

Technology Case Study in Storage Area Networks

A Master's Project

Presented to

Department of Computer and Information Sciences

In Partial Fulfillment

of the Requirements for the

Master of Science Degree

State University of New York

Institute of Technology

By

Ameya Pethe

May 2014

Technology Case Study in Storage area networks

Declaration

I declare that this project is my own work and has not been submitted in any form for another degree or diploma at any university or other institute of tertiary education. Information derived from the published and unpublished work of others has been acknowledged in the text and a list of references is given.

Ameya Pethe

05/15/2014

SUNYIT

**DEPARTMENT OF
COMPUTER AND INFORMATION SCIENCES**

Approved and recommended for acceptance as a thesis in partial fulfillment of the requirements for the degree of Master of Science in Telecommunications.

DATE

Dr. John Marsh

Thesis Advisor

Dr. Larry J. Hash

Mr. David Climek

Mr. Ronny Bull

Contents

List of figures	3
Project statement	1
1. Introduction	2
2. Background.....	4
3. Network components.....	8
4. Protocols.....	11
4.1 Fiber channel	12
4.1.1 Port types	12
4.1.2 FC topologies.....	13
4.1.2.2 Zoning.....	18
4.1.3 FC switches	20
4.1.3.1 Brocade switches	20
4.1.3.2 Cisco switches	23
4.2 FCoE.....	25
5. Storage arrays	27
5.1 Disk types	28
5.2 RAID protection	30
5.3 EMC arrays.....	35
5.3.1 VNX series	35
5.3.2 Symmetrix series	38

5.3.3 Isilon	39
5.4 HP arrays	40
5.4.1 MSA series	40
5.4.2 3par	41
5.4.3 StoreEasy series.....	42
5.5 IBM arrays.....	43
5.5.1 Storwise series.....	43
5.5.2 DS series.....	44
5.5.3 N series	46
6. Case Study	47
6.1 Pre-sales architecture planning.....	47
6.1.1 Capacity planning.....	50
6.2 Comparison and benchmarking solutions	53
6.3 Rack and floor planning	54
6.4 Implementation.....	56
6.4.1 VMAX 20K/40K implementation.....	56
6.4.2 VPLEX implementation.....	57
6.4.3 Migration	57
6.5 Post implementation.....	61
7 Conclusion.....	62
8. References	65

List of figures

Figure 1- Directly attached storage	5
Figure 2 - Network attached storage	6
Figure 3 - Storage area network	7
Figure 4 - Basic architecture of SAN	9
Figure 5- Point to point topology	14
Figure 6 - Arbitrated loop	15
Figure 7- Arbitrated loop with switched fabric	16
Figure 8 - Dual fabric storage area network	17
Figure 9 - Switched fabric	18
Figure 10- Edge core topology	20
Figure 11 - Traditional data center layout	26
Figure 12 - FCoE environment	26
Figure 13 - RAID 0	32
Figure 14 - RAID 1	32
Figure 15 - RAID 4	33
Figure 16 - RAID 5	34

Figure 17 - VNX gateway	38
Figure 18 - Existing infrastructure	49
Figure 19 - Refreshed architecture	51
Figure 20 - VMAX rack	55
Figure 21 - Rack at DR site	55
Figure 22 - Pre migration	59
Figure 23 - During cutoff	59
Figure 24 - Migration using vMotion	60

Project statement

This project is intended to serve as an introduction of storage area networks and case study of data center refresh of medium-sized healthcare office with EMC hardware.

1. Introduction

In today's world we need immediate access to data. The demand for networked data access has increased exponentially in the last 20 years^[13]. With that demand the importance and volume of networked data has also grown exponentially. The speed at which the data can be accessed has increased and with that the data has moved from individual workstations to a networked location.

Over the last decade there has been a trend to move mission critical data away from individual workstations to a centralized data center. A centralized data center removes the location constraint for accessing the data. If critical data is stored on individual servers, a failure will cause the data to be inaccessible. Today, mission critical applications are spanned over multiple servers for redundancy. With this topology, having the data in a central location allows the individual servers to better work with data. With the addition of virtualization, servers can be moved online from one physical server to another. If the data is centralized, it can be presented to all hosts in the cluster. This allows servers to move efficiently between hosts without losing access to the critical data. Many businesses in various industries like finance, airline, hospital, research, etc. depend on the speed and secure availability of their centralized data to function efficiently.

Increased demand for secure and reliable centralized storage has lead to development of new storage technologies. Today nearly all large companies and many medium size companies use a dedicated storage

network to access their data. They also use dedicated storage arrays to store their data reliably and safely. The combination of the storage network and storage arrays is called storage area network or SAN. A major advantage of SAN is in its reliability by reducing single points of failures. SAN also separates the storage traffic from the application traffic. This architecture increases performance of the storage as well as the application. This case study details various technologies and products involved in SAN. I start with the basics of SAN: dual fabric architecture, fiber channel protocol, storage arrays, disk types and protection levels. The paper ends with a design case study for a medium scale data center refresh.

The main references for this case study were training documentation ^{[5][8][16][18][19]} and best practice material^[17] from vendor support sites. The training material goes in depth about the product features, capabilities and implementation methods. EMC training material referenced covers the basics of SAN, the path from direct attached storage to enterprise level storage area networks. Brocade and Cisco training material explains the workings of fiber channel protocol as well as the alternatives such as FCoE, FCIP and iSCSI.

The best practices documentation consists of white papers with case studies exploring the effects of particular SAN configuration on the fabric and array side on other systems in the environment. These documents ensure that while designing the system as a whole, the effect of small configuration changes remains predictable on the entire environment. Referring those case studies, I designed my case study.

2. Background

In the late 90's storage systems evolved from server centric to information centric. Server centric architecture used was called direct attached storage or DAS. Directly attached storage evolved into network attached storage or NAS. The shortcomings of network attached storage were overcome in storage area networks or SAN.

Directly attached storage was the predominant storage system for all organizations until storage systems reduced in cost. Small organizations still use directly attached storage. As the name suggests directly attached storage consists of a bunch of disks connected to a server. This server runs the networked application. The advantage of directly attached storage is its low latency. Since there are no components between the server and the disks, there is negligible seek time for data. A major disadvantage of directly attached storage was that the storage capacity was not shared. This meant that a lot of storage space was wasted. Also time to provision new storage for growing applications was large. Directly attached storage also had multiple points of failure. If the server fails, access to the data stored in that server is lost. Figure 1 shows an example of directly attached storage. Each application server has its own storage system, which is not shared among the servers.

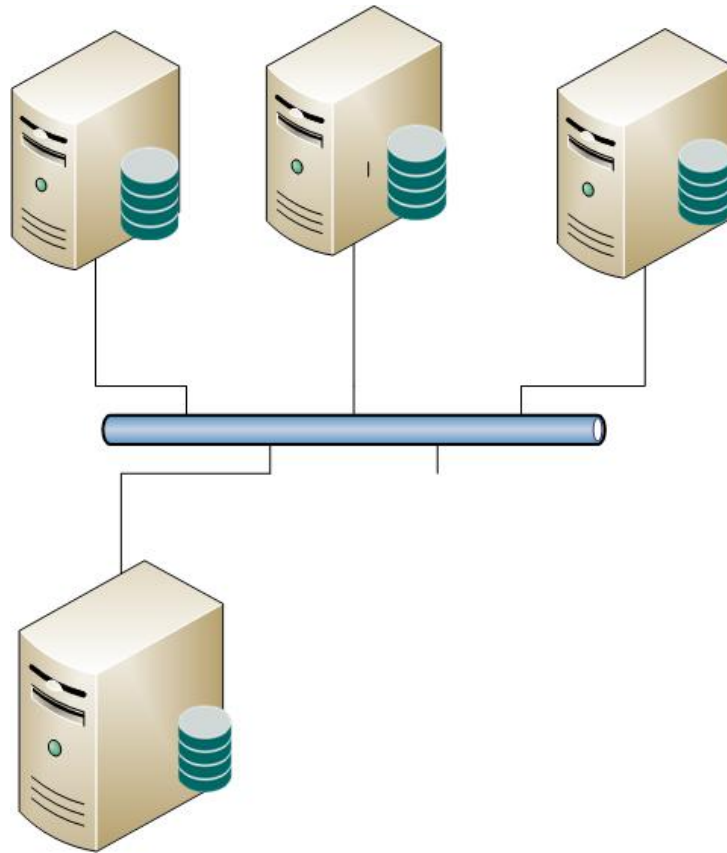


Figure 1- Directly attached storage

To overcome the wastage of valuable disk space and points of failure in DAS, network attached storage was introduced. Network attached storage is information centric. The application server and the storage system are two different entities. In NAS, the application server and the storage system are connected by the local area network. This means that all the devices on the network can access storage space on NAS. Network attached storage provides file level access to the networked devices over IP. The storage system in NAS usually contains its own operating system and can provide storage space to servers using three widely used protocols. Common internet file system (CIFS) which is used by Windows devices, network file system (NFS) which is used by Unix devices and file transfer protocol (FTP) which is used by Windows as well as Unix. The evolution of NAS enabled storage administrators to share the available disk space between different servers and utilize all available disk space. It also introduced low provision times and eliminated most points of failures. The major disadvantage of NAS is that it uses the

production network for storage access. This adds load to the network and may affect performance. Figure 2 shows an example of network attached storage. The storage system is shared among all application servers.

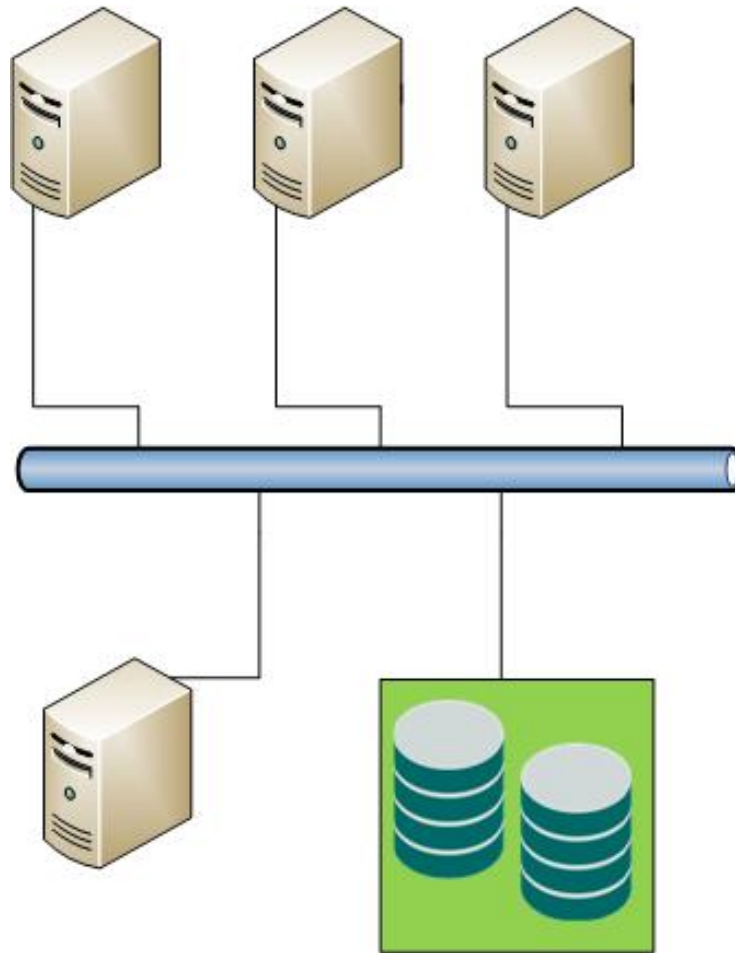


Figure 2 - Network attached storage

Storage area networks are an evolution of NAS where the storage network is independent of the production network. A disadvantage of network attached storage was that it provides file level access to the application servers. With the growing use of virtualization, the demand for block level access was needed. Another disadvantage of NAS is that it is bound to the physical location of the business in its use of the production network. Both these limitations were resolved by storage area networks. In SAN, the storage network is an independent fiber channel or SCSI based network. This allowed businesses to move

their storage off site. Access to the storage networks and the data within it is provided by gateways.

Figure 3 shows a typical storage area network.

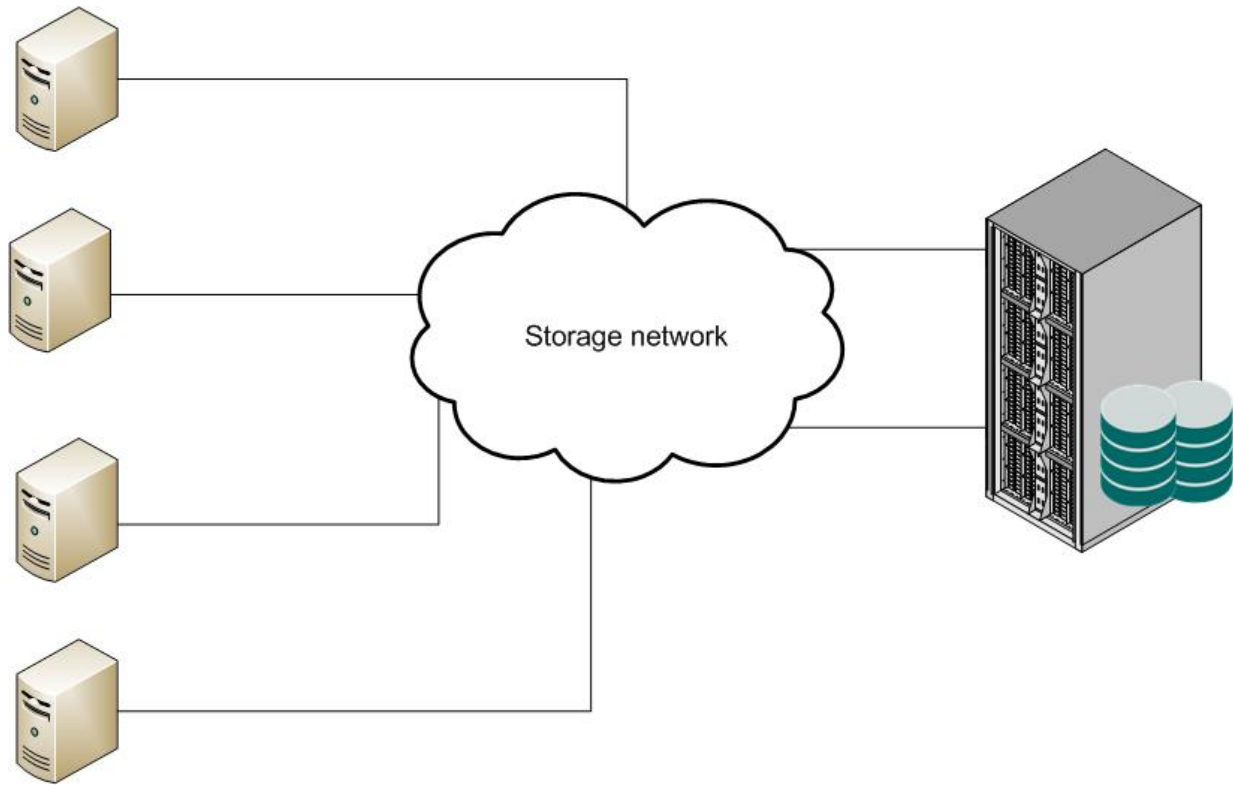


Figure 3 - Storage area network

Within the storage network, there are switches and storage arrays. The components are broadly classified into network components and storage arrays. At each level in the storage network there is redundancy. The design of a storage network is made such that the failure of one component will not result in inaccessible data. Within the storage array itself there is redundancy at the port, cache, and disk level. The chance of the entire network failing is very remote.

3. Network components

Network components in SAN are the switches and the protocols used within the network. Before investigating switches and protocols in depth, we review 2 important concepts - zoning and fabric. In storage networks the server requesting data from the storage system is known as the initiator and the storage array port on which the initiator is connected is called as the target. Figure 4 shows the basic architecture of a storage area network.

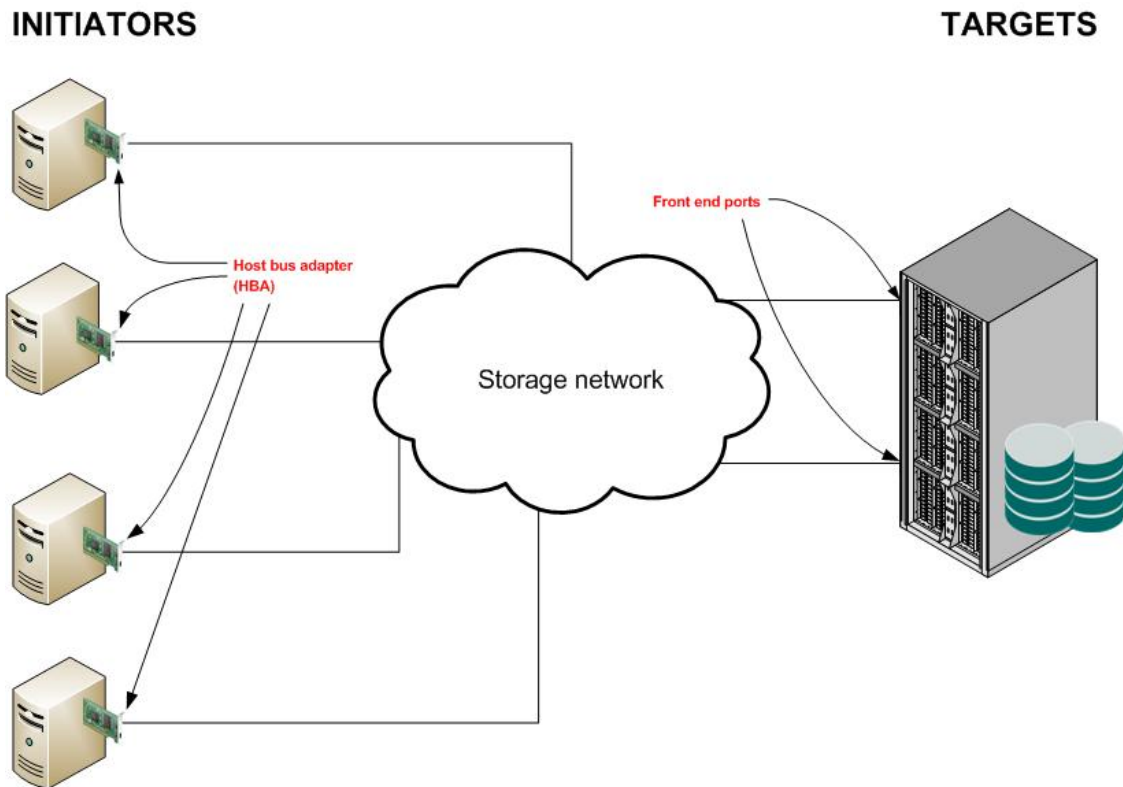


Figure 4 - Basic architecture of SAN

The network connectivity devices on server side are NIC, HBA or CNA. A network interface card (NIC) is used to connect the server to IP based networks. Host bus adapter (HBA) is used to connect the server to fiber channel networks. HBA is used to connect external storage using SCSI protocol with fiber channel as the propagation medium. Converged network adapter (CNA) has functionality of a HBA as well as a NIC. On the storage system side, the storage array has front end connectivity ports for SAN. They can be FC, Ethernet or iSCSI ports. FC ports are used to connect to the fiber channel network. Ethernet port is used for management of the storage array. iSCSI ports are Ethernet ports used for SCSI encapsulation over Ethernet. Newer low range and mid range arrays also support FCoE on their Ethernet ports.

In storage area networks each node in the network has a unique identifier. The identifier is known as world wide name (WWN). Like MAC address in IP networks, WWN is unique to a particular piece hardware. Each component in SAN has 2 world wide names. The entire node has a node world wide name (nWWN) and each port in that node has a port WWN (pWWN).

World wide name is an eight octet identifier. The first nibble is assigned by IEEE. The third, fourth and fifth octet is called organizationally unique identifier (OUI). The OUI is assigned by IEEE and identifies the manufacturer of the device.

4. Protocols

Protocols used in storage area network connectivity are fiber channel (FC), fiber channel over Ethernet (FCoE) and fiber channel over IP (FCIP). Fiber channel is used widely to connect different nodes within the data center. FCoE is replacing fiber channel in a lot of environments. The ease of using FCoE along with the existing ethernet network makes it attractive for smaller environments. FCIP is used to connect two distant datacenters where connection by FC over dark fiber is not feasible.

4.1 Fiber channel

Fiber channel (FC) is a networking protocol using optical fibers as propagation medium. It is a high speed networking protocol running at 4,8,10 or 16 Gbps. FC is the most widely used protocol in storage networks used along with SCSI as upper layer protocol. Fiber channel protocol was standardized as ANSI X3.230 in 1994.

Like OSI model for TCP/IP, fiber channel has its own 5 layer model. The layers are named FC0 to FC4. FC0 is the physical layer. It defines the specifications for cables and connectors. FC1 is the data link layer and contains the coding definitions for signals. The most common encoding format is 8B/10B encoding. FC2 is the network layer. It contains the protocols used in fiber channel networking. This layer defines zoning, fabric and RSCN protocol. FC3 defines the common services for the fiber channel network. This layer handles services like encryption and RAID algorithms. FC4 defines the upper layer protocols and their mapping to fiber channel networks. Mapping or encapsulations of upper layer protocols like SCSI are performed by this layer.

4.1.1 Port types

Fiber channel ports identify themselves on the fiber channel network according to their topologies and the role the node plays. Fiber channel ports are widely divided into F (fabric), N (node) and E (expansion) ports.

F port	Fabric port. F port is found on switches. F port identifies that the other end of the link is a node port.
FL port	Fabric loop port. FL port is found on switches. FL port identifies that the other end of the link is connected to node ports in arbitrated loop configuration.
N port	Node port. N port is found on hosts and storage devices in the fabric.
NL port	Node loop port. NL port is found on hosts and storage devices in the fabric. It identifies that the host or storage is connected using arbitrated loop topology.
E port	Expansion port. E port is found on switches. It identifies that the link connects two switches within the fabric with each other.
EX port	Expansion port. EX port is found on switches. It identifies that the link connects switches between two different fabrics. The fabrics do not merge and LSAN zoning is required for communication. When the port on the switch is EX port, the other end has to be a E port.
VE port	Virtual expansion port. VE ports are found in switches with FCIP compatibility. It acts as E port to connect switches within the fabric, but the link used is IP.
VEX port	Virtual expansion port. VEX ports are found in switches with FCIP compatibility. It acts as EX port to connect switches between two fabrics. The link between the ports is IP link. LSAN zoning is required for communication.

Table 1 - Port types

4.1.2 FC topologies

In fiber channel networks there are three possible topologies: Point to point, arbitrated loop and switched fabric. As the name suggests, in point to point topology the target and the initiator are connected with a single link between them. There are no switches in between them. Point to point topology can only support 2 nodes. Figure 5 shows an example of point of point to point topology. In the example a host is directly connected to a storage array. The host acts as initiator and the storage array acts as target. In a

similar way, two servers or two arrays can also be connected to each other. Both the ports act as N ports in this topology.



Figure 5- Point to point topology

Arbitrated loop topology connects the fiber channel initiators or targets in a topology similar to ring topology. In FC-AL topology, only two devices communicate with each other at one time. The maximum number of nodes allowed in the FC-AL is 127. Figure 6 shows an example of arbitrated loop architecture.

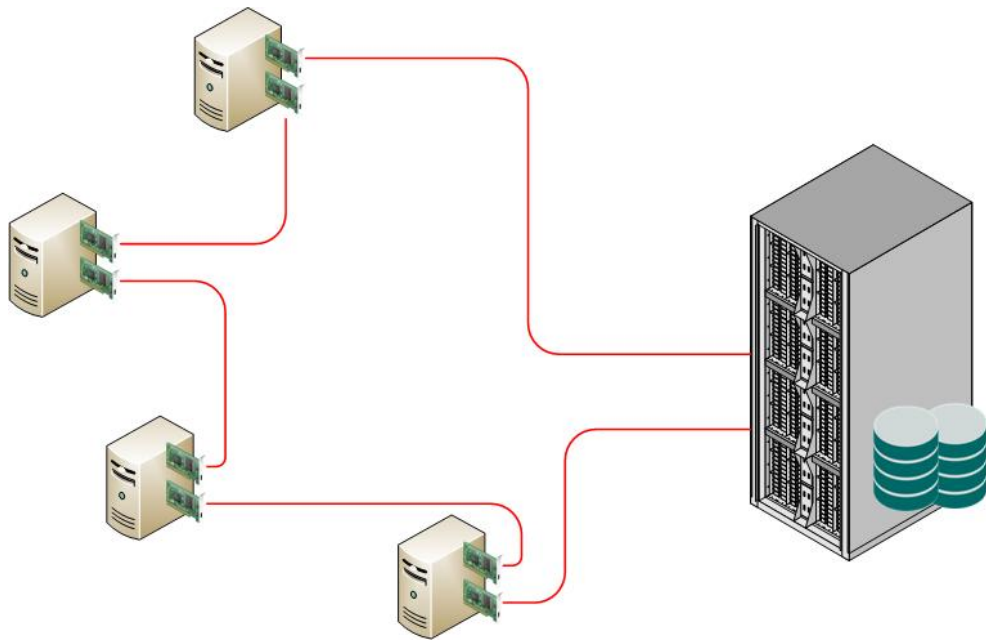


Figure 6 - Arbitrated loop

All the ports in FC-AL topology are configured as NL ports. An 8 bit addressing scheme is used to identify the target and the initiator. FC-AL topology was popular in the early days of SAN when switched fabric topology was expensive. In the current environment, FC-AL and FC-P2P topologies are used only in rare cases. Figure 7 shows one of the cases where arbitrated loop are used. The loop is connected to a fabric using 2 ports for each end of the loop. The fabric switch acts as a node in the loop topology. This topology is used to save ports on the fabric switches for non-critical servers.

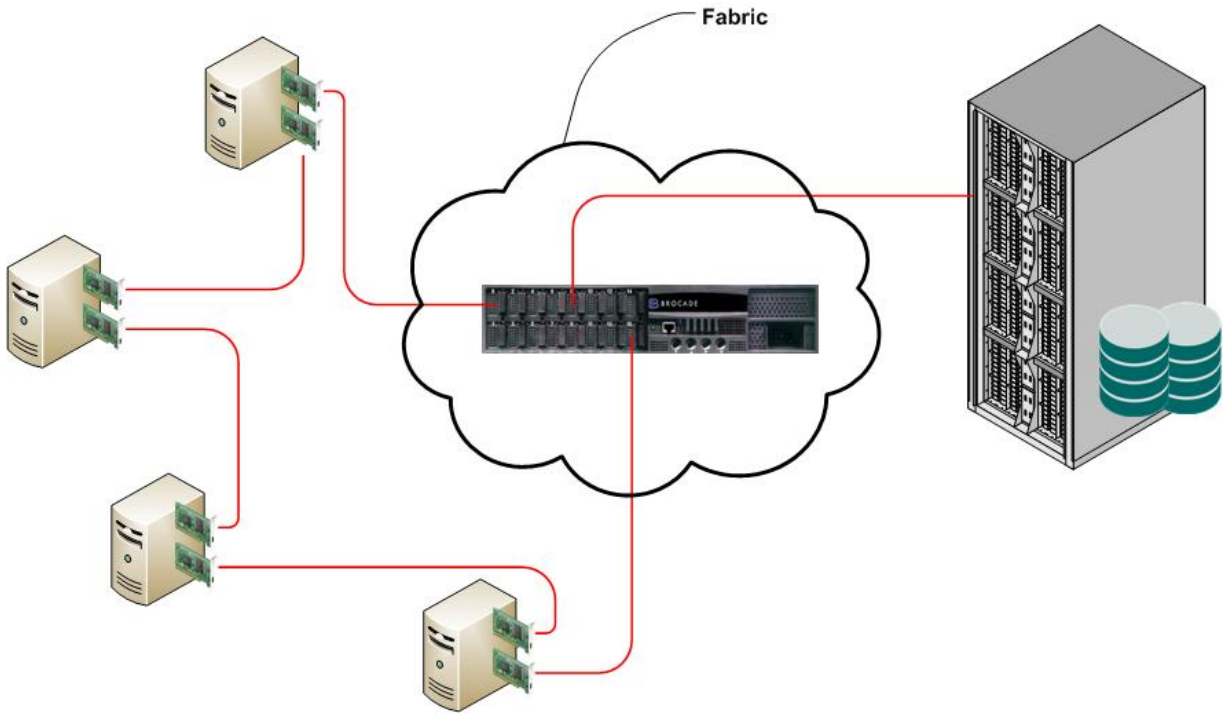


Figure 7- Arbitrated loop with switched fabric

4.1.2.1 Switched Fabric

Fabric is defined as a collection of switches sharing the same zoning. Good practice dictates that there should be at least two independent fabrics in the storage network for redundancy. Figure 8 shows the typical storage area network. It consists of two independent fabrics and multiple paths to the storage array. This architecture is used for redundancy and load balancing.

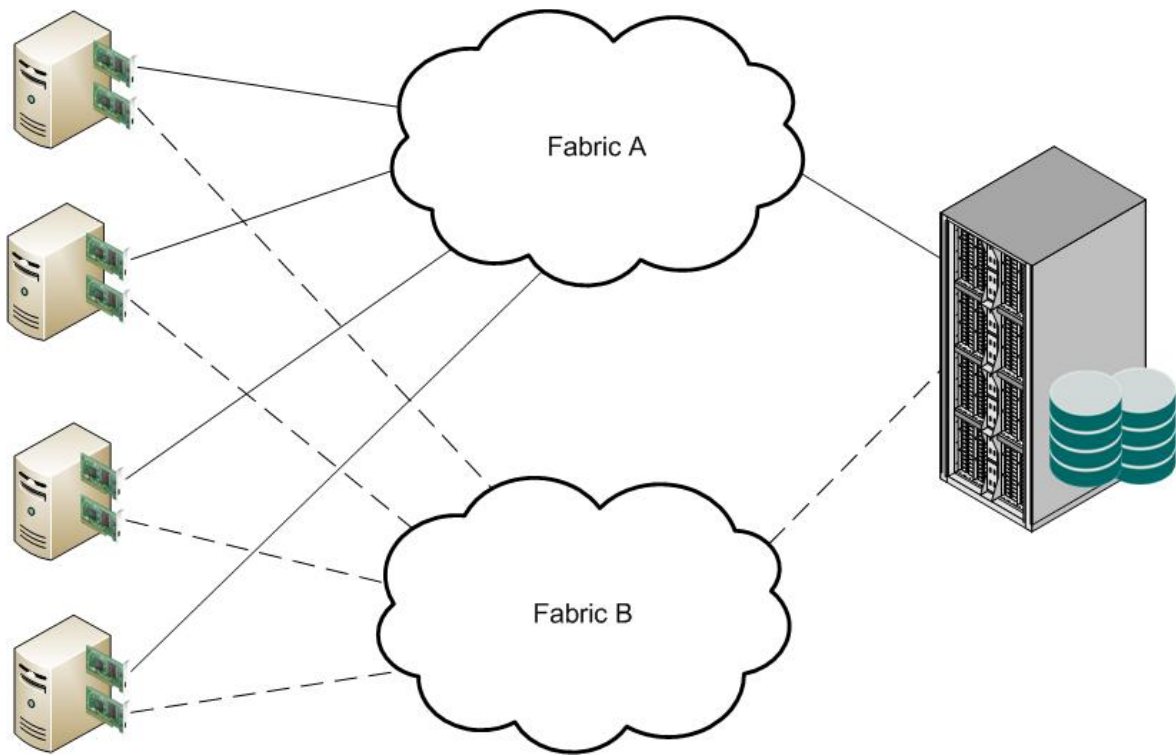


Figure 8 - Dual fabric storage area network

In a fabric, there is one switch that is selected as the primary switch. The primary switch maintains the fabric services and shares zoning information with the rest of the switches. Each switch in the fabric has a unique domain ID within that fabric. Domain ID is used when opting for hard zoning in the fabric. Hard zoning is defined as {domainID, Port; domainID, Port}

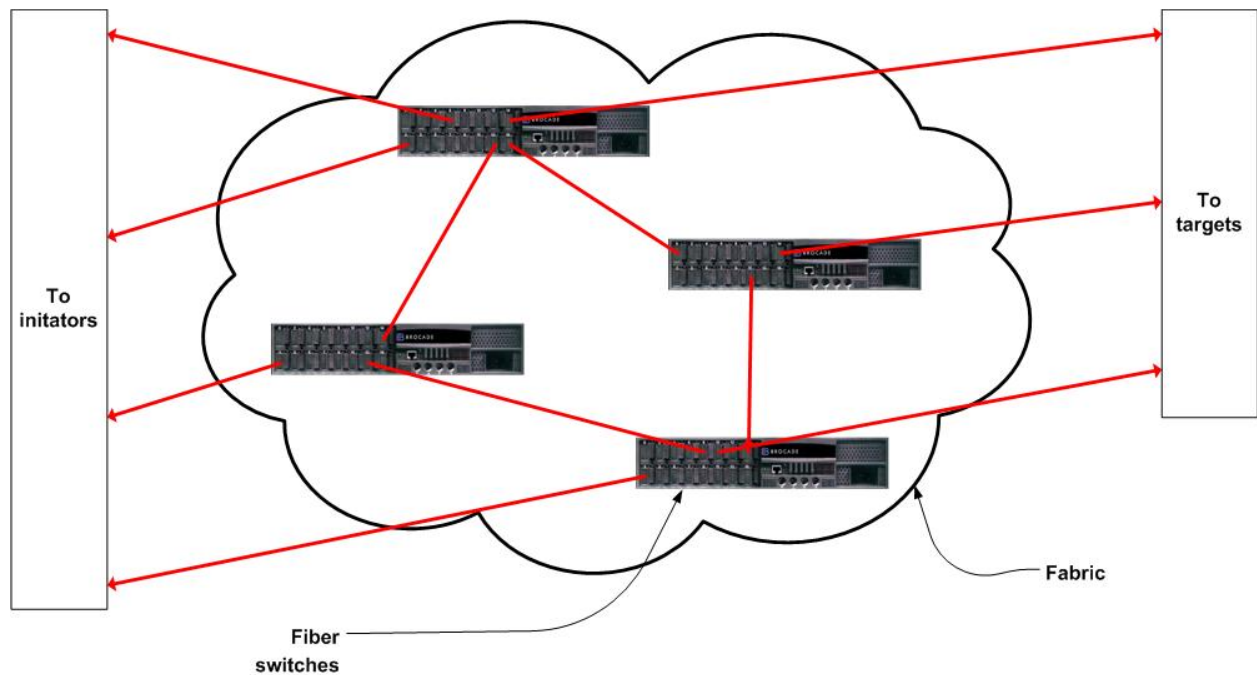


Figure 9 - Switched fabric

Switches within the SAN fabric are connected with each other using fiber channel links. Figure 9 shows a switched fabric. Topologies of fabrics are explained in later sections. There are 2 types of switches. Director class switches are powerful high end switches with enhanced processing capability. They are usually used as core switches. Departmental class switches are medium cost switches. Departmental switches are used as edge switches.

4.1.2 2 Zoning

A zone is a combination of initiators and targets. Zoning in SAN is done on the switch level. It is done to restrict access of a particular application server to the logical disks within the storage network. This increases the security of data within the network.

When a zone containing the initiator and target is done on a fabric, any storage device or application server outside of the defined zone is not visible or accessible to the application server. Best practices for zoning in SAN dictate that the zone should contain a single initiator and a single target. A zone is defined on the fabric as {initiators; targets}

Zoning in storage networks is classified into three types. Soft zoning utilizes the WWN of the application server HBA and the WWN of the storage array. An advantage of soft zoning is that it is switch independent. The application server or the storage array port can be connected to any port of any switch in the fabric. Hard zoning does not include the WWN's of the storage array port or the HBA of the application server. Hard zoning contains the switch ports on which the system is connected. The main advantage of hard zoning is that the storage admin can change the application server or the HBA cards without changing the zoning. The third type, mixed zoning, is a combination of hard and soft zoning. Table 2 lists some examples of soft, hard and mixed zoning

Soft zoning	{ 5001438002219756, 50000B0000C26240}
Hard zoning	{2,33; 2,196}
Mixed zoning	{3,45; 50060B0000C26240}

Table 2 - Zoning examples

LSAN or logical SAN zoning is done between 2 fabrics. This ensures communication between nodes on one fabric with nodes on another fabric. For LSAN zoning, the same zone definition must be present in both the fabrics connected together. In an LSAN zone, only one part of the host/initiator will be logged into the fabric.

Another concept to look into before we proceed is NPIV - Node Port ID virtualization. In traditional FC networks, a node port, i.e. a port on the host has 1 WWN for each HBA. LUN's are visible only to the WWN it is configured for. This becomes an issue when the server is running virtual machines. Each virtual machine contains its own virtual WWN.

To overcome this issue, NPIV was introduced. Using virtualization on the switchport, each port can now support upto 255 virtual HBA's connected to it. This ensures that the LUN's can be visible to a particular VM and not all the VM's connected to the HBA.

4.1.3 FC switches

Fiber channel switches and routers are broadly divided into two types depending on the number of ports available and the processing power of the switch. Director switches have more ports and processing power. Director switches are also called core switches. The other type of switches is called edge switches. They have fewer ports and are not as powerful as director switches.

To maximize the performance of the storage network, core and edge switches are mixed together in the fabric. Depending on how they are placed within the fabric, three broad topologies can be used: Edge-core, core-edge-core or core-edge. When describing the topology used, it is described from the host to the storage array. Therefore edge-core topology means that edge switches are exposed to the hosts and the core switches are exposed to the storage array. Figure 10 shows a simplified version of edge core topology.

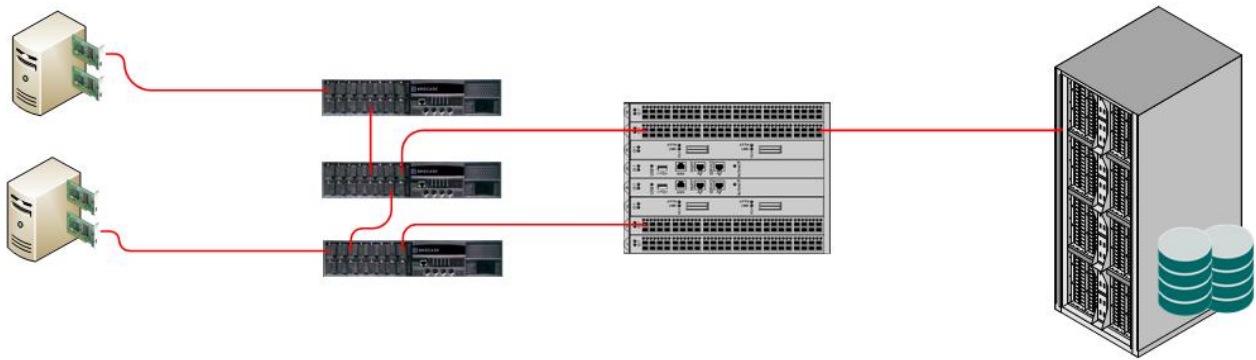


Figure 10- Edge core topology

Fiber channel switches are manufactured by two vendors: Brocade and Cisco. Both vendors can support each other's products connected together due to use of standardized channel. However the terminologies and special features of the two vendors are different. The following sections list out some of their products and features.

4.1.3.1 Brocade switches

Brocade offers a series of mid level departmental switches as well as more powerful director switches. There are a few concepts that are unique to Brocade switches. Brocade switches can aggregate

up to 8 fiber links together to create a trunked link. A trunked link aggregates the bandwidth of all the links together, thereby creating a larger pipe between switches for high bandwidth loads. A trunking license is required to enable this feature. A trunked link however uses available ports. To overcome this restriction in more powerful switches, an inter chassis link (ICL) is supported. An ICL acts the same way as a trunked link but uses dedicated ports. This reduces the loss of usable ports on the switch.

Brocade switches use a proprietary OS designed by Brocade called Fabric OS. Switch configuration is performed through an out of band Ethernet port. Out of band Ethernet management means that the switch uses a dedicated path for management. This makes the switch configuration independent of whether the fabric OS has completely booted on the switch.

Table 3 lists the available Brocade switches. All switches support trunking of 8 ports in a single port group. Brocade DCX series are the director switches. They contain a chassis and multiple hot pluggable modules. All Brocade switches support 3 classes of traffic: 2 (unencrypted), 3 (encrypted and unencrypted) or F (fabric maintenance). All Brocade switches also support NPIV, where 1 physical HBA can support up to 255 virtual HBA's.







Switch model	Number of FC ports	Fiber channel speed supported	Rack size	Aggregate bandwidth
Brocade 6505 	24	16 Gbps 8 Gbps	1U	384 Gbps
Brocade 6510 	48	16 Gbps 10 Gbps 8 Gbps	1U	768 Gbps
Brocade 6520 	96	16 Gbps 10 Gbps 8 Gbps	2U	1536 Gbps
Brocade 300 	24	8 Gbps 4Gbps	1U	192 Gbps
Brocade 5300 	80	8 Gbps 4Gbps	2U	640 Gbps
Brocade DCX 8510 Director 	384 (DCX 8510-8) 192(DCX 8510-4)	10 Gbps 8 Gbps	14U (DCX 8510-8) 8U (DCX 8510-4)	8.2 Tbps (DCX 8510-8) 4.1 Tbps (DCX 8510-4)
Brocade DCX	512 per chassis (DCX) 256 per chassis (DCX -4S)	8 Gbps 4 Gbps	14U (DCX) 8U (DCX-4S)	4.6 Tbps per chassis (DCX) 2.3Tbps per chassis (DCX-4S)

Table 3 - Brocade switches

4.1.3.2 Cisco switches

Cisco offers mid level departmental and more powerful director class switches. Cisco departmental switches include the MDS 9100 and the MDS 9200 series. They also offer MDS 9500 and MDS 9700 series director switches with expansion modules that the customer can configure according to need. Cisco departmental switches are unique in their numbering system. They offer 91xx switches where xx is 12, 24 or 48 according to the number of ports on the switch. MDS 9200 series switches are a new offering by Cisco for converged SAN solutions. They offer fiber ports for traditional SAN as well as FCoE and FCIP ports.

Cisco MDS can aggregate up to 16 fiber channel links to create a port channel. The port channel concept is similar to the trunking concept in brocade. Another unique feature to MDS switches is virtual SAN (VSAN) feature. VSAN is analogous to VLAN concept in ethernet fabrics. Using VSAN, the customer can isolate the SAN fabric. This allows easier management and setting of different QoS for different kinds of SAN traffic. Traffic between the VSAN's is isolated. In order to allow traffic to flow between two VSAN's an IVR (inter VSAN routing) license is required.

Table 4 lists the switches and their important features that cisco offer. All switches support encrypted and unencrypted traffic. All Cisco switches also support NPIV where a single physical HBA can be virtualized to represent up to 255 virtual HBA's.



Switch model	Number of FC ports	Fiber channel speed supported	Rack size	Aggregate bandwidth
MDS 9124 	24	4 Gbps 2 Gbps 1Gbps	1U	96 Gbps
MDS 9148 	48	8 Gbps 4 Gbps	1U	384 Gbps
MDS 9506 series director 	192	10 Gbps 8 Gbps 4 Gbps	7U	1.5 Tbps
MDS 9509 director	336	10 Gbps 8 Gbps 4 Gbps	14U	2.68 Tbps
MDS 9513 director	528	10 Gbps 8 Gbps 4 Gbps	14U	4.2 Tbps
MDS 9710 director 	384	16 Gbps 10 Gbps 8 Gbps 4 Gbps	14U	3 Tbps

Table 4 - Cisco switches

4.2 FCoE

Fiber channel over ethernet (FCoE) is an Ethernet network protocol that allows mapping of the fiber channel protocol. The fiber channel frames are encapsulated within Ethernet frames, allowing the fiber channel network to be extended to Ethernet networks. A converged network adapter (CNA) is required on hosts.

The main advantage of FCoE is the reduction in the number of ports required on hosts. Instead of having two Ethernet ports and two fiber channel ports, we can have just two CNA's. This topology reduces the cabling, power and cooling costs. The converged networked adaptor on the host distinguishes between pure Ethernet frames and encapsulated frames. On the network side, the FCoE switch converts the encapsulated FCoE frames to FC frames and forwards them on traditional FC network. The switch also filters out traditional Ethernet frames to the LAN network.

Figure 11 shows a traditional data center layout. The red lines represent fiber channel links. The blue lines represent Ethernet links. Figure 11 only shows the cabling for one of the minimum two fabrics. The number of cables and switches needed to connect the fiber channel switches to the fiber network as well as the Ethernet network is high. This topology is advantageous where there is an operational need to separate the two networks.

Figure 12 shows the same environment with FCoE enabled switches. The immediate advantage that can be seen is that less cabling and fewer network cards. The figure represents the main advantage of FCoE in small or medium environments - convergence of the fiber channel and ethernet networks.

An extension of FCoE is Fiber Channel over IP (FCIP). FCIP extends the fiber encapsulated traffic over internet. This allows fiber channel fabrics to extend or connect over large distances without leasing expensive dark fibers.

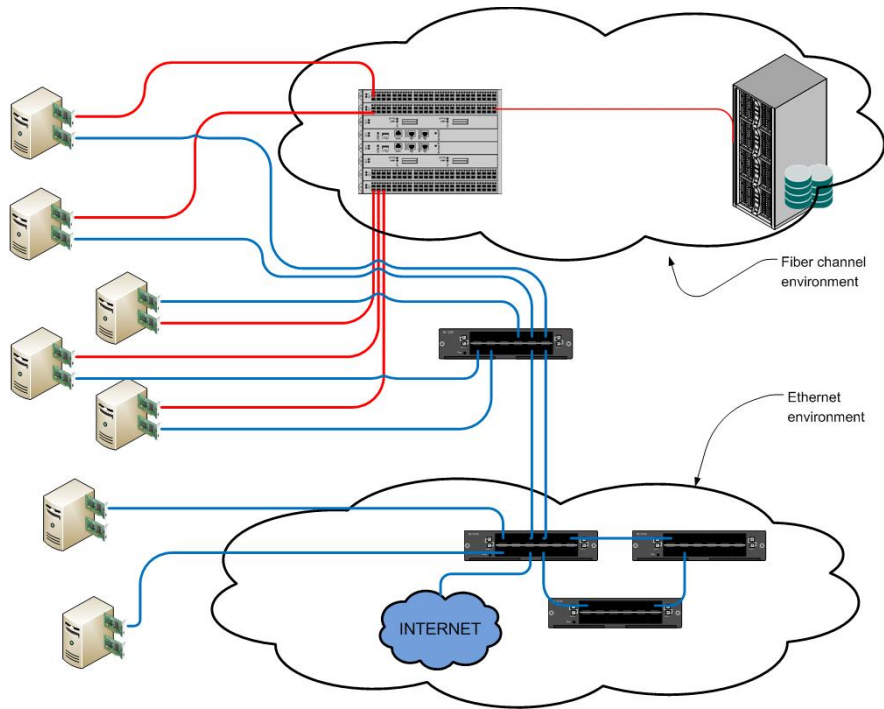


Figure 11 - Traditional data center layout

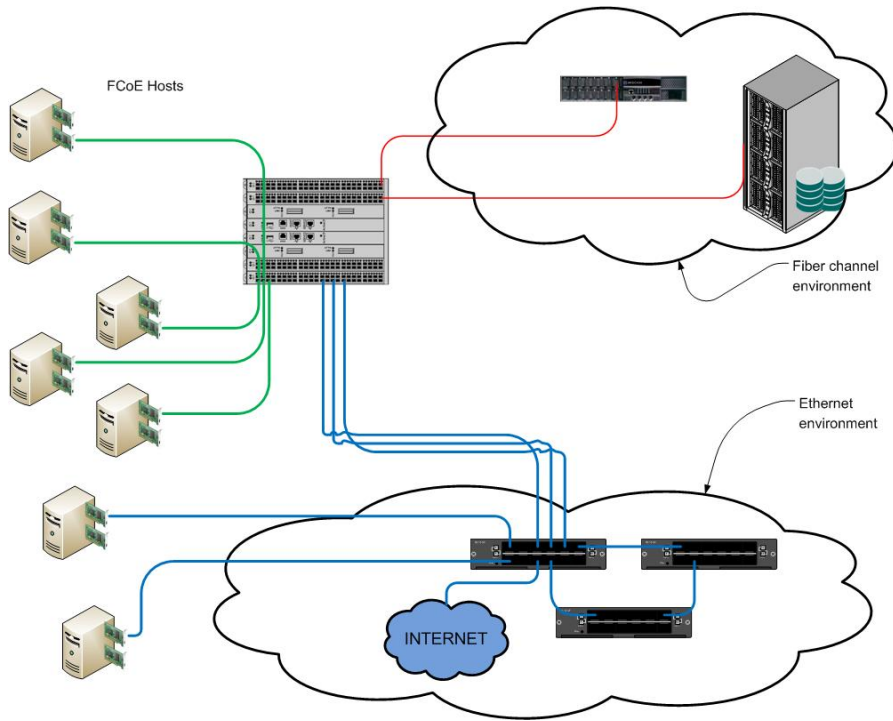


Figure 12 - FCoE environment

5. Storage arrays

Storage arrays consist of dedicated hardware that manages disk arrays. The controllers on a storage array process data requests from hosts and fetch the data. There are two main types of storage access levels: file access and block access.

File access, traditionally called NAS, allows the host to access data on the file level. The storage array creates a virtual server configured for CIFS (Windows) or NFS (Unix) storage. This storage is then shared on the network. File access storage arrays are separate due to the requirement for the array itself to host virtual servers for file shares. Block level access allows the host to access an individual blocks of storage. Physical drives are virtualized and blocks of data are presented to hosts as virtual disks, also called LUN's.

The LUN's show up as raw disks on the server. The server administrator is then free to format the disks as needed. Currently some operating systems support added protection on the server level. Windows servers can do striping, mirroring or concatenation of physical drives on the server side. Solaris ZFS adds software RAID functionality to the drives. Open source ZFS solutions are also available for UNIX based systems. All this is done on the server side and is not pertinent to this case study. On the storage side, care must be taken to ensure that server side protection does not add unnecessary overheads.

Storage arrays are highly resilient to any single point of failure. No single component can cause data loss or data to be inaccessible. The controllers are redundant. Depending on whether each of them is processing IO's the array can be active-active or active-passive. In active-active arrays, the LUN is mapped to both the controllers. Any controller can accept IO to the LUN. In active-passive arrays, the LUN is controlled by one controller only. All IO's to the LUN are processed by the controller owning the LUN. If the controller fails, the LUN ownership is trespassed to the non-owning controller. This non-owning controller then handles IO's to the LUN until the owning controller is back up.

All storage arrays have inbuilt redundant RAID controllers. The disks can be configured for RAID protection. Depending on the RAID level chosen to configure the LUN, the LUN may or may not be protected against drive failures.

5.1 Disk types

All data is stored on physical media called as a hard drive. The name hard drive emerged in the early days of computing. Volatile media such as random access memories were called 'soft' storage media. On the other hand non-volatile storage media was called 'hard' storage media. In the case of storage arrays, enterprise drives are used over commercial off the shelf disks. The specially manufactured enterprise drives give higher and predictable reliability and consistent response times over their lifetime.

In storage network planning, the main characteristic to look for in any enterprise disk technology is IOPS (input/output per second). Every application has different IOPS requirement for a particular response time. The disk needs to service that IOPS or the response time goes up and accumulates until the cache is full. If the cache gets full, the application crashes. The following equations are used when calculating what disk type needs to be used:

$$T_s = \text{Seek time} + \frac{0.5}{\left(\frac{\text{Disk RPM}}{60}\right)} + \frac{\text{Data block size}}{\text{Data transfer rate}} \quad \text{_____ (I)}$$

$$\text{IOPS} = \frac{1}{T_s} \quad \text{_____ (II)}$$

Equation I calculated the disk service time for a particular disk technology. Seek time is defined as the time taken to position the read/write head over the track on the disk. Data block size depends on the application accessing the disk. For most file systems, this value is defaulted to 256 kB. However care must be taken to verify this. Data transfer rate depends on the bus speed used to connect the disk. IOPS is the inverse of disk service time, as shown in equation II. Best practice dictates that the effective IOPS for the disk should be calculated as $IOPS_{effective} = IOPS * 70\%$. The 70% criterion is required since the disk performance starts to degrade at after exceeding 70% capacity.

When we have the required IOPS for the application, we can choose the appropriate disk type. Table 5 lists out the average IOPS per disk type.

Disk Type	Rotational speed (RPM)	IOPS
SATA (Serial ATA)	7200	80
NL-SAS (Near line SAS)		
FC (Fiber channel)	10000	140
SAS (Serially attached SCSI)		
FC	15000	180
SAS		
EFD (Enterprise flash disk)	<i>Solid state</i>	> 2000

Table 5 - Average IOPS per disk type

Once the disk type has been decided, the number of disks required for the application can be calculated using equations:

$$D_C = \frac{\text{Total capacity required}}{\text{Usable capacity of single disk}} \quad \text{--- (III)}$$

$$D_P = \frac{\text{IOPS generated at peak load}}{\text{IOPS serviced by single disk}} \quad \text{--- (IV)}$$

$$\text{Total disks required} = \max(D_C, D_P) \quad \text{--- (V)}$$

Equation III gives the number of disks required while taking into consideration the applications capacity requirement. Here the usable disk capacity is the formatted capacity of the disk. Equation IV gives the number of disks required to according while taking into consideration the applications performance requirement. The total number of disks required does not necessarily represent the actual number of disks. Equation V only gives the number of disks required while assuming that the disks are not RAID protected. Adding RAID protection will change the number of disks required as shown in the next section.

5.2 RAID protection

All data in the storage array is protected against drive failures. The main requirement in any well designed storage network is that there should never be a single point of failure. That single point of failure can be a disk, cable or entire components of an array. RAID or Redundant Array of Independent Disks is a technology pioneered by researchers at UC Berkley in 1987. RAID is standardized into 7 standard numbered levels. The most widely used are RAID 0, RAID 1, RAID 5, RAID 6 and nested RAID.

Each RAID level has IO penalties associated with them. These have to be taken into account when designing the system. An IO penalty means the effective IO operations performed in the back end. Depending on the RAID level, a single IO from the server may become 2, 4, or 5 IO's at the back end of the RAID controller. The IO penalty has to be considered when determining the number of disks required for performance. The application (or server) IO generated has to be multiplied by the IO penalty to get the actual IO at the disk level.

RAID 0 stripes the data across the disks in the raid group. RAID 0 is the only RAID level that offers no protection against disk failure. Each write is written to different disks as shown in Figure 13. The server sends the IO to the RAID controller. The RAID controller then splits the writes according to the stripe size. Each stripe size of the data will be written to a different disk. This configuration increases the read write performance of the data as more than one spindle is involved with each IO operation in

parallel. RAID 0 does not however provide any protection against disk loss. If a disk in the RAID group fails, all the data in the RAID group is inaccessible. RAID 0 has no IO penalty. It means that for every 1 application IO there is only 1 IO at the back end.

RAID 1 mirrors the data across the disks in the RAID group. A RAID 1 group must contain only 2 drives. The 2 drives are identical copies of each other with one disk as the primary disk and the second as the backup disk. If the primary disk fails the RAID controller will start active IO to the secondary disk. When the failed disk is replaced, the entire contents of the non-failed disk are copied to the new disk. This process happens in the background and is online. Figure 14 - RAID 1 shows the IO operation of a RAID 1 protected system. The IO penalty of a RAID 1 system is 2. It means that for every IO generated at the application level, there are 2 IO's at the disk level. These 2 IO's are for the 2 disks that are involved.

RAID 2 stripes the data into different drives and uses hamming code for error correction in dedicated disks. RAID 2 striping occurs per bit, with every bit getting written to different disks. Due to this architecture, all disks have to be simultaneously spun in sync to read a block of data. With data block sizes increasing, RAID 2 is not viable. There is no application or array which uses RAID 2 anymore.

RAID 3 stripes data into different drives and uses a dedicated parity disk. The parity disk contains the parity information of the data and is used to recover from errors. In RAID 3, the striping of data occurs per byte. Therefore a single block of data resides on more than one drive. In RAID 3 architecture, all the disks must spin to retrieve a block of data. Like RAID 2, there is little need for RAID 3. Only a handful arrays support RAID 3, other than that it's obsolete.

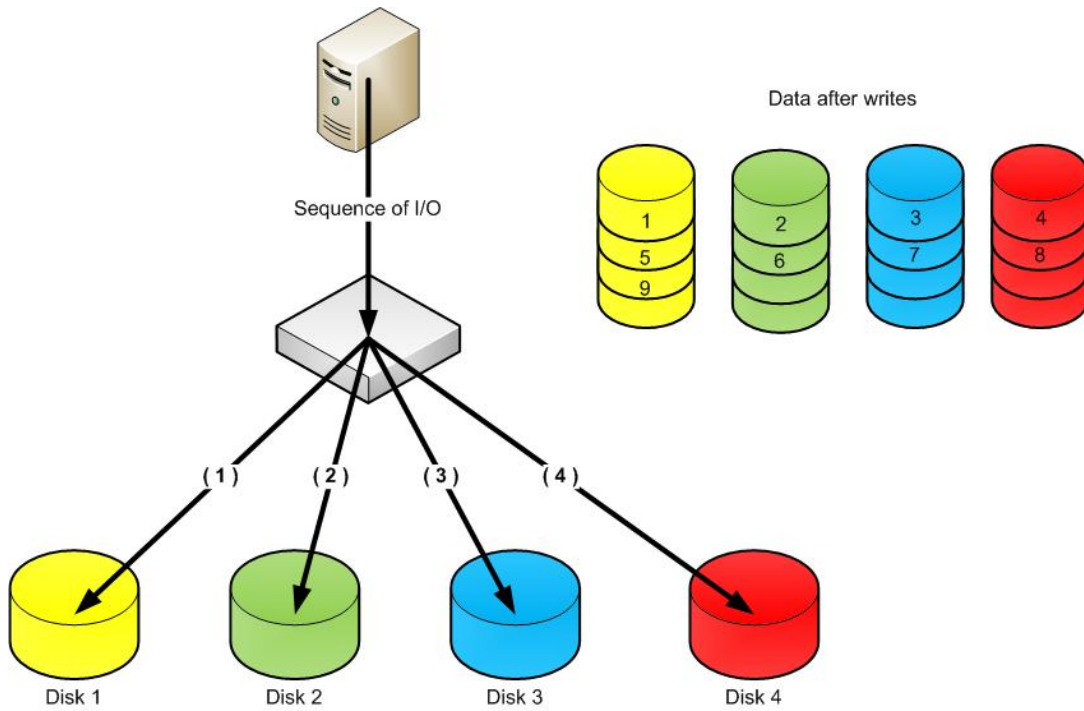


Figure 13 - RAID 0

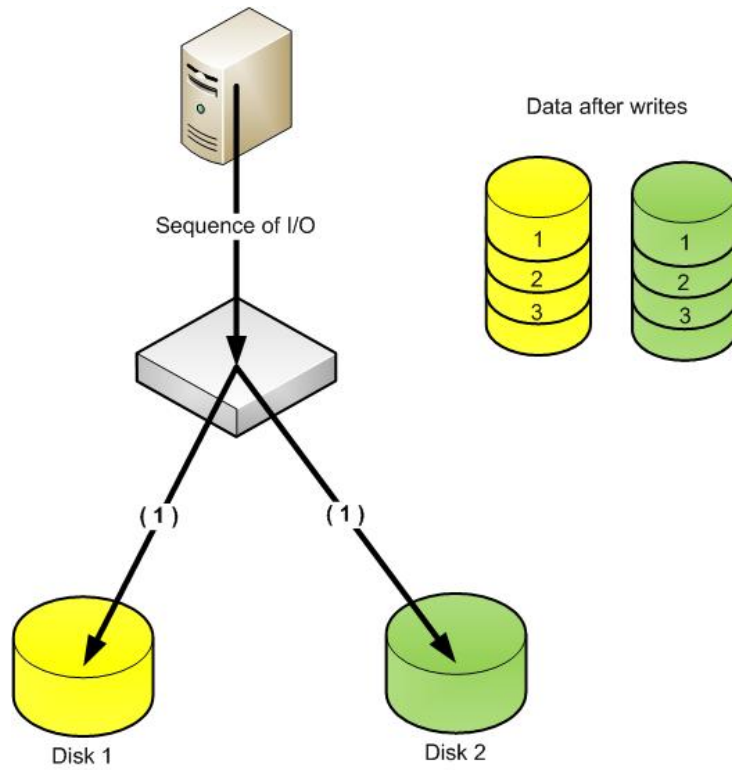


Figure 14 - RAID 1

RAID 4 stripes data into different disks and uses a dedicated parity drive. The biggest difference between RAID 3 and RAID 4 is in the stripe depth. RAID 4 stripes the data across the drives on block level. This ensures that each disk can service IO independently from one another. If disk 1 has the block that is being requested, only disk 1 has to spin to retrieve the data. The other disks are free to service IO's to the blocks that they are storing simultaneously. RAID 4 has an IO penalty of 4. Every IO results in 2 reads and 2 writes. The minimum number of drives needed for RAID 4 is 4- three disks to store data and one disk to store parity. Figure 15 shows the IO operations and disks states for RAID 4. This RAID level is rarely used currently. RAID 4 can recover from a single disk failure in the same RAID group. Until recently RAID 4 was widely used by Netapp. RAID 5 is more preferred over RAID 4.

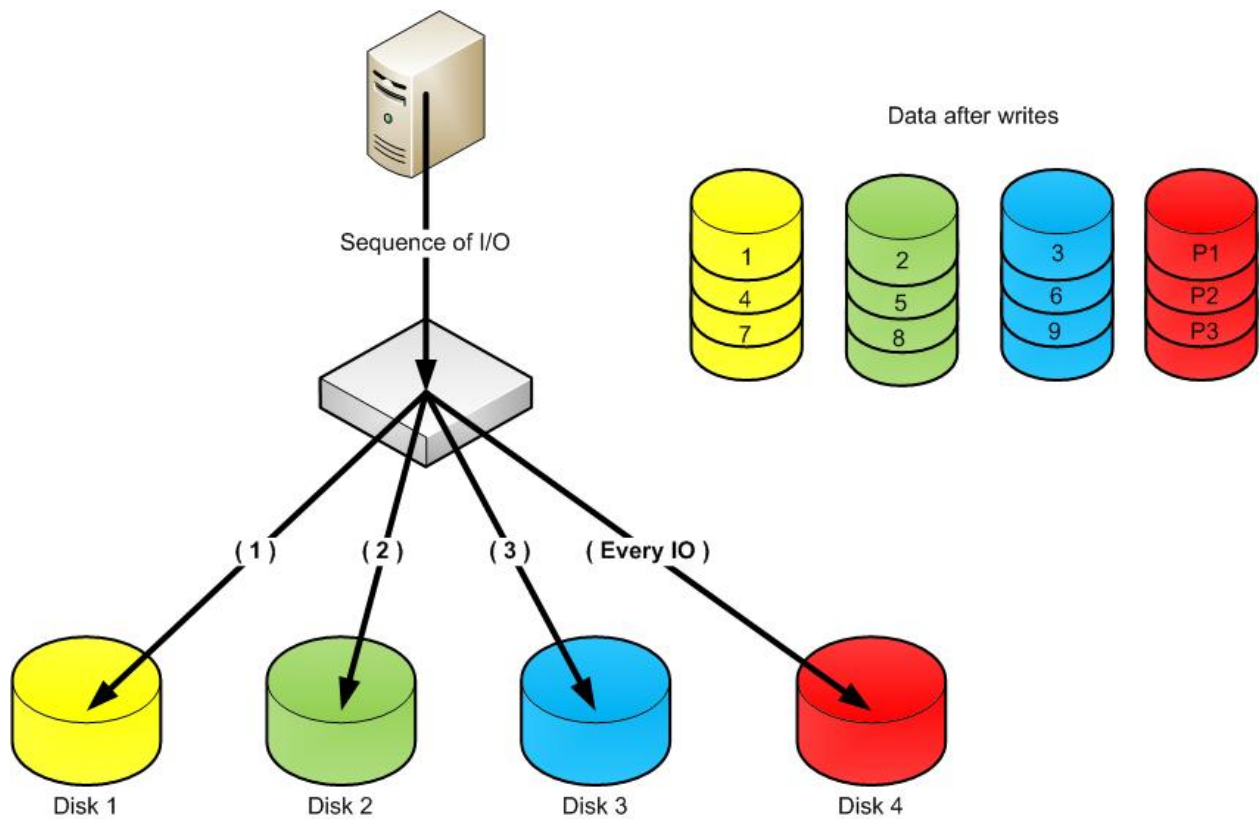


Figure 15 - RAID 4

RAID 5 stripes the data across drives at block level and uses distributed parity for protection. RAID 5 is very similar to RAID 4 in architecture. The only difference is that RAID 4 uses a dedicated parity disk

and RAID 5 distributes the parity across the drives in the RAID group. Figure 16 shows the IO operation of a RAID 5 subsystem. The minimum number of disks required to implement RAID 5 is four. Every write to a RAID 5 system results in 2 reads and 2 writes. Thus the IO penalty of a RAID 5 system is 4. RAID 5 can recover from a single disk failure in a disk group and is the most widely used RAID structure.

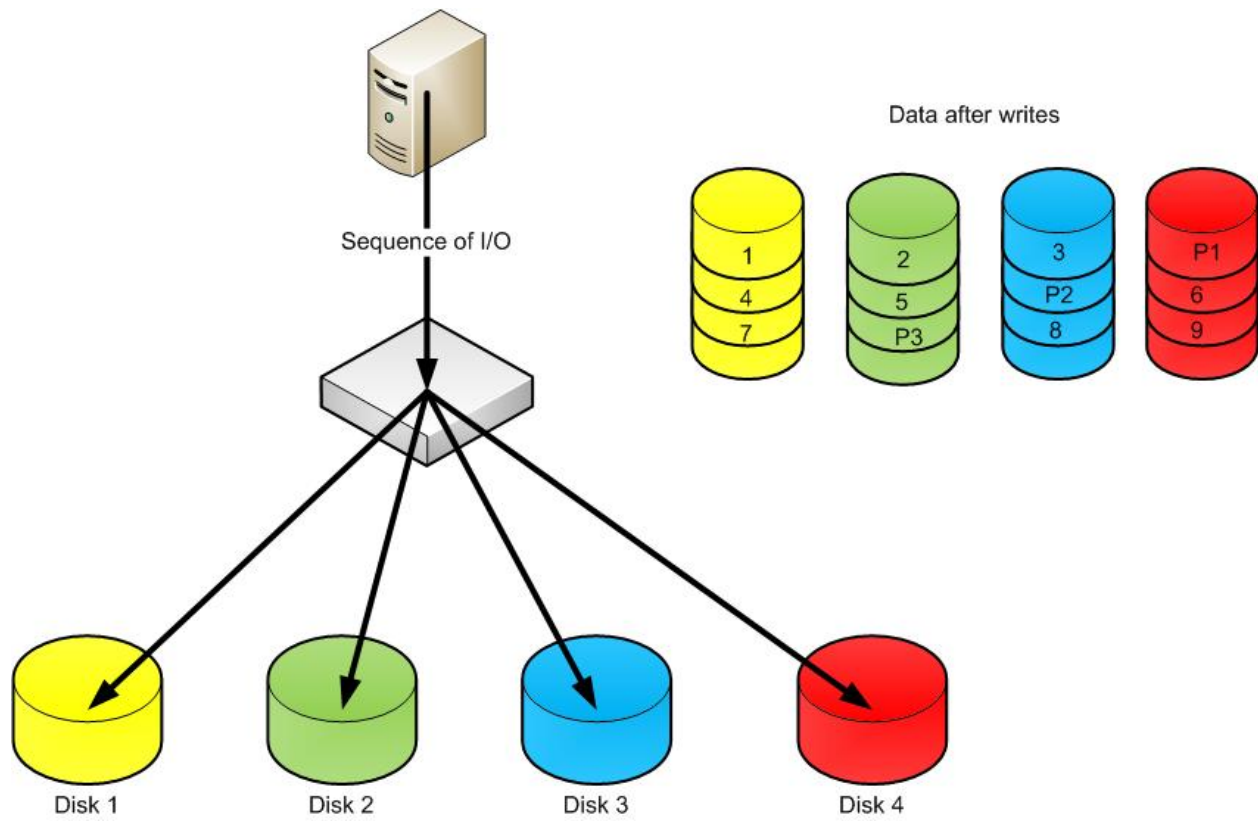


Figure 16 - RAID 5

RAID 6 uses block level striping with dual distributed parity. RAID 6 is just an extension of RAID 5 where there are 2 parity bits stored. This architecture allows RAID 6 to recover from 2 disk failures in the RAID group. RAID 6 requires a minimum of 5 disks due to the addition of another parity bit. RAID 6 has an IO penalty of 6. Each write from the application translates to 3 reads and 3 writes to the RAID subsystem. This RAID subsystem is very common where the rebuild time for the system is high and the MTBF (mean time between failure) of the drive is fairly low. This means that there is a high probability

that a second disk failure may occur in the RAID group while the data is being rebuilt. RAID is almost exclusively used in large SATA and NL-SAS drives where the rebuild times are in tens of hours. For other disks speeds, RAID 5 is a much better option over RAID 6 due to lower IO penalty and higher disk utilization ratio.

5.3 EMC arrays

EMC offers NAS, SAN and NAS gateway solutions. SAN arrays from EMC range from small and cheap unified VNXe to huge VMAX arrays geared to large data centers. NAS offerings include entry level unified VNXe to large Isilon arrays supporting big data.

5.3.1 VNX series

The VNX series is the low to mid range SAN and NAS offering from EMC. The VNX series includes VNX, VNXe, VNX-F, VNX-CA and VNX gateway.

The traditional VNX is a mid range storage array offering low cost, medium performance and easy setup and maintenance. VNX series supports both NAS and SAN; i.e. the VNX series can be used as block access as well as file access. If the VNX is configured only for block access it is known as VNX block. If the VNX is configured for file as well as block, it is known as VNX unified. The VNX 1 series was introduced in 2011 as a unified platform to replace the older block only Clariion and the file only Cellera. EMC upgraded their VNX range in 2014 with the introduction of the VNX2 series. It offered better processors and easier administration over the VNX1 series.

Model	5200	5400	5600	5800	7600	8000
Max Drives	125	250	500	750	1000	1500
Max data movers	3	4	4	6	8	8
Raw capacity in TB (fully loaded)	500	1000	2000	3000	4000	6000
Max SAN hosts	1024	1024	1024	2048	4096	8192
Max LUN size in TB	256	256	256	256	256	256
Max file system size in TB	16	16	16	16	16	16
Max file capacity per data mover	256	256	256	256	256	256

Table 6 - Comparison of VNX2 models

VNX arrays are managed mostly using the GUI tool unisphere. EMC also offers a CLI management tool for VNX with NaviCLI. Disaster recovery options for VNX include local snapshots (Snapview) and remote continuous replication (MirrorView). Table 6 lists out some of the important specifications of VNX2 series models. VNX2 series support online upgrade from one VNX model to another. That means that we can upgrade the VNX system from a 5200 to 7600 without any downtime required.

VNXe series offered by EMC is an entry level solution for small businesses. The VNXe series only offers NAS and iSCSI connectivity, fiber channel is not supported. It is highly scalable allowing the client to start small and expand rapidly. Table 7 lists out the VNXe models.

Model	3150 single	3150 dual	3300
Max Drives	50	100	150
Max data movers	1	2	2
Raw capacity in TB (fully loaded)	144	288	360
Max LUN size in TB	2	2	2
Max file system size in TB	16	16	16
Max file systems	128	256	512

Table 7 - Comparison of VNXe models

VNX-F is an all flash VNX array recently introduced. It uses a VNX 7600 engine to drive flash only disk array enclosures. The all flash VNX-F is geared towards applications demanding high IOPS and very low response times. Currently VNX-F is available in 4 configurations - 21TB, 35TB, 49TB and 96TB total raw capacity.

VNX gateway is a specialized VNX offering by EMC. VNX gateway leverages the existing SAN infrastructure and adds a NAS gateway. This enables small or medium businesses to use storage space on their existing EMC SAN arrays to support NAS file systems without buying NAS hardware. VNX gateway is offered in two models- VG2 with 256TB total file system size and VG8 with 1792TB total file system size. Figure 17 shows the use case of VNX gateway. The ethernet hosts can now use a NAS file system without the organization buying specialized and expensive NAS hardware.

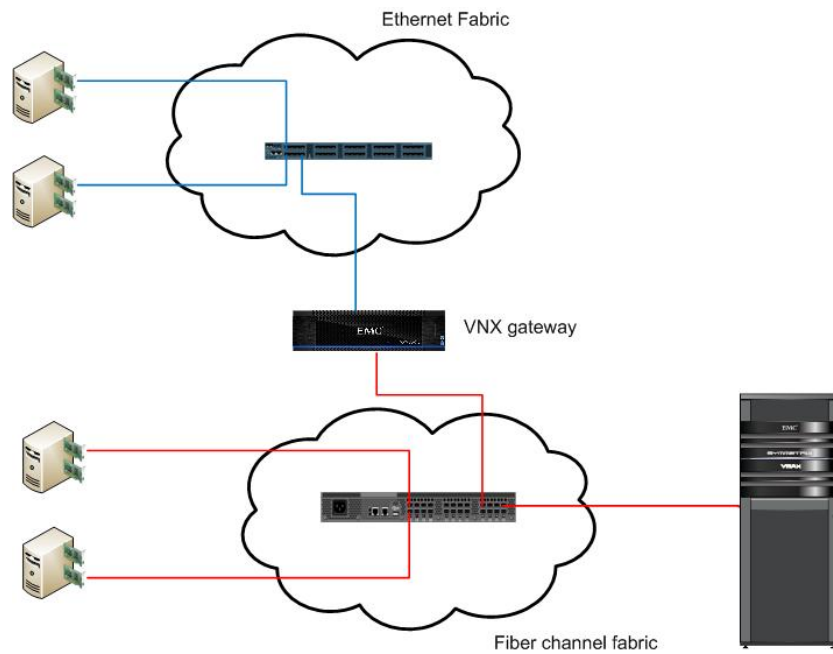


Figure 17 - VNX gateway

5.3.2 Symmetrix series

Symmetrix series is the high performance SAN only offering from EMC. Symmetrix series is geared towards large enterprise customers who demand high performance and storage space with a low response time and cost per TB. Symmetrix series includes the old DMX and the new VMAX.

Symmetrix storage arrays offer higher performance than the entry level VNX arrays due to the arrays having more engines. A symmetrix array can have up to 8 engines. Each engine has 2 directors for redundancy. The VMAX directors have a back end virtual matrix. This ensures that memory and CPU's in the directors are pooled, thereby offering better performance across the board. Current range in the symmetrix family includes 3 models- 10k, 20k and 40k.

VMAX 10k is an entry level symmetrix array with support of up to 4 engines. The biggest difference in 10k is that it only supports thin provisioning. The second major difference over 20k and 40k is that 10k is the only symmetrix array offering NAS compatibility.

VMAX 20k is a mid level symmetrix array with support of up to 8 engines. It delivers higher performance than 10k by having more number of CPU's per director. Unlike 10k, VMAX 20k supports thin provisioning as well as traditional thick provisioning.

VMAX 40k is the highest tier storage array offered by EMC in the symmetrix family. It has the largest number of CPU's per director within the Symmetrix range. The 40k array can support a huge back end storage space while keeping the response times low. 40k is heavily geared towards large enterprise SAN.

EMC symmetrix arrays can be configured and controlled using GUI in the form of unisphere. The preferred method for configuration with much deeper controls than unisphere is through CLI in the form of SYMCLI. Table 8 lists out the models offered by EMC in the symmetrix range.

Model	VMAX 10k	VMAX 20k	VMAX 40k
Number of engines	1 - 4	1 - 8	1 - 8
Max cache	512 GB	1 TB	2 TB
Max usable capacity	1536 TB	2067 TB	3874 TB
Virtual matrix BW	200 GB/s	192 GB/s	400 GB/s

Table 8 - Comparison of VMAX models

5.3.3 Isilon

Isilon is the high end NAS offering by EMC. Isilon is geared towards enterprise NAS where there is a requirement for a large file system. Isilon is able to provide large file shares with easy scale out capability. Isilon works on the cluster level with each cluster containing multiple nodes. Nodes can be added or removed from the cluster online with little configuration time.

Depending on the response times desired, Isilon offers three distinct node series. The high performance S series node is aimed at high IOPS file shares. These high IOPS workloads are usually encountered in simulators, image or video processing and web hosting industries. S series nodes offer capacity from 16 TB to 2 PB.

The mid range X series nodes offers a balance between IOPS and file system size. X series nodes can scale from 6 TB to 11 PB. The use case for Isilon X series is applications that are IOPS sensitive and require a low response time but do not warrant the higher performance S series nodes.

EMC also offers Isilon NL series nodes. These nodes are geared towards large storage capacity rather than IOPS. This makes the NL series ideal for storing large backups or archiving where IOPS is not mission critical. NL series nodes can scale from 256 TB to 20 PB.

All Isilon nodes run EMC propriety OneFS operating system. OneFS operating system allows for easier management of the Isilon environment. It can also span multiple nodes to provide a maximum of 20 PB per file system. Spanning multiple nodes helps in scalability as well as allows tiering of the storage data. Thus in a single share, highly accessed data can rest in S series nodes, whereas rarely accessed data can rest in NL series node. This ensures that the end user can keep cost/TB low.

5.4 HP arrays

HP offers entry level storage arrays, high performance storage arrays and file share storage arrays in addition to traditional backup and virtualization hardware. HP offers entry level MSA series of hardware. For high performance enterprise solutions HP offers the 3par series hardware. For NAS, HP offers StoreEasy 1000 as NAS storage array and StoreEasy 3000 as gateway NAS hardware.

5.4.1 MSA series

MSA series storage arrays offered by HP are the companies low end arrays. The main advantage of the MSA series hardware is their low cost and ease of management. Offerings from the MSA series include the low end MSA 1040 and the high end MSA 2040.

MSA 1040 is the cheaper array in the MSA series. It can only support SAS and MDL-SAS drives. There is no SSD support in MSA 1040. MDL-SAS drives stand for midline SAS, similar to NL-SAS offered by EMC. The advantage of the 1040 is that the controllers in MSA 1040 can be hot-swapped with MSA 2040 controllers. This ensures that the customer can upgrade from MSA 1040 to MSA 2040 online; i.e. without any downtime.

Table 9 lists out some of the specifications of MSA series hardware.

Model	MSA 1040	MSA 2040
Max number of drives	99	199
Drive type supported	SAS MDL-SAS	SSD SAS MDL-SAS
Max capacity	192 TB	384 TB
Max number of controllers	2	4
Cache per controller	4 GB	4 GB

Table 9 - MSA series

5.4.2 3par

3par is HP's enterprise level SAN offering. It is a highly configurable enterprise storage array offering with low response times, high IOPS and a high degree of configurability. 3par was an independent company until HP bought the company. Currently 3par offers three models. Midrange StoreServ 7000 series, StoreServ 7450 all flash array offering very high response times and high range StoreServ 10000 series.

Midrange StoreServ 7000 series offers two models. The 7000 series is aimed towards low response times and acceptable IOPS while balancing cost and drive capacities. 7000 series have 2-4 controllers depending on the model. All StoreServ 7000 models have hot pluggable IO modules. Table 10 lists out the StoreServ 7000 models. All models support a mix of SAS, MDL-SAS and SSD drives. Storage tiering is also supported.

Model	StoreServ 7200	StoreServ 7400
Controller nodes	2	2 or 4
Cache	24 GB	32 - 64 GB
Max number of drives	8 - 240	8 - 480
Max raw capacity	1.2 - 400 TB	1.2 - 1100 TB

Table 10 - Comparison of StoreServ 7000 models

StoreServ 10000 series is the high end offering by HP in the 3par range. The 10000 series is aimed towards large enterprise clients who demand low response times and high IOPS throughput. The 10000 series has a large number of controllers to deal with the high storage demand. Table 11 lists the 3par StoreServ 10000 models offered by HP. All 10000 series models support SAS, MDL-SAS and SSD drives with automatic storage tiering.

Model	StoreServ 10400	StoreServ 10800
Controller nodes	2 or 4	2, 4, 6 or 8
Cache	192 - 384 GB	192 - 768 GB
Max number of drives	16 - 960	16 - 1920
Max raw capacity	4.84 - 1100 TB	4.84 - 2200 TB

Table 11 - Comparison of StoreServ 10000 series

5.4.3 StoreEasy series

HP offers two models of StoreEasy series hardware as a NAS platform. The StoreEasy 1000 series are standalone NAS arrays, while the StoreEasy 3000 series are the NAS gateways.

The StoreEasy 1000 series is preconfigured by HP. The customer has to order the configuration which cannot be changed easily at the later state. This gives rise to many standard model numbers depending on the configuration ordered. StoreEasy 1440, 1540, 1640 and 1840 are preconfigured bundles

in the StoreEasy 1000 series. The 1000 series supports SATA, MDL-SAS and EFD drives as internal storage. The maximum storage for the StoreEasy 1000 series is 48 TB.

StoreEasy 3000 series gateway NAS nodes are offered by HP. The 3000 series nodes offer a NAS interface to the SAN storage in the network. This means that the 3000 series NAS gateways can use the existing SAN storage and present NAS shares on the Ethernet network.

5.5 IBM arrays

IBM was the earliest company to develop and deploy storage arrays. IBM currently offers a large range of SAN and NAS hardware. The SAN offerings include the DS (disk system) range of entry level and enterprise level arrays and Storwise range of entry level and mid range arrays. On the NAS side, IBM offers N series hardware along with unified Storwise products. The unified Storwise arrays offer both SAN and NAS functionality.

5.5.1 Storwise series

IBM storwise series is virtualized on cluster basis. Storewise series contains different model nodes that can be mixed into the cluster. Management of the array is done on the cluster. Each node in the cluster can add storage space and/or enhanced functionality.

Storewise series has entry level V3700 nodes, midrange V5000 and V7000 nodes for block level SAN access. The storewise series also has unified V7000 nodes that provide block level SAN and file level NAS access.

Model	V3700	V5000	V7000
Cache	8 GB	16 GB	64 GB
Max number of drives	120	336	960
Max capacity per cluster	240 TB	700 TB	1800 TB
Controller nodes	2	2, 4	2, 4, 6

Table 12 - Comparison of Storewise models

A storewise array is made up of controller nodes and storage nodes. The controller nodes interface with the SAN fabric. Depending on the level of performance desired, the array will contain a mixture of controller and storage nodes. The controller nodes cannot be mixed with each other i.e. a V3700 controller node cannot be mixed with a V7000 controller node. The entire array is controlled using GUI.

5.5.2 DS series

IBM offers the DS (disk storage) series as enterprise level SAN array. The DS series is highly configurable and offers much better IOPS and lower response times than the Storwise array series. The DS series is block only SAN storage array.

The midrange DS series arrays are intended for small enterprise clients with a requirement for higher IOPS than Storwise. The midrange DS series models are DCS 3500 and DCS 3700. In addition to these, expansion modules are offered. The expansion modules include high density storage modules as well as performance modules. Table 13 lists out the IBM midrange DS series arrays. For DCS 3700, the higher performance statistics require use of performance expansion modules.

Model	DS 3500	DS 3950	DS 5020	DCS 3700
Cache	1 GB or 2 GB	2 GB or 4 GB	4 GB	4 - 48 GB
Max number of drives	192	112	112	180 - 360
Max capacity	576 TB	224 TB	336 TB	1440 TB
Host interface supported	FC SAS iSCSI	FC iSCSI	FC iSCSI	SAS FC iSCSI
Disk technology supported	NL-SAS SAS SSD	FC SATA FC-SAS	FC FC-SAS SATA SSD	NL-SAS SAS SSD

Table 13 - Comparison of midrange DS series arrays

In addition to the midrange DS series, IBM offers enterprise DS series. The enterprise DS series is aimed at large enterprise clients having range of IOPS and response time requirements. The enterprise series offers very low response times and the ability to handle large IOPS while having a huge data storage capacity. In enterprise DS series IBM offers DS 8000 and XIV models.

Model	DS 8000	XIV
Cache	16 - 1024 GB	720 GB - 4 TB
Max number of drives	1536	180
Max capacity	3072 TB	325 TB
Host interface supported	FC	FC iSCSI

Table 14 - Comparison on DS series

5.5.3 N series

IBM offers N series for NAS file shares. All N series appliances support deduplication for saving the actual space required by the data. N series offers entry level arrays to enterprise level arrays along with gateway models. The NAS gateway array leverages existing SAN architecture to provide NAS file shares.

N3000 family is the entry level NAS offering by IBM. The N3000 family contains three models: N3240, N3220 and N3150. The N3000 family is intended towards small offices with need for a NAS platform. All the N3000 series nodes have an option for single or dual controllers with 6/12 GB cache depending on the customer need. N3000 series support data storage from 240 - 576 TB.

N6000 family is the mid level NAS platform in the N series. It contains two models: N6220 and N6250. The N6000 family is intended for mid level and small enterprise clients. N6000 series NAS platforms have up to 40 GB of cache and 2880 TB of data storage capacity.

N7000 series is the premier enterprise NAS platform for IBM along with SONAS array. Both arrays feature large storage capacity with low response times. The N7000 series contains traditional NAS array as well as a NAS gateway. The SONAS array is similar in performance to N7000 series. The major difference is that SONAS supports easy scale out for NAS shares.

All IBM NAS arrays support SAS, NL-SAS and EFD drives. Automated tiering software ensures that the data is stored efficiently. Frequently accessed data is promoted for storage to faster EFD drives while rarely accessed data is archived in high capacity but slow speed SATA drives.

6. Case Study

A technology refresh is a long drawn process. The steps involved include pre-sales architecture planning, comparison and benchmarking solutions, rack and floor planning, implementation, post implementation. The people involved in the tech refresh process are: storage architect, implementation engineers, SAN administrators and sales representatives from the company accounts team. For the case study, I worked as the implementation engineer. As implementation engineer, I was responsible for capacity planning, installation of arrays and SAN switches, migration planning, migrations and post migration.

6.1 Pre-sales architecture planning

Pre sales process includes the high level architecture planning of the client environment. The pre sales steps typically are: identify the customer environment, calculate IOPS for the environment, and plan for future scaling.

The current environment analysis typically takes a month. During that time, the network diagrams of the customer are studied. The month is also used to calculate the IOPS requirement that the environment needs to handle. Figure 18 shows the high level overview of the customer SAN environment. The

environment handles 2435 TB of data spread across 4 IBM arrays - two XIV, and two Storwise V7000; with an IBM SVC for management.

Dual fabric architecture is used for switching. This is mainly for redundancy purposes with the hosts cycling through each available path using round robin architecture. The fabric consists of Cisco 9513 switches - one per fabric. The server farm consists of six Cisco UCS chassis running 36 blade servers. The operating systems are VMware and Windows server 2008. The server farm is also connected to ethernet fabric at the datacenter.

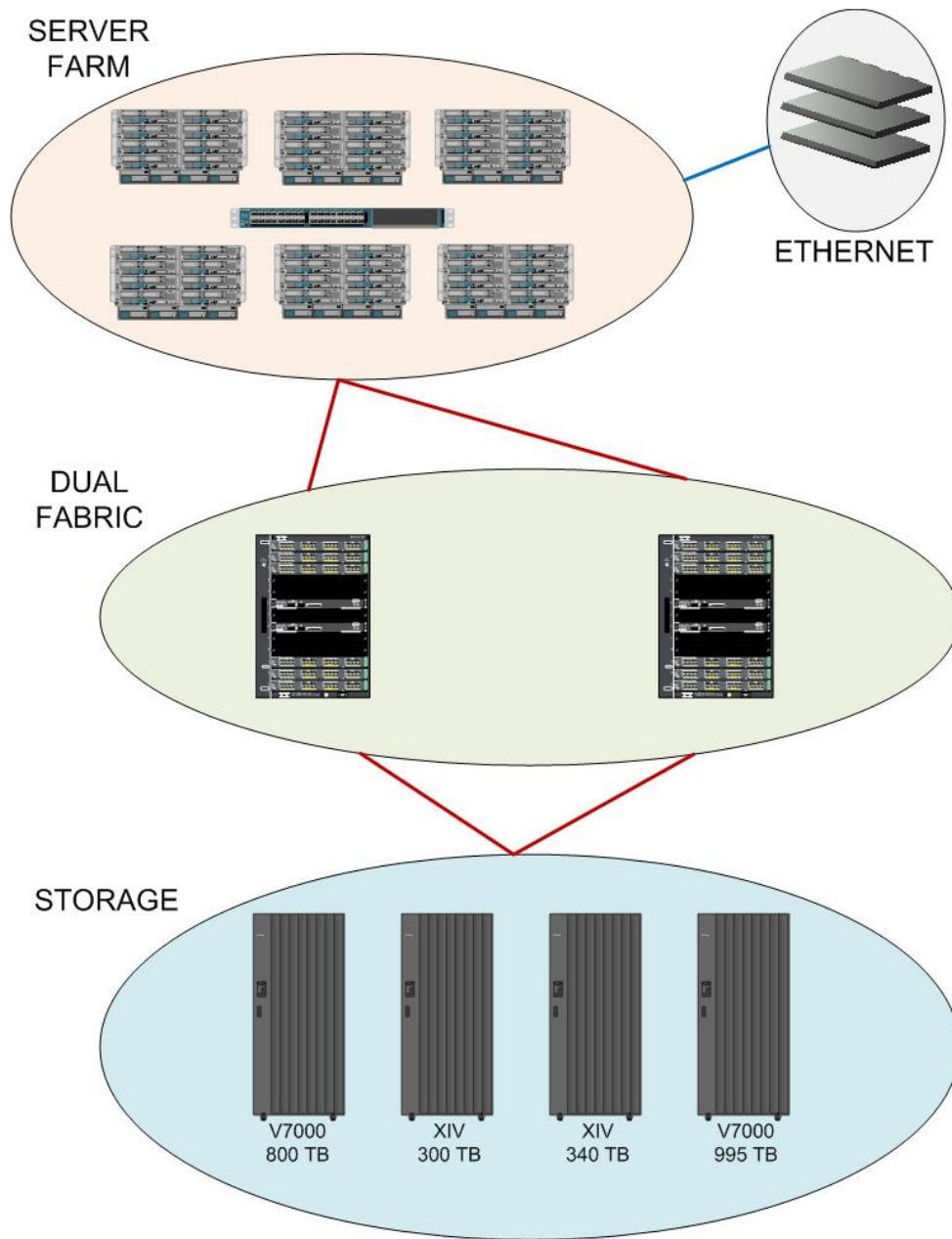


Figure 18 - Existing infrastructure

The entire environment was monitored for one month and the average IOPS was calculated. EMC has a tool called mitrend. The third party storage array monitoring tool was downloaded from the site and installed on the management server. The month long monitor file is then uploaded to the site for analysis. The output analysis list graphically the parameters required for planning. To analyze the servers, EMC grab utility was installed on the server. The grab utility gives a formatted output with all the server details.

Disk details, HBA WWN's and the capacity provisioned as well as the capacity used is listed. The database applications running on Windows machines were generating an average operating IOPS of 2800 each. The exchange server for email was running 4000 IOPS. The VM's were found to generate an average of 45 IOPS. The breakup of operating systems is listed in Table 15.

Operating system	VMware ESX 5.0	Windows server 2008
Number of machines	20	7
Virtualization	180 VM's	Not virtualized
IOPS requirement	150 VM's at 45 IOPS each 20 VM's at 2500 IOPS each 10 VM's at 2000 IOPS each	2 servers at 3600 IOPS each 5 servers at 175 IOPS each

Table 15 - Breakup of operating system infrastructure

For the IOPS load in the customer environment, one VMAX 40K array is needed. For lower RTO (recovery time objective), the VMAX 40K is put behind a VPLEX 4-engine cluster. For disaster recovery solution, the optimal solution is to metro the VPLEX to the DR site with two VMAX 20K behind it. Based on the discussion with the customer, the final high level solution for the data center refresh was finalized as shown in Figure 19.

The VMAX 40K at the production site is named VMXP1. The VMAX 20K's at the DR site are named VMXDR1 and VMXDR2. The VPLEX cluster at production site is called VPLP1, and the DR side is called VPLDR1.

6.1.1 Capacity planning

Capacity planning for the case study was individually performed by the author. After looking at the various needs of the customer and the analysis of the current environment, I came out with the number of drives that they need. The drive configuration was then verified and signed off by the storage architect.

The VMAX disk planning is done balancing the IOPS and the capacity requirement. The total IOPS that the environment needs to handle is calculated using Table 15. For the VM environment, the total IOPS needed are 76750, the Windows server needs to handle 8075. These are the peak IOPS in the environment considering all the applications are at full load, a case very rarely seen.

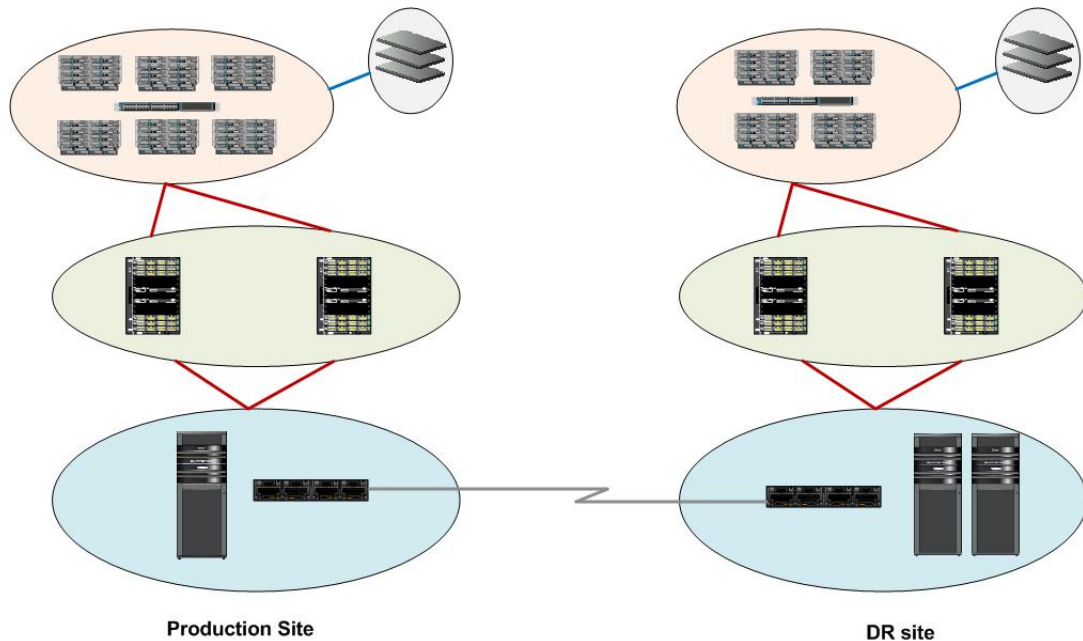


Figure 19 - Refreshed architecture

The drive distribution profile is planned to be 60% SATA, 30% FC and 10% EFD. The total capacity to plan for is 2435 TB. On analysis of the provisioned capacity, it was realized that the old arrays have been provisioned for 2435 TB, but only 1543 TB is used. Since we are using thin provisioning for the environment we have to plan for the actual data, not the provisioned data.

After discussion regarding capacity planning, a growth of 20% per year is envisioned. Therefore we have to plan for $1543 + 20\%$ TB. That comes out to 1852 TB. The per-tier capacity required comes to: 1112 TB in SATA drives, 556 TB in FC drives and 186 TB in EFD drives. This capacity is formatted capacity. RAID implementations will add space.

The best practice for RAID implementation is RAID-6 for SATA and RAID-5 for EFD and FC. RAID implementation will add capacity overhead. To calculate that overhead we need to figure out the size of the RAID groups and the RAID configuration to go for.

The analysis of the data broadly divided the data into 3 parts - index for database, database data and other data. The 50 TB of index data needs to be on higher performance drives to get maximum performance from the system. RAID-1 implementation on FC drives will give the best performance. The other two types of data are segregated for the purpose of tiering.

To calculate the number of drives for index data, we need to consider the current demand as well as future growth. From the 50 TB allocated for indexing purposes, 28 TB is used. Therefore considering their 20% growth for the next year, we get a final capacity requirement for 33.6 TB. The number of 600GB, 10K fiber channel (FC) drives required can be calculated as:

$$\text{No of disks required} = \frac{\text{Required capacity}}{\text{Capacity of each drive}} * \text{RAID overhead}$$

Thus the number of disks is (58 x RAID overhead). The RAID overhead for RAID-1 is 2. Therefore the total number of drives required is 116 drives.

To calculate the rest of the drives we go back to the distribution: 1112 TB in SATA drives, 556 TB in FC drives and 186 TB in EFD drives. 1112 TB in SATA drives comes to 371 x 3 TB SATA drives. The RAID overhead for RAID-6 (6+2) implementation is 1.25. Therefore the number of SATA drives required is 464 x 3 TB drives. For FC drives, 556 TB requirement comes to 949 x 600 GB drives. For RAID-5 (3+1) implementation, the RAID overhead is 1.25. Thus the total number of drives comes to 1187 x 600 GB drives. For the 186 TB flash drive requirement, we need 477 drives. For RAID-5 (7+1) implementation, the RAID overhead is 1.14. Therefore the number of EFD drives required comes to 544.

The total number of drives that VMAX 40K can support is 2400. The total number of drives we get according to our requirement is 2195 + 116 drives, which comes to a total of 2311 drives.

On the DR site, there is a need to archive critical data on the VMAX. For the DR site, there is a need for around 2500 TB data. Since the site is a DR site with reduced functionality, no EFD drives are planned to be installed. This will reduce the cost. For the DR site, distribution of 70% SATA and 30% FC was decided. After applying the same calculations as the production 40K system, the requirements were finalized as shown in Table 16.

System	VMXP1	VMXDR1	VMXDR2
# of SATA drives	464	365	365
# of FC drives	1303	800	800
# of EFD drives	544	-	-
Total number of drives	2311	1165	1165

Table 16 - Drive distribution

6.2 Comparison and benchmarking solutions

After providing the high level architecture design, the sales team comes out with a quote for all the components. The quote is given to the customer who then compares it to quotes from other companies. When selecting systems, three major parameters are considered: cost/TB, SLA and operational cost.

The cost to buy the system, including all drives, engines etc. is calculated. This is provided in the quote. This cost is then divided by the configured capacity of the system. This gives the cost/TB of the system.

SLA or service level agreement is the guaranteed uptime of the system. This guarantee is given by the manufacturer. The downtime cases considered in SLA is due to hardware faults. If the system suffers downtime due to power failure, disaster at the site or malicious actions by the staff, it is not covered. Some of the common SLA's are listed in Table 17.

SLA	Maximum downtime per year
Four 9's (or 99.99%)	52 minutes, 35 seconds
Five 9's (or 99.999%)	5 min, 15 seconds
Six 9's (or 99.9999%)	31 seconds

Table 17 - Common SLA's

The operational cost is calculated on a per year basis. The operational cost includes maintenance cost, cooling cost, cost to power the system, and the floor space cost per year.

A major part in getting the project was the unique nature of EMC VPLEX array. VPLEX adds a physical storage virtualization layer between the storage and servers. VPLEX supports EMC arrays as well as non-EMC arrays on the back end. This means that the next technology refresh that the customer does will be online, there will be no need for bringing the servers down. The design of the system, created by the storage architect was also found to be more robust than the competition.

Balancing out all the parameters and selecting the best system is done by the customer. Once the system and its architecture are finalized, the project proceeds to the next step.

6.3 Rack and floor planning

A rack is where the physical hardware is located. Rack sizes are standard and the space in racks is divided into rack units. Each rack unit is 1.75 inches. Racks typically come in 19 inch or 23 inch width. 40U, 19 inch rack is the most commonly used rack in data centers.

VMAX systems come pre-stacked in 19 inch 40U racks. The fully configured VMAX 40K - VMXP1 will have 8 racks. Figure 20 shows the VMAX racks without the front bezel. It consists of a system bay containing the engines and standby power supply. The standby power supply in the VMAX system can run for 3 minutes. The SPS does not power the drives, it only powers the engines. This time is used for cache vaulting and shutting down the engines. The system bay is flanked by storage bays. The storage bays have disk array enclosures (DAE). The drives are installed in DAE's. For the customer system we are using two types of DAE. 15 drive 3.5" are used for FC drives. 25 drive 2.5" are used for EFD and SATA.

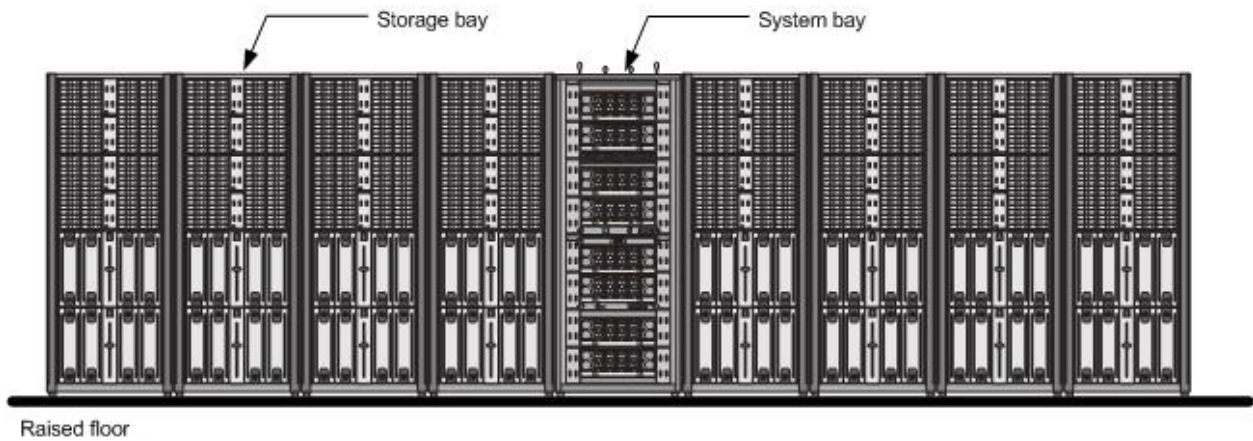


Figure 20 - VMAX rack

We have to plan for stacking VPLEX, Cisco UCS and the Cisco 9513 directors at the DR site. The hardware is installed in standard Cisco 19 inch 42U racks. Figure 21 shows the rack at DR site. The left rack consists of two 2U standby power supplies and four 6U Cisco UCS 5108 chassis in a Cisco 19 inch 42U rack. The center rack consists of two 2U standby power supplies and a 14U Cisco 9513 director switch in a Cisco 19 inch 42U rack. The right rack is for the four engine VPLEX node. This comes racked in a standard EMC 19 inch 40U rack.

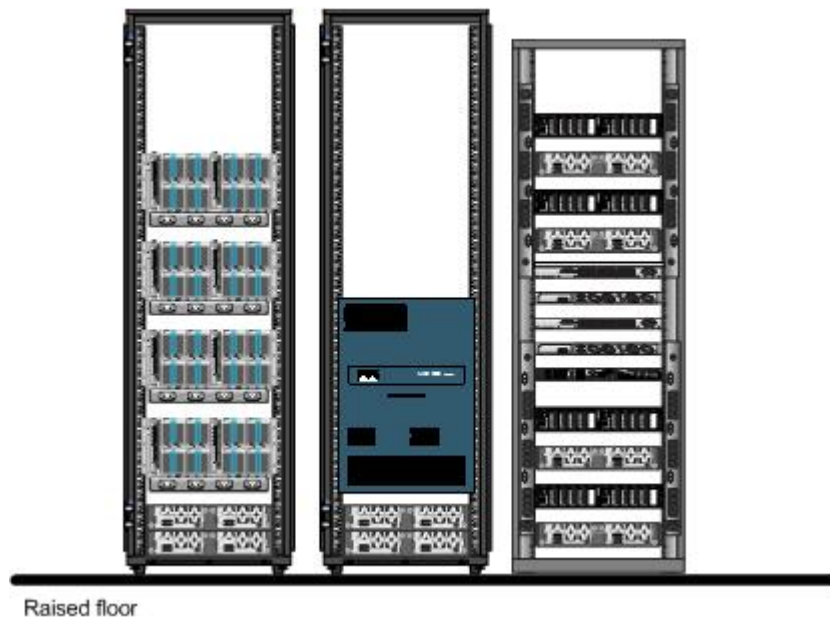


Figure 21 - Rack at DR site

Floor planning at the customer site is the most critical aspect of the implementation. When planning the space where the rack is being installed things like power, cooling, length of cables, etc have to be taken into account. Floor planning has to be done by the customer.

6.4 Implementation

After internal discussion, the implementation project was divided into four major parts: VMAX implementation, VPLEX implementation, UCS implementation and migration. As an implementation engineer I was entirely responsible for installation of the VMAX, VPLEX and the migrations. Cisco UCS had its own implementation engineer sent by Cisco. I was responsible for coordinating with the Cisco UCS engineer and was the point person of the implementation of case study for all future steps.

6.4.1 VMAX 20K/40K implementation

Before the VMAX system is brought online for configuration, the symmetrix BIN file needs to be loaded onto the box. The BIN file contains engine details and details of each drive. The BIN file is loaded onto the box by an EMC customer engineer using a tool called SYMMWIN. During the BIN file install, ESRS is also installed.

ESRS or EMC secure remote support is an application that is installed on a ESXi server. This application opens a secure connection between the data center and EMC. ESRS continuously monitors the environment and automatically contacts EMC support if particular problems are detected. This saves time and can avoid potential downtime or performance hits. ESRS needs to be installed at the customer site as the earlier environment is non-EMC.

After the BIN file is loaded, the VMAX array is zoned to the management server and the unisphere server. Usually these are located on the same hardware box. When zoning is complete, SYMCLI is installed on the management server. SYMCLI grants CLI access to the VMAX array. The licenses for features the customer is going to use are activated.

Once everything is setup, we start the implementation. Data devices are created out of the physical drives. The data devices are where the actual data is stored. Each thin device bound to a pool stripes data

across all the data devices in that pool. These data devices are then added to thin pools. VMAX does not support mixed thin pools. We create three different pools: SATA pool, FC pool and EFD pool.

After the thin pools are configured, FAST or federated automatic storage tiering is configured. FAST allows the data on LUN's to move between pools. This means that when the data is accessed regularly, it is bumped up to faster drives. In virtual provisioned environment, FAST works on sub-LUN extent level. The size of a single extent is 12kb.

6.4.2 VPLEX implementation

For the initial VPLEX setup, physical access to the rack is needed. Serial connection to the VPLEX management server is used to configure remote access port. This port connects the management server to the customer network. After the port is configured, we can access VPLEX from anywhere in the customers management subnet.

The backend connections between the VPLEX and the VMAX are zoned. Care is taken that no director exceeds 4 maximum active paths to VMAX. Four 80 GB volumes are created on the VMAX for metadata information and a 15 TB volume is created for logging.

Before further configuration of VPLEX, the latest GeoSynchrony code is checked. If the hardware is not running the latest code, it needs to be upgraded. The automated system setup for VPLEX is then run from the VplexCLI. Once VPLEX at both sites are up and running, the clusters are joined to create a metro cluster.

6.4.3 Migration

Migration of data from the old array to the new array is the most critical and time consuming aspect of any implementation. Before any migration, the entire process is planned in detail. Broadly the steps taken are: pre-migration planning, test environment migration, production environment migration, and post-migration.

Pre migration planning consists of listing detailed LUN information. All the LUN parameters like size, mirror configuration, etc is pulled from the old array on a per server basis. All the data is plugged

into a spreadsheet. The server information is then pulled. This information includes the WWN of the HBA's on the server, HBA firmware/driver details, hardware details, and operating system details. The operating system and HBA firmware/driver versions are matched with the support matrix for VPLEX. If the version is not supported, it is upgraded to a supported version. Zoning is then done between the old array and VPLEX and LUN's created on VMAX are presented and claimed on VPLEX.

After all the versions are verified, the old LUN's are encapsulated in VPLEX as shown in Figure 22. For the test environment, downtime is taken for the migration process. On the host level, 30 minutes of downtime is enough. To be on the safe side, 2 hours of downtime is approved. During downtime, the server is zoned to VPLEX and the encapsulated LUN is presented to the server as shown in Figure 23. The server is brought up online. The IBM multi-pathing software is uninstalled and EMC multi-pathing software (PowerPath) is installed. After the server is online, VPLEX data mobility is used to migrate data from the encapsulated old LUN to the new LUN. This process is online and invisible to the host.

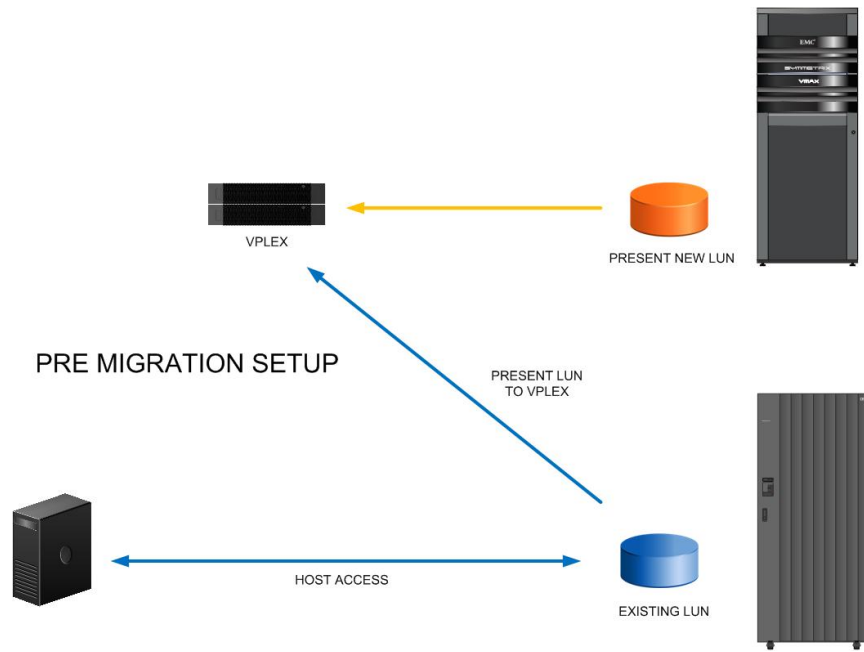


Figure 22 - Pre migraiton

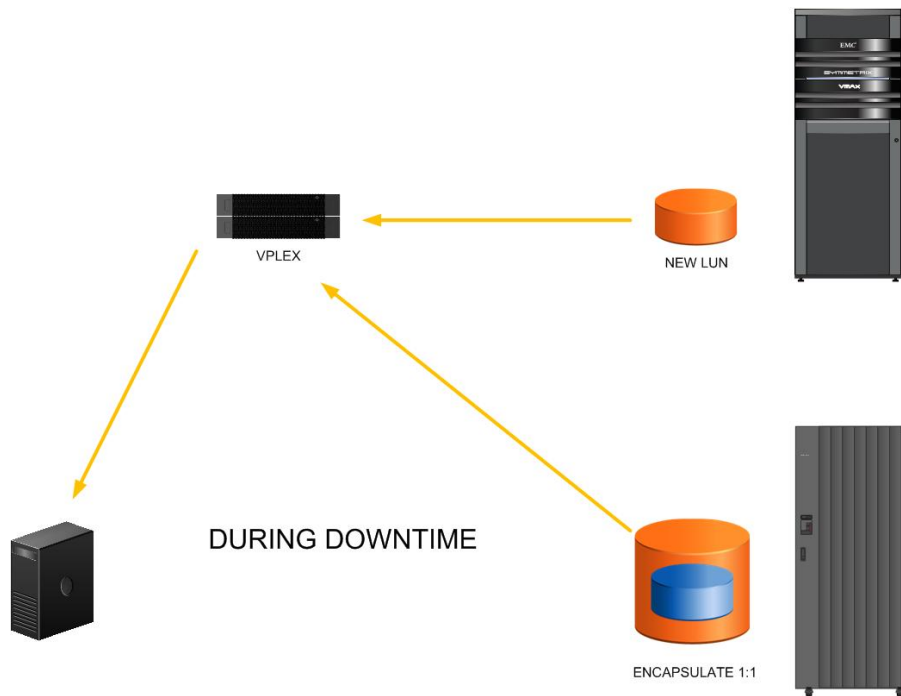


Figure 23 - During cutoff

This process is done for the windows servers in the environment. For the VMware servers in the environment, downtime is not required. PowerPath/VE is installed for ESXi servers. The actual migration is done online using the sVmotion VMware functionality. The VPLEX LUN is presented to vCenter and

claimed. The datastore now contains the old LUN as well as the new LUN. sVmotion is then used to transfer the old datastore to the new datastore online, without any downtime as shown in Figure 24.

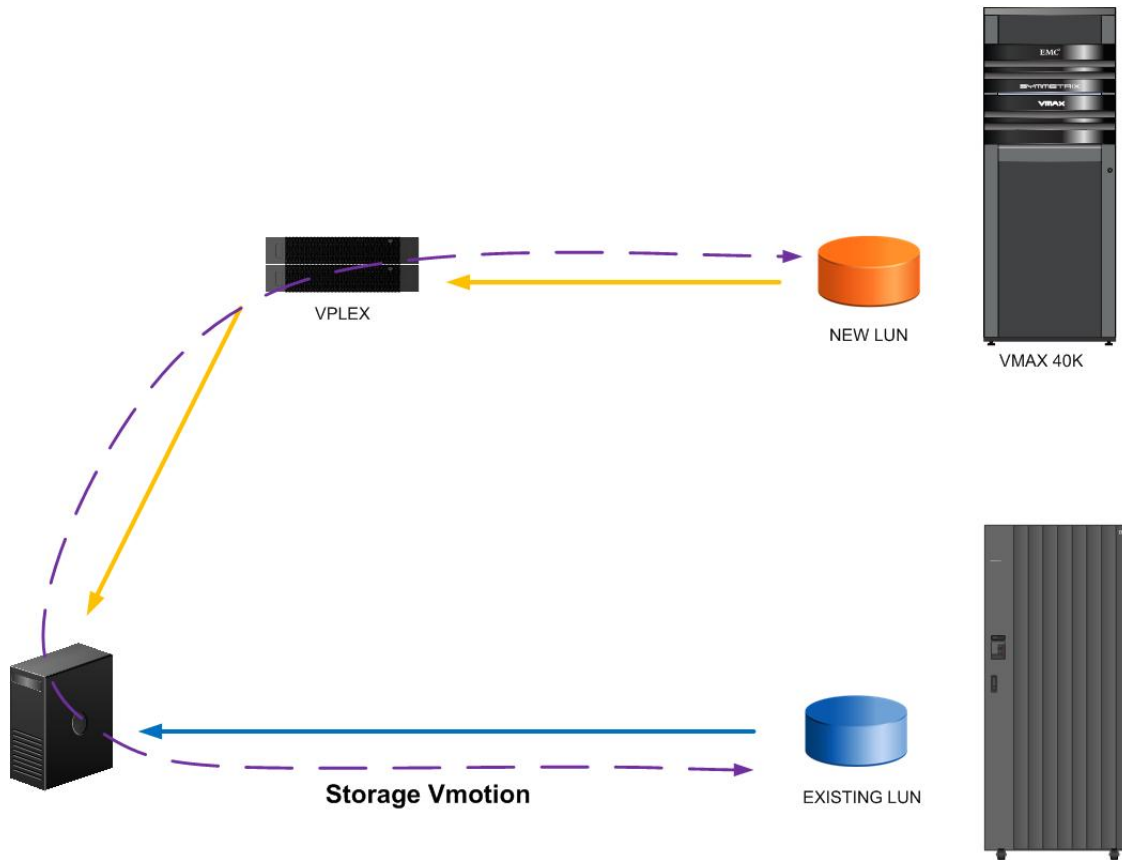


Figure 24 - Migration using sVmotion

After the test environment is migrated to VMAX, the environment is strained to the breaking point. This process reveals any bottlenecks that are present in the new environment. This process also gives the loaded response times. Using the information gathered from testing, bottlenecks are minimized and performance tuning is performed. After the customer is satisfied with the test environment performance, the actual production environment is migrated to the new array.

Post migration activities, performed for every server, include re-establishment of mirrors and backup schedule, reclaiming old storage and cleanup of old zones.

6.5 Post implementation

Post migration process involves support and knowledge transfer to the customer site. The environment is closely monitored for performance issues after migration. The volumes on VPLEX are converted to distributed volumes and the DR site is activated. Disaster scenarios are tested and DR functionality is verified.

The entire site is documented. All the installation parameters and LUN information is compiled and handed over. Best practices and common tasks are taught to the customer, documented and handed over.

7 Conclusion

In this case study I was involved with pre sales architecture and was entirely responsible for the implementation and migration activities. The end goal of this case study was to implement and migrate data off the old arrays to the new arrays within 4 months. I managed to implement and migrate the entire data to their new arrays within 2 months.

The main lesson learned from this case study is the importance of detailed planning and documentation. Planning activities down to the last detail helped to reduce any last minute surprises. Due to efficient planning, I was able to migrate the data and complete the implementation for this case study well before the deadline.

Documentation also plays an equally large part in successful implementation. The configuration changes and the best practices need to be thoroughly documented. Run-books have to be created for common jobs. Run-books list out the steps, commands and considerations to be followed for common SAN administration tasks. This ensures smooth operation of the environment without any major performance hit.

The drive technology is changing every year. New drive types and hardware are flooding the market. In recent years there is a push for flash storage. More all-flash arrays have been introduced and more are

in development. Flash drive development has accelerated also. Larger and cheaper flash drives are being introduced in the market.

For near-line storage, larger capacity drives are coming into the market. Recently drives up to 6 TB were introduced. Installing these drives into the array will require a simple code upgrade. One issue with the larger drives is the RAID rebuild times. Currently a 4 TB SATA drive takes around 20-25 hours to rebuild. If another drive fails during that time, the entire data on the RAID group is lost. Faster RAID controllers can mitigate the losses and reduce the RAID rebuild time.

The biggest change coming into SAN is virtualization. All companies manufacturing storage arrays are incorporating more and more virtual infrastructure into their arrays. Some arrays are even in the process of being offered in physical form with better SLA or in virtual form with slightly lower SLA. EMC is in the process of coming out with a virtual VPLEX appliance. Recently, they also announced new software defined storage. In Q3, EMC is releasing a new VMAX array. The new VMAX array has the capability to run virtual RecoverPoint and virtual VPLEX on the array itself. This bypasses the need to designate dedicated resources from the production VM pool.

With the advent of big data, the amount of block storage is reducing. Big data file systems like HDFS (hadoop) are being offered on NAS platforms alongside traditional CIFS and NFS. Currently the amount of file level storage is growing exponentially while block storage is growing at a constant rate. In storage networks, 2011 marked a milestone. In 2011, the amount of file storage crossed 50%.

With data growing at an exponential rate, storage networks have a bright future. All companies, big or small need to store data securely. Therefore more datacenters, small and large are being commissioned regularly. Storage networks will continue to grow in the future.

For future study in the field of storage networks, I will suggest a look into software defined SAN as well as the various virtualization technologies. Software defined SAN incorporates the functionality of SAN into DAS. The storage network is a pooled, virtual network between servers with direct attached storage. It does not give the redundancy and performance of a physical SAN, and is rarely used on the

enterprise level. Software defined SAN have a future in small environments where the cost of deploying a physical storage network is not justified over having the same functionality at a reduced SLA.

8. References

- [1] Phillips, B. (1998). "Have storage area networks come of age?" *Computer*, 31(7), 10-12.
- [2] Chidlow, S. (2003). "Storage area networks." *JISC Technology and Standards Watch report 03_07*. JISC: Bristol, UK. Available at: <http://www.jisc.ac.uk/whatwedo/services/techwatch/reports/horizonsscanning/hs0307.aspx>.
- [3] Thornburgh, R. H., & Schoenborn, B. (2000). "Storage Area Networks", Prentice Hall PTR.
- [4] "Brocade Certified Fabric Administrator Gen 5 Training," (2013). Retrieved , from <http://www.brocade.com/education/certification-accreditation/certified-fabric-administrator-gen-5/index>.
- [5] "Brocade Accredited FCoE Specialist Training" (2013). Retrieved from <http://www.brocade.com/education/certification-accreditation/accredited-FCoE-specialist/index.page>
- [6] "Brocade SAN Product documents" (2014). Retrieved , from <http://www.brocade.com/products/products-index.page>

- [7] “Brocade knowledge base” (2014). . Retrieved , from
<http://kb.brocade.com/kb/index?page=home>
- [8] Cisco product documentation (2014). Retrieved from
<http://www.cisco.com/c/en/us/products/storage-networking/index.html>
- [9] Cisco FC101 Training. (2014). Retrieved from
http://www.cisco.com/web/learning/le31/learning_learning_resources_home.html
- [10] EMC training: Symmetrix/VMAX (2014). Retrieved from *<http://www.education.emc.com>*
- [11] EMC training: VNX (2014). Retrieved from *<http://www.education.emc.com>*
- [12] EMC training: VPLEX (2014). Retrieved from *<http://www.education.emc.com>*
- [13] EMC training: Introduction to storage elements (2014). Retrieved from
<http://www.education.emc.com>
- [14] Symmetrix VMAX Series: Best practices for choosing the right protection type (2013). .
Retrieved from *<https://support.emc.com/kb/170635>*
- [15] Best practices for Symmetrix configuration. (2013). Retrieved from
<https://support.emc.com/kb/174725>
- [16] Design and Implementation Best Practices for EMC Symmetrix Federated Tiered Storage
(2014). Retrieved from
https://support.emc.com/docu40408_Design_and_Implementation_Best_Practices_for_EMCSymmetrix_Federated_Tiered_Storage.pdf
- [17] VPLEX Product Guide (2014). Retrieved from
https://support.emc.com/docu52652_VPLEX_Product_Guide.pdf
- [18] White Paper: EMC VPLEX Elements of Performance and Testing Best Practices Defined
(2012). Retrieved from
https://support.emc.com/docu44811_White_Paper:_EMC_VPLEX_Elements_of_Performance_and_Testing_Best_Practices_Defined.pdf
- [19] EMC Mitrend tool. (2014). Retrieved from *<https://emc.mitrend.com/portal/displayHome.do>*

- [20] HP product documentation (2014). Retrieved <http://www8.hp.com/us/en/products/disk-storage/index.html>
- [21] IBM product documentation (2014). Retrieved <http://www-03.ibm.com/systems/storage/san/>
- [22] Gibson, G. A., & Van Meter, R. (2000). "Network attached storage architecture", *Communications of the ACM*, 43(11), 37-45.
- [23] Farley, M. (2001) "Building storage networks", McGraw-Hill Professional.