What's New in VMware vSphere® 5.1 – Storage

VMware vSphere® 5.1

TECHNICAL MARKETING DOCUMENTATION
V 1.01 / UPDATED JUNE 2012

Table of Contents

Introduction

VMware vSphere® 5.1 brings many new capabilities that extend the benefits of vSphere 5.0, including a number of new, storage-related features. Many of these enable VMware® flagship products such as VMware vCloud® Director™ and VMware® View™. vSphere 5.1 storage features and enhancements also bring additional scalability, performance and troubleshooting capabilities. This paper focuses on the storage-specific features and enhancements that are available in vSphere 5.1.

The following topics will be discussed in this paper:

- VMware vSphere® VMFS-5 file-sharing enhancements
- Space-efficient sparse virtual disks
- All paths down (APD) condition–handling enhancements
- Storage protocol supportability improvements
- VMware vSphere® Storage APIs – Array Integration (VAAI)
- Boot from software Fibre Channel over Ethernet (FCoE)
- I/O device management (IODM)
- Solid-state disk (SSD) monitoring
- VMware vSphere® Storage I/O Control enhancements
- VMware vSphere® Storage DRS™ enhancements
- VMware vSphere® Storage vMotion® enhancements

This will provide a technical overview of new capabilities and enhancements regarding each of these new storage features.

VMware vSphere VMFS-5 File Sharing Enhancements

Prior to vSphere 5.1, the maximum number of hosts that could share a read-only file on a Virtual Machine File System (VMFS) was eight. This is a limiting factor for virtual desktop infrastructure (VDI) deployments using View, because the maximum number of hosts in a cluster that could share the same desktop base disk image file was also eight. It is also a limiting factor for deployments of vCloud Director, which uses linked clones for the fast provisioning of VMware vCloud® Director™ vApps™.

In vSphere 5.1, the maximum number of hosts that can share a read-only file on a VMFS is increased to 32. This means that linked clones deployed from a base image now can be hosted on any one of the 32 hosts sharing the datastore. This provides significant scalability for both View and vCloud Director. This feature applies only to hosts running vSphere 5.1 and higher on VMFS-5.

NOTE: VMware View 5.0 currently limits the number of hosts a pool can use to eight, even when using NFS for replica storage. VMware View 5.1 increases this host count to 32 for NFS. In a future release of VMware View, the host count will also be increased to 32 for VMFS. This implies that VMFS will be as scalable as NFS for future View deployments.

Space-Efficient Sparse Virtual Disks

Thin provisioning of storage addressed a major storage inefficiency issue by allocating blocks of storage to a guest operating system (OS) file system or database only as they were needed, rather than at the time of creation. However, traditional thin provisioning does not address reclaiming stale or deleted data within a guest OS, leading to a gradual growth of storage allocation to a guest OS over time.

With the release of vSphere 5.1, VMware introduces a new virtual disk type, the space-efficient sparse virtual disk (SE sparse disk). One of its major features is the ability to reclaim previously used space within the guest OS.

Another major feature of the SE sparse disk is the ability to set a granular virtual machine disk block allocation size according to the requirements of the application. Some applications running inside a virtual machine work best with larger block allocations; some work best with smaller blocks. This was not tunable in the past.

Space Reclaim

The new SE sparse disk implements a space reclaim feature to reclaim blocks that were previously used but now are unused on the guest OS. These are blocks that were previously written but currently are unaddressed in a file system/database due to file deletions, temporary files, and so on.

There are two steps involved in the space reclamation feature: The first step is the wipe operation that frees up a contiguous area of free space in the virtual machine disk (VMDK); the second step is the shrink, which unmaps or truncates that area of free space to enable the physical storage to be returned to the free pool.

Wipe

- Initiate a call to VMware Tools to scan the guest OS file system.
- Mark the unused blocks as free.
- Run the SCSI UNMAP command in the guest to instruct the virtual SCSI layer in the VMkernel to mark the blocks as free in the SE sparse disk.

Shrink

- SCSI device – VMware® ESXi™ issues an SCSI UNMAP command to the array.
- NFS device – ESXi issues an RPC call to TRUNCATE the file.

The wipe operation is initiated via an API call to VMware Tools. VMware Tools initiates a scan of the guest OS to find stranded space and mark the file system blocks as free. The first SCSI UNMAP operation is then run from within the guest OS, instructing the VMkernel as to which blocks can be reclaimed. The VMkernel captures these SCSI UNMAP commands and does not pass them through to the array. When the VMkernel detects which blocks are free, it uses its virtual SCSI layer to reorganize the SE sparse disk by moving blocks from the end of the disk to unallocated blocks at its beginning. This creates a contiguous area of free space within the VMDK. The shrink operation then sends either an SCSI UNMAP command (for SCSI disks) or an RPC TRUNCATE command (for NFS) to the array to free the space.

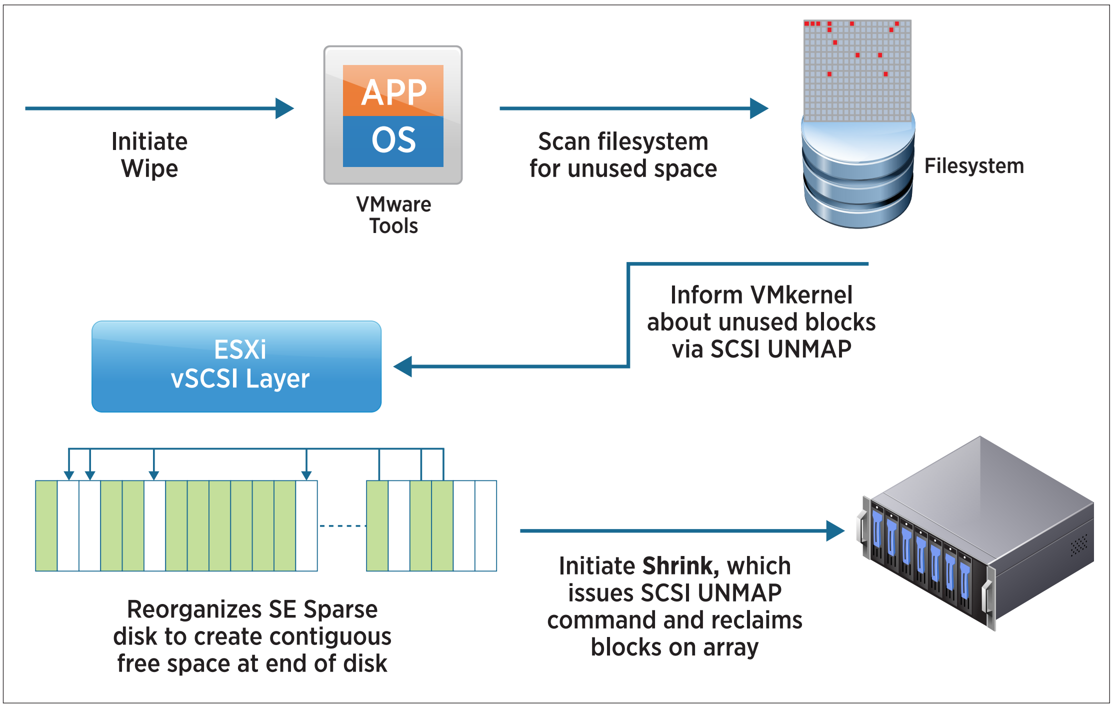

Figure 1. Wipe and Shrink Operation

The virtual machines require Hardware Version 9 (HWv9) to handle the SCSI UNMAP command in the guest OS. Earlier versions cannot handle this and will fail the operation.

New Grain Size

In vSphere 5.0, the default grain size/block allocation unit size for virtual machine disks on ESXi was 4KB. Redo logs, used by snapshots and linked clones, had a grain size of 512 bytes (one sector).

As mentioned previously, with the introduction of SE sparse disks, the grain size is now tunable and can be set based on the requirements of a particular storage array or application.

In the initial release of SE sparse disks in vSphere 5.1, the default grain size is set to 4KB. Specific VMware products and features that use the new SE sparse disk format for redo logs/linked clones will also use this new default grain size.

NOTE: Direct user tuning of the grain size is not exposed in vSphere 5.1.

SE Sparse Disk Initial Use Case

The scope of SE sparse disks in vSphere 5.1 is restricted to VMware View. VMware® View™ Composer can use *linked clones* for the rollout of desktops. Linked clones are read/write snapshots of a read-only parent desktop image. View benefits from the new 4KB grain size, which improves performance by addressing alignment issues experienced in some storage arrays with the 512-byte grain size used in linked clones based on the vmfsSparse (redo log) format. The SE sparse disk format also provides far better space efficiency to desktops deployed on this virtual disk format, especially with its ability to reclaim stranded space.

When deploying desktops with View Composer, a number of images/snapshots are created. In the vSphere 5.1 release, SE sparse disk format can be used by the View Composer for linked clones and subsequent snapshots of them. These images represent the majority of storage used in a View environment.

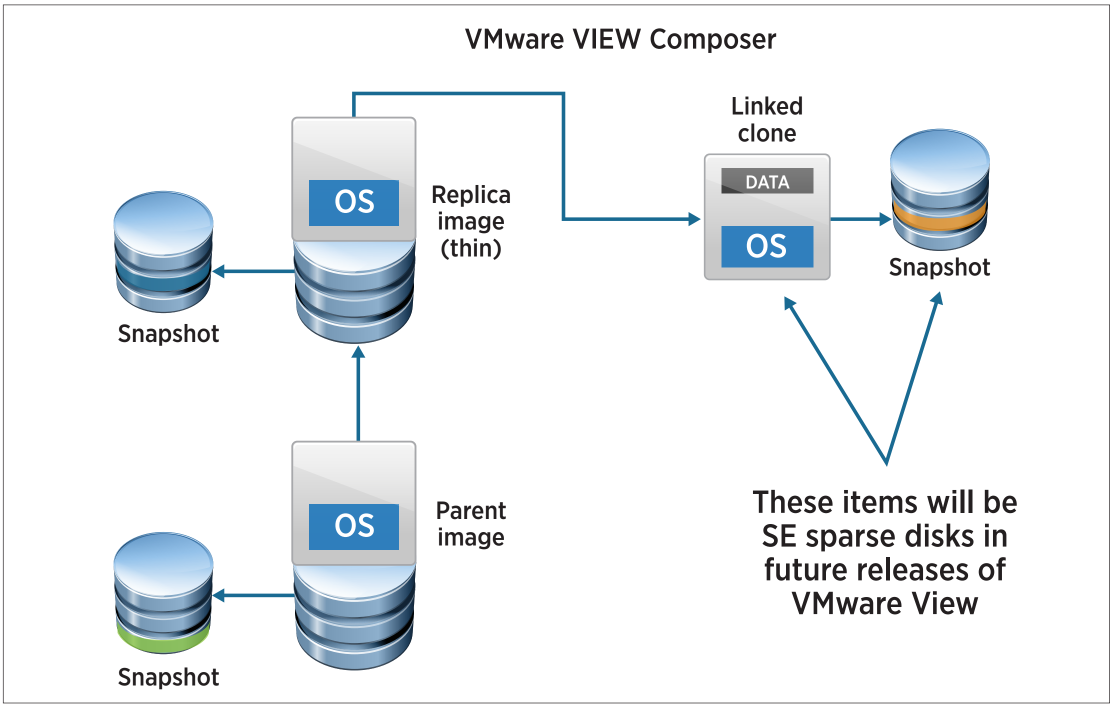

Figure 2. View Composer Use Case for SE Sparse Disks

All Paths Down

All paths down (APD) can occur on an ESXi host when a storage device is removed in an uncontrolled manner from the host, or if the device fails, and the VMkernel core storage stack cannot detect how long the loss of device access will last. One possible scenario for an APD condition is a Fibre Channel (FC) switch failure that brings down all the storage paths—or, in the case of an iSCSI array, a network connectivity issue that similarly brings down all the storage paths.

Over the previous several vSphere releases, VMware has made significant improvements in handling the APD condition. This is a difficult situation to manage because the ESXi host cannot detect whether it is a permanent device loss or a transient state. The biggest issue regarding APD is the effect on hostd, which is responsible for managing many ESXi host operations. It manages virtual machine operations such as creation, power on/off, VMware vSphere® vMotion®, and so on. It also manages LUN and VMFS volume discovery. If an administrator issues a rescan of the SAN, hostd worker threads will wait indefinitely for I/O to return from the device in APD. However, hostd has a finite number of worker threads, so if all these threads get tied up waiting for disk I/O, other hostd tasks will be affected. This is why a common symptom of APD is ESXi hosts disconnecting from VMware® vCenter™ because their hostd process has become hung.

A new condition known as permanent device loss (PDL) was introduced in vSphere 5.0. The PDL condition enabled the ESXi host to take specific actions when it detected that the device loss was permanent. The ESXi host can be informed of a PDL situation by specific SCSI sense codes sent by the target array.

vSphere 5.1 aims to improve APD and PDL handling in the following three ways:

- Handle more complex transient APD conditions. Do not allow hostd to become hung indefinitely when devices are removed in an uncontrolled manner.
- Enable VMware vSphere® High Availability (vSphere HA) to detect PDL and be able to restart virtual machines on other hosts in the cluster that might not have this PDL state on the datastore.
- Introduce a PDL method for those iSCSI arrays that present only one LUN for each target. These arrays were problematic, because after LUN access was lost, the target also was lost. Therefore, the ESXi host had no way of reclaiming any SCSI sense codes.

Host Resilience via APD Timeout

APD affects more than just virtual machine I/O. It can also affect hostd worker threads, leading to host disconnects from vCenter in worst-case scenarios. It can also affect vmx I/O when updating virtual machine configuration files. On occasion, VMware support has observed scenarios where the .vmx file was affected by an APD condition.

In vSphere 5.1, a new time-out value for APD is being introduced. There is a new global setting for this feature called **Misc.APDHandlingEnable.** If this value is set to 0, the current (vSphere 5.0) condition is used, i.e., permanently retrying failing I/Os.

If **Misc.APDHandlingEnable** is set to 1, APD handling is enabled to follow the new model, using the time-out value **Misc.APDTimeout.** This is set to a 140-second time-out by default, but it is tunable. These settings are exposed in the UI. When APD is detected, the timer starts. After 140 seconds, the device is marked as **APD Timeout.** Any further I/Os are fast-failed with a status of **No_Connect**, preventing hostd and others from getting hung.

If any of the paths to the device recover, subsequent I/Os to the device are issued normally, and special APD treatment concludes.

vSphere HA PDL Handling

Although not strictly a vSphere 5.1 storage feature, vSphere HA now has the ability to restart virtual machines that were terminated during a PDL condition. The objective is to start the affected virtual machines on another host that might not be in a PDL state for the shared storage device. vSphere HA does not migrate a virtual machine’s disks—in the event of a failure, it attempts to start a virtual machine on another host but on the same storage.

The **disk.terminateVMOnPDLDefault** option enables vSphere HA to terminate running virtual machines when the datastore on which they reside enters PDL. This option was added in vSphere 5.0, but there was an issue with vSphere HA in identifying whether a virtual machine had been terminated or shut down gracefully by an operator during a PDL. Because of this, virtual machines that had been gracefully shut down on a host that was in PDL could inadvertently be restarted on another host in the cluster.

This now has been addressed via the **das.maskCleanShutdownEnabled** option that was introduced in vSphere 5.0 Update 1. This option enables vSphere HA to differentiate between virtual machines that have been shut down gracefully during a PDL, and therefore should not be restarted, and virtual machines that have been terminated during a PDL, and therefore should be restarted.

PDL for iSCSI Arrays with Single LUN per Target

vSphere 5.1 extends PDL detection to those arrays that have only a single LUN per target. With vSphere 5.0, when LUN access was lost on these arrays, so was access to the target, so the ESXi host could not reclaim an SCSI sense code to check whether the device was in PDL.

With vSphere 5.1, for those iSCSI arrays that have a single LUN per target, an attempt is made to log in again to the target after a dropped session. If there is a PDL condition, the storage system rejects the effort to access the device. Depending on how the array rejects the efforts to access the LUN, the ESXi host can determine whether the device has been lost permanently (PDL) or is temporarily unreachable.

Storage Protocol Improvements

There have been a number of storage protocol improvements introduced into vSphere 5.1.

Boot from Software FCoE

In vSphere 4.1, VMware introduced the ability to install and boot an ESXi host via the software iSCSI initiator. vSphere 5.1 adds similar capabilities for the software FCoE initiator, which was introduced in vSphere 5.0. It enabled access to FCoE storage but did not support booting the ESXi host from FCoE storage. Booting a vSphere 5.0 host over FCoE required a dedicated FCoE hardware HBA.

VMware has seen a proliferation of diskless blade servers in the marketplace. And as customers move toward converged networks, FCoE is gaining traction. vSphere 5.1 enhances the software FCoE initiator to enable an ESXi 5.1 host to install to and boot from an FCoE LUN, using the software FCoE initiator rather than requiring a dedicated FCoE hardware adapter.

The configuration is quite simple because most of it is done via the option ROM of the network interface card with FCoE capabilities. The configuration parameters are set in the network interface card option ROM.

When using FCoE boot with a network adapter, an area is created in memory to expose information about the FCoE connection. A special VMkernel module (vmkfbft) accesses the configuration settings in these tables and stores them in VMkernel system information (VSI) nodes in the ESXi host. This enables the ESXi host to determine whether the attached device is bootable.

If the correct configuration parameters have been provided in the FCoE option ROM on the network adapter, FCoE LUN discovery should be available during the boot of the host, as illustrated in Figure 3.

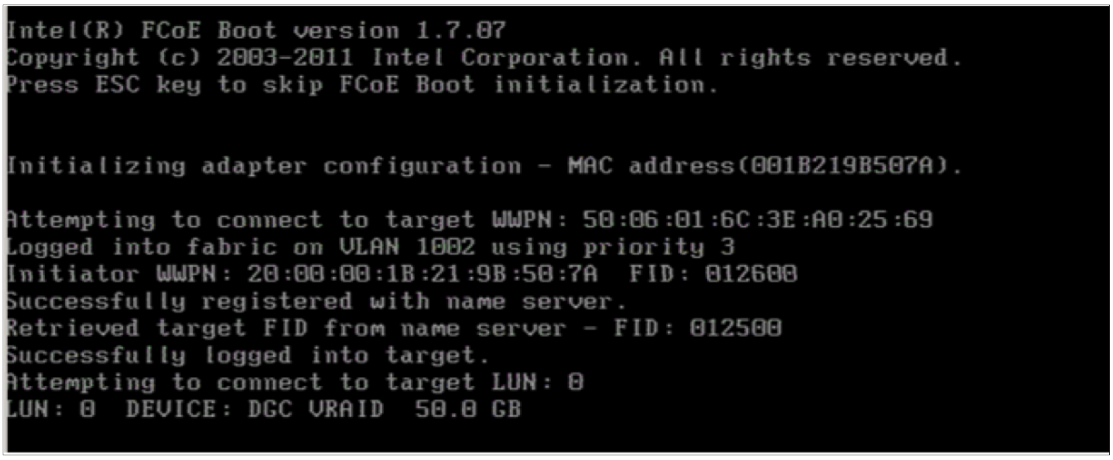

Figure 3. FCoE Boot Initialization

iSCSI Storage Driver Upgrade for Jumbo Frames

This enhancement adds jumbo frame support, including user interface support, to all iSCSI adapters—software iSCSI adapters, dependent hardware iSCSI adapters and independent hardware iSCSI adapters.

Independent hardware iSCSI adapters maintain their own networking and target configuration. They also provide a third-party IMA plug-in library as a management interface, e.g., the QLogic qla4xxx iSCSI driver. Dependent hardware iSCSI adapters require VMkernel networking configuration as well as VMware software iSCSI IMA plug-in and iSCSI daemon for session management, e.g., Broadcom bnx2i iSCSI driver.

For independent adapters, the MTU is set in the “Properties” dialog for the adapter. For software adapters and dependent adapters, the MTU is picked up from the underlying vmknic settings.

Support 16Gb FC HBAs

In vSphere 5.0, VMware introduced support for 16Gb FC HBAs, but they had to be throttled down to work at 8Gb. Now, vSphere 5.1 supports these HBAs running at 16Gb. However, there is no support for full, end-to-end 16Gb connectivity from host to array. To get full bandwidth, a number of 8Gb connections can be created from the switch to the storage array.

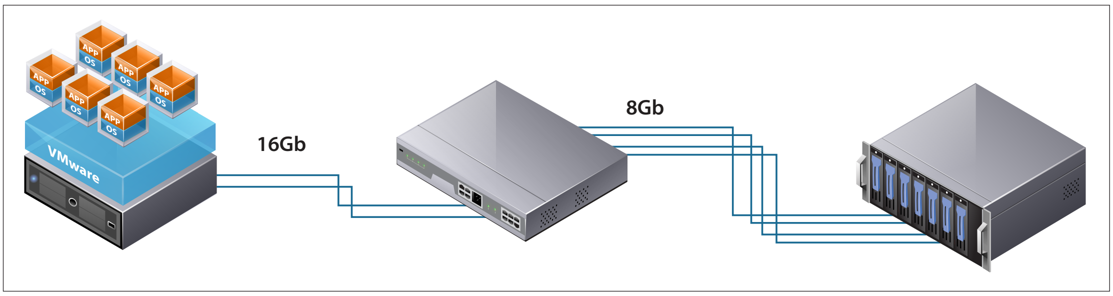

Figure 4. 16Gb HBA Support

VAAI

VAAI is the vSphere storage APIs for array integration. It enables certain storage operations to be off-loaded from the ESXi host to the storage array.

vSphere 5.0 introduced the off-loading of View desktop snapshots to the array via VAAI NAS primitives. vSphere 5.1 introduces additional VAAI NAS enhancements to enable array-based snapshots to be used for vCloud Director fast-provisioned vApps. This feature in vSphere 5.1 enables vCloud Director to off-load the creation of linked clones to the NAS storage array in a similar manner to how View does it in vSphere 5.0. When vCloud Director does a fast provision of a vApp/virtual machine, it transparently uses VAAI NAS to off-load the creation of the subsequent linked clones to the VAAI-supported storage array.

Similar to VAAI NAS support for View in vSphere 5.0, this feature also requires a special VAAI NAS plug-in from the storage array vendor.

Advanced I/O Device Management

Building on vSphere 5.0 enhancements that ease the storage-related demands on a vSphere administrator, vSphere 5.1 includes additional command capabilities for diagnosis of various storage protocol issues as supported by vSphere, including Fibre Channel, FCoE, iSCSI and SAS. The aim is to enable an administrator to determine whether a storage issue is occurring at the ESXi, HBA, fabric or storage-port level. The commands enable an administrator to examine critical events such as frame loss as well as to initiate various resets of the storage infrastructure.

There are a number of new namespaces in the vSphere 5.1 version of **esxcli.** There is also a new VMkernel module that instrumented drivers can call into and retrieve event caching information. For example, link-down and link-up messages from Fibre Channel are logged. The FC namespace also includes an option to perform a loop initialization primitive (LIP) reset to a given FC adapter on the system.

```
~ # esxcli storage san
Usage: esxcli storage san {cmd} [cmd options]
Available Namespaces:
   fc     IO Device Management operations to the FC adapters on the system.
   fcoe   IO Device Management operations to the FCoE adapters on the system.
   iscsi  IO Device Management operations to the SW iSCSI adapters on
          the system.
   sas    IO Device Management operations to the SAS adapters on the system.
```

SSD Monitoring

As SSDs become more prevalent, it is important to be able to monitor them from an ESXi host. Smartd is the SSD Self-Monitoring, Analysis and Reporting Technology (SMART) daemon on the ESXi 5.1 host. It runs every 30 minutes and makes API calls to gather disk information. The SMART features are also very useful, enabling insight into SAS and SATA SSD status, such as the media wearout indicator as well as the temperature and reallocated sector count. The reserved sector count should be about 100. But when a drive has issues, the SSD allocates from among the reserved sectors. When this figure starts to diminish, ESXi might start getting sector errors on the SSD, so the administrator must be aware of any use of the reallocated sectors.

The following table presents the output of a number of different SSD attributes, including the three mentioned previously:

```
~ # esxcli storage core device smart get -d t10.ATAxxxxxxxx
```

Parameter	Value	Threshold	Worst
Health Status	OK	N/A	N/A
Media Wearout Indicator	N/A	N/A	N/A
Write Error Count	N/A	N/A	N/A
Read Error Count	114	6	100
Power-on Hours	90	0	90
Power Cycle Count	100	20	100
Reallocated Sector Count	2	36	2
Raw Read Error Rate	114	6	100
Drive Temperature	33	0	53
Driver Rated Max Temperature	67	45	47
Write Sectors TOT Count	200	0	200
Read Sectors TOT Count	N/A	N/A	N/A
Initial Bad Block Count	100	99	100

VMware provides a generic SMART plug-in in vSphere 5.1, but disk vendors can provide their own SMART plug-in that enables additional, vendor-specific information to be retrieved about their particular brand of disk. The plug-ins are installed on the ESXi host in the directory **/usr/lib/vmware/smart_plugins**.

These events and statistics are *not* surfaced up into vCenter in vSphere 5.1. They are available only via the **esxcli** command line. There is a script called **smartinfo.sh** that gathers statistics from all disks attached to the host, whether SSD or not. The technical support log-gathering utility “vm-support” automatically collects SMART statistics for local and SSD devices using this script.

Storage I/O Control

The default latency threshold for storage I/O control (SIOC) is **30msecs.** Not all storage devices are equal, so this default is set to the middle of the range value. There are devices that hit their natural contention point earlier than others, for example, SSDs, for which the threshold should be decreased by the user. However, manually determining the correct latency can be difficult. This generates a need for the latency threshold to be automatically determined for each device. Rather than using a default/user selection for latency threshold, SIOC now can automatically determine the best threshold for a datastore.

The latency threshold is set to the value determined by the I/O injector (a part of SIOC). When the I/O injector calculates the peak throughput, it then finds the 90 percent throughput value and measures the latency at that point to determine the threshold.

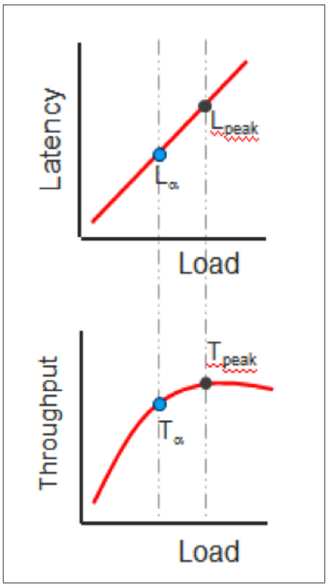

Figure 5. Automatic Latency Threshold

vSphere administrators can change this 90 percent to another percent value or they can continue to input a millisecond value if they choose.

Another enhancement to SIOC is that it now is turned on in "stats only" mode automatically. This means that interesting statistics that are presented only when SIOC is enabled are available immediately now. It also means that Storage DRS now has statistics in advance for new datastores being added to the datastore cluster.

VMware vSphere Storage DRS 2.0

Storage DRS and vCloud Director Interoperability

One of the major enhancements to Storage DRS (version 2.0) in vSphere 5.1 enables interoperability with vCloud Director. This essentially means that vCloud Director now can detect datastore cluster objects from Storage DRS and that Storage DRS can detect linked clones (on which fast-provisioned vApps are built). vCloud Director now uses Storage DRS for initial placement, space utilization and I/O load balancing of fast-provisioned vApps/linked clones.

With the detection of vCloud Director linked clones, Storage DRS selects a datastore that contains either the base disk or a shadow virtual machine (a copy of the base disk) when it must migrate a linked clone.

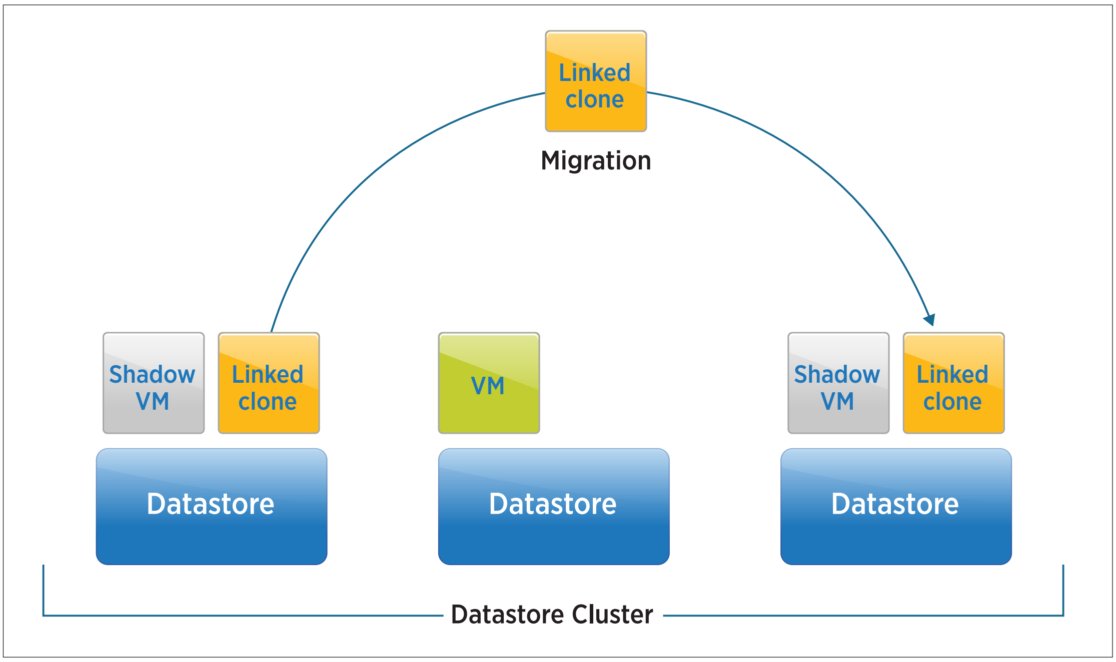

Figure 6. Storage DRS and Shadow Virtual Machines

If there is a cross-datastore linked clone configuration (created via vCloud Director APIs, for example), Storage DRS will not make a recommendation to place linked clones on the datastore that does not contain the base disk or a shadow virtual machine copy of the base disk.

Linked clones can be migrated between VMFS-3 and VMFS-5 file systems. There are a number of factors that come into the decision-making process when Storage DRS is determining where to migrate a linked clone. Factors such as the amount of data being moved, the amount of space reduction on the source and the additional amount of space required on the destination all are considered. The major decision is whether or not a shadow virtual machine of the base disk already exists on the destination.

Datastore Correlation

Datastore correlation refers to discovering whether two distinct datastores might be using the same set of disk spindles on the array. The purpose of this feature is to help the decision-making process in Storage DRS when determining where to move a virtual machine. For instance, there is little advantage to moving a virtual machine from one datastore to another if the two datastores are backed by the same set of physical spindles on the array. The datastore correlation detector can also be used for antiaffinity rules, making sure that virtual machine disks not only are kept on separate datastores but also are kept on different spindles on the back-end array.

The ability to detect datastore correlation was included in Storage DRS in vSphere 5.0, but it required VMware vSphere Storage APIs for Storage Awareness (VASA) support for the array.

vSphere 5.1 provides additional capability to the datastore correlation detector to use the I/O injector to determine whether a source and destination datastore are correlated, that is, using the same back-end spindles. It works by placing load on one datastore and monitoring the latency on another. If we see latency increase on both datastores when load is placed on one datastore, we can assume that the datastores are correlated. This enhancement allows datastore correlation detection on storage arrays where there is not specific VASA support.

VmObservedLatency

VmObservedLatency is a new metric in SIOC in vSphere 5.1. It replaces the datastore latency metric that was used in previous versions of SIOC. This new metric measures the time between receipt by VMkernel of the I/O from the virtual machine and receipt of the response from the datastore. Previously we measured the latency only after the I/O had left the ESXi host. Now we also are measuring latency in the host. This new metric is visible in the VMware vSphere® Client™ performance charts.

If all hosts using the datastore cluster are vSphere 5.1, Storage DRS uses this new metric, which closely represents the load on the datastore as seen by virtual machines.

Parallel Storage vMotions

vSphere 5.1 allows up to 4 parallel disk copies per Storage vMotion operation, where previous versions of vSphere used to copy disks serially. When you migrate a virtual machine with five VMDK files, Storage vMotion copies of the first four disk in parallel, then start the next disk copy as soon as one of the first four finish.

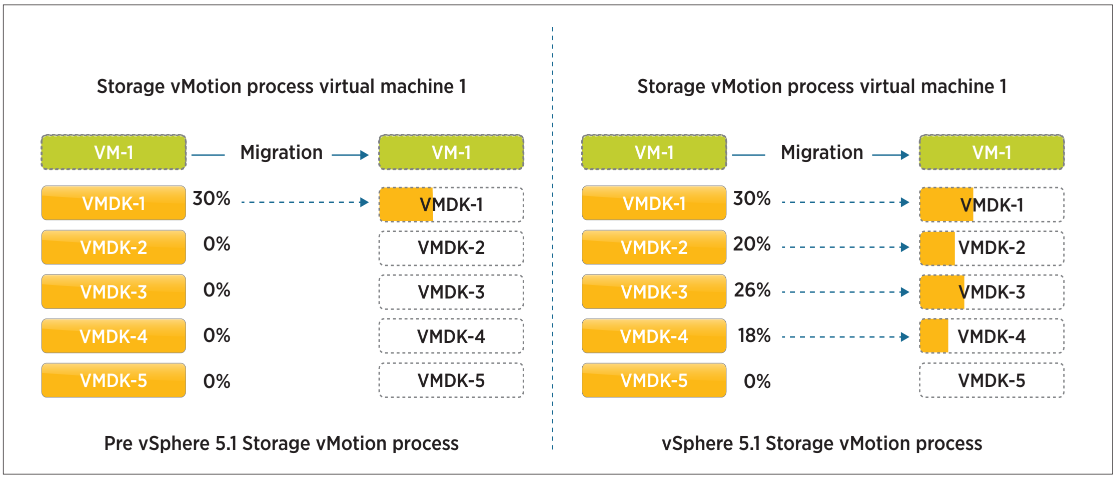

Figure 7. Storage vMotion

To reduce performance impact on other virtual machines sharing the datastores, parallel disk copies only apply to disk copies between distinct datastores. This means that if a virtual machine has multiple VMDK files on datastore A and B, parallel disk copies will only happen if destination datastores are C and D.

A fan out disk copy, in other words copying two VMDK files on datastore A to datastores B and C, will not have parallel disk copies. The common use case of parallel disk copies is the migration of a virtual machine configured with an anti-affinity rule inside a datastore cluster.

Conclusion

As mentioned in the introduction, one of the goals of storage enhancements in vSphere 5.1 is to increase the scalability of VMware View and VMware vCloud Director. This is accomplished in the following ways:

- Increasing the file share limit on VMFS-5 from 8 to 32, making VMFS as scalable a storage solution as NFS
- Making View desktops more space efficient using SE sparse disks
- Giving vCloud Director the ability to employ a VAAI NAS feature to off-load linked clones to use native snapshots on the array
- Introducing enhancements to storage protocols, such as 16Gb FC support, jumbo frame support for all iSCSI adapters, and boot from the software FCoE adapter, keeping VMware vSphere at the forefront of open storage
- Introducing new commands in I/O device management that aim to empower vSphere administrators from a storage monitoring and troubleshooting perspective
- Introducing enhancements to Storage DRS, storage I/O control and VMware vSphere Storage vMotion

About the Author

Cormac Hogan is a senior technical marketing architect in the Cloud Infrastructure Product Marketing group at VMware. His focus is on virtual storage in general as well as core VMware vSphere storage technologies in particular, including the VMware vSphere® Storage Appliance. He has been in technical marketing since 2011.

- Follow Cormac's blogs at http://blogs.vmware.com/vsphere/storage.
- Follow Cormac on Twitter: @VMwareStorage.

VMware, Inc. 3401 Hillview Avenue Palo Alto CA 94304 USA Tel 877-486-9273 Fax 650-427-5001 www.vmware.com