Internal Use Only

Installation and Configuration Guide Xyratex F54xxE – F6412E and VMware ESX Server

xyratex

Notices

The information in this document is subject to change without notice.

While every effort has been made to ensure that all information in this document is accurate, Xyratex accepts no liability for any errors that may arise.

TIP 106 | Issue 1.3 | March, 2009

Contents

Introduction

The purpose of this document is to describe the interoperability and configuration best practices to use when installing the Xyratex F54xxE series and F6412E Fibre Channel (FC) RAID Array storage products into VMware® ESX environments. This includes any known technical limitations with our partners' products of the date and versions reflected by the date of this paper. Additional information will be added to this document over time.

VMware Certification Program Overview

VMware has a certification program that certifies storage (as well as other system components). The results are updated periodically (*download PDF*).

Participation in the certification program is optional, but 3rd party products will not be listed in their SAN guide unless successfully qualified thru this program.

Overview

The VMware certification program currently tests in one or more of the following configurations, but access is based on partnership level and/or product type.

- Basic Connectivity — The ability of ESX Server hosts to interoperate with the storage array. This configuration does not allow for multipathing or any type of failover.
- Multipathing — The ability of ESX Server 3.x hosts to handle multiple datapaths to the same storage device(s).
  - Host Bus Adapter (HBA) Failover — ESX Server host with multiple HBAs connecting to one or more SAN switches. The server is robust to HBA and switch failure only.
  - Storage Port Failover — In this configuration, the ESX Server 3.x host is attached to multiple storage ports and is robust to storage port failures.
- Clustering Support — Clustering support applies to a limited set of configurations and products. For more details, refer to the VMware I/O Compatibility Guide (*download PDF*)
- Boot from SAN — The ability of ESX Server hosts to boot from a LUN stored on the SAN rather than a local disk.
- Direct Connect — The ability of the ESX Server host to directly connected to the array. HBA and Storage Processor Failover are supported, provided that there is no sharing of LUNs between multiple hosts. Clustering is not supported in this configuration.

Configuration Tested

The diagram below represents the typical test environment which is used as a baseline for the majority of the certification tests. It can be used for local or SAN boot, Single or Muti-Initiator, shared or dedicated datastores, or a number of other configurable scenarios that the automated test harness needs. Exact details can be provided by VMware.

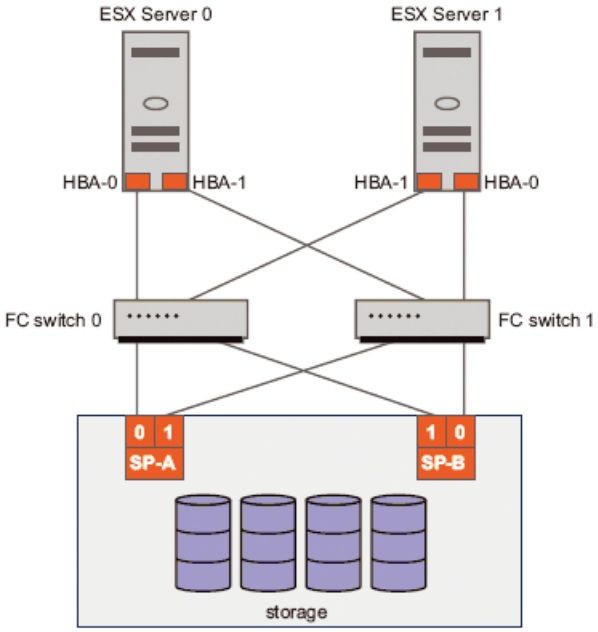


Figure 1: Typical VMware Configuration — Multiple-Server Dual-Fabric SAN

ESX0

- Boots ESX 3.5 from SAN (dedicated LUN)
- Contains eight Virtual Machines (VMs) with boot/data drives on shared LUN (datastore)
  - Two Windows Server 2003® — Buslogic® HBA
  - Two Windows Server 2003 — LSI® HBA
  - Two RedHat Linux — Buslogic HBA
  - Two RedHat Linux — LSI HBA

ESX1

- Boots ESX 3.5 from SAN (dedicated LUN)
- Contains eight VMs with boot/data drives on shared LUN (datastore)
  - Two Windows Server 2003 — Buslogic HBA
  - Two Windows Server 2003 — LSI HBA
  - Two RedHat Linux — Buslogic HBA
  - Two RedHat Linux — LSI HBA

Features Test

The below list is a summary of features and operations that are covered during the VMware verification. Exact details can be provided by VMware.

- VMware installation
- ESX Server boot from local drive
- ESX Server boot SAN drive
- Maximum LUN count
- Maximum LUN size
- Sparse LUNs
- VMFS Spanning
- Storage creation and usage
- Hot-add LUN to VM
- RAID controller online firmware (FW) upgrade
- RAID controller (active/inactive/stress) resets
- VM creations and usage with various operating systems (OS) types
- Data Integrity — ESX Server
- Data Integrity — VMs
- Reserve/Release with resets/reboots/HBAfail/SP_fails
- vHBA LSI under various OS types
- vHBA Buslogic under various OS types
- Datastore — shared
- Datastore — dedicated
- Single and Multi-Initiator use of storage
- Shared datastores between multiple VMs from multiple ESX Servers
- Multipathing with HBA failover
- Multipathing with FC cable failures
- Multipathing with Storage Processor (RAID Controller) failover
- Multipathing — recovery of failed components.

The tests below have been performed by Xyratex in addition to the VMWare certification:

- Create a DataStore on a Snapshot LUN
- Validate software RAID functionality by the VMs using virtual disks
- 2 ESX servers made part of a VMware HA cluster
- Expand the DataStore

- Verify the ESX server can communicate with the enclosure under test even if the LUN numbers are not in sequential order. This includes the absence of LUN 0
- Manually move VM to another ESX host using VMotion
- Confirm the operating modes of the HBA and enclosure such as FC speed and various different topologies
- Perform HA failover and DRS

Current Certifications

As of October 2008

	ESX 3.0	ESX 3.01	ESX 3.02	ESX 3.5	ESX 3.5i Embedded	EXS 3.5 Installable
F5402E		Notes: 1,3,4,6 Cert: Primary	Notes: 1,3,4,6 Cert: Inherited	Notes: 1,3,4,6 Cert: Inherited	Notes:1,3,4 Cert: Inherited	Notes: 1,3,4 Cert: Inherited
F5412E			Notes: 1,3,4,6 Cert: Primary	Notes: 1,3,4,6 Cert: Inherited	Notes:1,3,4 Cert: Inherited	Notes: 1,3,4 Cert: Inherited
E5412E				Notes: 1,3,4,6 Cert: Primary	Notes:1,3,4 Cert: Inherited	Notes:1,3,4 Cert: Inherited
F5404E				Notes: 1,3,4,6 Cert: Primary	Notes:1,3,4 Cert: Inherited	Notes:1,3,4 Cert: Inherited
F6412E				Notes: 1,3,4,6 Cert: Primary	Notes:1,3,4 Cert: Inherited	Notes:1,3,4 Cert: Inherited

Notes: (1) Basic connectivity
(2) Direct connect support
(3) Multipathing with HBA failover
(4) Multipathing with storage port failover
(5) Windows clustering support
(6) Boot from SAN

VMware ESX Server 3 Configuration Guide — *download PDF*

VMware ESX Server 3.5 Configuration Guide — *download PDF*

Best Practices for Xyratex RAID Usage with VMware

For Xyratex RAID storage use under VMware, please review and understand the below topics prior to deployment.

Planning

- Are all the software / hardware / firmware revisions at the latest support levels?
- How and where will the ESX server(s) boot from? Local/SAN?
- Where will the VMs boot from and virtual hard drives (vHD) be located?
- Will the VM swap location be on VM boot drives?
- Will the VM data drive(s) be on the VM boot drives?
- What performance do I need from the RAID storage?
- Do I need SAS or SATA drives in the RAID system?
- What will my cabling strategy be based on?
- Review SAN Multipathing cabling strategies

Configure Storage

- Reserve disk(s) as needed for hot spares.
- Using standard Xyratex configuration guidelines, configure the disk arrays and Logical Disks (LD) as desired.
- Reserve LDs (as needed) for snapshot usage.

Connectivity

- One must make sure that it can be presented to the HBA ports of the ESX servers for use.
- Review the RAID SAN/Port mapping, FC switch configuration/zoning, physical cabling, as well as possible HBA binding settings to insure correctness.
- Depending on usage, it is generally recommended to configure the RAID controller to SAN map the LUNs to only be visible to the ESX server FC ports that will use them.

Configure VMware

- Discovery — A LUN must be discovered by the ESX server before use. This may require a VMware rescan if the LUN was not visible to the ESX server at boot time.
- Multipathing — Xyratex storage is configured using VMware's fixed path strategy.
- Performance — Review the performance topics section for configuration recommendations.

Advanced Topics

- SAN Booting — Enable the HBA BIOS and set the WWN/LUN to boot from.
- High Availability (HA) Multipathing — This configuration requires that each ESX server has at least two supported FC HBAs (e.g. QLogic®, Emulex®), each connected to two independent (no Inter Switch Link) FC switches, and connected to a dual controller (storage processor) Xyratex RAID system. This will allow any of the four paths to the storage to be used.
- Monitoring RAID System — If in-band monitoring and management of the RAID system is desired, an additional FC host is needed on the SAN. If out-of-band monitoring and management will be used, a VM (or any host with network connectivity) can access the RAID controllers via Ethernet as needed.
- Replacing controllers and/or chassis — Xyratex storage is generally presented using a 'Configuration WWN' which ensures that a controller change will not impact the mappings on switch and server. In situations that do change the WWN/WWPN of the presented storage it may be required 'resignaturing' of the datastore before re-usage. Review VMware support procedures for more details.

Performance Topics

Summary

As with any system, the total performance is only as fast as the slowest component. This document is not intended to cover detailed performance tuning of either ESX or Xyratex RAID systems since each of these topics is discussed in other documents. However, it does cover some of the common causes of poorly performing systems with a technical discussion on the impacts and workarounds if they exist.

Common Issues

Below is a list of issues or situations that contribute to direct/indirect reduction of performance of a system. Also, included are scenarios that are often not taken into consideration when estimating what performance to expect from a given configuration. These may or may not apply to your specific situation.

- Improperly configured RAID system — RAID system performance can be impacted by a number of issues such as the number of drives per array, number of LDs (LUNs) per array, RAID level, Chunk/stripe size and cache optimization settings to name a few. Drive type also will impact performance based on differences in the native drive performance levels such as SAS (faster) vs. SATA (slower). Review the RAID system as a whole to establish an expected baseline performance level based on drive type, settings, and the type of I/O being generated. Benchmarking can be done on the array prior to VMware installation/usage to confirm expected performance.
- SATA Drives — In a virtualization environment where multiple VMs and/or ESX servers share a single LUN/Array, the combined I/O stream to the actual physical disks tends to be very much randomized (even though each VM individually may be performing sequential operations) due to the interleaved nature of the virtualized I/O. This randomness negatively impacts SATA drive performance more than other drive types (e.g. SAS or FC drives) due to their command queuing abilities, as well as the generally slower spindle speeds of SATA which result in a lower overall performance.
- Sharing of RAID Arrays and LDs — VMware allows datastores to be shared between multiple ESX servers, with each ESX server potentially having a large number of VMs each of which is performing I/O. Sequential I/O from the VMs, when shared between a large number of I/O threads from all the different sources, tends to get randomized which generally results in lower performance due to less effective caching and more seeking. Xyratex recommends the use of several smaller arrays instead of a single large array.
- Multiple ESX servers' share same FC path — In configurations where multiple ESX servers are SAN connected to the RAID storage and have more than one path to the storage, the default configuration is to use the same ordered sequence for path usage which results in all the servers converging on a single port of a single controller. This is independent of whether the ESX servers are sharing the same LUN or not.

- Accessing same LUN (datastore) via different RAID controllers — In a configuration where multiple ESX servers share a datastore (LUN), it is possible for one of the servers to access that LUN via a different controller to the other ESX servers. This can be due to an actual connectivity issue (bad HBA, SFP, cable, FC switch configuration/port, etc), or due to misconfiguration of the paths or multipathing setup (e.g. MRU instead of FIXED pathing). When this condition occurs, the RAID controllers still service the I/Os but are slower due to internal interaction mechanisms. Xyratex RAID controllers operate in what VMware calls Pseudo-Active/Active operations, which means that although a LUN is accessible from either or both controller(s), it is not optimal to do so.
- OS overhead on VMs — A VM that boots from the same datastore as it's vHD(s) will share its storage bandwidth between the VM's OS I/O overhead (e.g. memory paging/swap space, buffering, etc.) and the application(s) I/O running on that same VM. For example, a Windows Server 2003 VM running I/OMeter that reports 70MB/s may actually be generating 80MB/s of throughput from the RAID controller. The actual I/O rates (from a RAID controller perspective) can be viewed via the FC switch (e.g. on Brocade®, run portperfshow on the switch port that the RAID controller(s) are connected to),FC analyzer (if available), or even from the RAID controller statistics. Consider segregating the OS, swap, and/or data LUNs that the ESX server uses to provide optimal performance.
- vHBA Queue Depth — The VMware GUI allows a user to optimize the queue depths for both a vHBA as well as outstanding I/Os allowed for a VM using that vHBA. VMware documentation can provide more details, but may vary depending on many different variables.
- ESX Servers that boot from SAN — An ESX server that is configured to boot from SAN may be the same storage array that also hosts the datastores that the VMs boot from and perform I/O to. This means that VM I/O to the RAID system can be reduced by ESX storage intensive operations such as VM management operations (e.g. cloning) or ESX swap space operations.
- Overloaded VMware server — An ESX server by its nature hosts multiple VMs and as such divides its processor and resources between them, but there can be a point where there is a mismatch between the server's capabilities and the number of VMs that are running on it.

Xyratex Advanced Tuning Options

Each major release of Xyratex RAID controller firmware may have performance improvements or optimizations that can be tuned to 'potentially' increase performance for a given workload profile.

- Performance Tuning — From the StorView™ menu 'Advanced Settings/Performance Options', several configuration options are available to fine-tune VMware performance, but be careful to understand the costs and benefits before changing since improper setting can reduce performance.
  - Synchronize Cache Writes to Disk — Enabled (default).
  - Target Command Thread Balance — This option should be left disabled (default) for best VMware performance.

- Sequential Write Optimization — Generally a setting of 'Low' (default) works best for VMware since ESX I/O tends to be less sequential.
- Overload Management:
  - » Disabled — Disabling is not recommended and will prevent the RAID system from notifying the host system when it is overloaded resulting in I/O retries and errors.
  - » SCSI busy status — This setting works well with the exception of VMware excessive logging of them (may be corrected in a future release of VMware).
  - » Enabled (Task Set Full — TSF) — This is the controller default and works well for all VMware operations. Note: On VMware revs 3.5U3 and earlier, temporarily switch to Busy or disabled during some VMware admin operations (see below FAQ section for details).
  - » Busy/TSF Timeout — The default of four seconds is acceptable and changing does not improve performance levels.

- Array Creation Tuning — VMware does not alter the I/O from the VMs, so there is no single best configuration that can be given that will be optimal for all configurations. Be careful to understand the costs and benefits before changing since improper setting can reduce performance.

Baseline VM Performance

Example Performance Scenario for the F6412E

The below configuration outlines the topology used.

Hardware used

- ESX server — HP DL380G5 4GB RAM
  - HBA0 — Qlogic QLE2462
- FC Switch 0 — Brocade Silkworm
- Storage — Xyratex F5412E

Software used

- I/OMeter 2006.07.27
- Single Manager, Single Worker

Storage Configuration

- Dual controller, F6412E (FW3.5b0028)
- Single enclosure (no expansion)
- 12 HDD drives (Seagate® SAS HDDs)

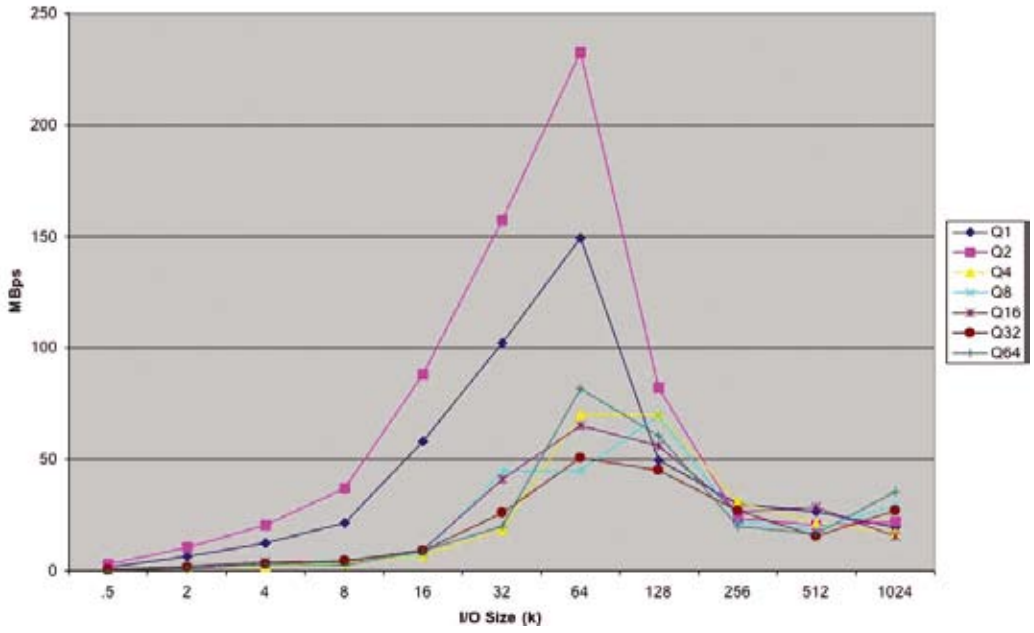
- Array 0, 1LD
  - Made up of drives one to six
  - RAID5 (5+1)
  - Contains LD 0 (256K chunk size)
  - LUNs presented out same as LD numbers, all ports
  - SAN mapping not used

VMware Configuration

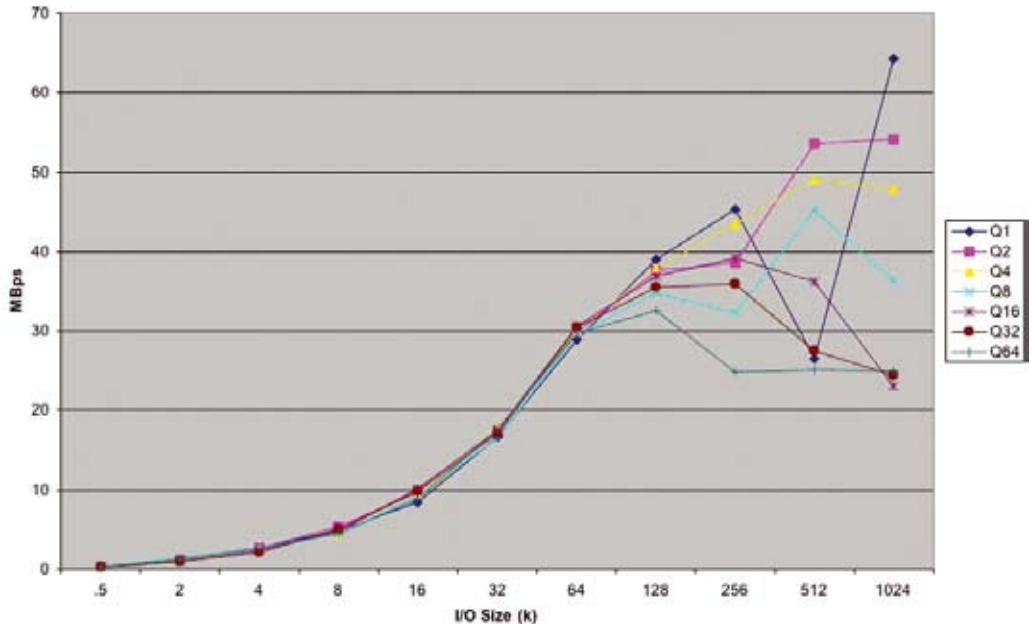
- ESX0
  - Boot ESX from internal SAS drive
  - FC storage presented as additional drive letter
  - VMs on shared datastore (LUN24) for boot and data drives

The following graphs illustrate the benchmark results of a single VM performing I/O to a single datastore (vHD). This shows the baseline performance level of a simple configuration that uses general default settings of the VM (Windows Server 2003), HBA, ESX server, file systems, and RAID storage. Optimizations can be made based on the types of I/O to increase these numbers. Note that the random performance numbers are shown here since multiple VM and/or ESX server consolidated loads tend to be random. Fewer VMs per datastore will allow sequential I/O to be less randomized and yield higher performance.

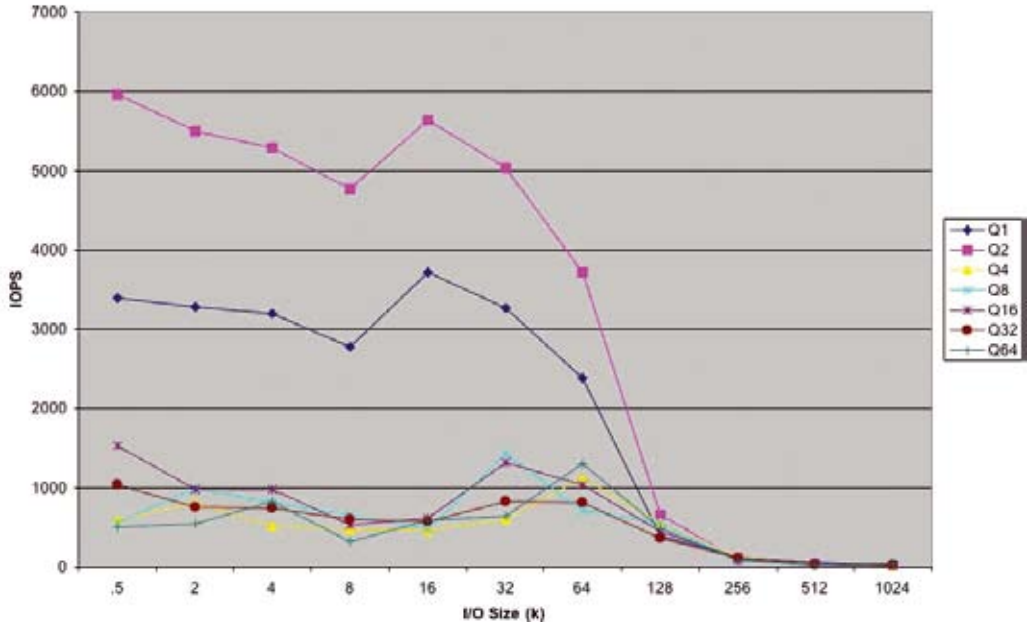
The delivered performance of the Xyratex RAID unit scales up to the advertised performance levels (for the types of I/Os generated) with the additional load of more VMs, ESX Servers and additional hosts.



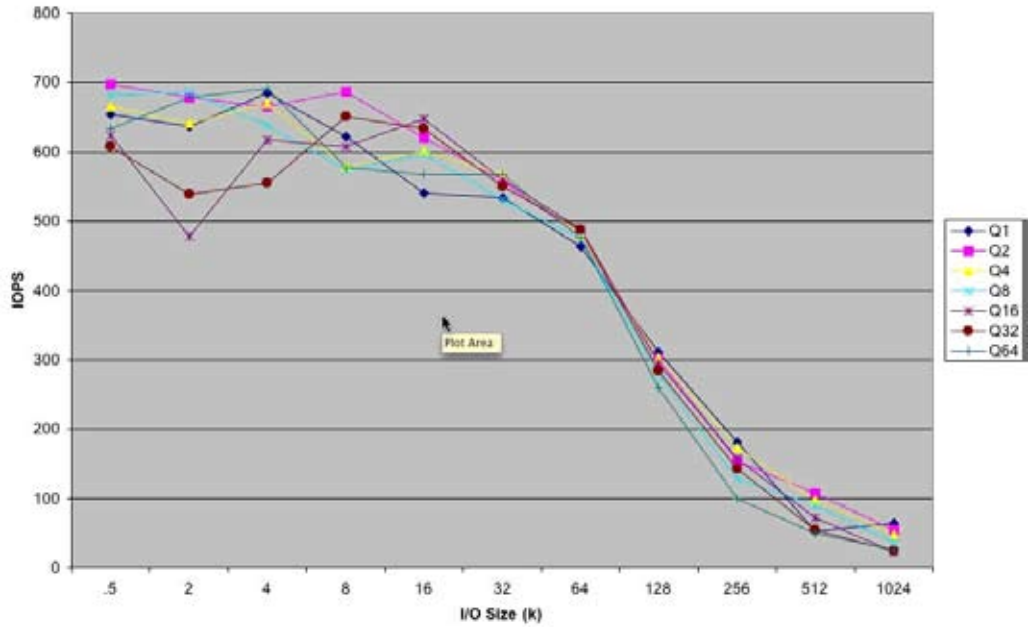
Random Read — MB/s



Random Write — MB/s



Random Read – IOPS



Random Write – IOPS

FAQ Topics

Does StorView Work in a VM?

Embedded StoreView (eSV) can be accessed from any VM running a supported OS/browser combination with proper Ethernet connectivity. Host-based StorView (inband) will not work in a VM since the storage is virtualized and presented via a vHBA.

SCSI Reservation Issues (Reservation Conflicts)

Technical summary — Reserve & release issues may be seen intermittently and not limited to our product, but impact any storage products that are used by VMware, because this is the mechanism that the VMware server uses to coordinate the sharing of a LUNs between physical servers and is the use of the SCSI reserve & release commands. A reserve is normally held for <100ms, but occasionally an ESX server will hold it for longer causing the other server(s) to receive excessive reservation conflicts returned for the I/O that they are trying to do to the same LUN (shared VMFS volume) at the same time. Note that 'some' ResConflicts are normal and expected and ESX only reports it when it receives more than sixty per I/O attempt.

Update — VMware released a December 2007 patch set for the various product versions that addresses some SCSI reservation issues.

Using LUNs Larger Than 2.2TB

The VMware website has reference material on proper use of and installation procedures for LUNs larger than 2.2TB.

Support for ESX 2.x

VMware versions prior to 3.0.1 have not been tested or certified with Xyratex RAID storage and are not recommended, but may work.

Support for ESX 3.0.x

VMware ESX 3.0.x revisions are certified and supported and generally work well with the below limitations:

- ESX 3.0.x had a known issues with reported SCSI reservation errors on heavily loaded systems. These issues were resolved with the ESX 3.5 release.

Support for ESX 3.5.x

VMware ESX 3.5 revisions are certified and supported and generally work well with the below limitations:

- ESX 3.5 has a known issue where some VMware administrative operations (e.g. datastore creation/initialization, cloning, suspending a VM, etc.) may fail if the ESX server receives a TSF reply to an I/O request from the storage. Note that this does NOT occur during normal VM I/O operations. This is a bug in ESX 3.5, updates one thru three, which is expected to be resolved in an upcoming release of 3.5U4 and ESX4 (both expected in January 2009). Until resolved in a future ESX release, the recommended Xyratex workaround is to temporarily change the controller's overload management setting to either disabled or return 'busy' during the datastore creation/initialize operation or retry the operation. Also, see the below note on ESX's 'busy' logging.
- VMware KB 1007041 discusses a VMware patch ESX350-200810201-UG that changes the QLogic driver to reduce logging of Debug SCSI under-run messages in the VMkernel log that have been reported when using storage that returns TSF.
- ESX 3.5 also has reports of VM Clone operations failing when TSF is enabled and seems related to the known datastore initialization issues and is resolved in the same way with the temporary workaround listed above. This issue is currently under investigation and may be resolved as part of the datastore initialization fix.
- ESX 3.5 has a known issue where replies of 'SCSI Busy' will be displayed in the ESX console event log. A 'Busy' is a normal SCSI communications flow control mechanism, and although ESX excessively logs them, they do not impact the normal operation of the ESX server or the integrity and/or performance of the I/O or data residing on the storage array.

Support for ESX 4.0

VMware ESX 4.0 has resolved the issue that occurred in VMware ESX 3.5

Additional Reference Materials

VMware ESX Server 3 Configuration Guide

VMware ESX Server 3.5 Configuration Guide

Download PDF's

Support

For more information or support issues please contact: info@xyratex.com

About Xyratex

Xyratex is the ultimate partner to the storage industry. We are a leading provider of enterprise-class data storage subsystems and storage infrastructure manufacturing equipment & automation solutions. Working with over 50 A-list companies, Xyratex ships over 14% of the world's external storage capacity, and 75% of all 3.5" drives are processed using Xyratex test systems. With unmatched expertise and a history of innovation and technological excellence, Xyratex delivers products which are high-performance, energy-efficient and extremely reliable.

For more information, please visit www.xyratex.com

Xyratex Headquarters
Langstone Road
Havant
Hampshire PO9 1SA
United Kingdom

Principal US Office
2031 Concourse Drive
San Jose, CA 95131
USA

UK HQ
T +44 (0)23 9249 6000
F +44 (0)23 9245 3654

USA Sales & Support
T +1 877 997 2839
T +1 877 XYRATEX

www.xyratex.com

ISO 14001: 2004 Cert. No. EMS91560