iSCSI Performance Options

Performance Solutions for iSCSI Environments

Toby Creek, Network Appliance | June 2005 | TR 3409

TECHNICAL REPORT

Network Appliance, a pioneer and industry leader in data storage technology, helps organizations understand and meet complex technical challenges with advanced storage solutions and global data management strategies.

Abstract

The iSCSI protocol has enabled information technology organizations to lower the cost of their storage area network (SAN) deployments. Cost is of little concern if the solution does not perform to expectations. This paper will examine some of the issues and options available to maximize performance in iSCSI environments.

Contents

1 Introduction

Since the iSCSI protocol was ratified in February of 2003, the ecosystem of products and solutions has grown at a steady pace. Deployments of iSCSI have grown with a more exponential curve. Significant growth in a technology attracts the attention of enterprise customers, who watch the market very closely for emerging technologies that offer them reduced cost, easier integration, and better service levels.

Performance is a major concern for these customers. Enterprise customers spend significant amounts of money on a variety of different projects, so it is imperative that they get the maximum value from any deployed solution. Maximum value is usually achieved by increasing the utilization of their investment. Performance enhancement of any solution means that they can increase utilization with the aim of deploying less hardware and therefore minimizing the capital investment required.

This document will explore the options and issues that must be considered when building a high-performance solution. Specific technical recommendations will be made for optimizing performance with Network Appliance™ storage systems.

2 General Network Design Recommendations

The foundation of a performance-oriented iSCSI solution is the storage network. An iSCSI storage network should consist of enterprise-class Gigabit Ethernet switches that support advanced networking features such as link aggregation, jumbo frames, and VLANs. The switches considered should be able to support full wire speed on its ports. Consequently, the switch's backplane should have sufficient bandwidth to support the anticipated peak traffic. For manageability and security, the storage network should be segmented from any front-end application or user networks, either by employing a VLAN or using a separate switch chassis.

To simplify the fabric and avoid network congestion, all switch and host ports in the SAN should be configured for the highest-speed full-duplex operation, overriding any autonegotiation functionality. Full-duplex operation allows the switch and host to exchange data bidirectionally at the same moment in time, as compared to half-duplex operation which requires that transmission occur in only one direction at a time. In half-duplex operation, simultaneous transmission is termed a "collision"; the packets are discarded and must be retransmitted. Half-duplex communication is required when the physical medium lacks enough wires to accommodate bidirectional signaling, such as coaxial cable, or when nonintelligent network equipment is used. Neither of these conditions should exist in a modern IP network designed to carry storage traffic.

3 Initiator and Target Technologies

A storage network consists of two types of equipment: initiators and targets. Initiators are data consumers, such as hosts. Targets are data providers, such as disk arrays or tape libraries. Initiators and targets, collectively referred to as "end-points," can be software, software with hardware assistance, or hardware. This section will examine the features and issues with each of these technologies.

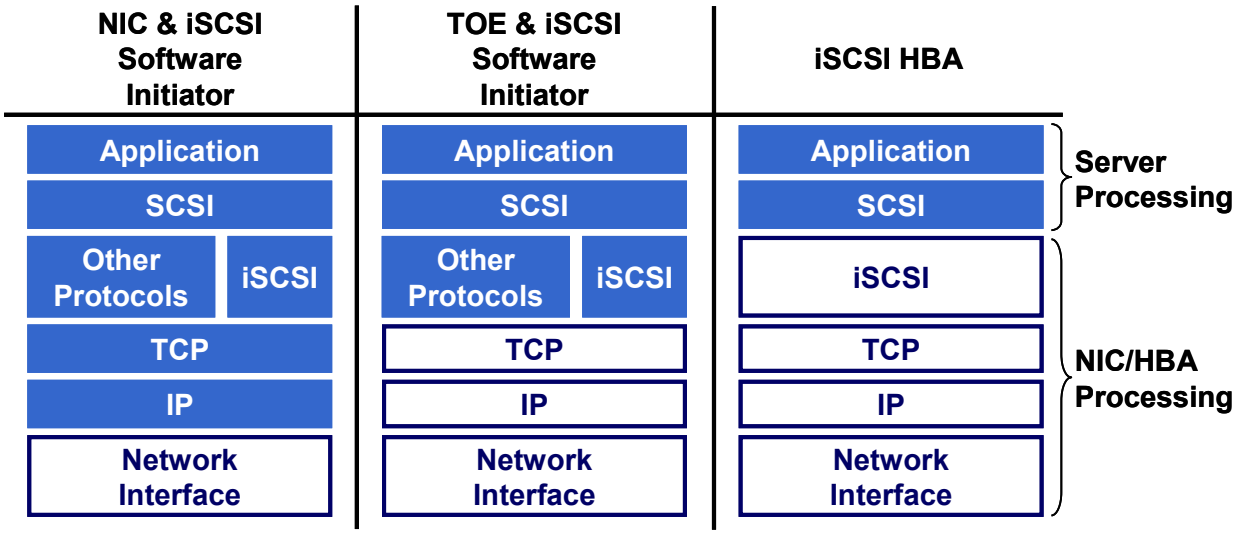

Figure 1) iSCSI end point technology comparison chart.

3.1 Software-only Solutions

Software initiators and targets are virtual SCSI adapters written as part of the operating environment. They use the host's CPU resources and network adapters to transfer data. Software end-points are easy to deploy and are either low-cost or free with the host operating system. The Network Appliance software iSCSI target is provided for free with the purchase of any other protocol license.

Software implementations can drive higher throughput than other implementations if sufficient host CPU resources are available. This is especially true of cases where smaller block sizes are used. Integration with the host operating system is usually very good, leveraging existing management tools and interfaces. Starting up a host from an iSCSI device is not possible using software initiators unless a prestartup execution environment exists. At a minimum, a DHCP server and some kind of file transfer protocol such as TFTP are required.

3.2 Software with Hardware Assistance

Hardware assistance in the context of an iSCSI software end-point generally comes in the form of a TCP Offload Engine, or TOE. With a TOE, the TCP stack processing, including framing, reordering of packets, checksums, and similar functions are offloaded to a dedicated card with its own network interface port. The TOE card may be a general-purpose card able to offload TCP traffic, or it may be restricted to just accelerating iSCSI traffic.

TOE adapters enjoy most of the benefits of software initiators with additional host CPU offload. TOE adapters may also support advanced networking features like link aggregation. Because the software initiator is still used on the host, integration with layered management applications is unaffected by the addition of the TOE hardware.

3.3 Hardware-only Solutions

Hardware adapters offload the TCP stack processing as well as the iSCSI command processing functions. The hardware adapter will look and function as a SCSI disk interface, just as a Fibre Channel

HBA does. The operating system has no knowledge of the underlying networking technology or interfaces. A separate management interface is used to configure the card's networking parameters.

Hardware-only solutions offload the largest amount of processing from the host CPU. Because they function as SCSI adapters, it is possible to start up from them if they provide the appropriate host BIOS interfaces and are recognized as a startup device. Advanced networking features may not be available due to lack of software visibility to the network functions in the card.

4 Jumbo Frames

Jumbo frame is a term applied to an Ethernet frame that carries more than the standard 1500-byte data payload. The most commonly quoted size for a jumbo frame is 9000 bytes, which is large enough for 8KB application data plus some amount of upper-layer protocol overhead.

Jumbo frames can improve performance in two ways:

1. Packet assembly/disassembly in high-throughput environments can be an intensive operation. A jumbo frame decreases the amount of packet processing operations by up to a factor of six.
2. The overhead associated with the Ethernet packet once prepared for transmission is a smaller percentage of a jumbo frame than a regular sized frame.

In testing, jumbo frames can result in a throughput increase of up to 30%.

Jumbo frames require the end points and all devices between them in the network to be configured to accept the larger packet size if they are not configured for them by default. This includes any network switching equipment. An example command to configure a Network Appliance storage system for jumbo frames is shown below:

```
netapp> ifconfig e9a mtusize 9000
```

The value for the maximum transmission unit size as configured should be added to the */etc/rc* file to be made persistent. Jumbo frame support can also be configured by running the `setup` command again.

Larger frame sizes can be used, but the 32-bit Ethernet CRC mechanism used to generate the 4-byte Frame Check Sequence (FCS) at the end of the frame begins to lose its effectiveness at around 12,000 bytes.

5 Link Aggregation

Link aggregation is the technique of taking several distinct Ethernet links and making them appear as a single link. Traffic is directed to one of the links in the group using a distribution algorithm. Availability is also enhanced, as each of the schemes presented in this section can tolerate path failure with minimal service disruption. This technology is referred to by many names, including *channel bonding, teaming,* and *trunking.* The term *trunking* is technically incorrect, as this term is used to describe VLAN packet tagging as specified in IEEE 802.1q.

The link aggregation technologies that will be examined here are IEEE 802.3ad/EtherChannel, active/active multipathing, and iSCSI multiconnection sessions.

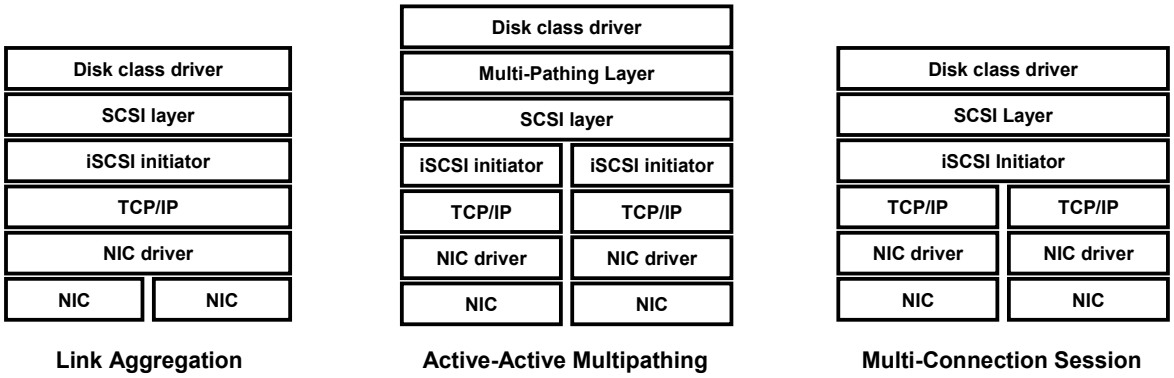

Figure 2) iSCSI architectures.bandwidth enhancement

5.1 IEEE 802.3ad Link Aggregation Control Protocol (LACP) and EtherChannel

IEEE standard 802.3ad and EtherChannel both specify a method by which multiple Ethernet links are aggregated to form a single network link. Packet traffic is distributed among the links using a hash function based on a variety of parameters, including source and destination MAC addresses. The paths are coalesced at the NIC driver layer or the operating system interface layer. A single IP address is assigned to the set of links on both the initiator and target.

Because the hash function used will always return the same value for a given initiator/target pair, the traffic distribution across an aggregated link may not be uniformly distributed. Link aggregation is best used to increase fan-in bandwidth to a target rather than a “large pipe” between a single initiator and target pair. A round robin algorithm would correct this problem, but it is not a valid algorithm according to the 802.3ad standard. EtherChannel and other nonstandard mechanisms do allow for the use of round robin distribution.

5.2 Active-Active Multipathing

Active-active multipathing requires a software multipathing layer to be inserted in the software stack above the SCSI layer. This layer of software takes the multiple device paths presented from the SCSI layer and coalesces them into single device paths, though there are multiple links to each device. This is accomplished by using SCSI inquiry commands to check for commonality in device paths, usually the LUN serial number. This abstraction prevents the operating system from trying to access multiple device paths as if they are individual devices, which would most likely result in data corruption.

Active-active multipathing can enhance the available bandwidth between an initiator and target, but this will depend on the implementation of the load balancing algorithm. Since SCSI does not have a robust sequencing mechanism, SCSI exchanges must always be delivered in sequence, which generally means they must use a single path for the exchange. Once the command frame has been placed into an IP packet, ordering is guaranteed, but the software must properly handle command frames before encapsulation takes place.

Active-active multipathing is most useful for situations where multiple HBAs are used, since HBAs do not generally present the network layers so that advanced networking features can be implemented.

5.3 Multiconnection Sessions

Multiconnection sessions, or MCS, work just as the name implies; for each iSCSI session, multiple connections are created. The number of allowed connections is negotiated during login and session creation. While it is possible to create multiple connections over a single physical interface, bandwidth enhancement requires that multiple physical interfaces be employed.

MCS must be implemented in both the initiator and target. Distribution of iSCSI traffic over the connections is not defined and therefore will be only as effective as the algorithm implemented.

6 Performance Impact of iSCSI Data Integrity Features

The iSCSI standard specifies additional features to detect data corruption. Digests supplement the TCP header checksums and Ethernet frame check sequence that were not considered robust enough for use with storage networks when the iSCSI standard was being developed. They are designed to detect bit errors and take action based on the Error Recovery Level (ERL) negotiated between the initiator and target at login time. The usefulness of digests when the ERL with random accessible storage is minimal.

Header digests are relatively inexpensive operations, as the 32-bit CRC algorithm need process only the 48-byte header of the iSCSI PDU. Performance impact of enabling header digests is on the order of 10% reduction in throughput versus no digest.

When compared to header digests, computing the data digest can be a much more expensive operation. The data segment of an iSCSI PDU can be significantly longer than the header portion of the PDU, resulting in more computation time during packet construction. Enabling data digests can reduce throughput from 30% to 70%, depending on the size of the underlying SCSI command frame.

7 Performance Impact of IPSec Signing and Encryption

For most deployments, if iSCSI storage networks are implemented on a nonrouted VLAN separate from front-end traffic, additional security measures above LUN masking and CHAP authentication are not usually needed. Extremely security conscious environments may opt to encrypt all of their storage network traffic. Since iSCSI is an IP-based protocol, all of the security features available to IP networks in general are applicable to IP storage networks.

IPSec is currently the most widely deployed encryption mechanism for IP networks. IPSec can utilize a number of different encryption algorithms, including the Data Encryption Standard (DES) and the Advanced Encryption Standard (AES). IPSec also has an Authentication Header (AH) mechanism that uses a hash function to ensure that static portions of the packet header are not manipulated in transit. IPSec uniquely identifies each party in an exchange with a certificate or private key issued by a Certificate Authority (CA) as well as a public key provided by the Public Key Infrastructure (PKI).

The encryption process of IPSec applied to iSCSI is a very computationally expensive process. When IPSec is implemented in software on the initiator and target, the overhead of IPSec processing can reduce throughput by as much as a factor of 10. Hardware assistance is strongly recommended for sites that wish to protect their traffic in-transit, but a performance penalty will still exist over nonencrypted traffic. Devices capable of wire-speed encryption for a Gigabit Ethernet network are currently prohibitively expensive for most environments.

8 Tuning Network Appliance iSCSI Target Settings

The iSCSI software target has a minimum of tunable parameters. The most important of these parameters is `iscsi.iswt.max_ios_per_session`. This setting has a default value that ranges from 32 to 128 based on the filer platform and is configurable up to a maximum value of 256. In certain instances, increasing the value may increase throughput if the storage processor is not fully utilized.

The option can be viewed and changed from the NetApp command line interface as shown below:

```
netapp> options iscsi.iswt.max_ios_per_session
iscsi.iswt.max_ios_per_session 32
netapp> options iscsi.iswt.max_ios_per_session 256
```

Setting a lower value can also be useful in environments where iSCSI and other protocols are used simultaneously. For example, a company uses a Network Appliance storage system with Fibre Channel attached hosts for its production environment and iSCSI connected hosts for development. Lowering the maximum I/Os per session will throttle the iSCSI traffic, preserving low response times for the production environment.

Note that changing `iscsi.iswt.max_ios_per_session` affects only new iSCSI sessions. Initiators that are already connected are not affected unless they start a new session.

9 Conclusion

Since its standardization, the number of performance solutions available for iSCSI has grown as rapidly as the deployment of iSCSI SANs. As the leader in iSCSI storage, Network Appliance has verified interoperability and performance with many vendors offering iSCSI products and solutions. The effectiveness of the solution will depend on the individual products chosen, their features, and the level of testing that has been done with Network Appliance storage systems. Specific performance expertise may be required during the design and implementation of the storage solution.

NetApp®

Network Appliance, Inc.
495 East Java Drive
Sunnyvale, CA 94089
www.netapp.com