What's New in VMware® vSphere™ 4.1 — Storage

VMware vSphere 4.1

WHITE PAPER

vmware®

Introduction

VMware® vSphere™ 4.1 brings many new capabilities to further extend the benefits of vSphere 4.0. These new features and enhancements to core capabilities in vSphere provide more performance optimization, and easier provisioning, monitoring and troubleshooting capabilities. This paper focuses on the storage-specific features and enhancements that are available in 4.1, and provides an overview of what new means exist for optimizing, monitoring and troubleshooting storage-related issues with both the vCenter and command-line interfaces provided in this release. Wherever possible, we will also provide use cases and requirements that may apply to the use of these new functions.

The topics to be covered in this paper are:

- Storage I/O Control
- vStorage API for Array Integration
- New performance monitoring screens
- Support for iSCSI hardware offload
- Support for 8Gb host–based adaptors

This will provide a technical overview of new capabilities as well as links to additional information about each of these new storage features.

Storage I/O Control

Storage I/O Control (SIOC) is a new feature introduced in vSphere 4.1 to provide I/O prioritization of virtual machines running on a cluster of ESX servers that access a shared storage pool. It extends the familiar constructs of shares and limits, which have existed for CPU and memory, to address storage utilization through a dynamic allocation of I/O queue slots across a cluster of ESX servers. When a certain latency threshold is exceeded for a given block-based storage device, SIOC will balance the available queue slots across a collection of ESX servers to align the importance of certain workloads with the distribution of available throughput. It can reduce the I/O queue slots given to virtual machines that have a low number of shares in the interest of providing more I/O queue slots to a virtual machine with a higher number of shares.

SIOC provides a means of throttling back I/O activity for certain virtual machines in the interest of other virtual machines getting a more fair distribution of I/O throughput and an improved service level. In Figure 1 the two business-critical virtual machines (online store and MS Exchange) are provided more I/O slots than the less important (data mining) virtual machine.

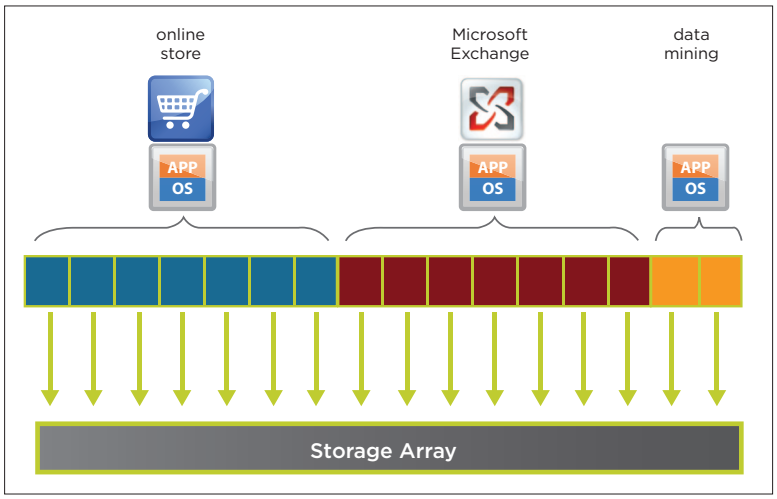


Figure 1.

For SIOC to engage in optimizing I/O to a given datastore there are two things that must be present.

1. The datastore must have this feature enabled. This is done by changing a property setting of that datastore.
2. There must be a sustained average latency detected across the hosts (ESX servers) that share that datastore. The default threshold is 30ms and can be modified through the advanced setting options for the datastore properties.

Once both of those conditions are met, SIOC engages in proactively managing the I/O queues across all ESX servers that share the datastore. It evaluates the percentage of I/O shares each virtual machine has relative to the total number of shares of all virtual machines accessing the datastore and will assign a relative number of I/O queue slots to ensure that high-priority virtual machines get more throughput and less latency than the lower-priority virtual machines. SIOC will throttle back the I/O slots on one ESX server in which a low-priority virtual machine might be the only workload running, to free up I/O queues on another server that might have several virtual machines running. See Figures 2 and 3.

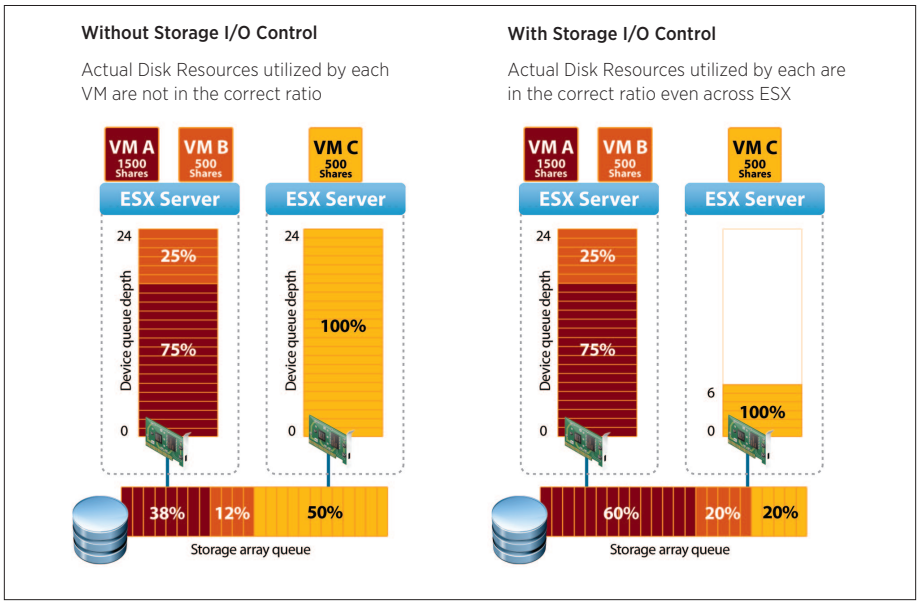


Figure 2. Without SIOC

Figure 3. With SIOC

SIOC provides a dynamic allocation mechanism that adjusts to changing conditions of a mixed workload. It leverages the I/O shares, which are set on the Virtual Machine Properties for each virtual disk, to distribute available I/O slots to ensure a quality of service is enforced not just at the host level, but across the collection of hosts that are sharing that datastore. This feature provides a new means for the vSphere administrator to achieve higher levels of consolidation with the confidence that shared resource pools will not result in low-priority workloads limiting the performance of higher-priority workloads. SIOC also benefits your virtual environment by providing I/O distribution fairness even when all virtual machines running on a cluster of ESX servers sharing a datastore have equal or default I/O shares.

SIOC works only with block-based datastores and datastores that reside on a single extent and are managed by one vCenter management server. More details about this feature are described in the technical white paper on SIOC concepts and deployment considerations as well as in the documentation. *Resource Management Guide, Chapter 4*.

vStorage API for Array Integration

The vStorage API for Array Integration (VAAI) is a new API for storage partners to leverage as a means to speed up certain functions that, when delegated to the storage array, can greatly enhance performance. This API is currently supported by several storage partners and requires these partners to release a special version of their firmware to work with this API. In the vSphere 4.1 release, this array offload capability supports three primitives:

1. Full copy enables the storage arrays to make full copies of data within the array without having the ESX server read and write the data.
2. Block zeroing enables storage arrays to zero out a large number of blocks to speed up provisioning of virtual machines.
3. Hardware-assisted locking provides an alternative means to protect the metadata for VMFS cluster-file systems, thereby improving the scalability of large ESX server farms sharing a datastore.

Full Copy

The ability to deploy a virtual machine will be greatly increased by use of the full copy primitive, as the process can be done either within the storage array or, for some storage vendors, between arrays where an xcopy function is supported. The process that took a few minutes will be accomplished in a matter of seconds and the traffic reduction on the ESX server will also lower the amount of CPU needed to perform this function. The benefit of this primitive gets much more interesting in a desktop infrastructure environment in which one might be doing hundreds of virtual machine deployments from templates.

With Storage vMotion, there is a similar reduction in the time it takes to migrate virtual machine's home, because copying no longer needs to go up to the ESX server and then back down to the array again. That frees up storage I/O and server CPU cycles.

Not only does this save time, but it also saves server CPU and memory, as well as network bandwidth and storage front-end controller I/O. Full copy enables as much as a 95 percent reduction in most of these metrics.

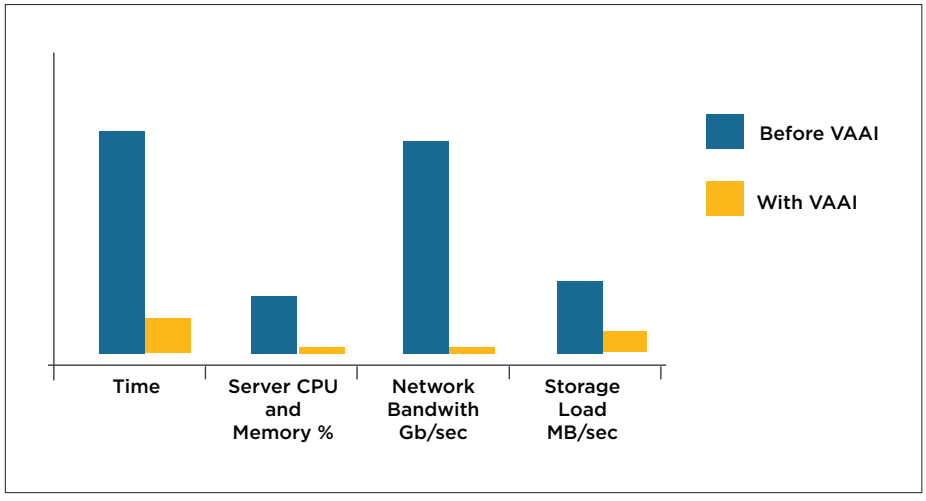


Figure 4.

Block Zeroing

To have the array complete a bulk zeroing out of a disk speeds up a standard initialization process. One use for block zeroing is to create a virtual disk as eager-zero thick in format. Without the block zeroing primitive, the command is not complete until the disk array has completed the zeroing process. For a very large disk this could take a long time. The block zeroing primitive, also referred to as "copy same," enables the disk array to return the cursor to the requesting service as though the process of writing the zeros has been completed and then finish the job of zeroing out those blocks without the need to hold the cursor until the job is done.

Hardware-Assisted Locking

The third primitive for VAAI in our 4.1 release is hardware-assisted locking. It provides a more granular means to protect the VMFS metadata than the SCSI reservations used before hardware-assisted locking was available. Hardware-assisted locking leverages a storage array atomic test and set capability to enable a fine-grained block-level locking mechanism. Simple things like moving a virtual machine, starting it, creating a new virtual machine from a template, taking snapshots or even stopping a virtual machine will result in VMFS having to allocate or return storage to or from the shared free-space pool. Although the VMFS use of the SCSI reservation locking the LUN does not often result in degradation of performance, the use of hardware-assisted locking provides a much more efficient means to avoid retries for getting a lock when many ESX servers share a single datastore.

Hardware-assisted locking enables the offloading of the lock mechanism to the arrays and does so with much less granularity than an entire LUN. So there is significant scalability that the VMware cluster can leverage without compromising the integrity of the VMFS shared storage-pool metadata.

Enabling vStorage API for Array Integration

By default these three primitives described above are not enabled upon install and must be enabled in the advanced settings for the ESX server, and they must have the correct array firmware loaded for this feature to work. Enabling or disabling these primitives is done through the advanced settings on the ESX servers.

Three settings under advanced settings:

DataMover.HardwareAcceleratedMove	- full copy
DataMover.HardwareAcceleratedInit	- block zeroing
VMFS3.HarwareAccelerated Locking	- hardware-assisted locking

More information can be found about this API in the *ESX Server Configuration Guide* (Chapter 9). It is listed in the index and table of contents as "storage hardware acceleration."

In the 4.1 release, the VAAI is supported for block-based storage only and requires the six storage partners to update firmware supporting this vStorage API for Array Integration.

Check the VMware storage HCL for a complete listing of storage vendors and storage models offering support for VAAI.

New Performance Metrics

New storage performance metrics are added in vSphere 4.1 to expand the troubleshooting and storage activity monitoring for both the command-line interface and the GUI. There are several new metrics for NFS devices to bring that more in line with the metrics for block-based storage, as well as some additional levels of visibility added to monitor storage throughput and latency at the virtual-machine and ESX-server levels.

Comprehensive performance statistics enable proactive monitoring to simplify troubleshooting host and virtual machine–storage levels. These new metrics offer varied usage scenarios:

- GUI for trending and user-friendly comparative analysis — real-time and historical trending (vCenter)
- Command line for scripting/drill-down for the ESX server — esxtop (for ESX) and resxtop (for ESXi)

Additional throughput and latency statistics are available for viewing all datastore activity from an ESX server, as well as for viewing the storage adaptor and path activity for a given ESX server.

At the virtual-machine level users can now also view the throughput and latency statistics for all the virtual disks or for all the datastores associated with the virtual machine.

There are also new performance charts added for datastores that are helpful in monitoring where SIOC is engaged and what queue depth is measured over time, as shown in the screen shot.

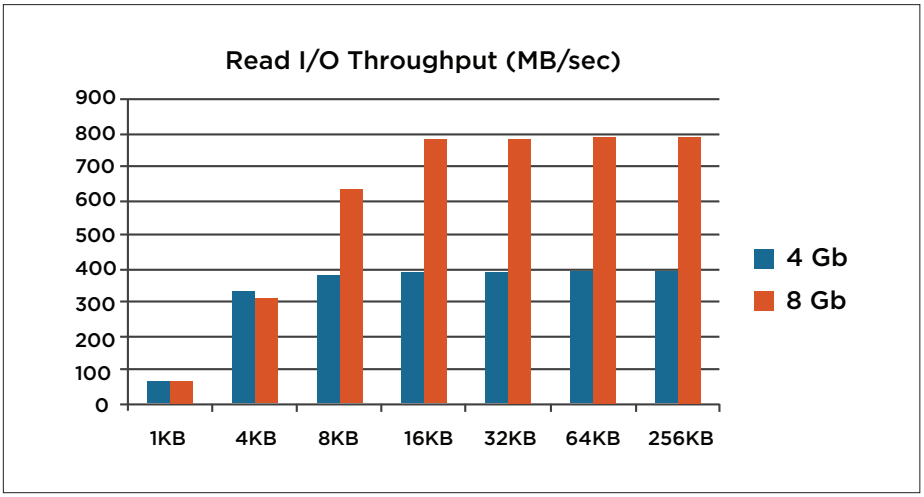
Support for iSCSI Hardware Offload

iSCSI has emerged as a high-performance networked storage technology that is widely used in many virtualization deployments. As server and enterprise-application customers strive to achieve density and computer-resource utilization objectives for their servers and enterprise applications, Broadcom's NetXtreme II iSCSI Hardware Based Acceleration (HBA) functionality, with support for VMware virtualization, provides the converged functionality needed in a virtualized-server environment by offering complete on-chip processing solutions that free up CPU resources, and increase bandwidth and performance.

The increase from 1Gigabit Ethernet (GbE) to 10GbE delivers increased storage-performance levels not previously achievable and provides sufficient bandwidth that permits multiple types of high-bandwidth protocol traffic to coexist on the same network. As a result, a server converges networking and storage onto the same network while lowering the total cost of ownership (TCO), or uses a dedicated network for data and for storage, thereby using the same equipment for multiple purposes. Broadcom iSCSI offload functionality enables on-chip processing of the iSCSI protocol (as well as TCP and IP protocols), which frees up host CPU resources at 10GbE line rates over a single Ethernet port. This functionality provides extended performance benefits that meet the demands of bandwidth-intensive applications requiring high-performance block storage I/O for VMware ESX, servicing all instances of the virtual machine.

Support for 8Gb FC Host-Based Adaptors (HBAs)

vSphere 4.1 adds full support for 8Gb FC from the ESX server to the storage array. Users can now deploy ESX servers with 8Gb HBAs with higher bandwidth for end-to-end FC SANs. With 8Gb support ESX effectively doubles the measured throughput with transfer size >8k, as shown in the graph.



The full list of partner HBAs that are supported in 4.1 can be found on the HCL.

Conclusion

vSphere 4.1 adds many new storage features to an already rich set of capabilities supported in vSphere 4.0. SIOC provides new means for shared storage pools to be optimized for I/O prioritization across many virtual machines running on a cluster of ESX servers. VAAI provides VMware storage partners a new means to offload workloads, making certain functions more efficient, saving both time and compute cycles. New performance screens in vCenter provide the vSphere administrator with new visibility and insight to track, troubleshoot and monitor usage, as well as adding support for SIOC. Support for iSCSI TCP/IP offload engine cards from Broadcom makes it more efficient to leverage iSCSI attached storage targets. And full support for 8Gb HBAs enables higher levels of throughput for FC-attached storage.

All of these capabilities help to increase and scale virtualization environments that are deployed on vSphere 4.1.

About the author:

Paul Manning is Storage Architect in the Technical Marketing group at VMware and is focused on virtual storage management. Previously, he worked at EMC and Oracle, where he had more than 10 years of experience designing and developing storage infrastructure and deployment best practices. He has also developed and delivered training courses on best practices for highly available storage infrastructure to a variety of customers and partners in the United States and abroad. He has authored numerous publications and has presented many talks on the topic of best practices for storage deployments and performance optimization.

vmware®

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