IP SAN Fundamentals: An Introduction to IP SANs and iSCSI

Updated April 2007

Table of Contents

Introduction: Why Build Storage Area Networks?

In a Direct Attached Storage (DAS) environment, more storage capacity than applications require is commonly purchased as insurance against running out of space: this is called over-provisioning.

Storage Area Networks (SANs) allow multiple heterogeneous hosts to share storage sub-systems. Disks are pooled behind an array controller; RAID volumes are created from the disks; the RAID volumes are carved up into volumes; and the volumes are presented to hosts. This reduces the need to over-provision resulting in: fewer sub-systems in the Data Center; higher utilization of the available storage; and savings on space, heat and power.

Fibre Channel SANs became popular in the late 1990s and are widely used today but, despite prices falling in recent years, the costs of a Fibre Channel Switch infrastructure, Fibre Channel Host Bus Adapters and training staff in the new skills required to manage a SAN remain a barrier to entry for many organizations and IT groups: IP based SANs offer many of the benefits of Fibre Channel SANs, but at a lower cost.

This document provides a high level technical overview of IP SANs and iSCSI, and positions IP SANs relative to Fibre Channel SANs and NAS.

Terminology

iSCSI

iSCSI (Internet Small Computer System Interface) is a data transport protocol used to carry block-level data over IP networks.

IP SAN

An IP SAN is a Storage Area Network that uses the iSCSI protocol to transfer block-level data over a network, generally Ethernet.

Initiator

In this document the term "initiator" is used interchangeably to refer to a server, host or device driver that initiates (i.e. begins) iSCSI command sequences.

Target

iSCSI targets break down iSCSI command sequences from initiators and process the SCSI commands. Examples of iSCSI targets are a disk or tape device with an iSCSI port and a NAS appliance with iSCSI target support.

Fibre Channel SAN

A Fibre Channel SAN is a Fibre Channel network over which the SCSI-FCP (Fibre Channel SCSI) protocol runs. See References for a resource that discusses Fibre Channel SANs in detail.

iSCSI Fundamentals

iSCSI Enabling Hosts

Software based iSCSI initiators are available for most operating systems including: the Solaris™ OS; Microsoft Windows; AIX, HP-UX; and Linux. These work with standard NICs.

iSCSI Host Bus Adapters (HBAs) are available to provide iSCSI support for some operating systems (or versions of operating systems) that do not have software initiators available. See reference (v) for a list of Solaris Ready iSCSI HBAs.

iSCSI Node Names

Every iSCSI initiator and target has a worldwide unique identifier called a node name. A target's node name is assigned by the manufacturer.

On a host, a worldwide unique node name is generated by the iSCSI initiator software when it is first enabled; the node name can also be set manually. A host's node name is used to represent all of the network ports for that host, i.e. it is unique to the host, not to an individual network port on the host.

There are two formats for iSCSI node names: Extended Unique Identifier (EUI) and iSCSI Qualified Name (IQN).

EUI's look very like Fibre Channel Worldwide Numbers. e.g. eui.02004567A425678A.

IQN's look like an unusual DNS name. An example of an IQN is iqn.1986-03.com.sun:01:e00000000000.44180a08.

The IQN format is proving to be more popular. Note that the domain name style information in an IQN node name is not derived from the host's actual domain but is defined by the company which wrote the initiator stack; the above example is from the iSCSI software initiator in the Solaris 10 OS.

iSCSI Log In

An iSCSI log in is the process of establishing an iSCSI session.

iSCSI Sessions and Portals

The association between an iSCSI initiator and an iSCSI target is known as an iSCSI session. To establish an iSCSI session the initiator logs into the target using the target's IP address and a TCP port number. This IP address and TCP port number pair is known as an iSCSI Portal.

iSCSI Discovery

iSCSI Discovery is the process by which an iSCSI initiator can learn which target iSCSI node names are available to it.

There are a number of different methods of discovery:

Static Configuration

The initiator is told the complete target name including portal addresses, etc. This information is configured manually.

Send-Targets

The initiator is told to query a discovery IP address. The initiator communicates with the discovery address to receive all the configuration data available to this initiator for that target, e.g. all of the volumes it has access to. This needs to be repeated for each target.

Internet Storage Name Service

For small IP SANs the methods described above will suffice. For larger IP SANs the Internet Storage naming Service (iSNS) removes the need to manually enter discovery information on each initiator by providing centralized naming services, iSNS is a large topic and is covered in a later section.

iSCSI Multipathing

Where iSCSI multipathing is required, there are a number of options available. Sun has published a Sun BluePrints™ document which discusses a range of iSCSI multipathing solutions; see References.

Internet Storage Name Service

The Internet Storage naming Service (iSNS) removes the need to manually enter discovery information on every initiator by providing centralized naming services for the IP SAN.

The iSNS server can be implemented on a range of platforms including network switches and as an application running on a host: Cisco implements an iSNS server in some of its SAN switches (which support Fibre Channel and iSCSI) and Microsoft offers a free iSNS server that runs on a server using Microsoft Windows. There are also plans to support an iSNS server in a future release of the Solaris 10 OS.

Initiators and targets must have iSNS client support built in to be able to work with an iSNS server. The iSCSI devices will need to be configured to register themselves with the iSNS server when they come online. When devices register they provide information about themselves to the name service (e.g. if they are a target or an initiator) and tell the iSNS server which events they wish to be notified of. Initiators then query the iSNS server for target information as part of the discovery process, and log in to the targets to find out more about them.

It is not always desirable that all initiators see all devices, and the iSNS Administrator can set up Discovery Domains that define which targets an initiator is given access to when it queries the iSNS server.

If a device is added or removed from the IP SAN a State Change Notification (SCN) is sent by the iSNS server to iSCSI nodes registered to receive them. For example, if a target goes offline initiators need to be notified so that they can flag LUNs as offline at the host OS level. Normally only initiators register to receive SCNs.

IP SANs are usually built using Ethernet switches which are not aware of iSNS. The iSNS server is aware of new devices joining the SAN as they register with it, but the service will not be notified by Ethernet switches of devices being removed. To ensure that it has an up to date picture of the IP SAN, the iSNS server periodically sends an Entity Status Inquiry (ESI) to all the registered devices to check that they are still present and if it finds that a device has gone offline it will send SCNs to the affected initiators.

Note that in a Fibre Channel SAN the Fibre Channel Simple Name Service runs on all the Fibre Channel switches in the SAN: there is no requirement to check if devices have been removed via direct polling, as the individual switch will see that a device connected to it has gone offline and will notify all the other switches in the Fabric of the event. See References for more on this topic.

LUN Masking

If a target is an array or NAS appliance with iSCSI support, many hosts may be initiating against it. A method of controlling access to the target's volumes is necessary, otherwise multiple hosts can discover and try to use the same volume and, with the exception of certain applications which support or require shared storage, data corruption would almost certainly result.

To achieve this, the Administrator maintains Access Control Lists (ACLs) on the iSCSI target which contain a list of the initiator node names that are permitted to access each iSCSI volume. Initiators cannot discover volumes that they have not been given access to. The volume (also known commonly as a LUN) is "masked": they cannot see it.

IP SAN Security

VLANs

Virtual Local Area Networks (VLANs) are the most common method of securing IP SANs. VLANs can be used to isolate iSCSI nodes from other devices on the network.

CHAP

The Challenge Handshake Authentication Protocol (CHAP) is used for authentication between iSCSI targets and iSCSI initiators.

CHAP can be Unidirectional or Bidirectional: using Unidirectional CHAP, an iSCSI initiator authenticates itself with an iSCSI target using a secret key (i.e. a password) known as the CHAP secret; using Bidirectional CHAP the target then also authenticates itself with the initiator using a second CHAP secret.

A RADIUS server can be use to simplify CHAP secret key management when using Bidirectional CHAP authentication (A RADIUS server is a centralized authentication service). While you must still specify the initiator's CHAP secret, you are no longer required to specify each target's CHAP secret on each initiator.

IPsec

IP Security (IPsec) is a set of protocols developed by the Internet Engineering Task Force (IETF) to support the secure exchange of packets at the IP layer. IPsec is deployed widely to implement Virtual Private Networks (VPNs).

IPsec can operate in Transport Mode or Tunnel Mode:

In Transport Mode, protection is provided all the way from the source to the destination. For iSCSI this would require that the initiator and the target support IPsec.

Tunnel mode provides gateway-to-gateway transmission security. This requires no special support in the iSCSI host driver or target. Data in transmission remains unprotected until it reaches a network gateway. Once at the gateway, it is secured with IPSec until it reaches the destination gateway. At this point, data packets are decrypted and verified. The data is then sent to the receiving host unprotected. Tunnel mode is often employed when data must leave the secure confines of a local LAN or WAN and travel between hosts over a public network such as the Internet.

iSCSI Host Bus Adapters

Figures of between 500 MHz and 1 GHz of CPU are often quoted as being required to drive an iSCSI software stack at line speed through a Gigabit Ethernet NIC. Most of these CPU cycles are actually consumed processing the TCP/IP stack transporting the data.

iSCSI HBAs process the TCP/IP and iSCSI commands using on-board custom chips, so off-loading the host's CPU(s). iSCSI HBAs provide the added convenience of presenting iSCSI volumes as standard SCSI: no native iSCSI support is required in the host OS.

Whilst the vast proportion of iSCSI users are happy using an iSCSI software stack, the CPU load generated by I/O intensive applications can become an issue in some cases; when this occurs, iSCSI HBAs can be used to resolve the problem.

See reference (v) for a list of Solaris Ready iSCSI HBAs.

Booting Over iSCSI

Until recently, the only way to support booting over iSCSI was to use iSCSI HBAs. Information about the host's iSCSI boot device is programmed directly into the HBA and, so far as the host OS, is concerned it is booting off a local SCSI disk.

Software-only based iSCSI boot solutions are now available for some OS that work with standard NICs. Microsoft announced software-only support for diskless booting of Microsoft Windows over iSCSI; see reference (iv). Solutions are available for Linux. In addition, there are plans to support this in a future release of the Solaris 10 OS.

iSCSI Targets and iSCSI Routers

The most popular iSCSI targets are NAS appliances; iSCSI support allows NAS appliances to provide block services as well the traditional file services. Other options are to add iSCSI connectivity to an existing Fibre Channel array or use an iSCSI Router to provide connectivity for iSCSI initiators to arrays in an existing Fibre Channel SAN.

iSCSI Routers route SCSI traffic between IP SANs and Fibre Channel SANs. They are typically used where an organization has a Fibre Channel SAN and wishes to give a number of hosts access to devices in that SAN without the expense of Fibre Channel HBAs.

For example, we wish to give iSCSI initiators access to volumes on a Fibre Channel array connected to a Fibre Channel SAN. We connect the iSCSI Router to the Fibre Channel SAN, present volumes from the array to the router over Fibre Channel, and the router then does the necessary protocol conversion between SCSI-FCP and iSCSI to present those volumes to hosts on the IP SAN.

Extending IP SANs Over Wide Area Networks

Extending Fibre Channel SANs over long distances requires specialized hardware to route Fibre Channel over IP network links.

iSCSI runs natively over IP networks and benefits from the TCP/IP flow control mechanisms and optimizations that allow them to work efficiently over long distances. This means that iSCSI initiators can connect directly to targets over an existing network infrastructure without any additional specialized equipment being required.

Positioning IP SANs with NAS and Fibre Channel SANs

NAS and IP SANs

NAS clients access files in file systems on a File Server or NAS appliance over an IP network using CIFS or NFS protocols.

iSCSI provides hosts with block-level access to data over an IP network. Most NAS appliances now support iSCSI, and can service iSCSI traffic on the same ports as they service CIFS and NFS.

Block-level access is more suitable for some applications than NFS or CIFS, particularly databases and some email systems: Microsoft does not support Microsoft Exchange or Microsoft SQL Server over CIFS or NFS, but does support them over iSCSI.

To provide an iSCSI LUN to an initiator an extent of storage on a NAS appliance is designated as a raw iSCSI volume. An iSCSI initiator sees this volume as a LUN and can create a file system on it. It is important to note that you cannot access the same storage or data through both iSCSI and CIFS or NFS, the NAS appliance has no visibility of the data in the file system built on the iSCSI volume by the client, and the raw iSCSI volume cannot be shared to CIFS or NFS clients.

Fibre Channel SANs and IP SANs

Fibre Channel SANs require a Fibre Channel switch infrastructure to be installed and all the hosts require Fibre Channel HBAs. This can be cost-prohibitive for smaller organizations and IT groups, so Fibre Channel SANs tend to be deployed in Data Centers where an organization's largest servers, storage sub-systems and the most important applications are housed; Core Data Centers generally can justify the costs and have the necessary skills to manage a Fibre Channel SAN.

IP SANs can be deployed at lower costs than Fibre Channel SANs: IP SANs run over existing IP networks; hosts can be connected via standard Network Interface Cards (NICs); and software iSCSI initiators provided as part of the host OS can be used to connect to iSCSI target devices. To a large degree, existing networking skills and tools can be used to manage an IP SAN, but it must not be overlooked that IT staff will need to become familiar with LUN management tasks on targets and hosts.

Making a Choice between an IP SAN or a Fibre Channel SAN

What if you have to make a decision about proposing or buying an IP SAN based solution? This section offers some areas to consider other than cost, and finishes with a checklist.

Support and Interoperability

Fibre Channel technology is mature and well understood by hardware and software vendors, so a broad level of support and interoperability is available.

IP SANs are a relatively new technology area, so support by hardware and software vendors and the breadth of interoperability are less well developed.

Performance

Bandwidth

Most Fibre Channel SANs in production today are built on 2 Gbit Fibre Channel. 4 Gbit Fibre Channel devices are now available, and 8 Gbit is planned.

Most organizations run a combination of 100 Mbit and 1 Gbit Ethernet networks. 10 Gbit Ethernet is available but is not widely implemented. Where 10 Gbit Ethernet is installed it tends to be used as a Data Center backbone, not for connections to individual hosts and especially not for connections to low cost servers: it is fair to assume that most IP SANs today will run over 100 Mbit and 1 Gbit networks.

Latency

Bandwidth is not everything; I/O latency is very important for some applications. Database log files are very latency sensitive for example.

In both Fibre Channel SANs and IP SANs, the locality of targets and initiators and the loading of the network contribute to I/O latency. A factor in the favor of Fibre Channel SANs is that they are dedicated to block I/O, an organization's Ethernet network will not be. Deploying a latency sensitive and/or I/O intensive application using iSCSI over an existing network may result in performance problems. Direct connection of the hosts to the iSCSI target or dedicated Ethernet switches or Ethernet segments for the IP SAN is an option in these cases.

IP SAN Checklist

The below list of questions may be useful when considering an IP SAN:

- Does my iSCSI target support the initiators I wish to connect to it?
- Does my application vendor support the chosen hardware and software combination that make up my proposed IP SAN?
- IP SANs can be deployed over existing infrastructures but I/O intensive applications will generate significant amounts of network traffic. Is there capacity for this in the existing network? Do I need a dedicated switch?
- Network latency is an issue for some applications. Is my application vendor happy with my network latencies? Do I need a dedicated switch?
- What solutions are available for host to target iSCSI multipathing?
- Do I need to consider iSCSI HBAs?
- Can I manually manage the relationships between my targets and initiators or do I need an iSNS server?
- If I want to build a HA Cluster: is Clustering with iSCSI attached storage supported for my application?
- Are there any benchmark results, reference architectures or cases studies that I should look at?

References

i. Sun BluePrints Document: Using iSCSI Multipathing in the Solaris™ 10 Operating System
http://www.sun.com/blueprints/1205/819-3730.pdf

ii. SAN Fundamentals: How Fibre Channel SANs Are Built, Secured and Managed
(on BigAdmin): http://www.sun.com/bigadmin

iii. Internet Storage Name Service (iSNS) - A Technical Overview
http://www.diskdrive.com/iSCSI/reading-room/white-papers/Nishan_iSNS_A_Technical_Overview.pdf

iv. Microsoft Announces Availability of Windows Storage Server 2003 R2 With OEM Partners
http://www.microsoft.com/presspass/press/2006/apr06/04-04SNWPR.mspx

v. Solaris Ready iSCSI HBAs
http://www.sun.com/io_technologies/index.html

For more information, please see the Storage Administration Site on BigAdmin:
http://www.sun.com/bigadmin/hubs/storage/