Technical Report

Windows Multipathing Options with Data ONTAP: Fibre Channel and iSCSI

Ryan Hardin, NetApp
June 2013 | TR-3441

Version 3.0

SAN Configurations for High Availability

This report provides a description of the various multipathing options available for iSCSI and Fibre Channel SANs on Microsoft® Windows® in conjunction with NetApp® Data ONTAP®. The pros and cons of each solution are discussed with the intention of helping the reader determine the best solution for the reader's particular environment.

TABLE OF CONTENTS

LIST OF FIGURES

1 Introduction

In order to have a highly available storage area network (SAN), steps must be taken so that no single failure results in an outage. In this report we look at redundancy in the links connecting hosts to storage systems and the options available to achieve a highly available SAN infrastructure.

2 Intended Audience

This paper is intended for system and storage architects designing iSCSI and FC (Fibre Channel) solutions with NetApp storage appliances. We assume that:

- The reader has a general knowledge of NetApp hardware and software solutions, particularly in the area of block access.
- The reader is familiar with block-access protocols such as Fibre Channel and iSCSI.

This paper highlights the functionality of NetApp Data ONTAP 8.2 and Microsoft Windows Server® 2012, but includes information about past versions of both Data ONTAP and Microsoft Windows Server where appropriate.

A complete list of linked references is included at the end of the document.

3 Multipathing

Multipathing is the ability to have multiple data paths from a server to a storage array. Multipathing protects against hardware failures (cable cuts, switches, HBAs, and so on) and can provide higher performance limits by utilizing the aggregate performance of multiple connections. When one path or connection becomes unavailable, the multipathing software automatically shifts the load to one of the other available paths. Multipathing is often split into two categories: active-active and active-passive. A multipathing solution is generally considered to be active-active when I/O for a single LUN travels multiple paths simultaneously.

3.1 Eliminating Single Points of Failure

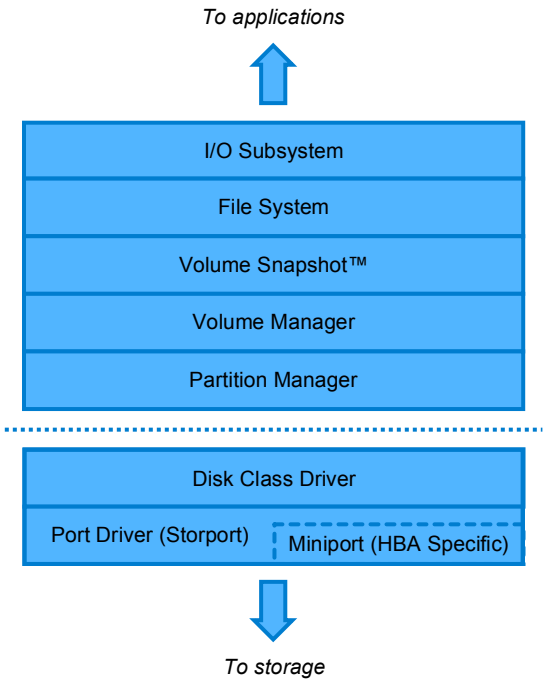
Electrical and mechanical components can always fail. The way to achieve high availability is by eliminating single points of failure in the environment. That way, when an individual component does fail, the overall system continues to be available to users. Any multipathing solution should utilize separate adapters, cables, and switches to avoid a single point of failure.

3.2 Windows Storage Stack

When an application writes data to disk, that data flows through the host-side storage stack and through its storage interconnect (for example, parallel SCSI, Fibre Channel, iSCSI, and so on) to the storage array.

Figure 1 illustrates the storage stack for Microsoft Windows Server 2012.

Figure 1) Microsoft Windows storage stack.



Multipathing is achieved by sophistication at some layer of the storage stack. The application writes to a single file system of raw device. The multipathing-capable layer receives the request and routes it to one of the underlying data paths. This routing is performed transparently to the other layers of the stack, both above and below the multipathing layer. There are various layers in which this split from a single to multiple paths can occur. Each option has its advantages and limitations.

3.3 Choosing the Right Multipathing Solution

Three supported multipathing solutions are described in this document: link aggregation, iSCSI with multiple connections per session (MCS), and MPIO.

Link aggregation can be used in conjunction with both MCS and MPIO and offers advantages and disadvantages that are described in detail in section 4.

When deciding between MCS and MPIO, NetApp recommends the use of MPIO for the following reasons:

- MCS is not supported with clustered Data ONTAP. Since MPIO can be used for clustered Data ONTAP or Data ONTAP running in 7-Mode, NetApp recommends using MPIO consistently across architectures.
- MCS is an iSCSI-only feature. Since MPIO can be used for FC and/or iSCSI, NetApp recommends using MPIO consistently across protocols.
- MCS is a feature for Windows only. Since MPIO can be used for all host operating systems supported by NetApp, NetApp recommends using MPIO consistently across host operating systems.

For detailed supportability information, refer to the NetApp Interoperability Matrix Tool.

4 Link Aggregation

One possible split point is at the NIC driver layer using TCP/IP link aggregation. Link aggregation is the technique of taking several distinct Ethernet links and making them appear as a single link. It is specified by the 802.3ad IEEE specification. Traffic is directed to one of the links in the group using a distribution algorithm. This technology is referred to by many names, including “channel bonding” and “teaming.” Link aggregation is not storage specific, and all types of network traffic can benefit from the multiple connections.

With the release of Windows Server 2012, host link aggregation with the iSCSI software initiator is supported. In Windows Server 2008 R2 and earlier, the Microsoft iSCSI software initiator does NOT support link aggregation on the host. Link aggregation on the storage side (a “VIF” in Data ONTAP 7G and an “ifgrp” in Data ONTAP 8.0 and later) is supported by both Microsoft and NetApp for all host OS versions.

Advantages:

- Transparent to all network protocols: the advantages of link aggregation are shared not just with iSCSI but also with other network traffic such as NFS, CIFS, and so on.
- Well-known, mature technique.

Disadvantages:

- Not supported on a host with the Microsoft iSCSI software initiator in Windows Server 2008 R2 and earlier.
- Aggregated interfaces must be connected to the same network, often the same switch or card within a switch, thus limiting the physical isolation of the multiple paths.
- Dependent on aggregation-capable drivers and switches.
- Because the load balancing is constrained to the load balance algorithm of the aggregated interface, which typically only uses a single physical link per destination IP or MAC, you will typically see more efficient link utilization when using MPIO for load balancing than link aggregation.
- In Windows, enabling MPIO activates additional retry mechanisms that are not present when using NIC teaming alone.

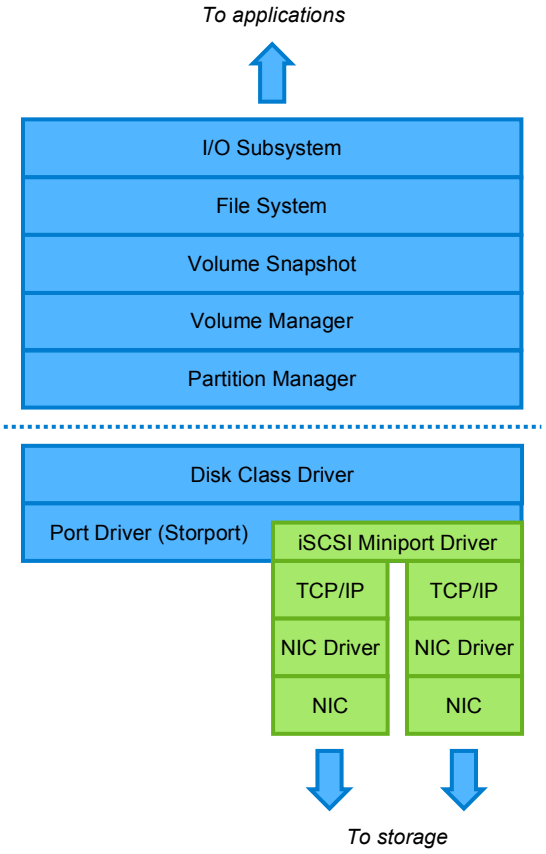
5 Multiple Connections per Session (MCS)

iSCSI sessions with multiple connections per session (MCS, or MC/S) are part of the iSCSI specification. They create multiple paths within a single iSCSI session using separate TCP connections. Both the iSCSI initiator (host) and iSCSI target (storage system) need to support MCS for them to be used. Current versions of NetApp Data ONTAP running in 7-Mode and Microsoft Windows support MCS. As of Data ONTAP 8.2 running in 7-Mode, the default maximum number of connections per session is 32. Refer to the NetApp Interoperability Matrix for the most up-to-date information regarding supported Data ONTAP and initiator releases.

While iSCSI with MCS is supported in certain 7-Mode environments, it is not supported in clustered Data ONTAP.

Figure 2 illustrates where iSCSI multiconnection sessions fit into the storage stack.

Figure 2) Microsoft Windows storage stack with MCS.



The Microsoft iSCSI initiator does support multiconnection sessions across a single path or multiple paths. SnapDrive® for Windows will work with preexisting iSCSI connections that have multiconnection sessions enabled, but will not create a multiconnection session–enabled connection and has no knowledge of those created manually. See the “Microsoft iSCSI Initiator User’s Guide” for instructions on setting up multiconnection session iSCSI connections.

iSCSI multiconnection sessions can be performed over a single target or initiator port or can utilize multiple ports on either end. If multiple target ports are used, all target interfaces for the connection must be in the same target portal group. By default, each interface is in its own target portal group. More details on iSCSI target portal groups can be found in the “Block Access Management Guide.”

Although technically possible, mixing iSCSI multiconnection sessions and MPIO multipathing styles to the same LUN is not supported by Microsoft or NetApp.

Advantages:

- Part of the iSCSI specification.
- No extra vendor multipathing software required.
- No dependency on aggregation-capable Ethernet infrastructure.

Disadvantages:

- Not supported in clustered Data ONTAP.
- Not manageable by SnapDrive iSCSI connection wizard.

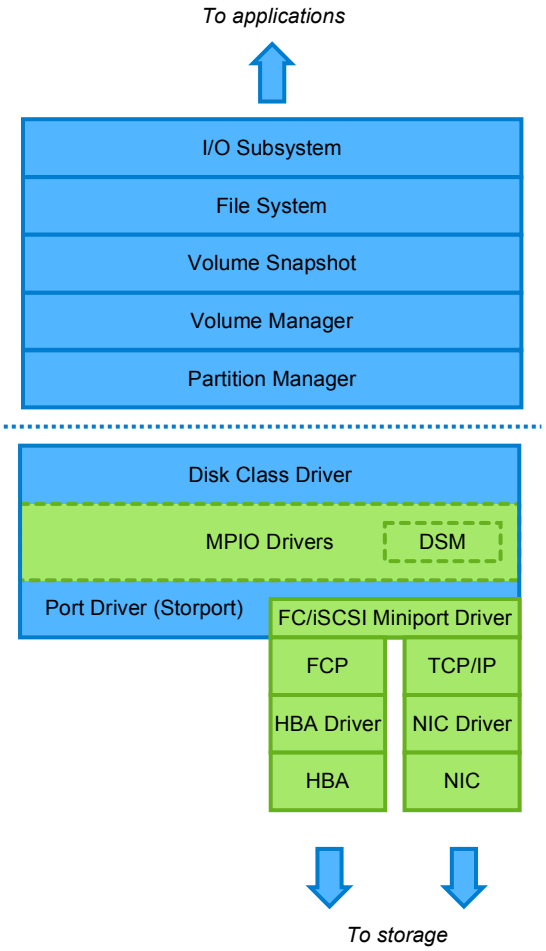
- Not supported for Microsoft software initiator boot (see the Windows Host Utilities Release Notes). Not supported when used in conjunction with MPIO.
- Load balance policy is set on a per-session basis; all LUNs in an iSCSI session share the same load balance policy.

6 Multipathed I/O (MPIO)

The classic way to do multipathing is to insert a separate multipathing layer into the storage stack. This method is not specific to any underlying transport and is the standard way to achieve multipath access to iSCSI, Fibre Channel, and even parallel and serial SCSI targets. There are multiple implementations of this type of multipathing on various operating systems. With Microsoft Windows, each storage vendor supplies a device-specific module (DSM) for its storage array. In addition, Microsoft also provides its own DSM (iSCSI only for Windows 2000 Server and Windows 2003 Server, both Fibre Channel and iSCSI for Windows Server 2008 and later).

Figure 3 illustrates how MPIO fits into the storage stack.

Figure 3) Microsoft Windows storage stack with MPIO.



Since MPIO occurs above the miniport driver layer, the MPIO driver only sees SCSI devices and does not know about the transport protocol. This allows Fibre Channel and iSCSI paths to the same LUN to be mixed. Since the protocols have different access characteristics and performance, NetApp recommends

that, if they are mixed, they be used in an active-passive configuration in which one takes over if the other fails.

Advantages:

- No dependency on aggregation-capable Ethernet infrastructure.
- Very mature implementation.
- Can mix protocols between paths (for example, iSCSI and Fibre Channel).
- Each LUN can have its own load balance policy.

Disadvantage:

- Extra multipathing technology layer is required.

6.1 Asymmetrical Logical Unit Access (ALUA)

Not all paths available to a LUN necessarily have equal access. In clustered Data ONTAP, one node owns the LUN, but ports on two or more nodes may provide access. In a NetApp HA pair running in 7-Mode, one node owns the LUN, but, in the case of Fibre Channel, ports on both nodes may provide access. Paths terminating at another node cross the cluster interconnect to reach the LUN. In clustered Data ONTAP, this traffic traverses the 10 Gigabit Ethernet cluster interconnect network. In Data ONTAP running in 7-Mode, this traffic traverses the HA cluster interconnect. Paths to ports on the owning node are called “primary” or “optimized.” Paths to ports on another node are commonly called “unoptimized,” “partner,” “proxy,” or “secondary.” Figure 4 illustrates direct and indirect paths in a two-node cluster, while Figure 5 illustrates a primary and partner path in 7-Mode. The colored lines are the paths over which data traffic flows.

Figure 4) Path failover in clustered Data ONTAP.

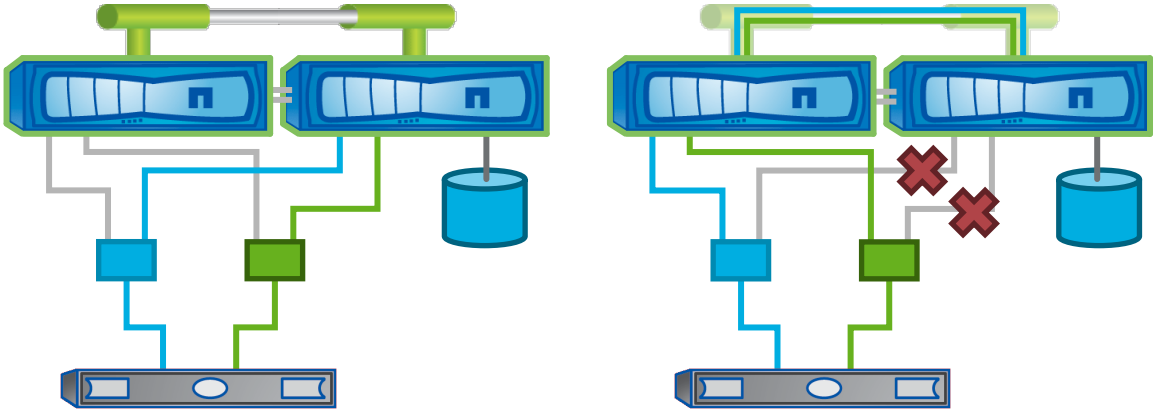
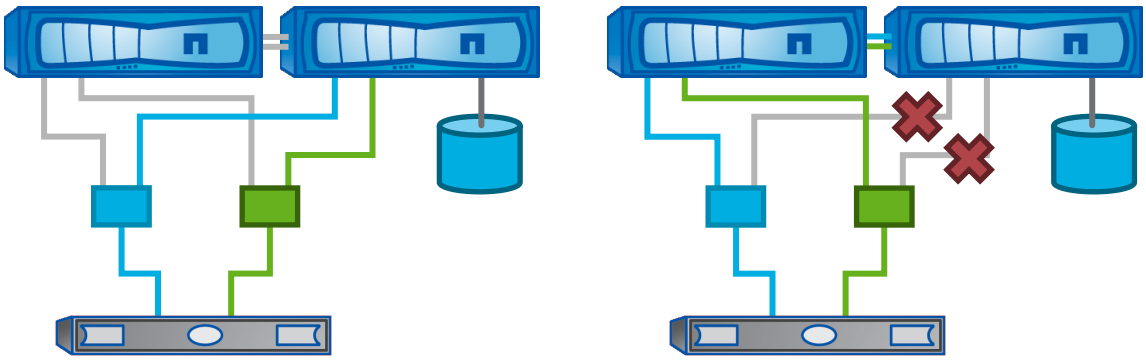


Figure 5) Fibre Channel path failover in Data ONTAP running in 7-Mode.



To make sure that data traverses the primary paths only, the host must communicate with the storage controller to determine which paths are primary and which are proxy. This has traditionally been done with vendor-specific multipathing software. A standardized method was added to the SCSI specification called Asymmetrical Logical Unit Access (ALUA) and was initially implemented in Data ONTAP 7.2 and Windows 2008. ALUA allows the initiator to query the target about path attributes, such as which paths are primary or secondary. As a result, multipathing software can be developed to support any array that uses ALUA.

In clustered Data ONTAP, ALUA must be enabled on the host to make sure of direct access to the LUN. ALUA is used for all SAN protocols, including Fibre Channel, FCoE, and iSCSI.

Unlike clustered Data ONTAP, ALUA is not supported for iSCSI connections in Data ONTAP running in 7-Mode. This is because there is no proxy path with iSCSI, and because link failover operates differently from Fibre Channel.

6.2 DSM Options

Three primary multipathing options are available for a Windows Server 2008 or newer host: the built-in Microsoft MPIO feature using the Microsoft DSM, the Data ONTAP DSM provided by NetApp, and Symantec™ Veritas™ Dynamic Multipathing (DMP). Windows Server 2003 hosts may also use the Data ONTAP DSM or the Microsoft iSCSI DSM.

Windows Server 2008 introduced a native MPIO feature that utilizes ALUA for path selection. It is enabled as a feature in Windows Server 2008 and supports both Fibre Channel and iSCSI. The standard set of load balance policies is available, including failover only, round robin, round robin with subset, least queue depth, and weighted paths. Windows Server 2008 R2 also adds the least blocks policy. The default policy for Fibre Channel connections is round robin with subset, and the default for iSCSI is failover.

If the Microsoft MPIO feature is used for LUNs connected over Fibre Channel, ALUA must be enabled on the Data ONTAP igroup to which its initiator connects. The command for this is `igroup set <igroup_name> alua yes`. Enablement should be done prior to the LUN being discovered by the host.

The NetApp Data ONTAP DSM provides standard load balance policies and adds an easy-to-use interface (both GUI and CLI). In addition, because it is a NetApp product, it is supported alongside the NetApp storage appliance, providing a single point of contact and easier troubleshooting. Windows Server 2003 hosts connected to a clustered Data ONTAP system with Fibre Channel or iSCSI must use the Data ONTAP DSM, since Windows Server 2003 does not include native multipathing components.

Symantec provides a multipathing solution called Veritas DMP. This provides a robust set of features and load balancing policies that are well suited in environments with strict performance requirements. Because Veritas DMP is also available for VMware®, Linux®, and UNIX®, using Veritas DMP can provide a homogeneous management interface across platforms with different operating systems.

Note: At the time of publication of this document, Symantec Veritas DMP is not supported with clustered Data ONTAP deployments. For supportability information, refer to the NetApp Interoperability Matrix Tool.

6.3 Load Balance Policies

When multiple paths to a LUN are available, a consistent method of utilizing those paths should be determined. This is called the load balance policy. There are six standard policies in Windows Server 2012, and they apply to MCS and MPIO:

- Failover only: Only one path is active at a time, and alternate paths are reserved for path failure.
- Round robin: I/O operations are sent down each path in turn.

- Round robin with subset: Some paths are used as in round robin, with the remaining acting as failover only.
- Least queue depth: I/O is sent down the path with the fewest outstanding I/Os.
- Least blocks: I/O is sent down the path with the fewest outstanding blocks.
- Weighted paths: Each path is given a weight identifying its priority, with the lowest number having the highest priority.

The Data ONTAP DSM adds the Auto Assigned policy:

- Auto assigned: The Auto Assigned policy is an “active-passive” policy. For each LUN, only one path is used at a time. If the active path changes to a passive path, the policy chooses the next active path. The Auto Assigned policy does not spread the load evenly across all available local paths.

When using clustered Data ONTAP, the Data ONTAP DSM 3.5 and later adds the ability to prioritize FC paths over iSCSI paths using the iSCSILeastPreferred registry value. The iSCSILeastPreferred parameter specifies whether the Data ONTAP DSM prioritizes FC paths over iSCSI paths to the same LUN. You might enable this setting if you want to use iSCSI paths as backups to FC paths.

By default, the Data ONTAP DSM uses ALUA access states to prioritize paths. It does not prioritize by protocol. If you enable this setting, the DSM prioritizes by ALUA state and protocol, with FC paths receiving priority over iSCSI paths. The DSM uses iSCSI optimized paths only if there are no FC optimized paths available.

This setting applies to LUNs that have a load balance policy of either least queue depth or round robin with subset.

Note: iSCSI connections in 7-Mode do not support ALUA and therefore cannot be used in a mixed protocol igroup with FC connections. For this reason, the iSCSILeastPreferred setting does not apply to 7-Mode environments.

For more information, see the Data ONTAP DSM for Windows MPIO Installation and Administration Guide.

7 iSCSI Network Design Recommendations

7.1 iSCSI Network Topologies

The iSCSI protocol is defined by RFC 3270, published by the Internet Engineering Task Force. A copy of the standard can be obtained from www.ietf.org/rfc/rfc3720.txt.

The first decision a customer needs to make is whether to run iSCSI traffic over a physically separate dedicated network. A dedicated iSCSI Ethernet infrastructure can include its own switches or VLANs. For smaller configurations, hosts can connect directly to single node storage systems using crossover cables.

Note: NetApp recommends that if using multiple paths or sessions with iSCSI, each path should be isolated on its own subnet.

On the storage system, the iSCSI service should be disabled on network interfaces that will not be used for iSCSI sessions. Once disabled, the service rejects subsequent attempts to establish new iSCSI sessions over that interface. This increases security by only allowing iSCSI connections on predetermined ports.

Starting in Data ONTAP 7.3, iSCSI access lists have been implemented to give more granular control. Access can be granted to specific initiators over specific storage interfaces.

NetApp recommends a network topology that minimizes the risk of unauthorized access to or modification of data as it traverses the network. You can limit access to data through the use of direct cabling, switched network environments, virtual LANs (VLANs), and dedicated storage network interfaces where appropriate.

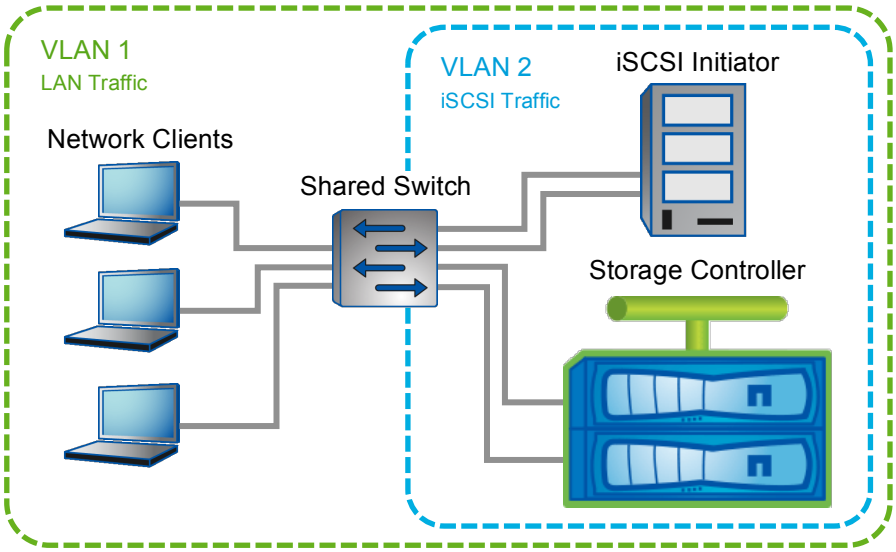
Three topologies can be used to design iSCSI networks. Each has advantages and disadvantages.

7.2 Shared Switched iSCSI Network

Shared configurations run both iSCSI and other Ethernet traffic over the same physical network. Because it is shared with other traffic and hosts, this option is less secure than a dedicated network; you should implement available security features to reduce exposure.

NetApp recommends that VLANs be used to segregate iSCSI from other network traffic in shared configurations. The VLAN provides some additional security and simplifies network troubleshooting. A NetApp storage system can be configured as a VLAN-aware device that processes VLAN tags, or the VLAN can be managed at the switch port level and be transparent to the storage system.

Figure 6) iSCSI topology with a shared switch.



Advantages:

- Link aggregation is possible if supported by the switch.
- Multiple switches can be used for redundancy.
- The number of hosts and storage systems is limited only by the available switch ports.
- The existing Ethernet switch infrastructure is utilized, saving money.
- Each LUN can have its own load balance policy.

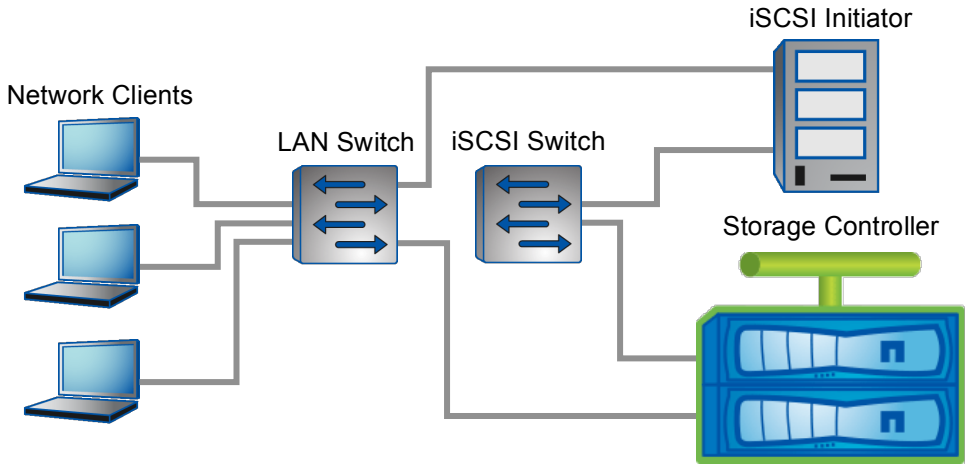
Disadvantages:

- Network bandwidth is shared across public LAN and iSCSI traffic unless initiators and targets are connected to the same switch.
- It requires switches capable of implementing VLANs.

7.3 Dedicated Switched iSCSI Network

In this configuration, Ethernet switches and cables are dedicated to carrying iSCSI traffic between iSCSI hosts and storage systems. This configuration is very similar to a Fibre Channel fabric in that only iSCSI and related traffic uses this dedicated infrastructure. There are additional costs for dedicated Ethernet equipment compared to running iSCSI traffic over the existing Ethernet infrastructure, but you gain security and performance improvements.

Figure 7) iSCSI topology with a dedicated switch.



Advantages:

- Very secure: iSCSI traffic is isolated from public LAN traffic.
- The full bandwidth of the link is available.
- Link aggregation is possible if supported by the switch.
- Multiple switches can be used for redundancy.
- The number of hosts and storage systems is limited only by available switch ports.
- You can use less expensive, unmanaged switches because VLANs are not needed.

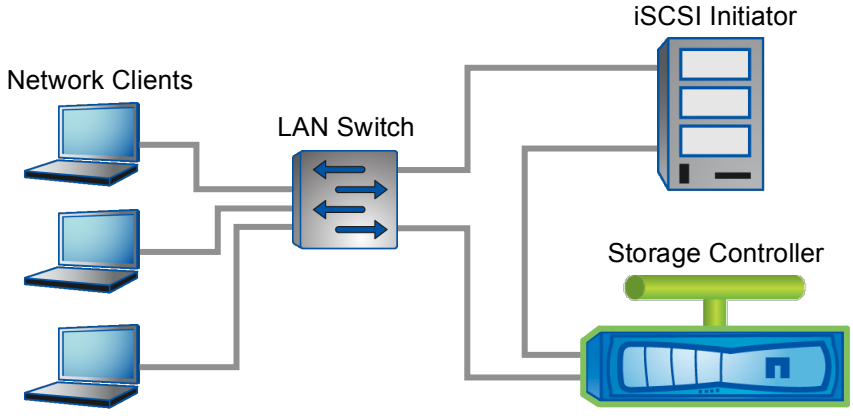
Disadvantages:

- One or more switches must be dedicated to the iSCSI network.
- Configuration and administration are more complex than direct connection.

7.4 Directly Connected iSCSI Network

The host is connected to the storage system using crossover cables. No Ethernet switches are involved. This is the most physically secure configuration and allows full bandwidth between the initiator and target.

Figure 8) Direct connect iSCSI.



Advantages:

- Low cost: no Ethernet switches required.
- Very secure: no chance of man-in-the-middle attack.
- Easy to set up and maintain.
- Full bandwidth of link is available.

Disadvantages:

- The number of initiators and/or paths is limited by the number of available network ports.
- Limited distance between initiator and target.
- Not supported with storage HA failover.

7.5 Using Jumbo Frames

By default, Ethernet sends up to 1,500 bytes of data in a single frame. This works well for applications that send small amounts of data, such as client applications. However, for transferring larger blocks of data, as is common in iSCSI, a larger frame size is more efficient. Because the size of the network headers is fixed, an increased frame size increases the ratio of SCSI payload to network overhead, allowing more data to be transmitted per request. This increases overall throughput and reduces the amount of network overhead that must be handled by the NIC and/or the CPU.

The term “jumbo frame” typically refers to Ethernet frames with bytes of data, although it technically applies to any size larger than 1,500 bytes. Unlike the standard frame size, there is no standard size for a jumbo frame. Each network device must typically be configured with the specific maximum transmission unit size that will be used. Therefore, each network device must support the same size for jumbo frames. NetApp storage systems support jumbo frames on all 1 and 10 Gigabit Ethernet interfaces.

7.6 Using Flow Control

For 1 Gigabit Ethernet connections, NetApp recommends setting flow control on network switch ports to “full” and all iSCSI targets and initiators to “send.” For 10 Gigabit Ethernet connections, NetApp recommends disabling flow control on all iSCSI targets, initiators, and network switch ports. For more information on flow control, see the Data ONTAP Network Management Guide.

7.7 NetApp Host Utilities

NetApp provides a SAN Host Utilities kit for every supported OS. This is a set of data collection applications and configuration scripts. These include SCSI and path timeout values and path retry counts. Also included are tools to improve the supportability of the host in a NetApp SAN environment, such as gathering host configuration and logs and viewing the details of all NetApp presented LUNs.

Note: When using the Data ONTAP DSM 3.5 and later, the Host Utilities kit is not required to set timeout values. The Data ONTAP DSM makes the same changes to the host configuration, including path timeout values.

8 Fibre Channel Fabric Design Recommendations

Of the multipathing topics we discussed, only MPIO is applicable to Fibre Channel. Three general topologies are available in a Fibre Channel environment:

- Direct attached: The initiator and 7-Mode target are connected directly by a cable.

  Note: Due to the use of NPIV and the architecture of clustered Data ONTAP, direct attached Fibre Channel connections are not supported with clustered Data ONTAP.

- Single fabric: All ports of the initiator and target connect to a single switch or fabric of switches.
- Multifabric: Some ports of the initiator and/or target connect to separate fabrics for redundancy.

These configurations are detailed in the “Data ONTAP SAN Configuration Guide” and include diagrams and supported topologies for different NetApp platforms.

As discussed before, NetApp recommends that any SAN solution use redundant components to reduce or eliminate single points of failure. For Fibre Channel this means utilizing multiple HBAs, switches/fabrics, and storage clustering.

References

The following references were used in this TR:

- Clustered Data ONTAP SAN Configuration Guide
  http://support.netapp.com/documentation/productlibrary/index.html?productID=30092
- Data ONTAP SAN Configuration Guide for 7-Mode
  http://support.netapp.com/documentation/productlibrary/index.html?productID=30092
- NetApp Interoperability Matrix Tool
  http://support.netapp.com/matrix/mtx/login.do

Version History

Version	Author(s)	Date	Document Version History
Version 3.0	Ryan Hardin	May 2013	Updates include clustered Data ONTAP, Windows Server 2012, Symantec Veritas DMP, multipathing recommendations, and new graphics and diagrams.
Version 2.1	Patrick Strick Richard Jooss	July 2010	

Refer to the Interoperability Matrix Tool (IMT) on the NetApp Support site to validate that the exact product and feature versions described in this document are supported for your specific environment. The NetApp IMT defines the product components and versions that can be used to construct configurations that are supported by NetApp. Specific results depend on each customer's installation in accordance with published specifications.

NetApp provides no representations or warranties regarding the accuracy, reliability, or serviceability of any information or recommendations provided in this publication, or with respect to any results that may be obtained by the use of the information or observance of any recommendations provided herein. The information in this document is distributed AS IS, and the use of this information or the implementation of any recommendations or techniques herein is a customer's responsibility and depends on the customer's ability to evaluate and integrate them into the customer's operational environment. This document and the information contained herein may be used solely in connection with the NetApp products discussed in this document.

Go further, faster®

www.netapp.com