use a non-preferred path with a possible performance penalty due to port or path imbalance.

Switch-SAN Failure

In the event of a switch failure in a high-availability SAN environment, all access is routed through the surviving switch. At the 2000 family storage system, I/O requests are automatically routed to their respective LUNs through whichever ports receive the I/O requests. However, I/O requests might now arrive through a reduced number of paths, which can result in a performance penalty at the server.

When using a fixed multipathing policy, upon recovery of the failed switch, all hosts resume access through their preferred paths, which increases the number of active independent paths. When using MRU multipathing policy, paths through the recovered switch remain unused unless a second failure occurs, forcing a path change from the server.

Storage Port Failure

In the event of a single storage port failure, the current active path to the server's LUNs becomes unavailable and all access is routed through the next available path by VMware virtual infrastructure environment multipathing and the 2000 family active-active design.

When using a fixed multipathing policy, upon recovery of the failed port, all hosts using the port resume access to storage through their preferred paths. When using MRU multipathing policy, paths through the recovered port remain unused unless a second failure occurs, forcing a path change from the server.

Storage Controller Failure

In the event of a failure of one of the two load-balanced storage controllers that also affects the ports, the current active paths to the LUNs managed by that controller become unavailable and all access is routed to ports on the remaining controller by VMware virtual infrastructure environment multipathing. VMware virtual infrastructure environments discover and use the newly available paths as described in an earlier section. When using a fixed multipathing policy, upon recovery of the failed controller, all hosts resume access to storage through their preferred paths. When using MRU multipathing policy, paths through the recovered controller remain unused unless a second failure occurs, forcing a path change from the server.

Storage Controller Microcode Upgrade

During a microcode upgrade on one of the two 2000 family storage controllers, all ports remain active and access is routed correctly by the storage system through the remaining controller. I/O for all LUNs transfers temporarily to the remaining controller's processor during the upgrade. Because only one storage controller is active, the ESX Server hosts can experience reduced performance during the upgrade process. However, the storage system does not report path failures and all ports continue to be active. When the controller is recovered after maintenance, LUN I/O is once again balanced across both controllers and performance at the servers returns to the pre-maintenance level.



Conclusion

New features in Hitachi Adaptable Modular Storage 2000 family systems improve the operational effectiveness and availability of applications deployed on current versions of VMware virtual infrastructure environments. In a farm environment, Hitachi Data Systems and VMware recommend the use of a fixed multipathing policy. Using a fixed multipathing policy maintains balanced performance after recovery from unexpected fault conditions and after scheduled storage maintenance. The use of fixed multipathing policy enables balanced access to ports and allows maximum performance by distributing throughput across multiple paths to storage.



Corporate Headquarters 750 Central Expressway, Santa Clara, California 95050-2627 USA

Contact Information: + 1 408 970 1000 www.hds.com / info@hds.com

Asia Pacific and Americas 750 Central Expressway, Santa Clara, California 95050-2627 USA □

Contact Information: + 1 408 970 1000 www.hds.com / info@hds.com

Europe Headquarters Sefton Park, Stoke Poges, Buckinghamshire SL2 4HD United Kingdom

Contact Information: + 44 (0) 1753 618000 www.hds.com / info.emea@hds.com

Hitachi is a registered trademark of Hitachi, Ltd., and/or its affiliates in the United States and other countries. Hitachi Data Systems is a registered trademark and service mark of Hitachi, Ltd., in the United States and other countries.

Hi-Track is a registered trademark and Universal Storage Platform is a trademark of Hitachi Data Systems Corporation.

Microsoft is a registered trademark of Microsoft Corporation.

All other trademarks, service marks and company names are properties of their respective owners.

Notice: This document is for informational purposes only, and does not set forth any warranty, express or implied, concerning any equipment or service offered or to be offered by Hitachi Data Systems. This document describes some capabilities that are conditioned on a maintenance contract with Hitachi Data Systems being in effect and that may be configuration dependent, and features that may not be currently available. Contact your local Hitachi Data Systems sales office for information on feature and product availability.

Hitachi Data Systems sells and licenses its products subject to certain terms and conditions, including limited warranties. To see a copy of these terms and conditions prior to purchase or license, please go to <http://www.hds.com/corporate/legal/index.html> or call your local sales representative to obtain a printed copy. If you purchase or license the product, you are deemed to have accepted these terms and conditions.

© Hitachi Data Systems Corporation 2008. All Rights Reserved.

WHP-302-00 DG October 2008