Integrating & Troubleshooting VMware with EMC's Midrange CLARiiON: Deep Dive

EMC Proven Professional Knowledge Sharing 2012

Jason L. Gates
Systems Engineer - Storage & Virtualization
Presidio Networked Solutions
jgates@presidio.com
http://www.linkedin.com/in/mrjasongates

Contents

Disclaimer: The views, processes or methodologies published in this article are those of the author. They do not necessarily reflect EMC Corporation's views, processes or methodologies.

Audience

This article is intended for system administrators, VMware engineers, and storage administrators. Readers are expected to be familiar with VMware ESX/ESXi hosts, basic CLARiiON® system operations, fabric switches, basic networking, VCenter Server, and EMC Navisphere/Unipshere Manager.

Overview

Many in the EMC Proven™ community have questions about best practices and the advanced skills that it takes to troubleshoot problems at the Systems Engineer level when it involves VMware and CLARiiON all in one solution/configuration. Problems do arise and resolving these problems takes an engineer who has a good understanding of the VMware hosts and expected behavior of the CLARiiON at the SCSI level.

What kind of problems arise? LUN trespassing, performance issues, connectivity, iSCSI and SCSI reservation issues. What happens when these type of problems occur? The storage administrator is pointing fingers at the VMware server and the VMware engineer is pointing fingers at the storage subsystem.

Common questions are, “What is the best path policy?” “What invokes a trespass?” “How does native multipath behave?” “What causes SCSI Reservations and how to resolve?” “What are best practices to configure iSCSI with the CLARiiON?” “How should my LUNs be carved out based on VM (Virtual machine) needs?” This article covers troubleshooting skills, configuration methods, and best practices anyone can utilize on the world’s number one mid-range storage sub-system and VMware’s leading virtualization software.

Troubleshooting Connectivity VMware ESX Version 4, 3.x with EMC CLARiiON

Most seasoned VMware administrators and storage administrators understand that ESX hosts access CLARiiON LUNs via the Fibre Channel protocol or iSCSI at the block level. Basics such as zoning, LUN masking, creating storage groups, registering hosts, and assigning host/servers to storage groups are steps that need to occur for storage to be available, which the majority of EMC professionals are comfortable with. What happens when all the above is done and there is still no storage available to our ESX hosts? Reboot? Rebooting is a common workaround for problems with storage but sometimes this is not an acceptable workaround, so we must dig deeper to really understand what is happening. I'm a strong supporter of using ESX command line to troubleshoot issues, so anytime there is a fibre channel (FC) connectivity problem, I use the command line to delve deeper.

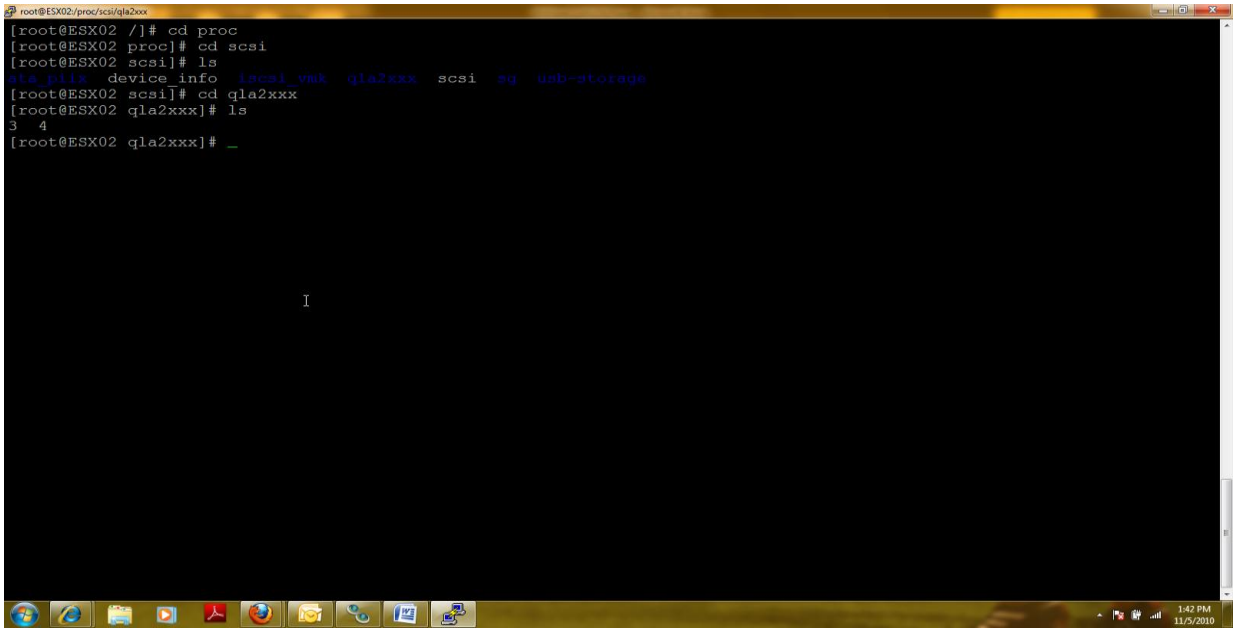


Figure 1 Using SSH as in this display, proc/scsi/qla2xxx is the directory that stores very useful information about the HBA. (In your environment, the HBA nomenclature could be different)

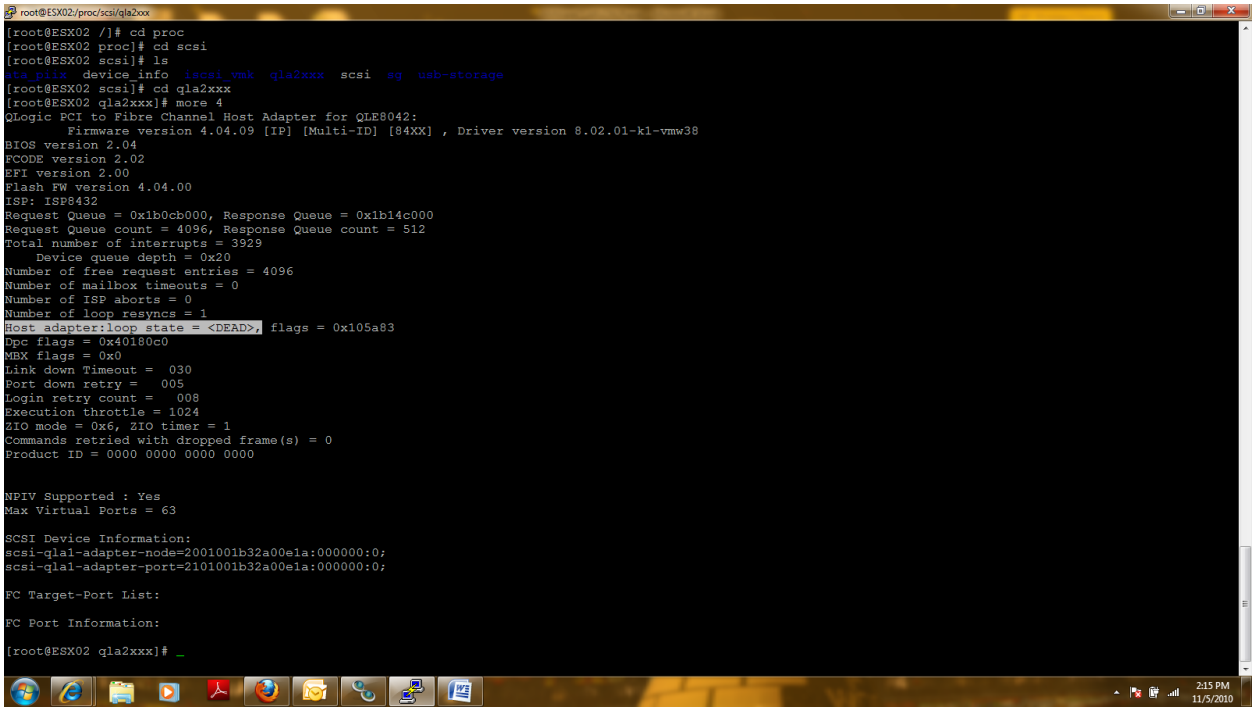


Figure 2 Using the command *more 4*, we see the loop state of this HBA is "Dead" This indicates a bad HBA, HBA cable, or perhaps a GBIC on the switch. Sometimes it is difficult to trace cables and wires so this is a quick way to confirm connectivity.

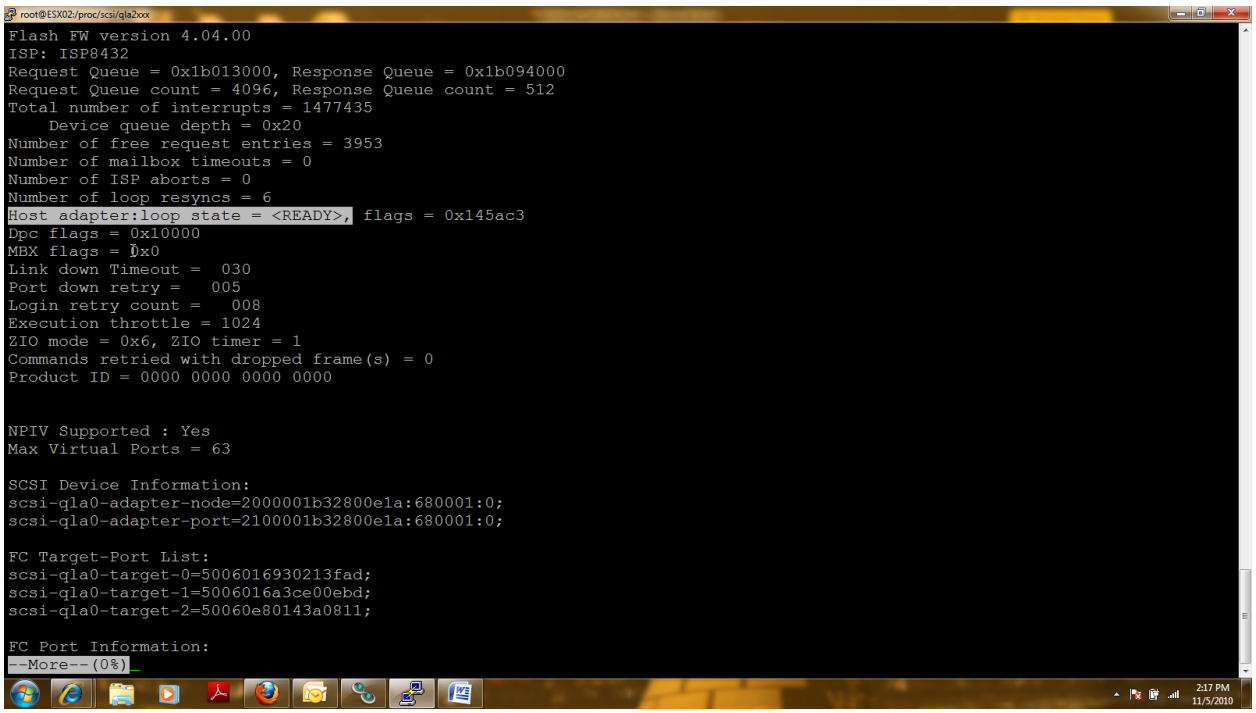


Figure 3 I replaced the HBA cable. Now, the Target-Port List has the WWPN of the CLARiiON listed, which indicates connectivity is OK. The state of the HBA is changed to "Ready".

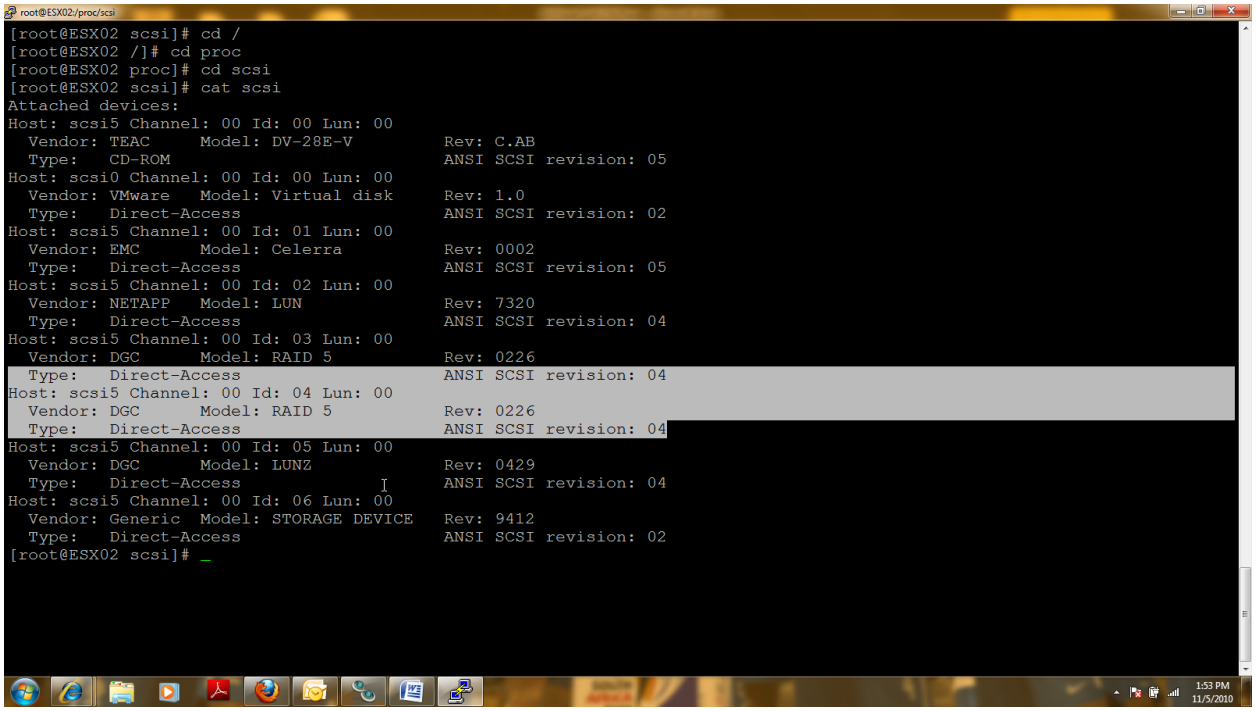


Figure 4 Connectivity is now intact. We can also confirm that SCSI devices are attached from proc/scsi and cat the file scsi. The highlighted entry is a CLARiiON LUN, because the vendor type is "DGC", (Data General Corporation.)

Configuring Multipathing and Failover for High Availability

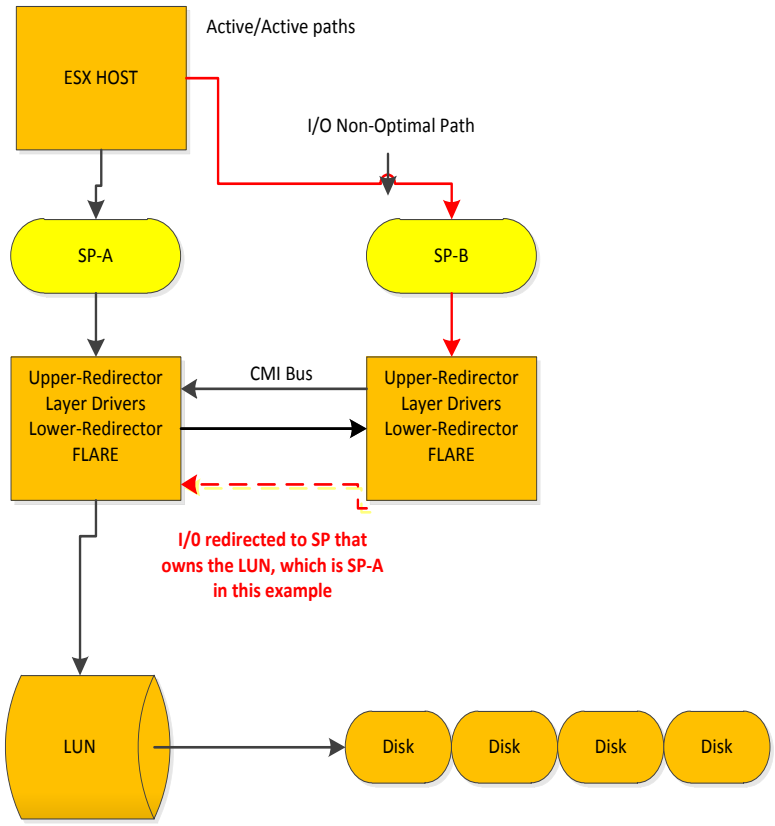
Multipathing and load balancing increase the level of availability for applications running on ESX servers. CLARiiON supports nondisruptive upgrade (NDU) operations for VMware's native failover software and EMC's popular failover software, PowerPath. The CLARiiON is an active/passive array and the policy for all paths should be set as MRU if using ESX 3 to avoid path thrashing. Path thrashing occurs when a server cannot access a LUN or access is very slow. In most cases two or more servers are attempting to access the LUN through different storage processors and as a result, the LUN is **never** truly available. Let's discuss this in detail.

The storage processors are independent computers that access shared storage and algorithms are used to determine how concurrent access is handled. When using the MRU policy, there is no concept of preferred path; in this case, the preferred path can be disregarded. The MRU policy uses the most recent path to the disk until this path becomes unavailable. If two ESX servers are connected, with path one from HBA1 to SPA, and path two from HBA0 to SPB, a single LUN configured as a VMFS volume can be accessed by multiple ESX servers. Using MRU, each HBA should be cross-zoned to each SP on the CLARiiON array. The ESX host will only use one active path for I/O and thus makes use of only one HBA. In this configuration, the

host will continue to use the active path until it finds a failure and then will failover to the alternate path (which could be via the same or different SP). Keep in mind, native failover is not dynamic I/O balancing software; a manual restore is needed to take the original path again or it will continue to use the alternate path even after the original path recovers.

With VSphere 4.0 and above, failover and mulipathing can be configured for Asymmetric Logical Unit Access (ALUA) on the CLARiiON array with FLARE 28 and higher. I highly recommend this setting. This configuration is quite different from the active/passive configuration. To understand mulitpathing behavior of VMware at the host level, we must understand expected behavior on the CLARiiON backend with active/active configuration. Asymmetric Active/Active introduces initiator Failover Mode (Failover mode 4,) where initiators are permitted to send I/O to a LUN regardless of which SP actually owns the LUN.

ALUA uses SCSI 3 commands that are part of the standard SCSI SPC-3 specification to determine I/O paths. Case in point, if I/O for a LUN is sent to an SP that does not own the LUN, that SP redirects the I/O to the SP that does own the LUN. The redirection is done through internal communication (upper-redirect) between the SP's; a LUN trespass is not needed and the redirection is transparent to the hosts. The diagram below illustrates ALUA behavior:



The following event code will be logged in the SP(s) event log:

"UpperRedirector 711b0000 IO is being forwarded to the other SP for processing"

Use of ALUA failover mode has additional benefits when combined with VSphere native multipathing software, providing automatic restore of LUNs to its preferred paths after a fault has been repaired. However, this only applies to "Fixed" policy configuration. This does not apply to "Round Robin" policy. So what is the major difference between Fixed policy and Round Robin? The Fixed policy supports auto-restore of failed paths but NO load balancing. The Round Robin policy performs load balancing but NO auto-restore of failed paths. The ALUA feature also eliminates path thrashing in Active/Active cluster configurations, and avoids unavailability when booting from a SAN LUN during a path failure.

ISCSI Troubleshooting

Internet Small Computer System Interface (iSCSI) is a cost-effective, easy-to-manage storage solution that many companies deploy into production. The iSCSI protocol allows clients (called *initiators*) to send SCSI commands to SCSI storage devices (*targets*). Configuring the software iSCSI initiator is straight-forward with both versions of ESX using VMware's iSCSI configuration guide. Most iSCSI issues are network related—poor connections, switch issues, hops, routing issues, and so forth. These kinds of problems can severely impact the read & write performance of storage attached to the ESX software through the iSCSI initiator.

If your network is congested or has very high periods of intensive I/O activities that negatively affect your ESX environment, a workaround could be to disable the delayed acknowledgement (ACK) feature on your VMWare server via a configuration option. ESX uses delayed ACK to increase efficiency in both the network and the hosts by sending less than one ACK acknowledgment segment per data segment received.

Congestion can manifest itself as a delay, timeout, or packet loss and during this "congestion" period the CLARiiON will only re-transmit **one** lost data segment at a time. The ACK feature coupled with the CLARiiON only re-transmitting one data segment would slow read performance to a halt in a congested network and frequent timeouts will be reported in the ESX kernel logs, virtual machine guest would timeout, and so on. To modify the delayed ack settings on a specific discovery target, do the following:

1) Select the **Static Discovery** tab
2) Select the **Server Address** tab
3) Click **Settings**
4) Click **Advanced**
5) Uncheck **DelayedAck**

Note: As a rule of thumb, this setting should be tested to confirm this is ideal for your applications and network before deploying into production.

SCSI Reservations

All VMware clusters use the SCSI protocol to manage disks on a shared bus. When a reserve command is issued by a host bus adapter (HBA), the command allows a HBA to obtain or maintain ownership of a SCSI device. A device that is reserved refuses all commands from all other host bus adapters except the one that initially reserved it, the initiator. In the VMware world, VMFS data store-level operations require a lock or reservation to ensure there is no data corruption**,** because two hosts cannot write to the same device at the same time; hence, the need for reservations.

So what happens when host A has a reservation on LUN 1 and host B attempts to access LUN 1? The answer is a "reservation conflict" is reported by Host B. A reservation conflict means that host B is trying to access the LUN and is not being allowed to, because host A has already reserved the LUN for its own access. Some administrators believe that the CLARiiON creates the SCSI reservations. This is incorrect; reservations are not created by the array itself. The CLARiiON simply responds to the commands sent by the ESX hosts. When the CLARiiON fails to release the reservation, the request may not have come through (hardware, firmware, pathing problems, and so forth). For example, virtual machines may experience I/O failures and blue screens due to too many SCSI reservation conflicts. To resolve reservations conflicts, the following should done or considered:

1) Confirm all HBA drivers are up-to date.
2) Determine which HBA has the lock; remove the HBA cable from fabric switch, wait 10 seconds, then reconnect. This forces a Port Login and Fabric Login which would clear the reserve.
3) Do not place Exchange, SQL, or Oracle VM's on the same LUN(s), because these types of applications change data frequently.

4) Reduce heavy amount of VM snapshots.
5) Confirm failover settings are correct on all hosts accessing shared storage.
6) As a last resort, the Storage Processor that owns the LUN may need to be rebooted or reboot the ESX host that placed the reservation on the LUN.

Reservation conflicts should be expected in an ESX clustered environment, but a high number of conflicts could cause unexpected outages. I strongly suggest keeping all systems—including the CLARiiON—on latest codes and patches.

LUN Layout Considerations

When designing LUN layouts, each situation is different; however there are some basic guidelines to follow. Select RAID 5 protection on Fibre Channel drives for boot volumes and servers such as DNS or large file servers because low I/O activity is expected from these type of virtual machines. RAID 5 writes require parity updates to each disk; however each disk can be read from independently.

SATA-II drives should be used for storing archived data and could be RAID 5 or RAID 1 protected.

For write intensive application data volumes and logs of databases, use RAID 1/0 on fast Fibre Channel drives such as 146GB 15k or flash drives. Also in a RAID 1/0 group, I would advise to span the drives across two separate buses if feasible. The purpose of creating the RAID Group this way is to place data and mirrors on two separate enclosures. In the event of an enclosure failure, the other enclosure would still be online and maintain access to the data or the mirrored data. See the diagram below:

Order of Disks Example for Raid 1/0 six disk RAID Group

First Data1 1_0_0

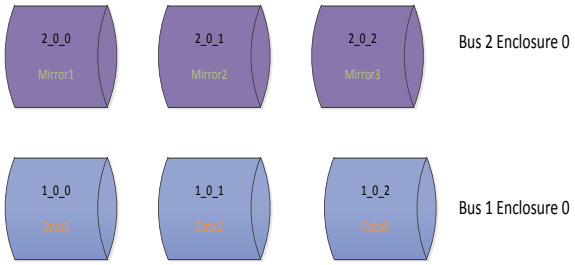
Second Mirror1 2_0_0

Third Data2 1_0_1

Fourth Mirror2 2_0_1

Fifth Data3 1_0_2

Sixth Mirror3 2_0_2



Note: VMware Vmfs-3 volumes are already aligned to 64KB during creation. However, if using Windows 2003/2000, the volume will need to be aligned at the virtual machine level using diskpart. This is not required on Windows 2008. The CLARiiON formats disks in blocks of 128 per disk, which is equivalent to a 64 KB of data that is written to a disk from write cache.

Conclusion

EMC's mid-range storage leader, CLARiiON, and VMware provide a robust, industry-leading Information Lifecycle solution. These technologies complement each other well when configured properly using best practices. I sincerely hope that this article will be a great asset to EMC Proven Professionals and the community in general when troubleshooting common problems in a VMware/EMC CLARiiON environment.

References

EMC CLARiiON Integration with VMware ESX Server - White Paper

EMC's Host Connectivity Guide for VMware - Best Practices Guide

Fibre Chanel SAN Configuration Guide -
www.**vmware**.com/pdf/vsphere4/r40/vsp_40_san_cfg.pdf

ISCSI SAN configuration Guide -
www.**vmware**.com/pdf/vsphere4/r40/vsp_40_**iscsi_san**_cfg.pdf

Performance for VSphere - www.**vmware**.com/pdf/Perf_Best_Practices_vSphere4.0.pdf

Next Generation Best Practices for Storage and VMware - http://virtualgeek.typepad.com