

# Optimizing the Hitachi Adaptable Modular Storage 2000 Family in vSphere 4 Environments

Best Practices Guide

*By Henry Chu*

*February 2011*

## Feedback

Hitachi Data Systems welcomes your feedback. Please share your thoughts by sending an email message to [SolutionLab@hds.com](mailto:SolutionLab@hds.com). Be sure to include the title of this white paper in your email message.

# Table of Contents

<b>Solution Components .....</b>	<b>4</b>
Hitachi Adaptable Modular Storage 2000 Family.....	4
Hitachi Dynamic Provisioning Software .....	5
vSphere 4 .....	5
<b>Storage Configuration.....</b>	<b>6</b>
Redundancy .....	6
Zone Configuration.....	7
Host Group Configuration.....	8
Hitachi Dynamic Provisioning Software .....	10
vStorage API for Array Integration .....	18
Scalability Considerations .....	23
Disk Alignment and RAID Stripe Size .....	24
<b>ESX Host Configuration.....</b>	<b>25</b>
Multipathing .....	26
Queue Depth .....	26
Metrics to Monitor .....	27
SCSI Reservations .....	28
VMkernel Advanced Disk Parameters .....	28
<b>Conclusion .....</b>	<b>30</b>
<b>Appendix A — Test Environment.....</b>	<b>33</b>
<b>Appendix B — Viewing Queue Depth and VM I/O Latency .....</b>	<b>36</b>
Viewing Queue Depth at Adapter Level .....	36
Viewing Queue Depth at Device Level.....	37
Viewing Queue Depth at Virtual Machine Level.....	39
Monitoring Virtual Machine I/O Latency .....	40

# Optimizing the Hitachi Adaptable Modular Storage 2000 Family in vSphere 4 Environments

## Best Practices Guide

Today's demanding economic and IT environment makes it more important than ever that IT administrators provide high-performance, scalable, highly available and easy-to-manage infrastructures. To meet those business-critical objectives, many organizations are virtualizing more of their applications.

When you virtualize with VMware vSphere 4, you can optimize your IT infrastructure by choosing a Hitachi Adaptable Modular Storage 2000 family storage system. The 2000 family's SAS-based drive architecture delivers industry-leading price and performance for midrange storage systems. The 2000 family offers symmetric active-active controllers that can automate many complex tasks, reducing management costs. The 2000 family's symmetric active-active controller architecture, when combined with vSphere's round robin multipathing policy, dramatically simplifies the setup and management of your virtualized environment because the workload is automatically balanced across host bus adapters (HBAs), SAN fabrics and the storage system's front-end Fibre Channel ports. It also eliminates single points of failure within the SAN.

The Hitachi Adaptable Modular Storage 2000 family also supports VMware's vStorage API for Array Integration (VAAI). VAAI is a set of primitives that allow IT organizations to offload processing for certain data-related services to VAAI-supported storage systems, such as the Hitachi Adaptable Modular Storage 2000 family. Doing so can enable significant improvements in virtual machine performance, virtual machine density and availability in vSphere 4.1 environments. Moving these functions to a storage system offers many benefits, but also requires the use of a highly available, scalable, high performance storage system like the 2000 family.

VMware vSphere 4 has many storage technologies that enable you to create a robust environment that improves resource utilization, reduces management costs and increases application uptime. When you combine your vSphere and 2000 family solution with Hitachi Dynamic Provisioning software and VMware's Dynamic Resource Scheduling (DRS), your resource utilization rates will improve.

This document describes best practices for deploying the 2000 family with vSphere 4. It is written for storage administrators, vSphere administrators and application administrators who are charged with managing large, dynamic environments. It assumes familiarity with SAN-based storage systems, VMware vSphere and general IT storage practices.

## Solution Components

The following sections describe the key components used in the Hitachi Data Systems lab when developing these best practice recommendations.

### Hitachi Adaptable Modular Storage 2000 Family

The Adaptable Modular Storage 2000 family systems are the only midrange storage systems with the Hitachi Dynamic Load Balancing Controller that provide integrated, automated hardware-based front to back end I/O load balancing, thus eliminating many complex and time-consuming tasks that storage administrators typically face. This ensures I/O traffic to back-end disk devices is dynamically managed, balanced and shared equally across both controllers. The 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.

The active-active Fibre Channel ports mean the user does not need to be concerned with controller ownership. I/O is passed to the managing controller through cross-path communication. Any path can be used as a normal path. The Hitachi Dynamic Load Balancing controllers assist in balancing microprocessor load across the storage systems. If a microprocessor becomes excessively busy, the LU management automatically switches to help balance the microprocessor load. Table 1 lists some of the differences between the 2000 family storage systems.

**Table 1. Hitachi Adaptable Modular Storage 2000 Family Overview**

<i>Feature</i>	<i>Adaptable Modular Storage 2100</i>	<i>Adaptable Modular Storage 2300</i>	<i>Adaptable Modular Storage 2500</i>
Maximum number of disk drives supported	159	240	480
Maximum cache	8GB	16GB	32GB
Maximum attached hosts through Fibre Channel virtual ports	1,024	2,048	2,048
Host port options	<ul style="list-style-type: none"><li>• 8 Fibre Channel</li><li>• 4 Fibre Channel</li><li>• 4 Fibre Channel + 4 iSCSI</li></ul>	<ul style="list-style-type: none"><li>• 16 Fibre Channel</li><li>• 8 Fibre Channel</li><li>• 8 Fibre Channel + 4 iSCSI</li></ul>	<ul style="list-style-type: none"><li>• 16 Fibre Channel</li><li>• 8 iSCSI</li><li>• 8 Fibre Channel + 4 iSCSI</li></ul>
Back-end disk drive connections	16 x 3 Gb/s SAS links	16 x 3 Gb/s SAS links	32 x 3 Gb/s SAS links

For more information about the Adaptable Modular Storage 2000 family, see the [Hitachi Data Systems Adaptable Modular Storage 2000](#) family web site.

## Hitachi Dynamic Provisioning Software

On Hitachi Adaptable Modular Storage 2000 family systems, Hitachi Dynamic Provisioning software's thin provisioning and wide striping functionalities provide virtual storage capacity to eliminate application service interruptions, reduce costs and simplify administration, as follows:

- Optimizes or “right-sizes” storage performance and capacity based on business or application requirements.
- Supports deferring storage capacity upgrades to align with actual business usage.
- Simplifies and adds agility to the storage administration process.
- Provides performance improvements through automatic optimized wide striping of data across all available disks in a storage pool.

The wide-striping technology that is fundamental to Hitachi Data Provisioning software dramatically improves performance, capacity utilization and management of your environment. By deploying your virtual disks using DP-VOLs from Dynamic Provisioning pools on the 2000 family, you can expect the following benefits in your vSphere environment:

- A smoothing effect to virtual disk workload that can eliminate hot spots across the different RAID groups, reducing the need for VMFS workload analysis by the VM.
- Significant improvement in capacity utilization by leveraging the combined capabilities of all disks comprising a storage pool.

## vSphere 4

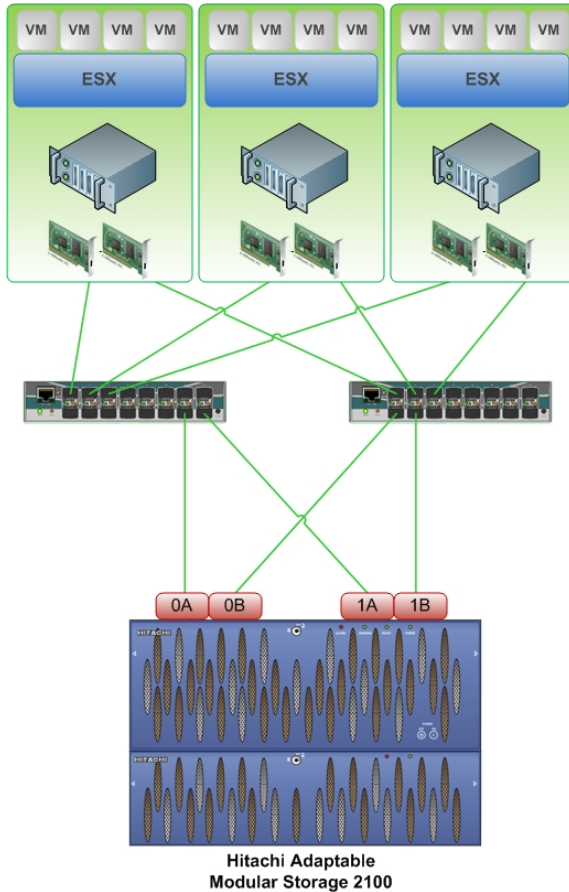
vSphere 4 is a highly efficient virtualization platform that provides a robust, scalable and reliable infrastructure for the data center. vSphere features like the Distributed Resource Scheduler, High Availability, Fault Tolerance provide an easy to manage platform.

Use of ESX 4's round robin multipathing policy with the symmetric active-active controllers' dynamic load balancing feature distributes load across multiple host bus adapters (HBAs) and multiple storage ports. Use of VMware Dynamic Resource Scheduling (DRS) with Hitachi Dynamic Provisioning software automatically distributes loads on the ESX host and on the storage system's back end. For more information, see VMware's vSphere web site.

For more information, see the [Hitachi Dynamic Provisioning software datasheet](#).

# Storage Configuration

The following sections describe configuration considerations to keep in mind when optimizing a 2000 family storage infrastructure to meet your performance, scalability, availability, and ease of management requirements. Figure 1 provides an overview of the storage infrastructure described in this best practices guide.



**Figure 1**

## Redundancy

A high-performance, scalable, highly available and easy-to-manage storage infrastructure requires redundancy at every level.

To take advantage of ESX's built-in multipathing support, each ESX host needs redundant HBAs. This provides protection against both HBA hardware failures and Fibre Channel link failures. Figure 1 shows that when one HBA is down with either hardware or link failure, another HBA on the host can still provide access to the storage resources. When ESX 4 hosts are connected in this fashion to a 2000 family storage system, hosts can take advantage of using round robin multipathing algorithm where the I/O load is distributed across all available paths. Hitachi Data Systems recommends a minimum of two HBA ports for redundancy.

Fabric redundancy is also crucial. If all of your ESX hosts are connected to a single Fibre Channel switch, failures on that switch can lead to failures on all of your virtual machines stored on the storage system. To prevent a single point of failure in the fabric, deploy at least two Fibre Channel switches or a director class switch. On the ESX host, one HBA is connected to one switch while the second HBA is connected to the other switch. Hitachi Data Systems recommends using at least two Fibre Channel switches.

Without redundancy, a single point of failure within the storage system exposes all of your virtual machines to that single point of failure. The 2000 family provides redundancy with two storage controllers where each storage controller can have two to eight Fibre Channel ports. Storage controllers with eight ports contain two Fibre Channel interface boards (four ports per Fibre Channel interface board). The first four ports (0A, 0B, 0C, and 0D) are on one Fibre Channel interface board and second set (0E, 0F, 0G, and 0H) are on the second Fibre Channel interface board. To take advantage of this, connect each storage controller to two or more Fibre Channel switches. If your storage controller contains two Fibre Channel interface boards, use one port from each Fibre Channel interface board. This protects against Fibre Channel link failures and storage controller failures. If one link fails, another link is available. If one storage controller fails, another storage controller is available.

---

**Key Best Practice** — Connect at least two Fibre Channel ports per storage controller to the Fibre Channel switches, one from each Fibre Channel interface board.

---

When a 2000 family storage system is connected in this fashion along with redundant connections on the ESX 4 hosts, the necessary pieces are in place to allow the full use of the round robin multipath algorithm and the dynamic load balancing active-active controllers. Fibre Channel connections configured as shown in Figure 1 have four paths per LU, which provides performance and scalability. The fully redundant paths at each level ensure availability.

## Zone Configuration

Zoning divides the physical fabric into logical subsets for enhanced security and data segregation. Incorrect zoning can lead to LU presentation issues to ESX hosts, inconsistent paths, and other problems. Two types of zones are available, each with advantages and disadvantages:

- **Port** — Uses a specific physical port on the Fibre Channel switch. Port zones provide better security and can be easier to troubleshoot than WWN zones. This might be advantageous in a smaller static environment. The disadvantage of this is ESX host's HBA must always be connected to the specified port. Moving an HBA connection results in loss of connectivity and requires rezoning.
- **WWN** — Uses nameservers to map an HBA's WWN to a target port's WWN. The advantage of this is that the ESX host's HBA can be connected to any port on the switch, providing greater flexibility. This might also be advantageous in a larger dynamic environment. However, the disadvantage is the reduced security and adds more complexity in troubleshooting.

Zones can be created in two ways, each with advantages and disadvantages:

- **Multiple initiator** — Multiple initiators (HBAs) are mapped to one or more targets in a single zone. This can be easier to setup and reduce administrative tasks, but this can introduce interference caused by other devices in the same zone.
- **Single initiator** — Contains one initiator (HBA) with single or multiple targets in a single zone. This can eliminate interference but requires creating zones for each initiator (HBA).

When zoning, it's also important to consider all the paths available to the targets so that multipathing can be achieved. Table 2 shows an example of a single-initiator zone with multipathing.

**Table 2. Single Initiator Zoning with Multipathing Example**

<i>Host</i>	<i>Host HBA Number</i>	<i>Director Zone Name</i>	<i>Storage Port</i>
ESX 1	HBA 1 Port 1	ESX1_HBA1_1_AMS2K_0A_1A	0A
			1A
ESX 1	HBA 2 Port 1	ESX1_HBA2_1_AMS2K_0E_1E	0E
			1E
ESX 2	HBA 1 Port 1	ESX2_HBA1_1_AMS2K_0A_1A	0A
			1A
ESX 2	HBA 2 Port 1	ESX2_HBA2_1_AMS2K_0E_1E	0E
			1E
ESX 3	HBA 1 Port 1	ESX3_HBA1_1_AMS2K_0A_1A	0A
			1A
ESX 3	HBA 2 Port 1	ESX3_HBA2_1_AMS2K_0E_1E	0E
			1E

In this example, each ESX host has two HBAs with one port on each HBA. Each HBA port is zoned to one port on each controller with single initiator and two targets in one zone. The second HBA is zoned to another port on each controller. As a result, each HBA port has two paths and one zone. With a total of two HBA ports, each host has four paths and two zones.

Determining the right zoning approach requires prioritizing your security and flexibility requirements. With single initiator-zones, each HBA is logically partitioned in its own zone. Problems in the fabric caused by one HBA do not affect other HBAs. In a vSphere 4 environment, many storage targets are shared between multiple hosts. It is important to prevent the operations of one ESX host from interfering with other ESX hosts. Industry standard best practice is to use single-initiator zones.

## Host Group Configuration

Configuring host groups on the Hitachi Adaptable Modular Storage 2000 family involves defining which HBA or group of HBAs can access a LU through certain ports on the controllers. The following sections describe different host group configuration scenarios.

### *One Host Group per ESX Host, Standalone Host Configuration*

If you plan to deploy ESX hosts in a standalone configuration, each host's WWNs can be in its own host group. This approach provides granular control over LU presentation to ESX hosts. This is the best practice for SAN boot environments, because ESX hosts do not have access to other ESX hosts' boot LUs. However, this approach can be an administration challenge because keeping track of which host has which LU can be difficult. In a scenario when multiple ESX hosts need to access the same LU for vMotion purposes, the LU must be added to each host group. This operation is error prone and might lead to confusion. If you have numerous ESX hosts, this approach can be tedious.

## One Host Group per Cluster, Cluster Host Configuration

Many features in vSphere 4 require shared storage, such as vMotion, DRS, High Availability (HA), Fault Tolerance (FT) and Storage vMotion. Many of these features require that the same LUs are presented to all ESX hosts participating in these cluster functions. If you plan to use ESX hosts with these features, create host groups with clustering in mind.

---

**Key Best Practice** — Place all of the WWNs for the clustered ESX hosts in a single host group.

---

## Host Group Options

On a 2000 family storage system, host groups are created using Hitachi Storage Navigator Modular 2 software. In the **Available Ports** box, select all ports. This applies the host group settings to all the ports that you select. Choose **VMware** from the **Platform** drop-down menu. Choose **Standard Mode** from the **Common Setting** drop-down menu. In the **Additional Settings** box, uncheck the check boxes. These settings automatically apply the correct configuration.

Figure 2 shows the correct settings for ESX hosts.

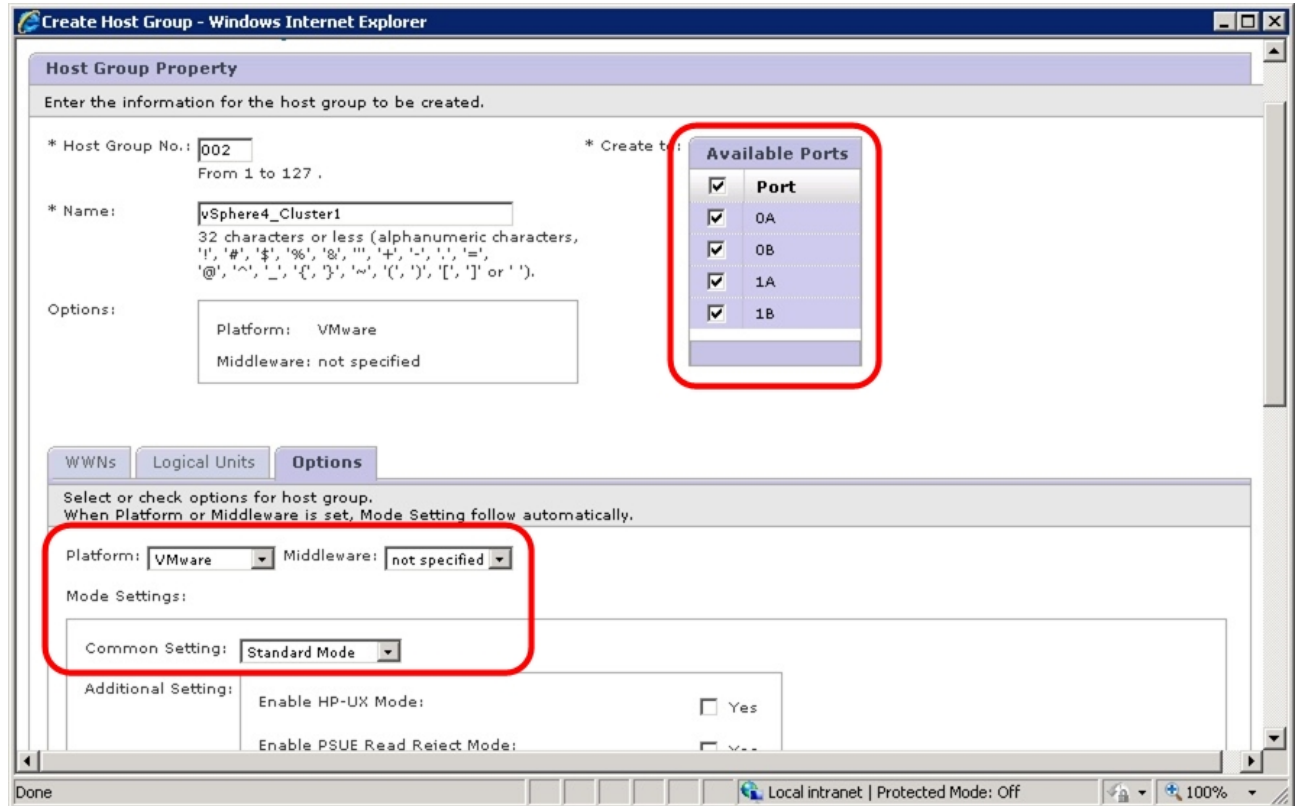


Figure 2

## Hitachi Dynamic Provisioning Software

The following sections describe best practices for using Hitachi Dynamic Provisioning Software with vSphere 4. For more information, see the [Using VMware vSphere 4 with Hitachi Dynamic Provisioning Software on the Hitachi Adaptable Modular Storage 2000 Family Best Practices Guide](#) white paper.

### *Dynamic Provisioning Space Saving and Virtual Disks*

Two of vSphere's virtual disk formats are thin-friendly, meaning they only allocate chunks from the Dynamic Provisioning pool as required. Thin and zeroedthick format virtual disks are thin-friendly, eagerzeroedthick format virtual disks are not. The eagerzeroedthick format virtual disk allocates 100 percent of the DP-VOLs space in the Dynamic Provisioning pool. While the eagerzeroedthick format virtual disk does not give the benefit of cost savings by over provisioning of storage, it can still assist in the wide striping of the DP-VOL across all disks in the Dynamic Provisioning pool.

When using DP-VOLs to overprovision storage, follow these best practices:

- Create the VM template on a zeroedthick format virtual disk on non-VAAI enabled environment. When used with VAAI, create the VM template on an eagerzeroedthick format virtual disk. When deploying, select the **Same format as source** radio button in the vCenter GUI.
- Use eagerzeroedthick format virtual disk in VAAI environments.
- Use the default zeroedthick format virtual disk if the LUN is not on VAAI-enabled storage.
- Using Storage vMotion when the source VMFS datastore is on a Dynamic Provisioning LU is a Dynamic Provisioning thin friendly operation.

Keep in mind that this operation does not zero out the VMFS datastore space that was freed by the Storage vMotion operation, meaning that Hitachi Dynamic Provisioning software cannot reclaim the free space.

### *Virtual Disk and Dynamic Provisioning Performance*

To obtain maximum storage performance for vSphere 4 when using the 2000 family storage, follow these best practices:

- Use eagerzeroedthick virtual disk format to prevent warm-up anomalies. Warm-up anomalies occur one time, when a block on the virtual disk is written to for the first time. Zeroedthick is fine for use on the guest OS boot volume where maximum write performance is not required.
- Use at least four RAID groups in the Dynamic Provisioning pool for maximum wide striping benefit.
- Size the Dynamic Provisioning pools according to the I/O requirements of the virtual disk and application.
- When larger Dynamic Provisioning pools are not possible, separate sequential and random workloads on different Dynamic Provisioning pools.
- For applications that use log recovery, separate the logs from the database on different Dynamic Provisioning pools. This separates the sequential and random workloads and can also help protect data. In the rare case where a dual hardware failure causes the corruption or loss of a Dynamic Provisioning pool, the logs are available for recovery.

### Virtual Disks on Standard LUs

Zeroedthick and eagerzeroedthick format virtual disks provide similar performance after the warm-up period required by the zeroedthick virtual disk on standard LUs. Either virtual disk format provides similar throughput after all blocks are written at least one time, however, zeroedthick initially shows some write latency and lower write throughput.

When deciding whether to use zeroedthick or eagerzeroedthick format virtual disks, keep the following considerations in mind:

- If you plan to use vSphere 4 Fault Tolerance on a virtual machine, you must use the eagerzeroedthick virtual disk format.
- If minimizing the time to create the virtual disk is more important than maximizing initial write performance, use the zeroedthick virtual disk format.
- If maximizing initial write performance is more important than minimizing the time required to create the virtual disk, use the eagerzeroedthick format.

### Distributing Computing Resource and I/O Loads

Hitachi Dynamic Provisioning software can balance I/O load in pools of RAID groups. VMware's Distributed Resource Scheduling (DRS) can balance computing capacity in CPU and memory pools. When you use Hitachi Dynamic Provisioning software with VMware DRS, CPU, memory and storage I/O loads are distributed by combining them within resource pools. VMware DRS combines CPU and memory into a DRS resource pool and Hitachi Dynamic Provisioning software combines multiple RAID groups into a Dynamic Provisioning pool. Figure 3 shows how DRS aggregate resources into a DRS resource pool.

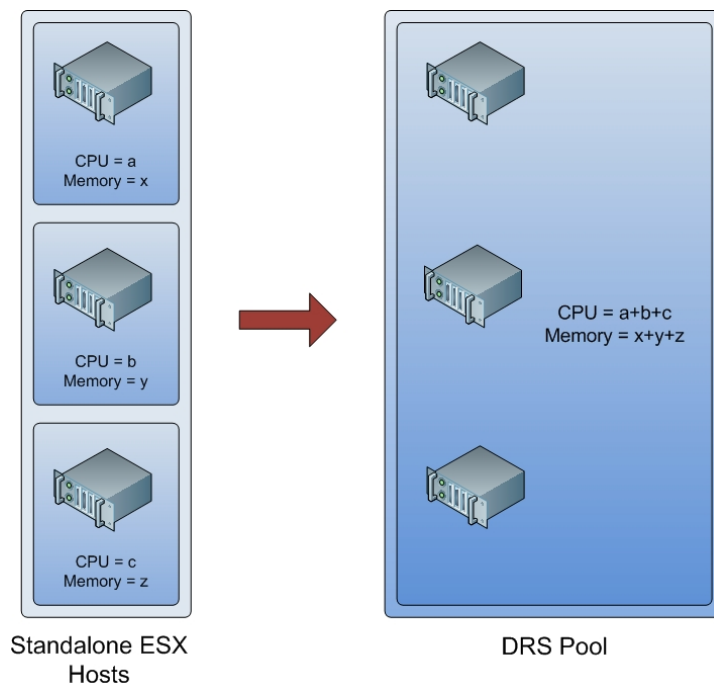
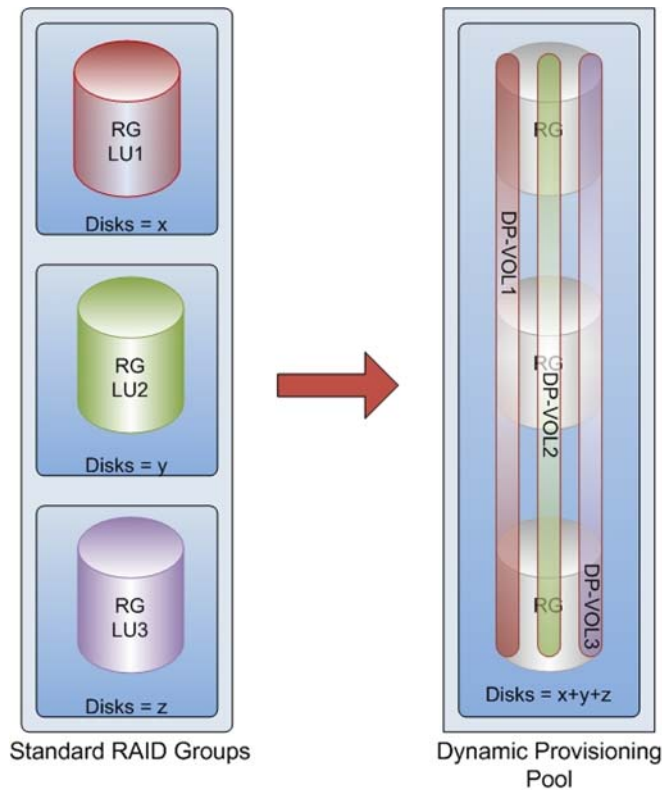


Figure 3

In Figure 3, each standalone ESX host contains an amount of CPU and memory resources. When ESX hosts are configured in a DRS cluster, the resources are aggregated into a pool. This allows the resources to be used as a single entity. A virtual machine can use resources from any host in the cluster, rather than being tied to a single host. DRS manages these resources as a pool and automatically places virtual machines on a host at power-on and continues to monitor resource allocation. DRS uses vMotion to move virtual machines from one host to another when it detects a performance benefit or based on other optimization decisions.

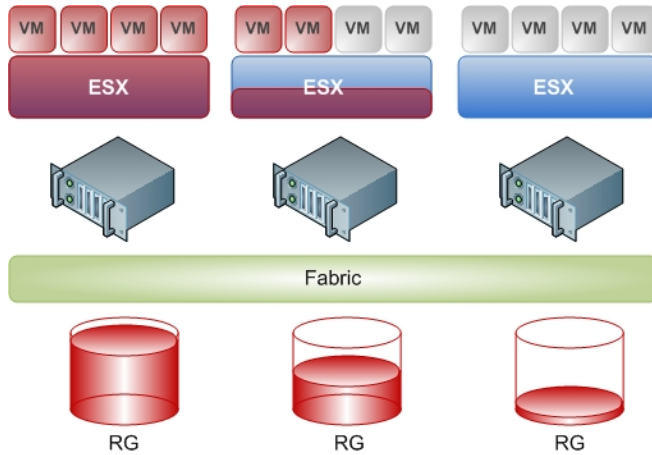
Figure 4 shows how Hitachi Dynamic Provisioning software aggregates disks in to a Dynamic Provisioning pool.



**Figure 4**

Similar to how DRS aggregates computing resources into a pool, Hitachi Dynamic Provisioning software aggregates all the allocated disks into a Dynamic Provisioning pool. This allows you to treat all the disks from each RAID group as a single entity. A single standard RAID group is analogous to a standalone ESX host. LUs created on a standard RAID group can only take advantage of the performance offered by the disks within that RAID group; they are bound by their RAID group. This is similar to how a virtual machine is bound to a single ESX host in a standalone configuration. When RAID groups are created in a Dynamic Provisioning pool, LUs are no longer tied to a single RAID group, similar to how a virtual machine is no longer tied to a single ESX host in a DRS cluster. LUs can span all the RAID groups and disks in the Dynamic Provisioning pool.

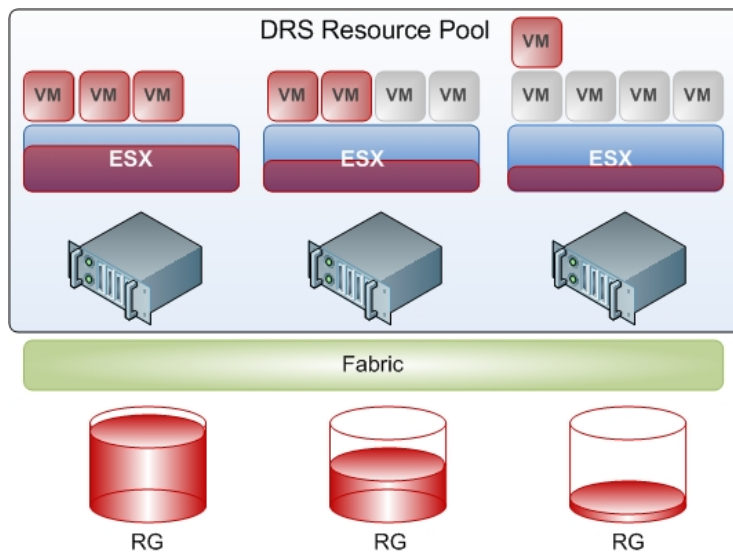
Figure 5 shows how standalone configurations with stand RAID groups can have unbalanced loads; note that resource utilization is shown in red.



**Figure 5**

With standalone ESX hosts, computing resources might be heavily used by certain virtual machines. These virtual machines might also be underperforming because the host can no longer provide any more resources; the host becomes a limiting factor. Other ESX hosts in the same farm might be moderately used or might be sitting idle. Meanwhile, the disk utilization in a RAID group might be at its performance limit handling the I/Os from the virtual machines. Other RAID groups might be moderately used or lightly used. In this scenario, heavy imbalance of resource utilization exists on both the host side and the storage side. Utilization of resources changes throughout the day and only careful monitoring and manual administration can balance these loads.

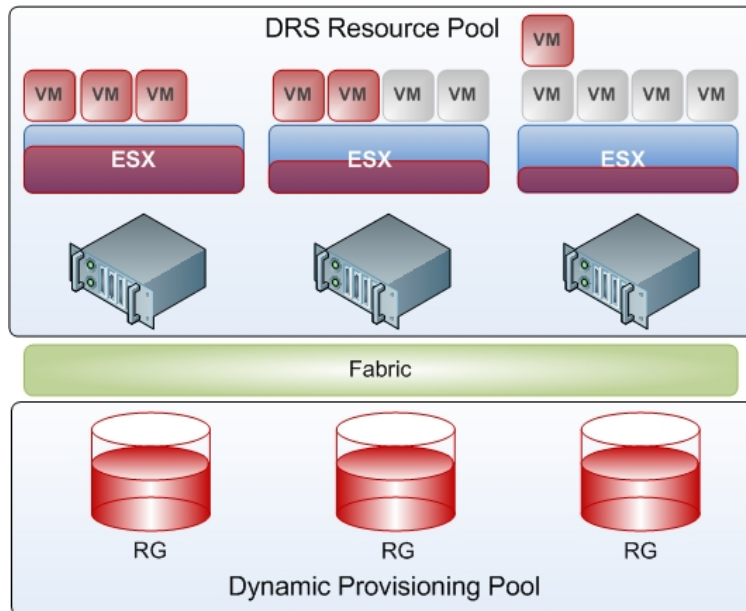
Figure 6 shows how DRS can distribute loads on the ESX hosts; note that resource utilization is shown in red.



**Figure 6**

With ESX hosts configured in DRS cluster, computing resources are managed as pool. DRS automatically use vMotion to move virtual machines to other hosts to evenly balance utilization or for performance benefits. Monitoring and placement of the virtual machines is done automatically by DRS; no manual administration is required. However, when used with standard RAID groups, a heavy imbalance still exists on the storage system. To balance the loads on the storage system, manual monitoring is required and storage vMotion can be used to migrate virtual disks to other RAID groups when necessary.

Figure 7 shows how DRS with Hitachi Dynamic Provisioning software can distribute loads on ESX hosts and the RAID groups; note that resource utilization is shown in red.



**Figure 7**

When RAID groups are configured in a Dynamic Provisioning pool, all the disks from each RAID group can be treated as a single entity. Dynamic Provisioning volumes (DP-VOLs), sometimes known as Dynamic Provisioning LUs, can span multiple RAID groups in the Dynamic Provisioning pool. This distributes the I/O across all the disks the pool. By combining the use of VMware DRS, Hitachi Dynamic Provisioning software, round robin multipathing and dynamic load balancing controllers, computing resources and I/O load are distributed for better performance and scalability with virtually no need for manual administration.

### *Hitachi Dynamic Provisioning with Dynamic Resource Scheduling Engineering Validation*

The following sections describe the Hitachi Data Systems test environment and test results. These tests demonstrate that use of VMware DRS with Hitachi Dynamic Provisioning software automatically distributes loads on the ESX host and on the storage system back end. This provides performance and scalability without the need for manual administration.

---

**Key Best Practice** — Use Hitachi Dynamic Provisioning software with DRS to automatically distribute load on both the ESX host and the storage system.

---

### Standalone ESX Hosts with Standard RAID Groups

This test determined how computer resources and I/O loads are distributed with standalone ESX hosts on standard RAID groups and LUs. Figure 8 shows each ESX host's CPU utilization in a standalone configuration.

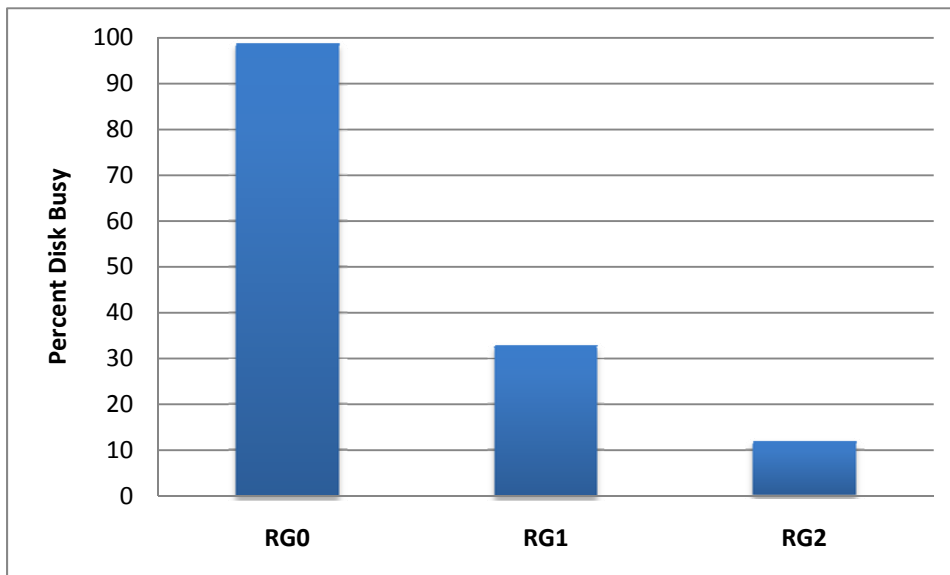
Name	State	Status	% CPU	% Memory	Memory Size
x4600-01.aselab	Connected	Alert	95	23	16383.63 MB
x4600-02.aselab	Connected	Nor...	52	22	15871.94 MB
x4600-03.aselab	Connected	Nor...	2	24	15871.94 MB

**Figure 8**

Standalone hosts running with this kind of load experienced significant imbalance. The CPU utilization of host x4600-01 is nearly at 100 percent, while the other hosts have plenty of CPU resources available.

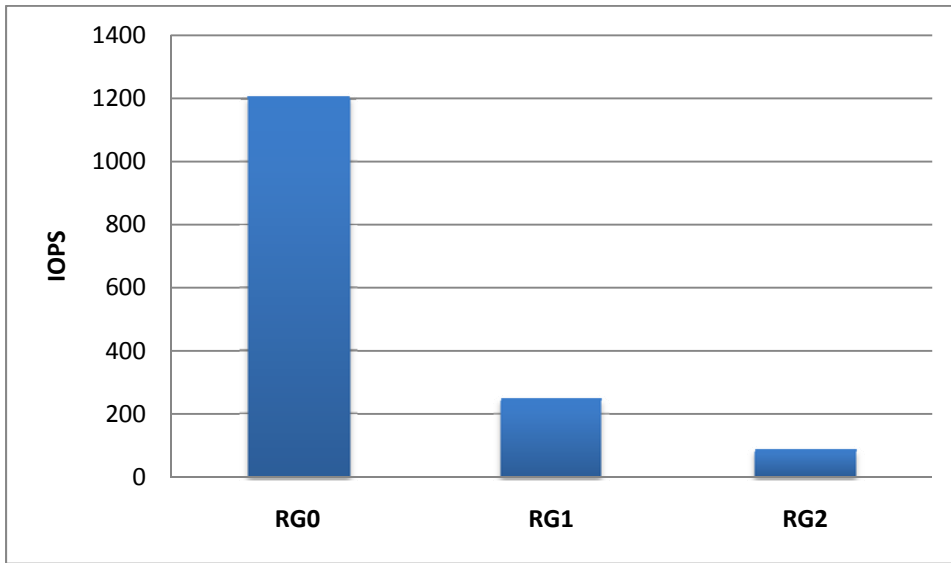
The standard RAID group's disk activity and the LU I/O were monitored and analyzed. The storage end, showed a significant imbalance or a "hot spot." RG0's disks are nearly 100 percent utilized processing the I/Os. This RAID group can no longer sustain any more virtual machines. Other RAID groups have plenty of performance available to process I/O.

Figure 9 shows the standard RAID groups' disk busy rates.



**Figure 9**

Figure 10 shows the standard LU's IOPS



**Figure 10**

*ESX Hosts in DRS Cluster with Standard RAID Groups*

In this test, the standalone hosts were put into a DRS cluster with no change to the storage. Figure 11 shows the change in CPU utilization on the ESX hosts.

Name	State	Status	% CPU	% Memory	Memory Size
x4600-01.aselab	Connected	✓ Nor...	77	21	16383.63 MB
x4600-02.aselab	Connected	✓ Nor...	49	26	15871.94 MB
x4600-03.aselab	Connected	✓ Nor...	28	32	15871.94 MB

**Figure 11**

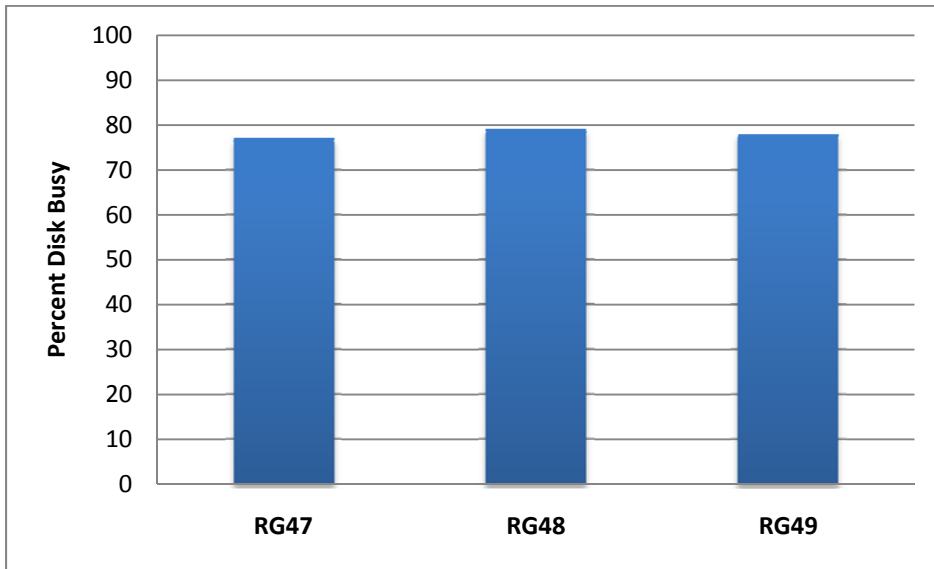
After the hosts were configured in a DRS cluster, virtual machines were automatically moved to other hosts using vMotion, which resulted in an improved CPU utilization. A single host was no longer approaching 100 percent CPU utilization. CPU resources began to be utilized on the previously idle host.

However, imbalanced loads still existed on the storage system. Because no change was made to the storage system, the disk busy percent and the distribution of the IOPS for each LU remained the same, as shown in Figure 10 and Figure 11.

*ESX Hosts in DRS Cluster with DP-VOLs*

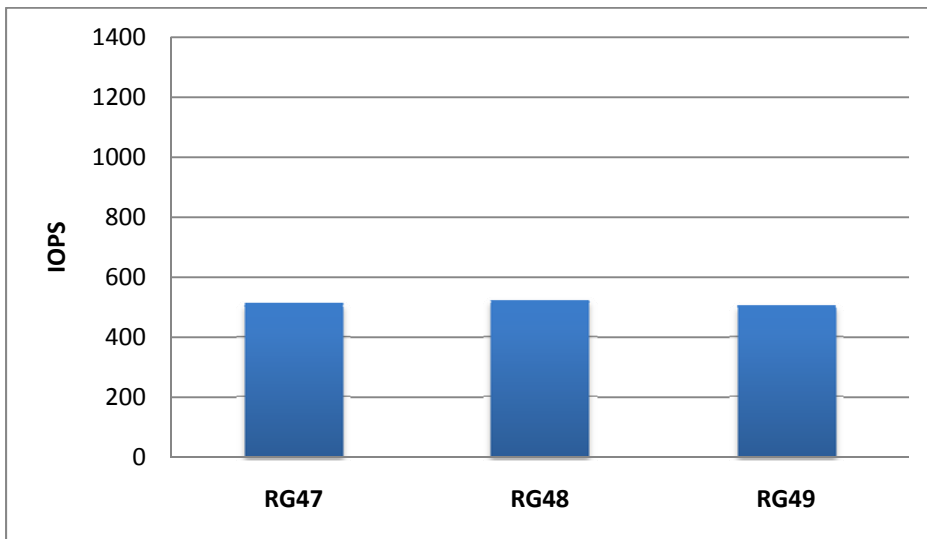
The virtual machines were migrated with Storage vMotion to DP-VOLs. The Dynamic Provisioning pool consisted of the same number of disks, type of disks, RAID level and number of RAID groups as the standard RAID groups. The disk activity and the DP-VOLs' I/O were monitored and analyzed on each RAID group in the Dynamic Provisioning pool, as shown in Figure 12 and Figure 13.

Figure 12 shows the Dynamic Provisioning RAID groups' disk busy rates.



**Figure 12**

Figure 13 shows the Dynamic Provisioning RAID groups' IOPS.



**Figure 13**

After migrating the virtual machines with Storage vMotion to DP-VOLs, the load distribution is nearly identical across all the RAID groups. Virtual machines previously limited to running in a single RAID group now have additional RAID groups to process I/O offered by the Dynamic Provisioning pool. Because virtual machines are no longer bound by a single RAID group, Hitachi Dynamic Provisioning software allows the environment to scale without additional management.

## vStorage API for Array Integration

The following sections describe best practices for using the VAAI with the Hitachi Adaptable Modular Storage 2000 family. VAAI is supported by the Hitachi Adaptable Modular Storage 2000 family.

For more information, see [Advantages of Using VMware VAAI with the Hitachi Adaptable Modular Storage 2000 family Lab Validation Report](#).

### *Storage Hardware Acceleration for Common vSphere Operations*

In the vSphere 4.1 release, the VAAI storage system offload capability supports three primitives:

- **Full copy** — Enables the storage system to make full copies of data within the storage system without having the ESX host read and write the data. The constant read and write operation is offloaded to the storage system. This results in a substantial reduction in provisioning times.
- **Block zeroing** — Enables the ESX host to offload zeroing operations to the storage system without the host having to issue redundant commands. This allows storage systems to zero out a large number of blocks to speed provisioning of virtual machines.
- **Hardware-assisted locking** — Enables the ESX host to offload locking operations to the storage system. This also provides a granular LUN locking method to allow locking at the logical block address level without the use of SCSI reservations or the need to lock the entire LUN from other hosts.

To maximize the hardware capabilities of the Adaptable Modular Storage 2000 family, enable VAAI for the following objectives:

- Reduce provisioning times:
  - For deploying virtual machines from templates
  - For cloning virtual machines
  - For deploying VMware Fault Tolerance compatible virtual machines
  - For deploying eagerzeroedthick virtual disks
- Reduce HBA I/O load when deploying from templates or cloning virtual machines
- Increase the number of virtual machines per LUN, allowing for use of larger LUN size.
- Reduce SCSI reservation conflicts
- Reduce time required for large-scale vMotion migrations

To ensure VAAI's full copy primitive is fully leveraged in the Adaptable Modular Storage 2000 family, follow these best practices:

- Source and destination VMFS volumes must have the same block size. For example, if you clone a virtual machine from a VMFS volume formatted with 1MB block size to a VMFS volume formatted with 8MB block size, the cloning process cannot take advantage of the full copy primitive.
- When cloning from raw disk mapping (RDM) disk type, ensure that the destination disk is also a RDM disk.
- When cloning a virtual machine, ensure that the source virtual disk is not of sparse type or hosted type.
- When cloning a virtual machine, ensure that no snapshots of it exist.
- Ensure that source and destination VMFS volumes are on the same storage system.

### *Hardware Accelerated Thin Provisioning*

Using Hitachi Dynamic Provisioning software with VAAI enhances thin provisioning by making eagerzeroedthick virtual disks thin provisioned at the storage level. Table 3 lists the results when each virtual disk type is provisioned in a standard RAID group, in a Dynamic Provisioning volume without VAAI, and in a Dynamic Provisioning volume with VAAI.

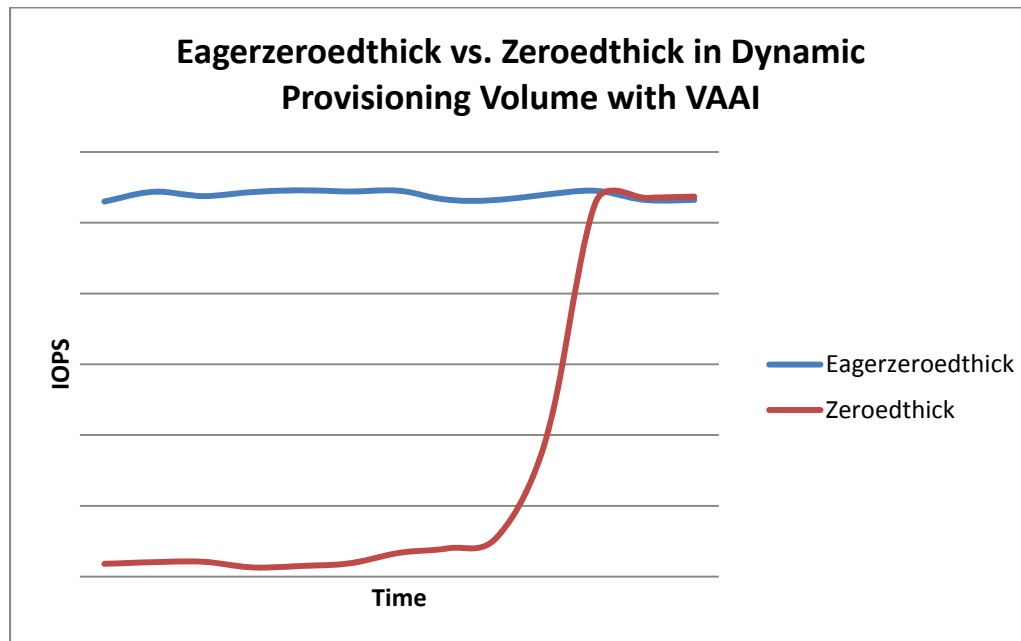
**Table 3. Disk Type Provisioning Results**

<i>Virtual Disk Type</i>	<i>Standard RAID Group</i>	<i>Dynamic Provisioning Volume without VAAI</i>	<i>Dynamic Provisioning Volume with VAAI</i>
Thin	Thin provisioned	Thin provisioned	Thin provisioned
Zeroedthick	Thick provisioned	Thin provisioned	Thin provisioned
Eagerzeroedthick	Thick provisioned	Thick provisioned	Thin provisioned

An eagerzeroedthick virtual disk provisioned in a Dynamic Provisioning pool results in fully provisioned virtual disk. However, when provisioned with VAAI enabled, eagerzeroedthick virtual disk is thin provisioned. Provisioning any type of virtual disks in a Dynamic Provisioning volume with VAAI enabled results in all virtual disk types being thin provisioned.

Thin and zeroedthick virtual disks can experience warm-up anomalies because the blocks have not been pre-zeroed. Virtual disks that are not pre-zeroed require zeroes to be written before the first write to a block. With eagerzeroedthick virtual disks, the blocks are pre-zeroed during provisioning. After all the blocks in a zeroedthick virtual disk are zeroed, performance is the same as an eagerzeroedthick virtual disk.

Figure 14 shows an example comparison between an eagerzeroedthick and a zeroedthick virtual disk in a Dynamic Provisioning volume with VAAI enabled.



**Figure 14**

A zeroedthick virtual disk experiences warm-up anomalies before the virtual disk reaches optimal performance. An eagerzeroedthick virtual disk experiences no warm-up anomalies and reaches optimal performance from the start. Although an eagerzeroedthick virtual disk is thin provisioned in a Dynamic Provisioning volume with VAAI enabled, it does not suffer any performance penalty.

When converting a thin or zeroedthick virtual disk to an eagerzeroedthick virtual disk in a Dynamic Provisioning volume with VAAI enabled, the resulting virtual disk is fully provisioned. To take advantage of the thin provisioning, zero page reclaim must be used. This process reclaims unused space from the eagerzeroedthick virtual disk, resulting in a thin provisioned virtual disk.

---

**Key Best Practice** — Use eagerzeroedthick virtual disks in a Dynamic Provisioning volume with VAAI.

---

### *Enabling vStorage API for Array Integration*

To enable VAAI for use with the 2000 family and vSphere 4.1, your environment must meet the following requirements:

- Hosts must be running ESX 4.1.
- Hitachi Adaptable Modular Storage 2000 family storage system must be at microcode version 0890/B or later
- Hitachi Storage Navigator Modular 2 software version must be 9.03 or later

In addition to the standard host groups settings, enable the **Enable Unique Extended COPY Mode** check box and the **Enable Unique Write Same Mode** checkbox in the **Additional Settings** pane, as shown in Figure 15. The **Enable Unique Extended COPY Mode** setting corresponds to the full copy primitive while the **Enable Unique Write Same Mode** setting corresponds to block zeroing primitive.

Note that no host group settings are required for the hardware-assisted locking primitive; hardware-assisted locking is always available provided your environment meets the requirements listed in this section.

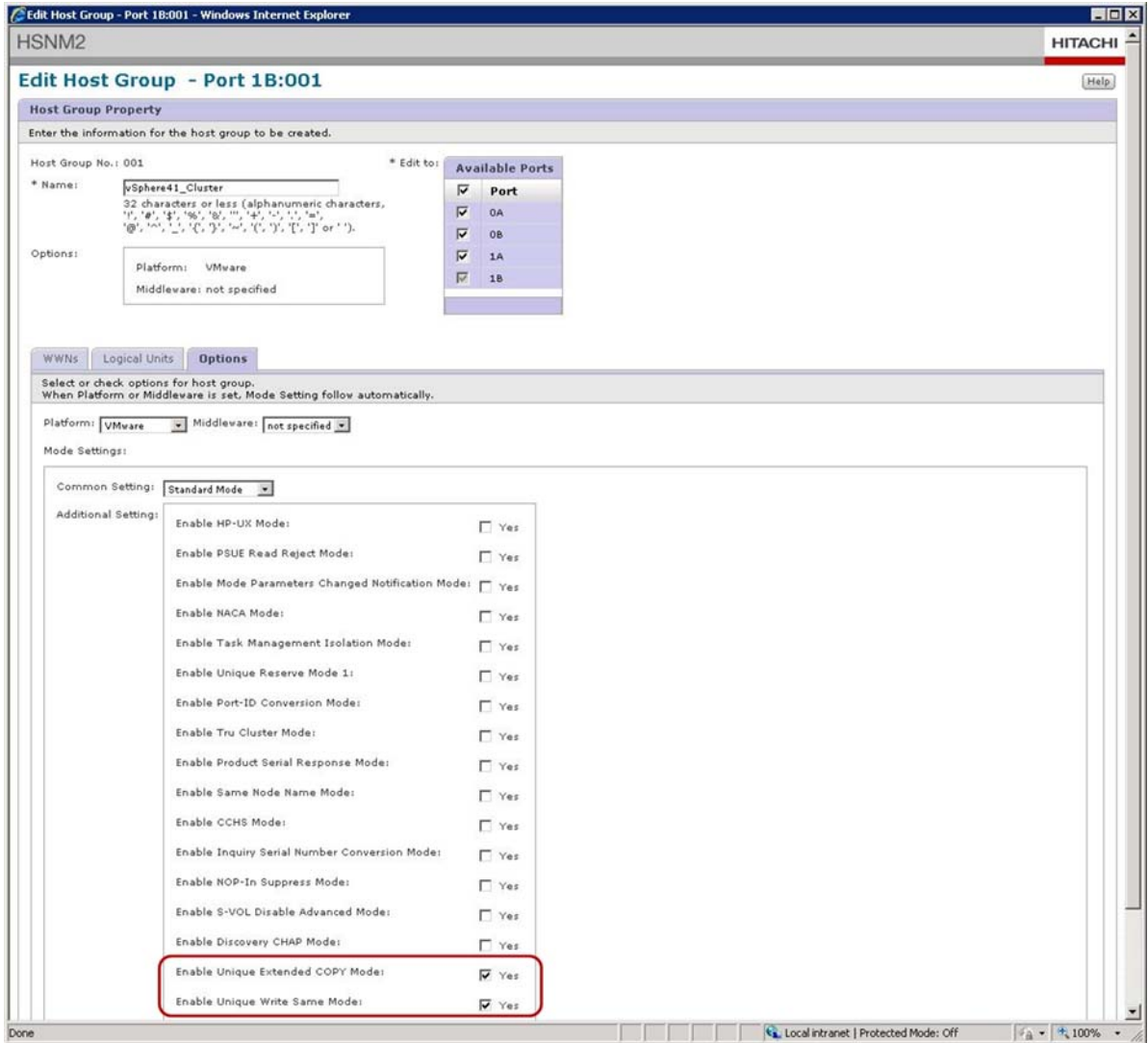


Figure 15

Figure 16 shows the ESX 4.1 configuration settings for VAAI enablement for the full copy primitive and the block zeroing primitive. The `DataMover.HardwareAcceleratedMove` setting corresponds to the full copy primitive and the `DataMover.HardwareAcceleratedInit` setting corresponds to the block zeroing primitive. Set both values to 1.

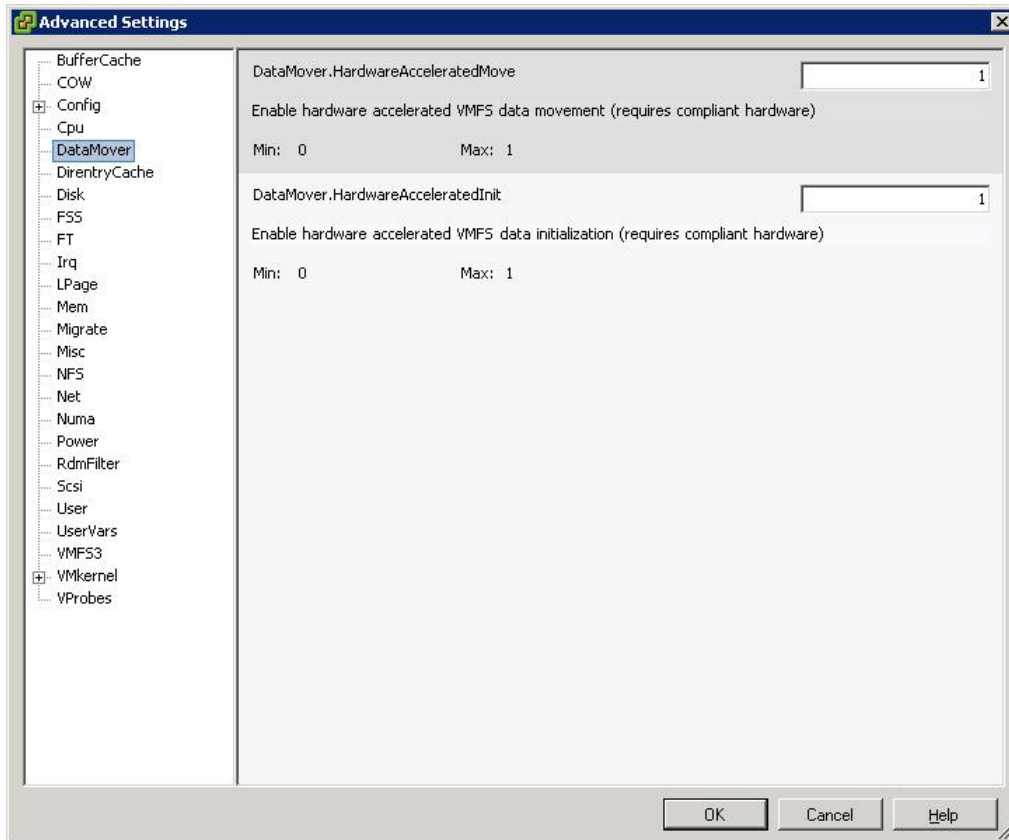


Figure 16

Figure 17 shows the setting that corresponds to the hardware-assisted locking primitive, VMFS3. HardwareAcceleratedLocking. Set this value to 1.

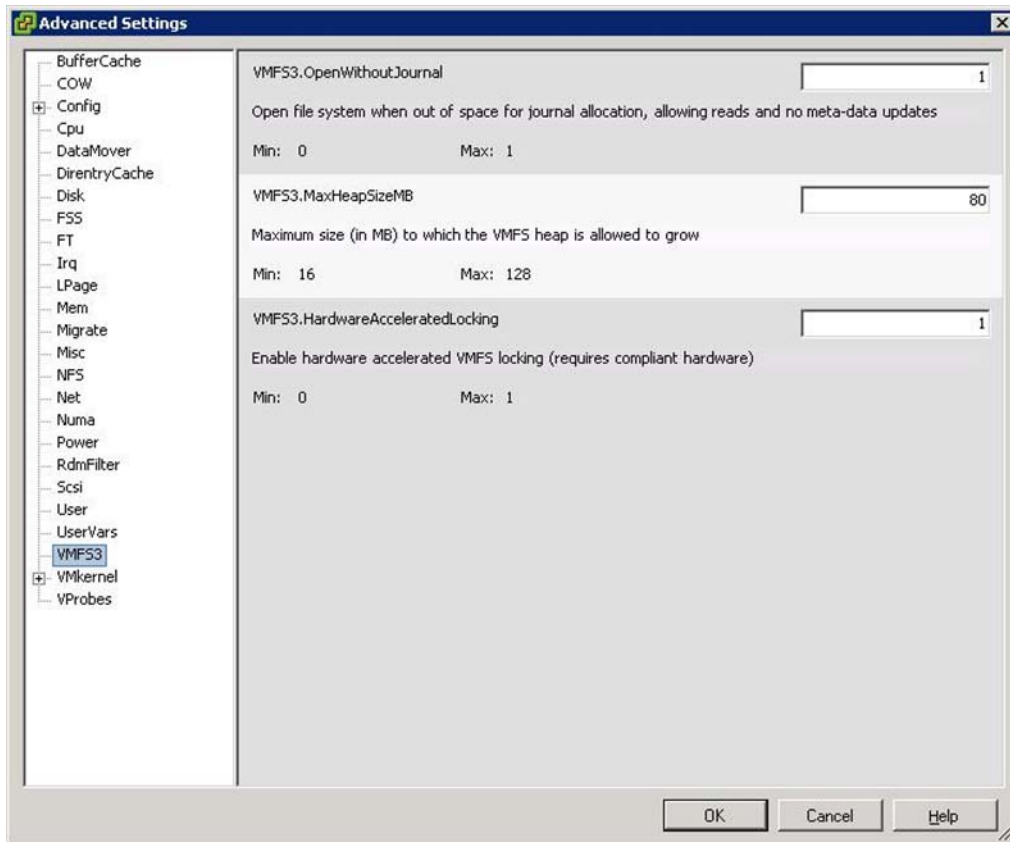


Figure 17

## Scalability Considerations

No “one size fits all” storage design exists. When designing a scalable environment, you must understand the following characteristics of your environment:

- LU size
- Number of virtual machines per LU
- Level of acceptable performance
- Size of virtual disks
- Number of virtual disks per virtual machine
- Type of workload
- Type of physical disks
- Size of physical disks
- RAID group type
- RAID group configuration

Changes to one characteristic can lead to changes in another characteristic. For example, changing the type of physical disk you choose affects the performance of your environment.

Follow these best practices when sizing your vSphere 4 environment:

- Configure for performance first, then capacity. The number of disks required to meet performance requirements might be greater than the number of disks required to meet capacity requirements. When you use Hitachi Dynamic Provisioning software, adding more RAID groups can yield more performance and capacity at the same time. RAID groups can be added to Dynamic Provisioning pools without disruption.
- Aggregate application I/O requirements, but take care not to exceed the capability of the RAID group. This is less of a concern with Hitachi Dynamic Provisioning software because Dynamic Provisioning pools contain multiple RAID groups. You can increase performance by adding RAID groups to the pool.
- Make configuration choices based on I/O workload. You can determine I/O workload by monitoring application I/O loads for a period of time that contains a full cycle of business demands. Another approach is to generate synthetic workloads.
- Distribute workloads to other RAID groups. An application with high I/O load can affect performance and can create hot spots. Consider moving these virtual machines to another RAID group or LU with Storage vMotion. This is less of a concern with Dynamic Provisioning software because loads are distributed across multiple RAID groups.

## Disk Alignment and RAID Stripe Size

Properly aligned storage is important for optimized I/O. Improper alignment can lead to incurring additional I/O when accessing data. A properly aligned system must be aligned at the guest operating system's file system level, ESX's VMFS, and on the storage system. The Adaptable Modular Storage 2000 family has a default stripe size of 256K, while VMFS3 has a starting block of 128. They are aligned along the 128K boundary. If the VMFS volume is upgraded from ESX 2.x, the VMFS volume might not be aligned. Previous versions have a starting block of 63, which is not aligned.

To see if the VMFS is properly aligned, issue the following ESX command:

```
fdisk -l u
```

The output might be similar to this:

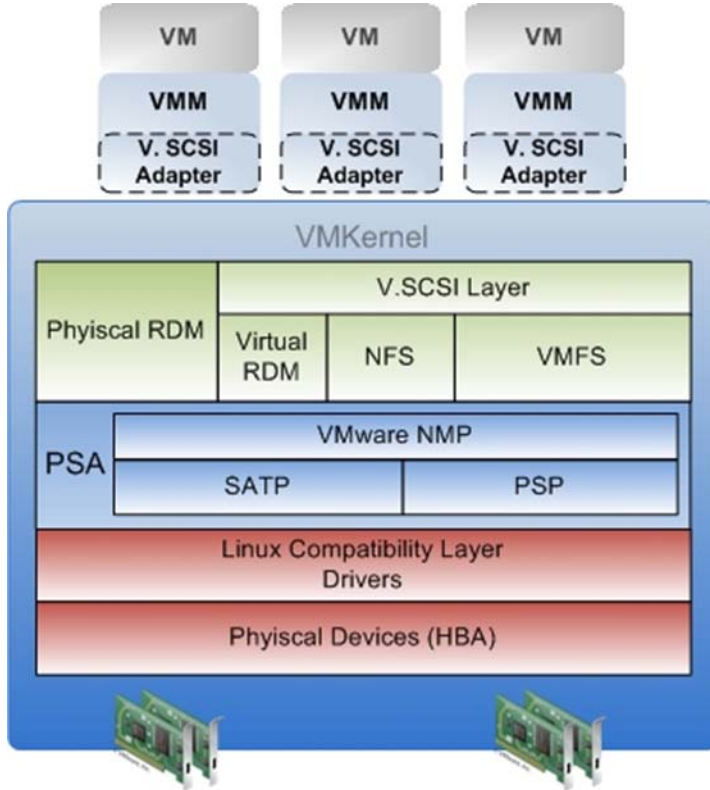
Device	Boot	Start	End	Blocks	Id	System
/dev/sdn1		128	2247091874	1123545873+	fb	VMware VMFS

The Start value of 128 indicates an aligned partition. A Start value of 63 indicates that the partition is not aligned. If the VMFS is not properly aligned, consider migrating the VMs to another LU and recreating the volume. If this is not an option, see VMware's [Recommendations for Aligning VMFS Partitions](#).

With Windows 2008, newly created partitions are properly aligned. New partitions created with previous versions of Windows operating system are not aligned by default. When a partition that was created on earlier versions of Windows is attached to Windows 2008, it carries the same partition properties as when it was created. To align Windows partitions, see Microsoft's [Disk Partition Alignment Best Practices](#).

# ESX Host Configuration

Configuring ESX 4 requires some understanding how I/O flows from a virtual machine through the VMkernel then to the physical disks. Figure 18 shows the I/O stack of ESX 4 with a 2000 family storage system.



**Figure 18**

When an application issues an I/O, the guest OS's virtual adapter driver passes the I/O to the virtual SCSI adapter in Virtual Machine Monitor (VMM). The I/O is then passed to the VMkernel. At this point, the I/O can take different route based on the kind of virtual disk the virtual machine uses. If it uses virtual disks on VMFS, the I/O is passed through the virtual SCSI layer, then through the VMFS layer. The I/O is issued to the pluggable storage architecture (PSA) to determine what path the I/O to be sent. I/O is then passed to the HBA driver queue. When the I/O is received by the storage system, the I/O is processed by the controller, through the SAS controller, then finally the physical disks.

## Multipathing

ESX 4 uses PSA, which allows for third-party storage vendor multipathing software. Figure 18 shows that when I/O is issued through the PSA, the Native Multipathing Plug-in (NMP) calls the Path Selection Plug-in (PSP) assigned to the storage device. PSP determines to which path the I/O is to be sent. If the I/O is complete, NMP reports its completion. If the I/O is incomplete, the Storage Array Type Plug-in (SATP) is called to interpret the error codes and activate paths if necessary. PSP then resets the path selection. When the 2000 family of storage is used with ESX 4, NMP is automatically configured with the following plug-ins:

- **SATP** — Default active-active storage system type (VMW\_SATP\_DEFAULT\_AA)
- **PSP** — Fixed path policy (VMW\_PSP\_FIXED)

However, the 2000 family can also support the round robin multipathing policy (VMW\_PSP\_RR). Round robin rotates through all available paths distributing the I/O load across the paths.

Path management on asymmetric active-active storage systems requires that all ESX hosts set the preferred path to the asymmetric active-active storage system so that communication is to the proper storage processor. In addition, performance considerations for Fibre Channel ports and storage processor can be time consuming to troubleshoot because of interactions between ESX hosts using the same logical unit (LU). None of this is necessary on the 2000 family. Because the 2000 family uses symmetric active-active controllers, all ports on the controller can be used in a round-robin fashion.

All the necessary PSA plug-ins for optimal performance and redundancy are built into ESX 4. No third-party PSA plug-in is necessary.

---

**Key Best Practice** — To maximize the capabilities of the Hitachi Adaptable Modular Storage 2000 family's active-active symmetric controllers, use the round robin path policy.

---

In addition, the 2000 family's dynamic load balancing controller distributes the I/O loads between two controllers by shifting backend I/O load for LUs to the underused controller. This operation is independent of which Fibre Channel ports are accessing the LU.

For more information, see the [Advantages of Active-active Controllers in vSphere 4 Environments](#) white paper.

## Queue Depth

Hitachi Data Systems recommends an LU queue depth of 32 for SAS drives and 16 for SATA drives. The procedure to change queue depths differs depending on the HBA vendor or type of HBA driver.

The default LU queue depth of Emulex and Qlogic drivers are 30 and 32 respectively. To change queue depths for Emulex and Qlogic drivers, see the VMware Knowledge Base article, "[Changing the Queue Depth for QLogic and Emulex HBAs.](#)"

When changing LU queue depths, you also typically change the `disk.SchedNumReqOutstanding` ESX advanced parameter. However, Hitachi Data Systems recommends the default value of 32 on ESX 4 for this parameter. This parameter affects the number of outstanding commands to a target when competing virtual machines exist.

Monitoring queue depth is an important part of monitoring your environment or troubleshooting performance problems. When the queue depth is exceeded, I/O is queued in the VMkernel. This can increase I/O latency for the virtual machines. `esxtop` or `resxtop` utility can be used to monitor queue depth on ESX 4 at the storage disk's adapter, device and virtual machine level.

For more information about using `esxtop`, see Appendix B.

## Metrics to Monitor

Table 4 lists important metrics to monitor.

**Table 4. `esxtop` Storage I/O metrics.**

<i>Metric</i>	<i>Description</i>
AQLEN	Maximum number of ESX VMkernel active commands that the adapter driver is configured to support (storage adapter queue depth)
LQLEN	Maximum number of ESX VMkernel active commands that the LU is allowed to have (LU queue depth)
ACTV	The number of commands in the ESX VMkernel that are currently active for a LU
QUED	The number of commands in the ESX VMkernel that are currently queued for a LU
%USD	Percentage of queue depth used by ESX VMkernel active commands for a LU
CMDS/s	Number of commands issued per second for a device
READS/s	Number of read commands issued per second for a device
WRITES/s	Number of write commands issued per second for a device
MBREAD/s	Megabytes read per second for a device
MBWRTN/s	Megabytes written per second for a device
DAVG/cmd	Average device latency per command in milliseconds
KAVG/cmd	Average ESX VMkernel latency per command in milliseconds
GAVG/cmd	Average Guest OS latency per command in milliseconds
QAVG/cmd	Average queue latency per command in milliseconds

For more information about using `esxtop`, see Appendix B.

## SCSI Reservations

SCSI reservation conflicts can cause I/O performance problems and limit access to storage resources. This can occur when multiple ESX hosts access a shared VMFS volume simultaneously during certain operations. The following operations use SCSI reservations:

- Creating templates
- Creating virtual machines either new or from template
- Running vMotion
- Powering on virtual machines
- Growing files for virtual machine snapshots
- Allocating space for Thin virtual disks
- Adding extents to VMFS volumes
- Changing the VMFS signature

Many of these operations require VMFS metadata locks. Experiencing a few SCSI reservations conflicts is generally acceptable; however, best practice is to minimize these conflicts. The following conditions can affect the number of reservation conflicts:

- The number of virtual machines per VMFS volume
- The number of ESX hosts accessing a VMFS volume
- The use of virtual machine snapshots

To greatly reduce SCSI reservation conflicts, enable VAAI. This allows you to leverage the hardware-assisted locking primitive by offloading locking to the storage system and provide block level locking instead of relying solely on LUN level SCSI reservations.

In addition, follow these best practices on non-VAAI enabled environments to minimize SCSI reservation conflicts:

- Do not run VMware Consolidated Backups (VCBs) on multiple virtual machines in parallel to the same VMFS volume.
- Run operations that require SCSI reservations to the shared VMFS volume serially.

## VMkernel Advanced Disk Parameters

You must fully understand an application's I/O workloads and test all changes in a controlled environment before changing these parameters. Tuning for better performance for one type of workload might decrease performance for other types of workload. ESX hosts generally experience a wide variety of workloads and tuning must accommodate those.

---

**Key Best Practice** — Use the VMkernel's default advanced parameters.

---

In general, the default parameters are sufficient for wide variety of workloads.

Table 5 lists the advanced parameters that are most likely to affect performance.

**Table 5. Advanced Disk Parameters**

<i>Parameter</i>	<i>Default Value</i>	<i>Description</i>
Disk.BandwidthCap	4294967294	Limit on disk bandwidth (KB/s) usage.
Disk.ThroughputCap	4294967294	Limit on disk throughput (IO/s) usage.
Disk.SectorMaxDiff	2000	Distance in sectors in which I/O of a VM is considered "sequential." Sequential I/O is given higher priority to get the next I/O slot.
Disk.SchedQuantum	8	Number of consecutive requests from one world (VMs).
Disk.SchedNumReqOutstanding	32	Number of outstanding commands to a target with competing worlds (VMs).
Disk.SchedQControlSeqRegs	128	Number of consecutive requests from VM required to raise the outstanding commands to the maximum.
Disk.SchedQControlVMSwitches	6	Number of switches between commands issued by different VMs required to reduce outstanding commands to SchedNumReqOutstanding.
Disk.DiskMaxIOSize	32767	Maximum disk read/write I/O size before splitting (in KB).

## Conclusion

This white paper describes best practices for deploying the 2000 family with vSphere 4. Table 6 lists best practices for optimizing Hitachi Adaptable Modular Storage 2000 family storage systems for vSphere 4 environments.

**Table 6. Best Practices for Hitachi Adaptable Modular Storage 2000 Family**

<i>Description</i>	<i>Best Practices</i>
Distributing loads	Use Hitachi Dynamic Provisioning software with VMware Dynamic Resource Scheduling to distribute loads on the storage system and ESX hosts.
Redundancy	Use minimum of two HBA ports.
	Use at least two Fibre Channel switches or director class switch.
	Use two Fibre channel ports per storage controller connected to the Fibre Channel switches, one from each Fibre Channel interface board if storage controller has eight ports.
Zone	Use single-initiator zones.
Host groups	For standalone ESX hosts or boot from SAN configurations, use one host group per ESX host.
	For clustered ESX hosts use one host group per cluster.
Dynamic provisioning space saving and virtual disk	Create the VM template on a zeroedthick format virtual disk if the LUN is not on VAAI-enabled storage.
	Use the eagerzeroedthick format virtual disk if the LUN is on VAAI-enabled storage.
	Use the default zeroedthick format virtual disk if the LUN is not on VAAI-enabled storage.
Virtual Disk and Dynamic Provisioning performance	Use eagerzeroedthick virtual disk format to prevent warm-up anomalies.
	Use at least four RAID groups in the Dynamic Provisioning pool for maximum wide striping benefit.
	Size the Dynamic Provisioning pools according to the I/O requirements of the virtual disk and application.
	When larger Dynamic Provisioning pools are not possible, separate sequential and random workloads on different Dynamic Provisioning pools.
	For applications that use log recovery, separate the logs from the database on different Dynamic Provisioning pools.
Virtual disks on standard LUs	If minimizing the time to create the virtual disk is more important than maximizing initial write performance, use the zeroedthick virtual disk format.
	If maximizing initial write performance is more important than minimizing the time required to create the virtual disk, use the eagerzeroedthick format
Scalability	Configure for performance first, then capacity.
	Aggregate application I/O requirements, but take care not to exceed the capability of the RAID group.
	Make configuration choices based on I/O workload.
	Distribute workloads to other RAID groups.

<i>Description</i>	<i>Best Practices</i>
VAAI	Enable VAAI to offload common vSphere operations for reduced provisioning times and reduced HBA I/O.

Table 7 shows the best practices for optimizing vSphere 4 for the Hitachi Adaptable Modular Storage 2000 family.

**Table 7. Best Practices for vSphere 4**

<i>Item</i>	<i>Best Practice</i>
Multipathing policy	Use round robin (VMW_PSP_RR).
Queue depth	Set LU queue depth to 32 for SAS drives.
	Set LU queue depth to 16 for SATA drives.
	Use default value of 32 for ESX VMkernel advanced parameter Disk.SchedNumReqOutstanding.
Minimize SCSI reservation conflicts	Reduce the number of virtual machines per VMFS volume.
	Reduce ESX hosts accessing a VMFS volume.
	Minimize use of virtual machine snapshots.
	Avoid running VMware Consolidated Backups (VCBs) on multiple virtual machines in parallel to the same VMFS volume.
	Run operations that require SCSI reservations to the shared VMFS volume serially
	Enable VAAI's hardware-assisted locking feature to reduce SCSI reservation conflicts.
VMkernel advanced parameters	Use default VMkernel advanced parameters.
Scalability considerations	Configure for performance first, then capacity.
	Aggregate application I/O requirements, but take care not to exceed the capability of the RAID group.
	Make configuration choices based on I/O workload.
	Distribute workloads to other RAID groups.

Following these best practices helps to ensure that your vSphere 4 and Hitachi Adaptable Modular Storage 2000 infrastructure is robust, offering high performance, scalability, high availability, ease of management, better resource utilization and increased uptime.

Hitachi Data Systems Global Services offers experienced storage consultants, proven methodologies and a comprehensive services portfolio to assist you in implementing Hitachi products and solutions in your environment. For more information, see the Hitachi Data Systems Global Services [web site](#).

Live and recorded product demonstrations are available for many Hitachi products. To schedule a live demonstration, contact a sales representative. To view a recorded demonstration, see the Hitachi Data Systems Corporate Resources [web site](#). Click the **Product Demos** tab for a list of available recorded demonstrations.

Hitachi Data Systems Academy provides best-in-class training on Hitachi products, technology, solutions and certifications. Hitachi Data Systems Academy delivers on-demand web-based training (WBT), classroom-based instructor-led training (ILT) and virtual instructor-led training (vILT) courses. For more information, see the Hitachi Data Systems Academy [web site](#).

For more information about Hitachi products and services, contact your sales representative or channel partner or visit the Hitachi Data Systems [web site](#).

## Appendix A — Test Environment

The following tables describe the test environment used to demonstrate how using Hitachi Dynamic Provisioning software with DRS can distribute loads without manual administration.

Table 7 lists the hardware used in the Hitachi Data Systems lab.

**Table 7. Hardware Resources**

<i>Hardware</i>	<i>Description</i>	<i>Version</i>
Hitachi Adaptable Modular Storage 2100	Dual controllers 4 x 4GB Fiber Channel ports, 2 per controller 8GB cache memory, 4GB per controller 60 x 146GB, 15K RPM, SAS disks (30 used)	0870/C-M
Brocade 48000 Director	Director-class SAN switch with 4Gb Fiber Channel ports	FOS 5.3.1a
Sun X4600-M2	4 x Dual-Core AMD Opteron Processor 8220 2.8 GHz, 16GB RAM, equipped with 2 x Emulex LPe11002 4Gb	ESX 4.0 Update 1
HP DL385	4 x Dual-Core AMD Opteron Processors 2.8GHz, 8GB RAM	vCenter 4.0 Update 1

Table 8 lists the ESX resources used in testing.

**Table 8. ESX Resources**

<i>Feature</i>	<i>Description</i>
Number of VMs per LU	4
Multipathing enabled	VMware round robin on ESX
Load generators	VDBench (I/O), Prime95 (CPU)

Table 9 lists virtual machine resources used in testing.

**Table 9. VM Resources**

<i>Feature</i>	<i>Description</i>
OS	Windows 2008 R2 Enterprise
Memory	2GB RAM
CPU	1 vCPU
Load generators	VDBench, Prime95

Table 10 lists RAID group configuration details.

**Table 10. RAID Group Configuration**

<i>RAID Group Type</i>	<i>Physical LU Size (GB)</i>	<i>Number of RAID Groups</i>	<i>Number of LUs</i>	<i>RAID Group Level</i>
Standard	584	3	3 LU (1 LU per RAID group)	RAID-5 (4D+1P)
Dynamic Provisioning	584	3 (in Dynamic Provisioning pool)	3 LU (3 DP-VOLs)	RAID-5 (4D+1P)

Table 11 lists ESX 4.0 and virtual machine configuration details.

**Table 11. ESX 4.0 and VM Configuration**

<i>ESX</i>	<i>Number of VMs per LU</i>	<i>Number of VMs with CPU Load Generator</i>	<i>I/O Profile</i>	<i>I/O Rate (IOPS)</i>
Host 1	4	4	75% read, 25% write, 8K size, 8 threads	2 VMs I/O rate = 400 2 VMs I/O rate = 200
Host 2	4	2	75% read, 25% write, 8K size, 8 threads	2 VMs I/O rate = 100 2 VMs I/O rate = 20
Host 3	4	0	75% read, 25% write, 8K size, 8 threads	4 VMs I/O rate = 20

Figure 19 shows the architecture used for this test.

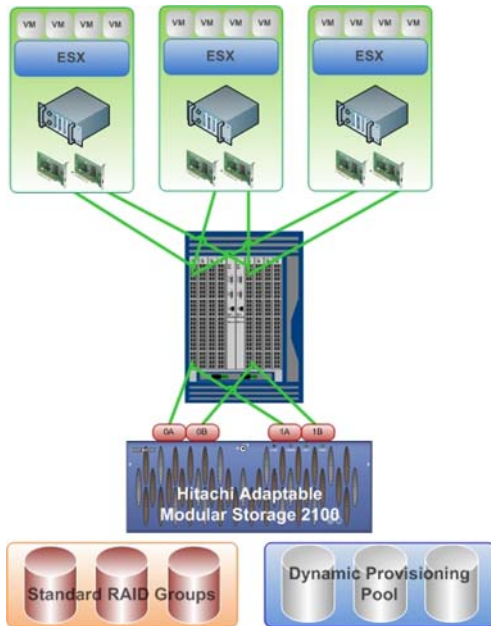


Figure 19

## Appendix B — Viewing Queue Depth and VM I/O Latency

The following procedures describe how to use `esxtop` to view queue depth at various levels and to monitor virtual machine I/O latency.

### Viewing Queue Depth at Adapter Level

To view queue depth at the adapter level, follow these steps:

1. Run `esxtop` in interactive mode.
2. Enter the following command:

**d**

`esxtop` switches to the storage disk adapter view.

3. Enter the following command:

**P**

`esxtop` displays a **Adapter to expand at path level** prompt.

4. Enter the adapter number in the form of `vmhba<#>`.

An expanded view of each path for the `vmhba` appears. The `LQLEN` column shows the LU queue depth. The `ACTV` column shows number of commands currently active in the VMkernel. `QUED` shows the number of commands queued in the VMkernel. The `%USD` column shows the percent of the LU queue depth used actively by the VMkernel.

```
12:40:20am up 4 days 7:04, 144 worlds; CPU load average: 0.02, 0.02, 0.02
```

ADAPTR	CID	TID	LID	NCHNS	NTGTS	NLUNS	AQLEN	LQLEN	WQLEN	ACTV	QUED	%USD	LOAD	CMDS/s	READS/s
vmhba0	-	-	-	1	2	14	1014	-	-	-	-	-	-	0.00	0.00
vmhba1	0	0	0	1	1	1	1014	32	-	0	0	0	0.00	0.00	0.00
vmhba1	0	0	1	1	1	1	1014	32	-	0	0	0	0.00	0.00	0.00
vmhba1	0	0	2	1	1	1	1014	32	-	0	0	0	0.00	0.00	0.00
vmhba1	0	0	3	1	1	1	1014	32	-	8	0	25	0.25	335.49	250.92
vmhba1	0	0	4	1	1	1	1014	32	-	0	0	0	0.00	0.00	0.00
vmhba1	0	0	5	1	1	1	1014	32	-	0	0	0	0.00	0.00	0.00
vmhba1	0	1	0	1	1	1	1014	32	-	0	0	0	0.00	0.00	0.00
vmhba1	0	1	1	1	1	1	1014	32	-	0	0	0	0.00	0.00	0.00
vmhba1	0	1	2	1	1	1	1014	32	-	0	0	0	0.00	0.00	0.00
vmhba1	0	1	3	1	1	1	1014	32	-	0	0	0	0.00	398.92	302.78
vmhba1	0	1	4	1	1	1	1014	32	-	0	0	0	0.00	0.00	0.00
vmhba1	0	1	5	1	1	1	1014	32	-	0	0	0	0.00	0.00	0.00

5. Check to see if LQLEN, ACTV, QUED columns appear in the results.

If these columns do not appear, queue statistics are not enabled. Follow these steps to enable queue statistics:

- a. Enter the following command to display the current field order selection menu:

**f**

- b. Enter the following command to enable queue statistics:

**f**

## Viewing Queue Depth at Device Level

To view queue depth at device level, follow these steps:

1. Run esxtop in interactive mode.
2. Enter the following command:

**u**

esxtop switches to storage disk device view.

3. Enter the following command:

**L**

esxtop displays a Change the name field size prompt.

4. Enter a value of the device name field.

A value of 37 might be sufficient to see the full names of the devices.

5. Enter the following command:

**e**

esxtop displays a **Device to expand/rollup** prompt.

6. Enter the name of a device as listed in the DEVICE column.

An expanded view of a storage device appears. DQLEN shows the device's queue depth. WQLEN shows the world queue depth. Schedulable entities in the VMkernel are referred to as worlds, where worlds can be a virtual machine. ACTV shows number of commands currently active in the VMkernel. QUED shows the number of commands queued in the VMkernel. %USD show the percent of the device queue depth used actively by the VMkernel.

```
12:41:52am up 4 days 7:05, 144 worlds; CPU load average: 0.02, 0.02, 0.02
```

DEVICE	PATH/WORLD/PARTITION	DQLEN	WQLEN	ACTV	QUED	%USD	LOAD	CMDS/s	READS/s
naa.60060e80104	4290	32	32	8	0	25	0.25	1565.12	1174.13
naa.60060e80104	4096	32	32	0	0	0	0.00	0.59	0.00
naa.60060e80104	4109	32	32	0	0	0	0.00	0.00	0.00
naa.60060e80104	4110	32	32	0	0	0	0.00	0.00	0.00
naa.60060e80104	4113	32	32	0	0	0	0.00	0.00	0.00
naa.60060e80104	4235	32	32	0	0	0	0.00	0.00	0.00
naa.60060e80104	4171	32	32	0	0	0	0.00	0.00	0.00
naa.60060e80104	4111	32	32	0	0	0	0.00	0.00	0.00
naa.60060e80104	4114	32	32	0	0	0	0.00	0.00	0.00

7. Check to see if LQLEN, ACTV, QUED columns appear in the results.

If these columns do not appear, queue statistics are not enabled. Follow these steps to enable queue statistics:

- a. Enter the following command to display the current field order selection menu:

**f**

- b. Enter the following command to enable queue statistics:

**f**

- c. Enter the following commands to sort fields:

- **r** — Sort by reads per second
- **R** — Sort by MB reads per second
- **w** — Sort by writes per second
- **T** — Sort by MB writes per second

## Viewing Queue Depth at Virtual Machine Level

To view queue depth at virtual machine level, follow these steps:

1. Run `esxtop` in interactive mode.

2. Enter the following command:

**v**

`esxtop` switches to storage disk virtual machine view.

3. Enter the following command:

**e**

`esxtop` displays **VM to expand/rollup** to world prompt.

4. Enter the GID of a virtual machine as listed under the GID column.

An expanded view of a virtual machine statistics displays.

5. Enter the following command:

**i**

`esxtop` displays a **World to expand at device level** prompt.

6. Enter the world ID of the virtual machine's VMM.

An expanded view of the virtual machine's device displays. **DQLEN** shows the device's queue depth. **WQLEN** shows the world queue depth. Schedulable entities in the VMkernel are referred to as worlds, where worlds can be a virtual machine. **ACTV** shows number of commands currently active in the VMkernel. **QUED** shows the number of commands queued in the VMkernel. **%USD** show the percent of the device queue depth used actively by the VMkernel.

```
12:34:47am up 4 days 6:58, 144 worlds; CPU load average: 0.02, 0.02, 0.02
```

ID	GID	NAME	DEVICE	NWD	NDV	DQLEN	WQLEN	ACTV	QUED	%USD	LOAD	CMDS/s
2	2	system	-	3	-	-	-	0	0	0	0.00	0.00
6	6	helper	-	19	-	-	-	0	0	0	0.00	0.00
10	10	console	-	1	-	-	-	0	0	0	0.00	10.90
21	21	vm03	-	4	-	-	-	0	0	0	0.00	0.00
22	22	vm04	-	4	-	-	-	0	0	0	0.00	0.00
4285	23	vmware-vmx	-	4	1	-	32	0	0	0	0.00	0.00
4301	23	vcpu-0:vm01	-	4	1	-	32	0	0	0	0.00	0.00
4305	23	vcpu-1:vm01	-	4	1	-	32	0	0	0	0.00	0.00
4290	23	vmm0:vm01	naa.60060e8	4	1	32	32	0	0	0	0.00	1543.91
24	24	vm02	-	4	-	-	-	0	0	0	0.00	0.00

## Monitoring Virtual Machine I/O Latency

To view virtual machine I/O latency, follow these steps:

1. Run `esxtop` in interactive mode.
2. Enter the following command:

**v**

`esxtop` switches to storage disk virtual machine view. `DAVG/cmd` shows the average device latency in milliseconds per command. `KAVG/cmd` shows the average ESX VMkernel latency in milliseconds per command. `GAVG/cmd` shows the average Guest OS latency in milliseconds per command. `GAVG/cmd` is the summation of the `KAVG/cmd` and `DAVG/cmd`. `QAVG/cmd` shows the average queue latency in milliseconds per command. An increase in `KAVG/cmd` might be an indication that the LU queue depth might be exceeded where the I/O is now queued in the VMkernel. Consider adjusting the queue depth or storage vMotion the virtual machine to another volume.

```
9:31:12pm up 36 days 3:55, 135 worlds; CPU load average: 0.02, 0.03, 0.03
```

ID	GID	NAME	DAVG/cmd	KAVG/cmd	GAVG/cmd	QAVG/cmd
2	2	system	0.00	0.00	0.00	0.00
6	6	helper	0.08	0.01	0.09	0.00
10	10	console	4.90	0.01	4.92	0.01
20	20	vmware-vmkauthd	0.00	0.00	0.00	0.00
139	139	seoracle	1.91	0.01	1.91	0.00

3. Check to see if `DAVG/cmd`, `KAVG/cmd`, and `GAVG/cmd` columns appear in the results.

If these columns do not appear, latency statistics are not enabled. Follow these steps to enable queue statistics:

- a. Enter the following command to display the current field order selection menu:

**f**

- b. Enter the following command to enable latency statistics:

**j**

If the latency statistics view is truncated, disable other statistics by toggling their fields from the current field order selection menu.

 **Hitachi Data Systems Corporation**

---

*Hitachi is a registered trademark of Hitachi, Ltd., 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. All other trademarks, service marks and company names mentioned in this document are properties of their respective owners.*

*Notice: This document is for informational purposes only, and does not set forth any warranty, expressed or implied, concerning any equipment or service offered or to be offered by Hitachi Data Systems Corporation*

© Hitachi Data Systems Corporation 2011. All Rights Reserved. AS-053-02 February 2011

**Corporate Headquarters**

750 Central Expressway,  
Santa Clara, California 95050-2627 USA  
[www.hds.com](http://www.hds.com)

**Regional Contact Information**

**Americas:** +1 408 970 1000 or [info@hds.com](mailto:info@hds.com)  
**Europe, Middle East and Africa:** +44 (0) 1753 618000 or [info.emea@hds.com](mailto:info.emea@hds.com)  
**Asia Pacific:** +852 3189 7900 or [hds.marketing.apac@hds.com](mailto:hds.marketing.apac@hds.com)