IP SAN BEST PRACTICES

PowerVault MD3000i Storage Array

www.dell.com/MD3000i

TABLE OF CONTENTS

Table of Contents

INTRODUCTION .... ERROR! BOOKMARK NOT DEFINED.
OVERVIEW ISCSI .... ERROR! BOOKMARK NOT DEFINED.
IP SAN DESIGN 4
BEST PRACTICE - IMPLEMENTATION 5
REDUNDANCY 5
SECURITY 6
IP SAN NETWORK CONFIGURATION 7
IP SAN OPTIMIZATION 8
SUMMARY 11

INTRODUCTION

The intent of this document is to provide guidance for optimizing an IP SAN environment utilizing the Dell MD3000i storage array. The best practices within this document are recommendations to provide a fault tolerant, high performance environment to maximize the capabilities of an MD3000i SAN. The recommendations may be applied according to the requirements of the environment in which the installed storage array or arrays are utilized, and not all best practices may be applicable to all installations. The best practices in this paper are focused on Dell Inc. technology based solutions.

iSCSI OVERVIEW

iSCSI is a block-level storage protocol that lets users create a storage network using Ethernet. iSCSI uses Ethernet as a transport for data from servers to storage devices or storage-area networks. Because iSCSI uses Ethernet, it doesn't suffer from some of the complexity and distance limitations that encumber other storage protocols.

The iSCSI protocol puts standard SCSI commands into TCP and sends those SCSI commands over standard Ethernet. An iSCSI SAN consists of servers - with an iSCSI host bus adapter (HBA) or network interface card (NIC) - disk arrays and tape libraries. Unlike other SAN technologies, iSCSI uses standard Ethernet switches, routers and cables, and the same Ethernet protocol deployed for communications traffic on LANs (TCP/IP). It can take advantage of the same type of switching, routing and cabling technology used for a LAN.

Because iSCSI uses SCSI commands, relying on Ethernet only to transport the SCSI commands, operating systems see iSCSI-connected devices as SCSI devices and are largely unaware that the SCSI device resides across the room or across town.

Most components inside these iSCSI devices are very familiar to network professionals, including RAID controllers and SCSI or Fibre Channel drives. The only added feature is the iSCSI protocol, which can be run on standard NICs in software or on specialized iSCSI silicon or HBAs that off-load the TCP/IP and iSCSI protocol.

ISCSI is built using two of the most widely adopted protocols for storage (SCSI) and networking (TCP). Both technologies have undergone years of research, development and integration. IP networks also provide the utmost in manageability, interoperability and cost effectiveness.

IP SAN DESIGN

For an IP SAN, the network infrastructure consists of one or more network switches or equivalent network equipment (routers, switches, etc.). For the purpose of this paper, it is assumed that the network has at least one switching or routing device. While it is possible to connect an MD3000i array to hosts without utilizing a network, directly connecting hosts to arrays is not within the scope of this paper. An IP SAN therefore consists of one or more hosts, connected to one or more storage arrays through an IP network, utilizing at least one switch in the network infrastructure.

There are several factors that need to be kept in mind when designing an IP SAN. The importance of these factors will depend on the specific implementation of the IP SAN. These factors include and are not limited to:

1. Redundancy: If data availability is required at all times, a fault tolerant IP SAN should be considered.
2. Security: Depending on your IP-SAN implementation, different security mechanisms can be taken into consideration. This includes dedicated networks, CHAP, array passwords, etc.
3. Network Infrastructure: Components of the network infrastructure like NICs, HBAs, switches, cabling, routing, etc. can affect IP SAN performance and maintainability.
4. Optimization: Depending on the application, various elements of your IP SAN can be tuned for improved performance. Some of these include the ability to use hardware offload engines, jumbo frames, etc.

BEST PRACTICE - IMPLEMENATATION

There are many ways to implement an IP SAN based on need, available resources and intended application. For instance one important but easily overlooked item that can improve the manageability of your IP SAN implementation is to assign a consistent and representative naming scheme to the storage arrays. This is especially useful if the SAN has more than one storage array attached. The "blink array" feature of the MD Storage Manager can be used to correctly identify each array physically.

Some of the general implementation guidelines will be described below. However, one should note that these are general guidelines and may not benefit all applications.

Redundancy

Redundancy in general is having a second set of hardware and communication paths so that if one piece of hardware on one path breaks down there is a second path that can be utilized. In an IP SAN this can be done with a second controller in the Array and by using two different switches for the iSCSI network. The diagram below is a simplified diagram of doing this with a Dell PowerVault MD3000i; the desciptions following provide further discussion of the benefits in doing this.

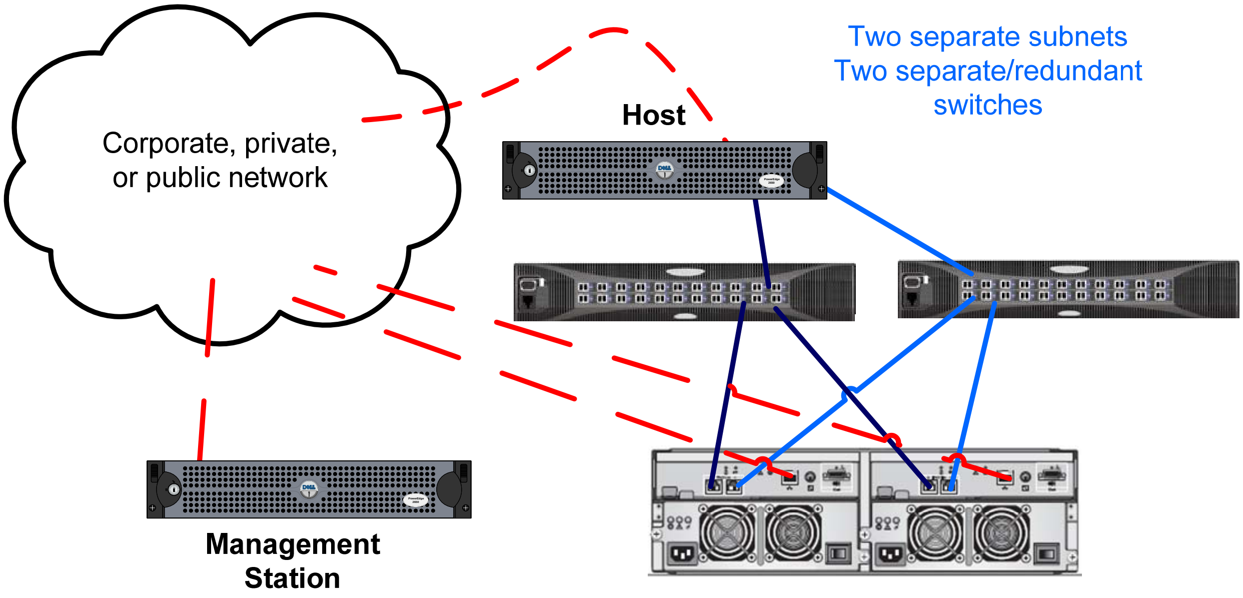
Two separate subnets
Two separate/redundant switches
Host
Corporate, private, or public network
Management Station

Diagram 1: Fully Redundant MD3000i Config

Physical Network Infrastructure: A fully redundant IP SAN is characterized by multiple physical independent iSCSI data paths between the hosts and the array. Each data path should be on a separate subnet.

iSCSI configuration: In the case of an iSCSI target like the MD3000i, it is recommended to establish multiple sessions to the storage subsystem from each host. It is recommended to set up one session per port from each of the network cards to each RAID controller module. This method allows one session to restart if a link goes down while not affecting any of the other links.

RAID: An appropriate RAID level should be chosen based on your applications. RAID 1 or higher provide some level of redundancy that will be useful in the case of failed physical disks. Each RAID level works best with certain applications and this should be taken into consideration while configuring the MD3000i.

Power: Each redundant component of the data path should be on a separate power source. This will ensure that even if one component fails due to a power issue, the alternative path will continue to work. In the same way two power supplies of the MD3000i should be connected to separate power sources.

Security

The optimal way of ensuring data security on an IP SAN is by implementing an isolated physically independent network for the iSCSI data traffic. Besides better security, another benefit of having an isolated network dedicated for storage traffic is the avoidance of networked traffic congestion with other non-storage traffic using the network.

VLAN: If physically isolated iSCSI networks are not feasible then VLANs can help to seperate iSCSI traffic from the general traffic in the network. It is recommended you turn on VLAN Tagging. The MD3000i array supports VLAN tagging. A port can either transmit all tagged IP packets or all non-tagged IP packets.

Note: VLAN must be enabled throughout the entire iSCSI SAN from the NICS, Switches, and iSCSI ports, otherwise, behavior may be inconsistent. To simplify troubleshooting initial deployments, make sure that NICs, switches, and MD3000i storage arrays are fully operational before enabling the VLAN feature solution wide.

CHAP: To have secure access between your host and array, target and mutual CHAP authentication should be enabled on the host(s) and storage array(s). Standard CHAP password guidelines should be followed for best security.

It is highly recommended to set a password on all devices with your IP SAN. It is advisable to use a strong password that meets standard IT guidelines.

IP SAN Network Infrastructure

Some of the general implementation guidelines will be described below. However, one should note that these are general guidelines and may not benefit some applications.

General Network Practices: Make sure the category rating for the cables used are gigabit Ethernet compliant. (CAT5e, CAT6) Design your network to have the least amount of hops between the array(s) and the host(s). This will greatly reduce your failure points, simplify your manageability, and reduce latency and complexity of your network architecture (particularly in the area of redundancy). Managed switches are recommended because they provide advance features to help you optimize and maintain your network for your application. It is recommend you use auto-negotiation only, since gigabit ethernet networks are designed to always have auto-negotiation enabled. If a particular application requires a specific speed/duplex mode, this must be done by changing the advertisement options of the switch.

Spanning-Tree Protocol: It is recommended that you disable spanning-tree protocol (STP) on the switch ports that connect end nodes (iSCSI initiators and storage array network interfaces). If you still decide to enable STP on those switch ports, then you should turn on the STP FastPort feature on the ports in order to allow immediate transition of the ports into forwarding state. (Note: PortFast immediately transitions the port into STP forwarding mode upon linkup. The port still participates in STP. So if the port is to be a part of the loop, the port eventually transitions into STP blocking mode.)

Note: PowerConnect Switches default to RSTP (Rapid Spanning Tree Protocol) an evolution in STP that provides for faster Spanning tree convergance and is preferable to STP

> **Note**: The use of Spanning-Tree for a single-cable connection between switches or the use of trunking for multiple-cable connections between switches is encouraged.

TCP Congestion avoidance: TCP Congestion Avoidance is an end to end flow control protocol that will limit the amount of data sent between a TCP sender and a TCP transmitter. This protocol uses a sliding window to size the data being sent to the TCP reciever. This protocol starts with a small segment size and keeps increasing with each acked segment sent, until a segment is dropped. Once it is dropped TCP starts this over again.

Ether Flow Control: Dell recommends that you enable Flow Control on the switch ports that handle iSCSI traffic. In addition, if a server is using a software iSCSI initiator and NIC combination to handle iSCSI traffic, you must also enable Flow Control on the NICs to obtain the performance benefit. On many networks, there can be an imbalance in the network traffic between the devices that send network traffic and the devices that receive the traffic. This is often the case in SAN configurations in which many hosts (initiators) are communicating with storage devices. If senders transmit data simultaneously, they may exceed the throughput capacity of the receiver. When this occurs, the receiver may drop packets, forcing senders to retransmit the data after a delay. Although this will not result in any loss of data, latency will increase because of the retransmissions, and I/O performance will degrade.

Note: PowerConnect Switches default to Flow Control being off.
The MD3000i will autoconfigure to the switch when Flow control is turned on.

Unicast Storm Control: A traffic "storm" occurs when a large outpouring of packets creates excessive network traffic that degrades network performance. Many switches have traffic storm control features that prevent ports from being disrupted by broadcast, multicast, or unicast traffic storms on physical interfaces. These features typically work by discarding network packets when the traffic on an interface reaches a percentage of the overall load (usually 80 percent, by default).

Because iSCSI traffic is unicast traffic and can typically utilize the entire link, it is recommended that you disable unicast storm control on switches that handle iSCSI traffic. However, the use of broadcast and multicast storm control is encouraged. See your switch documentation for information on disabling unicast storm control

Jumbo Frames: Dell recommends that you enable Jumbo Frames on the switch ports that handle iSCSI traffic. In addition, if a host is using a software iSCSI initiator and NIC combination to handle iSCSI traffic, you must also enable Jumbo Frames on the NICs to obtain the performance benefit (or reduced CPU overhead) and ensure consistent behavior.

Note: Jumbo Frames must be enabled throughout the entire iSCSI SAN from the NICS, Switches, and array ports, otherwise, behavior may be inconsistent. To simplify troubleshooting initial deployments, make sure that NICs, switches, and MD3000i storage arrays are fully operational before enabling jumbo frames.

IP SAN Optimization

When designing your IP SAN you have to look at various factors in your network and the actual application you are using. There are some general rules that can be used when designing your IP SAN. In order to maximize the data throughput of your storage arrays, all data ports need to be utilized. If your application is IO intensive, utilizing an iSCSI offload NICs is recommended. Consider manually balancing your virtual disk ownership so that no single controller is processing an excessive amount of I/O relative to the other controller.

The MD3000i supports active/active controllers, with each controller being able to simultaneous process IO. The asymmetric design of the controllers means that a virtual disk (LUN) is owned by a controller and all IO access to the virtual disk is only possible through the owning controller. To take advantage of both the controllers for IO access, virtual disks can be distributed among the controllers. Virtual disk ownership can be modified to balance IO access so as to balance utilization of both controllers. With a host configured for redundant access, if a host loses IO access to a virtual disk through its owning controller, the failover drive will execute ownership transfer from one controller to the other and resume IO access through the new owning controller.

The following figure illustrates the active/active asymmetric architecture of the MD3000i. The configuration consists of two virtual disks (Virtual Disk 0 and Virtual Disk 1), with Virtual Disk 0 owned by Controller 0 and Virtual Disk 1 owned by Controller 1. Virtual Disk 0 is assigned to Host 1 and Virtual Disk 1 assigned to Host 2.

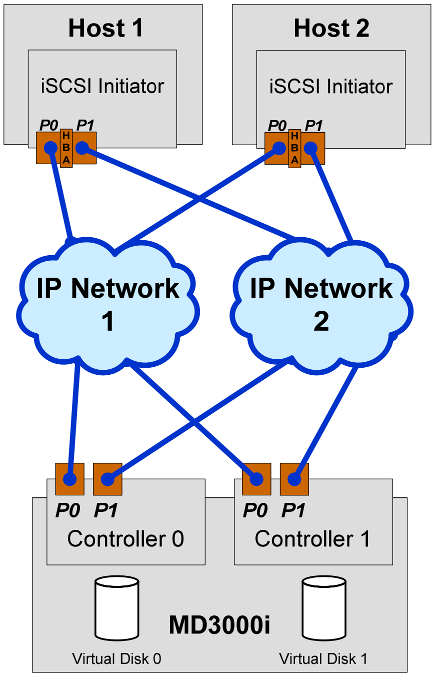

Diagram 2: MD3000i Controller Configuration

Virtual disk ownership defined by the asymmetric architecture ensures that Host 1 accesses Virtual Disk 0 through Controller 0 and Host 2 accesses Virtual Disk 1 through Controller 1.

Bandwidth Aggregation: With the MD3000i you can have two Ethernet ports from one host connected to one controller and the badwidth will be aggegated. Set up the MD3000i iSCSI driver with a Round Robin Que, this will aggregate all the packets being sent to that controller placing them on each link therefore doubling the available bandwidth.

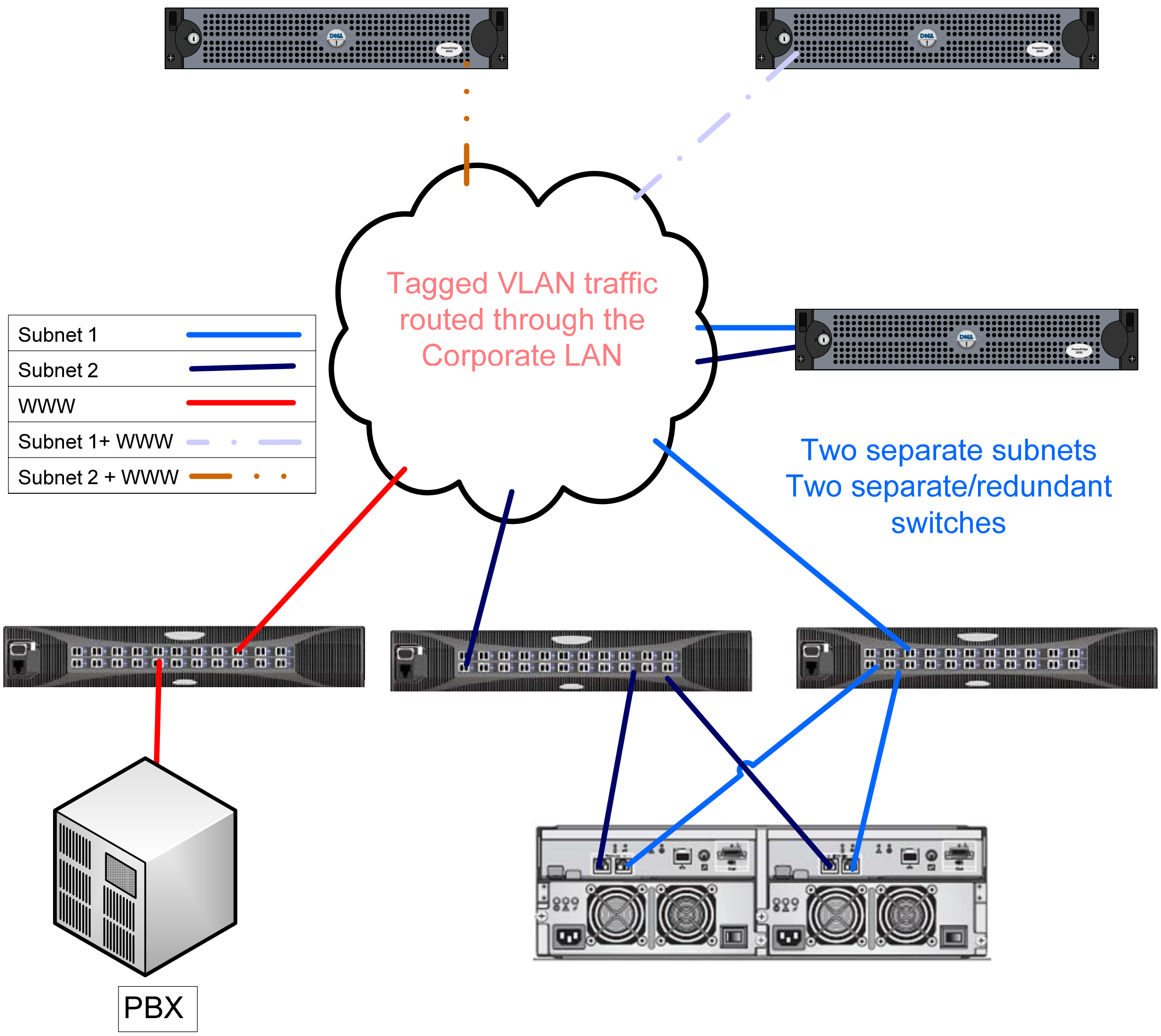
Subnet 1
Subnet 2
WWW
Subnet 1+ WWW
Subnet 2 + WWW
Tagged VLAN traffic routed through the Corporate LAN
Two separate subnets Two separate/redundant switches
PBX

Diagram 3: MD3000i in a Network

Examine your network architecture to ensure there is no bottleneck in the network between the host and array. There are some things we talked about under security that also will help to optimize your IP SAN. Using separate switches to physically isolate the iSCSI data traffic, and using VLANs with FastPort turned on.

Layer 2 Optimization: When setting up the VLAN through your nework, VLAN tagging can be helpful in routing the iSCSI Data Traffic through your network. You can then set priority within the VLAN, but you have to look at all your traffic to determine priorities. If for example your VOIP traffic runs through the same VLAN you need to ensure that voice quality is not hurt, plus you need to look at general internet traffic versus iSCSI and VOIP.

Layer 3 Optimization: Differentiated Services (DiffServ) gives a good method for managing your traffic. Some switches have a proprietary implementation of this that is called Quality of

Service (QoS). DiffServ uses the Differentiated Services Code Point (DSCP) to distinguish between service levels of each IP connection. These service level agreements are on a Per Hop Basis (PHB), as such within the internal corporate network traffic flows can be predictable but once a WAN link leaves the company the Service agreements are no longer valid. There are four levels normally used with DiffServ.

1. Default PHB—which is typically best-effort traffic
2. *Expedited Forwarding* (EF) PHB—for low-loss, low-latency traffic
3. *Assured Forwarding* (AF)—behavior group
4. *Class Selector* PHBs - which are defined to maintain backward compatibility with the IP Precedence field.

In order to choose what service level to use you have to examine the needs of the applications connected to the Array. For instance if you have your hosts set up to iSCSI boot, or are using Virtualization to “hide” the array and the guest OS is booting off a C: drive that is actually on the array you must select **EF** as the data must get there and if there is much delay the host will lock up. On the other land you may want all your traffic coming in from the WWW set to the lowest possible class of **AF** so it doesn’t affect your critical data.

SUMMARY

An IP SAN is a flexible, easy to deploy and use storage solution for businesses of all sizes. By following the practices recommended in this whitepaper and using regular IT best practices you can have a highly reliable, flexible data storage solution. Remember it is important to design and build out your corporate network with the IP SAN in mind, as your data needs grow so will your data traffic. By following the recommendations in this white paper you will be in a much better position to deal with those changes.