iSCSI Best Practices (Microsoft, EMC, VMware) - Summarized

This document summarizes the best practices of iSCSI usage as described in the following guides from MS, EMC, and VMware.


iSCSI Users Guide .doc


iSCSI_design_deploy .pdf


vi3_35_25_iscsi_san _cfg.pdf


Iscsi_300-002-262_A 02_elccnt_0.pdf


Iscsi_300-003-805_A 06_elccnt_0.pdf

MS & iSCSI –

Exchange:

- Microsoft Exchange Server can store its program files, mailboxes, public folders, logs and other data on iSCSI disk volumes in both cluster and non cluster configurations.
- Keep the Exchange disks in a separate pool on the array.

SQL Server – Microsoft SQL Server can store its program files, logs and other data on iSCSI disk volumes in both cluster and non cluster configurations

NOTE: *iSCSI initiator does not support Dynamic disk volumes and NIC teaming.*

iSNS – Microsoft iSNS Server is a Microsoft Windows service that processes iSNS registrations, de-registrations, and queries via TCP/IP from iSNS clients, and also maintains a database of these registrations. *iSNS servers can be clustered.*

iSCSI Boot:

- Windows can be booted off of an iSCSI disk however; the iSCSI boot initiator will disable all kernel mode code paging. Additionally the pagefile must **not** be located on an iSCSI disk.
- Windows Server 2003 can be booted from a SAN using either an FC HBA or an iSCSI HBA

iSCSI Best Practices:

- Deploy on fast networks – at least a GigE or better network
- Ensure physical security
- Use strong passwords for all accounts
- Use CHAP authentication because that ensures each host has its own password. Mutual CHAP authentication is even better. Use One Way or Mutual CHAP
- Use iSNS for discovery
- Segregate iSCSI SANs from LAN traffic
- Use IPsec
- Use Access Control or LUN masking

Networking Best Practices for iSCSI:

- Use non blocking switches and set the negotiated speed on the switches.
- Disable unicast storm control on iSCSI ports. Most switches have unicast storm control disabled by default.
- Enable Flow Control on network switches and adapters; flow control ensures a receiver can make the sender pace its speed and is important in avoiding data loss.
- Ensure spanning tree algorithm for detecting loops is turned off on the ports used for iSCSI traffic.
- Segregate SAN and LAN traffic. iSCSI SAN interfaces should be separated from other corporate network traffic (LAN). Servers should use dedicated NICs for SAN traffic. Deploying iSCSI disks on a separate network helps to minimize network congestion and latency. Additionally, iSCSI volumes are more secure when… Segregate SAN & LAN traffic can be separated using port based VLANs or physically separate networks.
- Configure additional Paths for High Availability; use either Microsoft MPIO or MCS (multiple connections per session) with additional NICs in the server to create additional connections to the iSCSI storage array through redundant Ethernet switch fabrics.
- Unbind File and Print Sharing from the iSCSI NIC – on the NICs which connect only to the iSCSI SAN, unbind File and Print Sharing.
- Use Gigabit Ethernet connections for high speed access to storage. Congested or lower speed networks can cause latency issues that disrupt access to iSCSI storage and applications running on iSCSI devices. In many cases, a properly designed IP-SAN can deliver better performance than internal disk drives.
- iSCSI is suitable for WAN and lower speed implementations including replication where latency and bandwidth are not a concern.
- Use Server class NICs. It is recommended to use NICs which are designed for enterprise networking and storage applications.
- Use CAT6 rated cables for Gigabit Network Infrastructures. For 10Gigabit implementations, Cat-6a or Cat-7 cabling is usually required for use with distances over 55 meters.
- Use Jumbo Frames if supported in your network infrastructure. Jumbo Frames can be used to allow more data to be transferred with each Ethernet transaction and reduce the number of frames. This larger frame size reduces the overhead on both your servers and iSCSI targets. For end to end support, each device in the network needs to support Jumbo frames including the NIC and Ethernet switches.

Redundancy & Load Balancing:

There are two technologies supported with the MS iSCSI software initiator to enable redundancy and load balancing:

- Multiple connections per session (MCS) – Multiple connections per session (MCS) support is defined in the iSCSI RFC to allow multiple TCP/IP connections from the initiator to the target for
the same iSCSI session. This is iSCSI Protocol specific. In this way I/O can be sent over either TCP/IP connection to the target. If one connection fails another

connection can continue processing I/O without interruption to the application. Note that not all iSCSI targets support MCS.
- Microsoft MPIO support

There are a number of things to consider when choosing to use MCS or Microsoft MPIO for multipathing.

- If your configuration uses hardware iSCSI HBA then Microsoft MPIO should be used.
- *If your target does not support MCS then Microsoft MPIO should be used*. Most iSCSI target arrays support Microsoft MPIO.
- *If your target does support MCS and you are using the Microsoft software initiator driver then MCS is the best option*. There may be some exceptions where you desire a consistent management interface among multipathing solutions and already have other Microsoft MPIO solutions installed that may make Microsoft MPIO an alternate choice in this configuration.
- *If you need to specify different load balance policies for different LUNs then Microsoft MPIO should be used*.
- If you are using Windows XP or Windows Vista, MCS is the only option since Microsoft MPIO is only available with Windows Server SKUS.

***NOTE**: There does not exist a mechanism within the iSCSI protocol to determine whether a target is active/active or active/passive.*

Load Balance Policies:

- **Fail Over Only**: No load balancing is performed. There is a single active path and the rest of the paths are standby paths. The active path is used for sending all I/O. If the active path fails then one of the standby paths is used. When the formally active path is reconnected it will become active and the standby path that was activated would return to standby.
- **Round Robin:** All paths are active paths and they will be used for sending I/O in a round robin fashion.
- **Round Robin with a subset of paths:** A set of paths are configured as active and a set of paths are configured as standby. I/O is sent in a round robin fashion over the active paths. If all of the active paths fail then one of the standby paths is used. If any of the formerly active paths become available again then the formerly active paths are used and the standby path that was activated becomes a standby path again.
- **Weighted Path:** Each path is assigned a weight and I/O will be sent on the path with the lowest weight. If the path with the lowest weight fails then the path with the next lowest weight will be used.
- **Least Queue Depth:** This is only supported by MCS. The path that has the fewest number of requests queued is the one where the I/O is sent.

***NOTE**: Windows does not support disks that have been formatted to anything other than a 512byte block size. Block size refers to the low level formatting of the disk and not the*

cluster or allocation size used by NTFS. Be aware that using a disk with a block size larger than 512 bytes will cause applications not to function correctly. You should check with your iSCSI target manufacture to ensure that their default block size is set to 512 bytes or problems will likely occur.

EMC & iSCSI –

A Look At EMC iSCSI Storage Systems (CLARiiON):

- EMC supports Microsoft Windows® 2000 and Microsoft Windows Server™ 2003 servers that run the native iSCSI Microsoft-certified driver for NICs. Supported devices include both onboard NICs in Microsoft-certified servers and PCI-based NICs that are Microsoft-certified.
- EMC supports Microsoft Windows 2000 and Microsoft Windows Server 2003 servers that use QLogic QLA4010 (optical) or QLA4010C (copper) HBAs and drivers.
- You cannot mix NICs and HBAs in the same server, even if they are connected to different storage systems.
- You must ***not*** connect a single server to both a CLARiiON® Fiber Channel storage system and an iSCSI storage system.
- Servers with HBAs and servers with NICs can connect to the same storage system.
- A single server can connect to CLARiiON CX-Series iSCSI storage systems and Symmetrix® iSCSI storage systems when a common network configuration, common failover software, and common driver support for both platforms exists.
- A single server can connect to CLARiiON AX-Series iSCSI storage systems, and through IP-to-FC switches, to CLARiiON AX-Series Fibre Channel storage systems when a common network configuration, common failover software, and common driver support for both platforms exists.
- A single server can connect to CLARiiON CX-Series iSCSI storage systems, and through IP-to-FC switches, to CLARiiON CX-Series Fibre Channel storage systems when a common network configuration, common failover software, and common driver support for both platforms exists.
- Using the CLARiiON Open Systems Configuration Guide (OSGC) definition of fan-in (server to storage system), you can connect a server to a maximum of four storage systems.
- Using the EMC Support Matrix (ESM) definition of fan-in (storage-system ports visible to a single initiator), you can connect an initiator to a maximum of 16 storage-system ports, but no more than four storage systems. The connection to the storage system must be 1-gigabit copper (RJ45).
- EMC does ***not*** support 10/100 NIC connections that are connected directly to the storage system, except for those connected to the management ports.
- Direct connections must be either with or 10/100/1000 NICs (operating at 1 gigabit) or QLA4010C HBAs. Ethernet crossover cables must be used with NICs to direct the server to the storage system.
- Using the OSCG and ESM definitions of fan-out (initiators per SP port), you can connect a maximum of 128 initiators to a CX-Series iSCSI SP port.

- If your service will *not* use iSNS, you must configure target storage-system addresses manually on the server initiators.
- You must configure server names and passwords manually on the iSCSI storage system. If you want authentication, you must use CHAP (Challenge Handshake Authentication Protocol).
- A CX-Series iSCSI storage system has two front-end (data) iSCSI ports per storage processor.
- EMC supports up to four HBAs or four NICs in one server that connects one CX-Series iSCSI storage system.
- Currently you cannot boot a Windows system using an iSCSI disk volume that the Microsoft iSCSI Software Initiator provides. The only currently supported method for booting a Windows system using an iSCSI disk volume is with a supported HBA.
- Microsoft iSCSI Software Initiator does not support dynamic disks.
- Microsoft iSCSI Initiator version 1.05a supports iSCSI Windows Server 2003 Cluster environments with a maximum of two nodes.
- The Microsoft iSCSI Initiator default configuration ignores multiple NICs on the same subnet. When multiple NICS are on the same subnet, use the **Advanced** button in the **Log On to Target** dialog box of the Microsoft iSCSI Software Initiator UI to associate a specific NIC with a specific SP port.
- ***Do not use*** Microsoft iSCSI Software Initiator to control the QLogic HBAs. QLogic's SANsurfer utility is the only supported interface to HBAs.
- A CX-Series iSCSI storage system does not support Microsoft iSCSI Software Initiator version 1.05a configured for MPIO, CRC/Checksum Data digest, or Header digest.
- **Supported Configurations:**
    - Servers' with Single NIC/HBA & 1 Subnet
    - Servers' with Multiple NICs/HBAs & 1 Subnet
    - Servers' with Multiple NICs/HBAs & Multiple Subnets
    - Servers with Multiple NICs/HBAs & Direct Connections
    - Multiple NICs/HBAs to Multiple Subnets, Routed or Independent (Including Direct Connections) is supported.

NOTE: *A high hop count can also contribute to performance degradation. Performance anomalies can also result for reasons associated with the various inherent TCP/IP flow control algorithms such as delayed ACK, slow start, and Nagle.*

NOTE: *You can use an iSCSI analyzer to perform protocol analysis of traffic flowing into and out of any suspect port on the storage system.*

NOTE: *You must install the Microsoft iSCSI Software Initiator because the Navisphere Server Utility uses it to configure iSCSI connections. You must install the* ***Initiator Service*** *option of the Microsoft iSCSI Software Initiator because the QLogic driver requires it.*

NOTE: *PowerPath iSCSI is no longer available for CX3 series and CX series storage systems.* ***PowerPath 4.5.1 or earlier*** *Do not select* ***Microsoft MPIO Multipathing***

***Support for iSCSI**. Do not select **Microsoft MPIO Multipathing Support for iSCSI** or **Software Initiator**.*

NOTE: *You can improve the performance of any NICs that will be used primarily for iSCSI traffic rather than general network traffic by changing the network settings so that NICs immediately acknowledge incoming TCP segments. If you are running a version of the Navisphere Server Utility that is earlier than 6.24.1.4.0, you need to manually modify the TCP/IP registry settings, as described below, to improve performance. If you are running Navisphere Server Utility version 6.24.1.4.0 or later, the system will prompt you to change these settings when you configure the network parameters for your NICs (set up iSCSI connections).*

NOTE: *When you remove an iSCSI target, the specified target and all other targets on the storage system will be removed. If you want to remove a specific target but not all targets on the storage system, you must use the Microsoft Software Initiator.*

VMware & iSCSI –

- VI3 uses single connection for a session
- At present, the VMware software initiator does **not** support jumbo frames. And until 10 gigabit Ethernet is supported by the VMware software initiator, the performance benefit of using jumbo frames would be minimal.
- The software-initiator iSCSI implementation leverages the VMkernel to perform the SCSI to IP translation and does require extra CPU cycles to perform this work. As a result, software iSCSI can reduce overall server performance when CPUs are under heavy load.
- Don't use VMware Consolidated Backup over iSCSI
- Best practice is to have a dedicated LAN for iSCSI traffic and **not** share the network with other network traffic. It is also best practice **not** to oversubscribe the dedicated LAN.
- VMkernel has a single routing table for all its VMkernel Ethernet interfaces
- Make sure both the VMotion and IP Storage network and the service console port connection have appropriate IP addresses and are routed properly to the array.
- The VMware VMkernel IP networking stack has been extended to handle the following functions:
    - iSCSI as a virtual machine datastore (new in ESX Server 3)
    - NFS as a virtual machine datastore (new in ESX Server 3)
    - NFS for the direct mounting of ISO files, which are presented as CD-ROMs to virtual machines
    - Migration with Vmotion
- Make sure both the VMotion and IP Storage network and the service console port connection have appropriate IP addresses and are routed properly to the array.
- The IP address that you assign to the service console during installation must be different from the IP address that you assign to VMkernel's IP stack from the **Configuration** > **Networking** tab of the Virtual Infrastructure Client. The NFS and iSCSI functions must be configured together. They always share the same IP address, gateway, netmask, and other parameters. They are connected to the same

virtual switch and, therefore, to the same physical Ethernet adapter. Before configuring software iSCSI for the ESX Server host, you need to open a firewall port.

- **Metadata Updates –** A VMFS holds files, directories, symbolic links, RDMs, and so on, along with corresponding metadata for these objects. Metadata is accessed each time the attributes of a file are accessed or modified. These operations include, but are **not** limited to:
    - Creating, growing, or locking a file.
    - Changing a file's attributes.
    - Powering a virtual machine on or off.

  CAUTION After you create a new VMFS volume or extend an existing VMFS volume, you must rescan the SAN storage from all ESX Server hosts that could see that particular volume (LUN). If this is not done, the shared volume might become invisible to some of those hosts.
- **Levels of Indirection –** If you're used to working with traditional SANs, the levels of indirection can initially be confusing.
    - You cannot directly access the virtual machine operating system that uses the storage. With traditional tools, you can monitor only the VMware ESX Server operating system (but **not** the virtual machine operating system). You use the VI Client to monitor virtual machines.
    - Each virtual machine is, by default, configured with one virtual hard disk and one virtual SCSI controller during installation. You can modify the SCSI controller type and SCSI bus sharing characteristics by using the VI Client to edit the virtual machine settings. You can also add hard disks to your virtual machine.
    - The HBA visible to the SAN administration tools is part of the ESX Server system, **not** the virtual machine.
    - Your ESX Server system performs multipathing for you. Multipathing software (such as PowerPath) in the virtual machine is **not** supported (and not required).
- **Choosing Larger or Smaller LUNs:**

Plan how to set up storage for your ESX Server systems before you perform installation.

- **One large LUN or many LUNs with a single VMFS volume spanning all LUNs**: You might want fewer, larger LUNs for the following reasons:
    - More flexibility to create virtual machines without asking the SAN administrator for more space.
    - More flexibility for resizing virtual disks, doing snapshots, and so on
    - Fewer LUNs to identify and manage

- **Many LUNs with one VMFS volume on each LUN:** You might want more, smaller LUNs for the following reasons:

- Less contention on each VMFS because of locking and SCSI reservation issues.
- Different applications might need different RAID characteristics.
- More flexibility (the multipathing policy and disk shares are set per LUN).

NOTE: *You can divide your datacenter into servers that are best configured with fewer, larger LUNs and other servers that use more, smaller LUNs.*

NOTE: *You can boot from a SAN only with ESX Server 3 and with hardware iSCSI.*

NOTE: *If you plan to use NIC teaming to increase the availability of your network access to the iSCSI storage array, you must turn off port security on the switch for the two ports on which the virtual IP address is shared. The purpose of this port security setting is to prevent spoofing of IP addresses. Thus many network administrators enable this setting. However, if you do not change it, the port security setting prevents failover of the virtual IP from one switch port to another and NIC teaming cannot fail over from one path to another. For most LAN switches, the port security is enabled on a port level and thus can be set on or off for each port.*

NOTE: *SCSI reservations are held during metadata updates to the VMFS volume. ESX Server uses short - lived SCSI reservations as part of its distributed locking protocol.*

NOTE: *VMware recommends that you load balance virtual machines over servers, CPU, and storage. Run a mix of virtual machines on each server so that not all experience high demand in the same area at the same time.*

NOTE: *Whether a virtual machine can run management software successfully depends on the storage system in question.*