Dell EqualLogic Best Practices Series

EqualLogic iSCSI SAN Concepts for the Experienced Fibre Channel Storage Professional

A Dell Technical Whitepaper

Storage Infrastructure and Solutions Engineering

Dell Product Group

January 2012

Contents

Acknowledgements

This whitepaper was produced by the PG Storage Infrastructure and Solutions of Dell Inc.

The team that created this whitepaper:

Mike Kosacek, Tony Ansley and Camille Daily

We would like to thank the following Dell team members for providing significant support during development and review:

Hector Valenzuela, Richard Hou, and Richard Golasky

Feedback

We encourage readers of this publication to provide feedback on the quality and usefulness of this information by sending an email to SISfeedback@Dell.com.

SISfeedback@Dell.com

1 Overview

As data storage has become central to any successful computer based solution, advanced technologies for connecting computing resources to data have matured over the last four decades. From early storage connection protocols that utilized parallel data transmission such as SCSI and ATA that connected storage devices directly to individual computing resources through today's modern distributed storage networking solutions, everything has revolved around one key goal: provide access to data in ever faster, more secure, and less expensive ways.

Storage networks provide several unique benefits over direct attached storage including:

- Increased throughput
- Improved flexibility
- Greater scalability
- Centralized management
- Multi-host clustering

While storage networking – the use of a shared storage and communications infrastructure – has been around for some time, it became increasingly popular with the advent of the Fibre Channel (FC) standard in the early to mid-1990s. Since then FC has been the predominant shared storage network protocol in use… that is, until the IP SCSI (iSCSI) storage networking protocol was developed in the early 2000's.

The FC and iSCSI protocols both use an updated version of the original SCSI protocol adopted in the mid 1980's for directly attaching storage devices to individual computing resources. Even so, both Storage Area Networking (SAN) standards use different methods of transmitting these SCSI commands between the computing resources and the storage resources. This paper provides an overview of the iSCSI protocol and concepts, as they pertain to the Dell EqualLogic PS Series of virtual storage arrays, for storage administrators with FC storage experience.

Fibre Channel over Ethernet (FCoE) is another storage protocol for carrying SCSI commands, and has features similar to both FC and iSCSI. However, FCoE requires additional features such as Data Center Bridging (DCB) and 10 Gb Ethernet to properly support it and detailed discussion of these concepts are outside of the scope of this paper. This paper will focus on the similarities and differences between FC and EqualLogic iSCSI SAN concepts.

While some of the concepts are similar for both FC and iSCSI protocols, there are differences in terminology and capability that must be understood to successfully design and manage an iSCSI solution such as the Dell EqualLogic PS Series SAN. The goal of this paper is not necessarily to convince anyone that is well entrenched in a FC storage environment into migrating to iSCSI, but simply to help those that are considering the move, or have already made that decision, and to enable a smooth adoption of EqualLogic iSCSI.

2 Audience

This paper is intended for the experienced storage networking administrator or SAN architect who is planning to purchase or already has purchased a Dell EqualLogic iSCSI storage solution. This paper assumes that the reader has prior experience in, and a thorough understanding of traditional FC storage area network architecture and its related concepts. For further definition of EqualLogic-specific terms, see Appendix A.

3 Fibre Channel and iSCSI basic comparison

The roots of both the FC and iSCSI protocols are the same – SCSI. Both protocols were developed to improve upon this well-established standard, and to allow it to work with modern hardware technologies. Whereas SCSI was designed as a parallel, bus architecture, FC and iSCSI both transmit data serially in frames or packets. FC and iSCSI both encapsulate SCSI commands in these frames or packets to retain a level of backward compatibility with operating systems and applications that rely on block-based storage.

Both FC and iSCSI protocols are block-based storage protocols and both allow connection of hosts and storage devices together, to create a Storage Area Network or SAN. The SAN allows administrators to more effectively manage and share storage resources between various systems, instead of creating "islands" of storage that are dedicated to a single host.

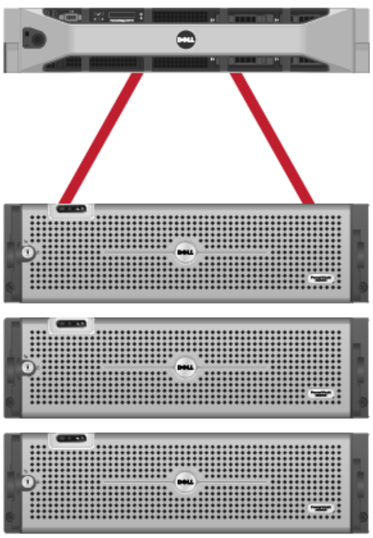

Figure 1 - Traditional Direct Attach Storage Island

FC has internal mechanisms for automatic addressing and discovery, whereas iSCSI relies on TCP/IP addressing and iSCSI has its own methods for discovery.

FC requires a dedicated switch fabric (switch or switches) that understands the FC protocol and how to transport FC frames to and from the target. Host Bus Adapters (HBAs) and storage device ports are connected to the fabric via approved copper or optical cabling. Today FC supports speeds up to 8 Gb/sec, with plans to increase this speed in future versions.

iSCSI, on the other hand, leverages or builds upon Ethernet and TCP/IP standards –and therefore can make use of existing 1 Gb/sec and 10 Gb/sec Ethernet hardware technologies. In addition to dedicated HBAs that support iSCSI offload, many standard Ethernet network interface cards (NICs) can be used with a software iSCSI initiator to provide the same functionality. However, it is important to understand that not all Ethernet switches and NICs are created equal, and while most will provide the basic functionality required for connectivity, not all will provide the performance required. This will be discussed in more detail later in the paper.

4 Fabric architecture comparison

4.1 Network congestion and prevention

FC utilizes a credit or token method of dealing with congestion, or more accurately of preventing congestion altogether. When end points (hosts and storage) initiate communication, they exchange operating parameters to inform each other about their capabilities and decide how much data they can send to each other. Each device can only send data if it has "credit". In simpler terms, the receiving device must have enough buffer space to store the data from the sender. This system ensures that FC frames are not typically dropped in normal operation.

Ethernet relies on congestion avoidance algorithms built into TCP/IP and Flow Control capabilities between switches and end points. Without Flow Control enabled, one or more devices could cause another to start dropping packets when its buffer is filled or near full depending on implementation. With Flow Control enabled, PAUSE frames are sent to manage packet flow rate in hopes of avoiding congestion or the overload of a switch port or end device. However, some packets may still be dropped and when packets are dropped, TCP/IP will attempt to resend or re-transmit those missing packets or packet segments. Occasional retransmissions are normal for TCP/IP; however continual retransmitting can have a significant impact on performance.

For Flow Control to function properly, all devices on the network (or SAN) must enable and support it. Otherwise, packets may be dropped or ignored by that device. EqualLogic PS Series array ports are always automatically configured to negotiate flow control. Switches and host ports may require manual configuration depending on the device.

Another recent technology for preventing congestion on Ethernet networks is the use of Data Center Bridging (DCB). DCB introduces Priority Flow Control (PFC) and Enhanced Transmission Selection (ETS) to help create a "loss-less" Ethernet environment. DCB does require hardware support so it is most often found in newer models of 10 Gb Ethernet switches. FC over Ethernet (FCoE) devices require features found in DCB to function properly. For proper support of iSCSI devices, an iSCSI Type-Length Value (TLV) is required to allow the device to identify and separate iSCSI packets. While an in-depth discussion of DCB is out of the scope of this paper, detailed information can be found in *Creating a DCB Compliant EqualLogic iSCSI SAN with Mixed Traffic*, which is available on delltechcenter.com.

http://www.delltechcenter.com/page/Creating+a+DCB+Compliant+EqualLogic+iSCSI+SAN+with+Mixed+Traffic

4.2 Packet and frame size

A typical FC frame is around 2148 bytes (up to 2112 bytes in the data field). Ethernet packets are typically around 1518 bytes; however, if "Jumbo Frames" are configured, that allows for a MTU (Maximum Transmission Unit) of up to 9000 bytes. Depending on the device, the Jumbo Frame size may be 9000 or slightly higher, usually depending on whether the packet header information is included in that value. If Jumbo Frames are enabled, all devices in the path must support Jumbo Frames, or the packets may be dropped or discarded. Check the specifications on each device for the manufacturer's recommendation.

EqualLogic PS Series array will always attempt to communicate using Jumbo Frames, however if the attempt to create an iSCSI session fails, then it will revert to standard MTU and try the connection again. Connections for replication sessions are independent of other SAN activity, so if the SAN is capable of Jumbo Frames, but WAN connectivity for replication is not, the storage will automatically use the appropriate frame size for each connection. Dell recommends the use of Jumbo Frames for optimal SAN performance.

4.3 SAN topologies and interconnects

FC switches are typically arranged in dual-fabric designs, where a pair of switches is used to make two independent storage area networks that provide redundant and usually load-balanced connections to hosts and storage arrays. If switches are connected together (such as to expand the number of available ports in a fabric), then an ISL (Inter-Switch Link) is created on Expansion-Ports (E-ports) between the switches. Usually an ISL consists of two or more links to provide for redundancy and added bandwidth. Fabric Shortest Path First (FSPF) accommodates multiple switch environments and selects optimal paths based on bandwidth and utilization.

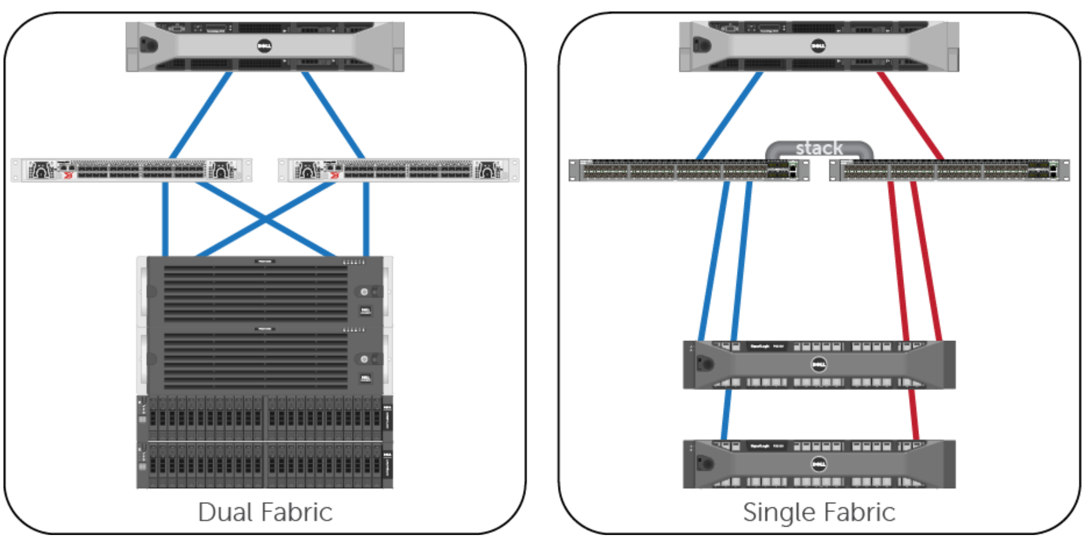

Figure 2 - Dual vs. Single Fabric Design

All network ports of the EqualLogic array must be able to communicate with every other array that is part of the same group, therefore they must all reside on the same (Layer 2) network. The simplest example of this is a single Ethernet switch, however to provide redundancy, a minimum of two switches are usually recommended (Figure 2 – Single Fabric). However, when two or more switches are used they still must operate as a single flat network and not as a dual-fabric design as found with many FC SANs (Figure 2 – Dual Fabric).

Some Ethernet switches have stacking abilities, where a dedicated (usually proprietary) interface is connected between two or more switches, allowing them to be managed as a single switch and also to forward traffic as if it was a single switch. Non-stackable switches can be connected by configuring standard ports or higher-speed uplink ports as a Link Aggregation Group (LAG) – although sometimes these are also referred to as ISL's. Some switches also offer virtual-chassis features to allow linking several switches together to operate as one. Of course high-end chassis based switches use a high-speed switching backplane to connect line-cards or blade modules together, and allow expansion by adding more cards or modules to the chassis.

When links are created between Ethernet switches, whether it be a LAG or a single port, careful attention must be given to Spanning-Tree configuration. When multiple links are detected, Spanning-Tree Protocol (STP) will block or disable one or more links to avoid a loop (where a switch has more than one path to another switch). This design can sometimes cause optimal links to be blocked and even reduce the bandwidth available between switches. Assigning cost to the links allows the network administrator to determine which link or links are blocked by default. Configuring a LAG allows two or more links to be treated as a single link by STP so that the aggregate bandwidth is achieved.

The FC protocol is also confined for local use. That is, it is not capable of routing or travelling over a WAN as is. For FC networks to travel longer (than the maximum of 10 Km) distances, either specialized extenders must be used, or the FC packets must go through a routing or bridging device. FC over IP (FCIP) is another example of how FC is encapsulated in TCP/IP so that it can be used over longer distances.

Since iSCSI natively travels over Ethernet and TCP/IP networks, it does not require specialized hardware; however careful consideration must be placed on architecting any long-distance SAN connection. For example, connecting two sites together for iSCSI (or FC over IP) replication may require significantly more bandwidth than had been used previously by other non-storage WAN activity. Also, network connections that are used for long-distance replication should be selected with latency as low as possible and that higher latency WAN traffic is isolated from the SAN through methods like VLANs, tunnels, firewalls or routing filters.

5 Host adapters/interface card comparison

5.1 FC host bus adapters

In the FC architecture, each host has (preferably) at least two FC HBAs, which provide the hardware and logic to connect the host server to the FC SAN. A number of vendors produce FC HBAs with varying capabilities including speed, number of ports, connector types (optical, copper, etc.), bus interfaces (PCI-X, PCI-e), and buffer memory sizes. The FC HBA contains a task-specific processor responsible for SERDES (serial/de-serial operations) and frame processing, moving data between the host bus and FC connections, and is usually independent of host processors.

5.2 iSCSI network interface cards

iSCSI SANs are connected using either a simple Ethernet NIC (in conjunction with a software iSCSI initiator) or a dedicated iSCSI HBA. If a standard Ethernet NIC is used with a host-based software iSCSI initiator, it shares or borrows host-processor cycles to move data and to process TCP packets. More advanced Ethernet NICs offer offload processing capabilities -- such as TCP Offload Engine (TOE) or iSCSI Offload Engine (iSOE) -- to offload the packet processing from the host. Ethernet cards also vary in their implementation and various models offer larger internal buffers or different connectivity options (copper vs. optical network connections or PCI-e vs. PCI-X host-bus connectivity).

Purpose-built iSCSI HBAs are also available. These are similar to FC HBAs in that they contain a dedicated processor that handles all iSCSI and usually TCP packet processing. Dedicated iSCSI HBAs generally cost more than standard Ethernet NICs, but may offer additional features such as iSCSI boot ROMs that enable boot from SAN capability.

An iSCSI HBA can be selected that is specifically designed for iSCSI traffic, or an Ethernet NIC can be selected and used with a host-based software iSCSI initiator. The *Validated Component List for EqualLogic PS Series SANs* provides detail of which interface cards are tested and proven to work with EQL iSCSI arrays. Dell recommends evaluating these options prior to making an investment in a particular NIC or iSCSI HBA.

Note: The *Validated Component List* can be downloaded from delltechcenter.com: http://www.delltechcenter.com/page/EqualLogic+Validated+Components

6 Storage array architecture comparison

6.1 Fibre Channel arrays

FC storage systems have been around for over a decade and there are numerous choices when selecting a FC environment. While these systems range from large, highly scalable models to smaller FC arrays, many of them are monolithic designs. By that, we mean that they consist of a controller unit (or units) that contain storage processors, cache memory, and front-end (for host/switch connectivity) and back-end interfaces (for disks). The controller unit attaches to a disk enclosure, or usually several disk enclosures.

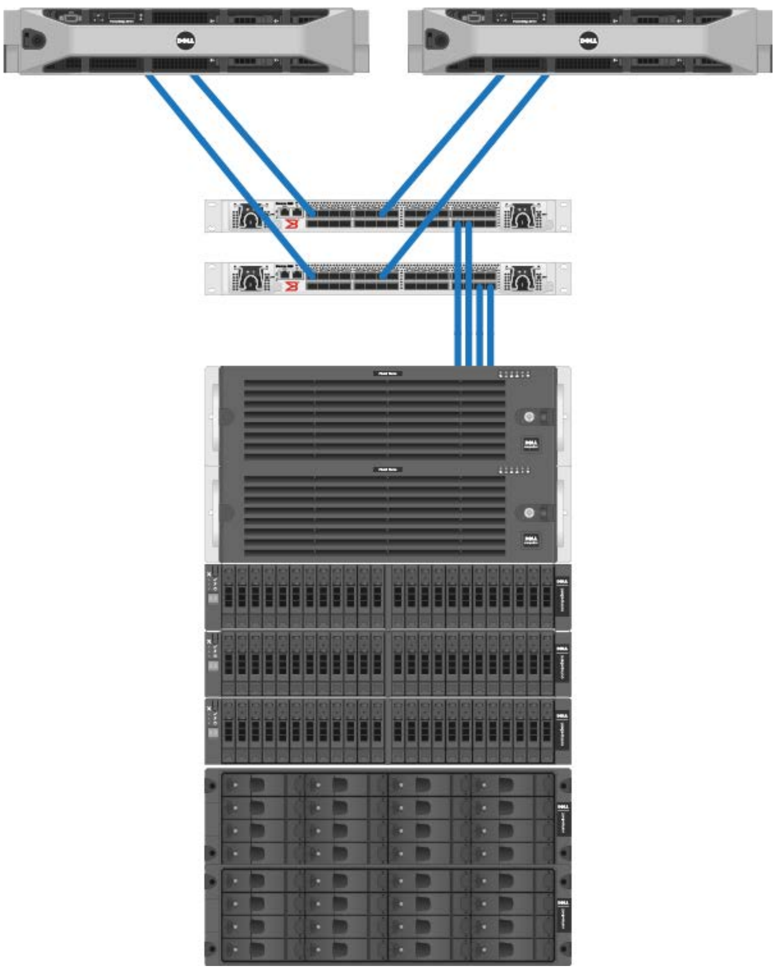

Figure 3 - Monolithic Storage Architecture

When more storage capacity is needed, another disk enclosure is added to the system and new volumes (or LUNs) can be allocated from those drives. If I/O capacity is required, the same thing happens – or at least until the controller unit has reached its maximum potential. Sometimes this limit is based on I/O processing limits (CPU), or there may be limitations on the number of disks an array can support. At this point, a decision must be made to either add another controller unit (and additional drives) or to replace the older unit with a newer, more scalable model, and migrate any existing data.

An FC system may have a mix of drive types (such as FC, SAS, SATA, and SSD disk), or it may only have a single type. Depending on the type of array, it may have multiple ports for front-end and back-end connectivity. Some may even support additional connectivity options (such as SAS, iSCSI, etc.) or NAS features (for support of NFS and CIFS protocols over Ethernet). Typical systems are built around redundant hardware components throughout and cache memory that is protected against power loss.

Many FC storage systems traditionally require the administrator to make choices about disk configuration and placement. For example, a group of disks may be configured as part of a RAID set, a RAID policy assigned to the set of disks, and then logical volumes are created and assigned to hosts. Mapping of a volume is usually made as a logical unit (or LUN) that is presented by the controller units or targets.

6.2 EqualLogic PS Series arrays

An EqualLogic PS Series array uses the iSCSI protocol over standard Ethernet switching. While it is technically possible to deploy an EqualLogic storage solution on an existing Local Area Network (LAN), it is recommended to isolate iSCSI storage traffic to a dedicated storage network. Creating a dedicated iSCSI SAN network, with its own dedicated bandwidth, allows for more deterministic performance similar to that of a dedicated FC SAN.

Like most FC storage systems, EqualLogic PS Series arrays also contain redundant power and controllers, and protected cache memory. However, one of the key differences in the EqualLogic architecture is its ability to scale both capacity and performance as additional arrays are added. Each PS Series array contains disk, cache memory, network controller ports, and CPUs so when the storage system is expanded, adding another array adds more of each of these resources. The "PS" stands for "Peer Storage", which is a reference to how the members communicate with one another.

As arrays are added to the group, data is automatically spread across multiple members of the storage group. A typical volume may extend across 1-3 members in a PS Series group. Mapping of volumes is virtualized, and each volume is presented to the host as a separate target with a single LUN (LUN0) that is the volume.

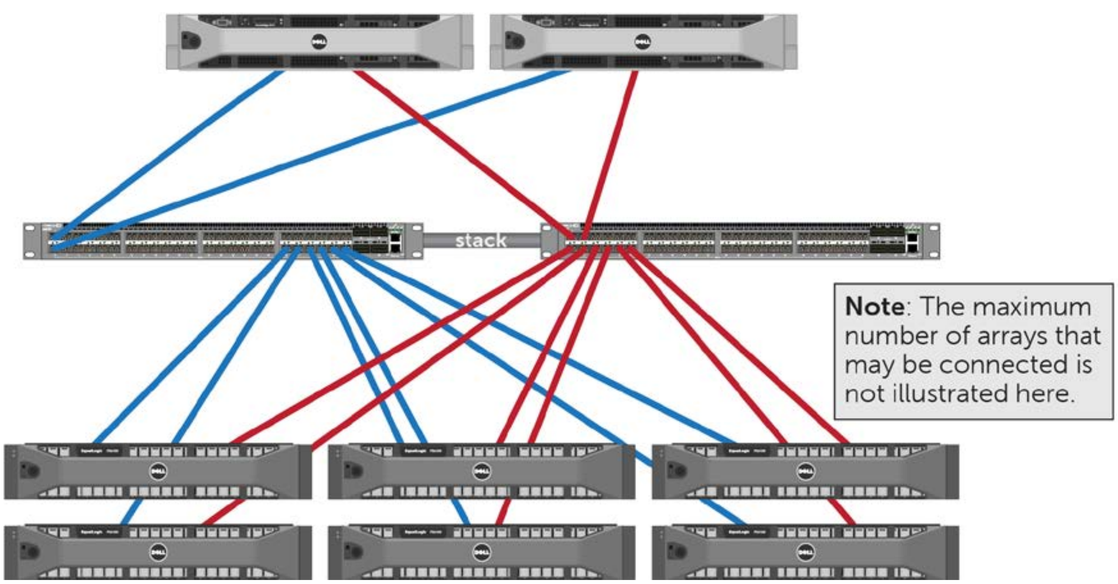

Figure 4 - EqualLogic scale-out architecture

With EqualLogic PS Series storage each array contains a number of disks that are usually of the same type – one exception being "hybrid" arrays which contain a mix of SAS and SSD disks. When an array is added to a pool, all of the disks in that array are consumed and automatically added to the pool as available space. When a RAID policy is assigned, it is assigned to all disks in that array (enclosure) except for those automatically designated as hot spares. The pool is sub-divided into "pages" and when volumes are created, a number of pages are allocated or reserved for that volume until the capacity is met. A pool may contain up to 8 members (arrays) which may all be configured for the same RAID level, or the pool may contain multiple RAID levels. A group may contain up to 16 arrays total.

Note: A minimum of two arrays are required in a pool to allow for multiple RAID levels to be used in the pool.

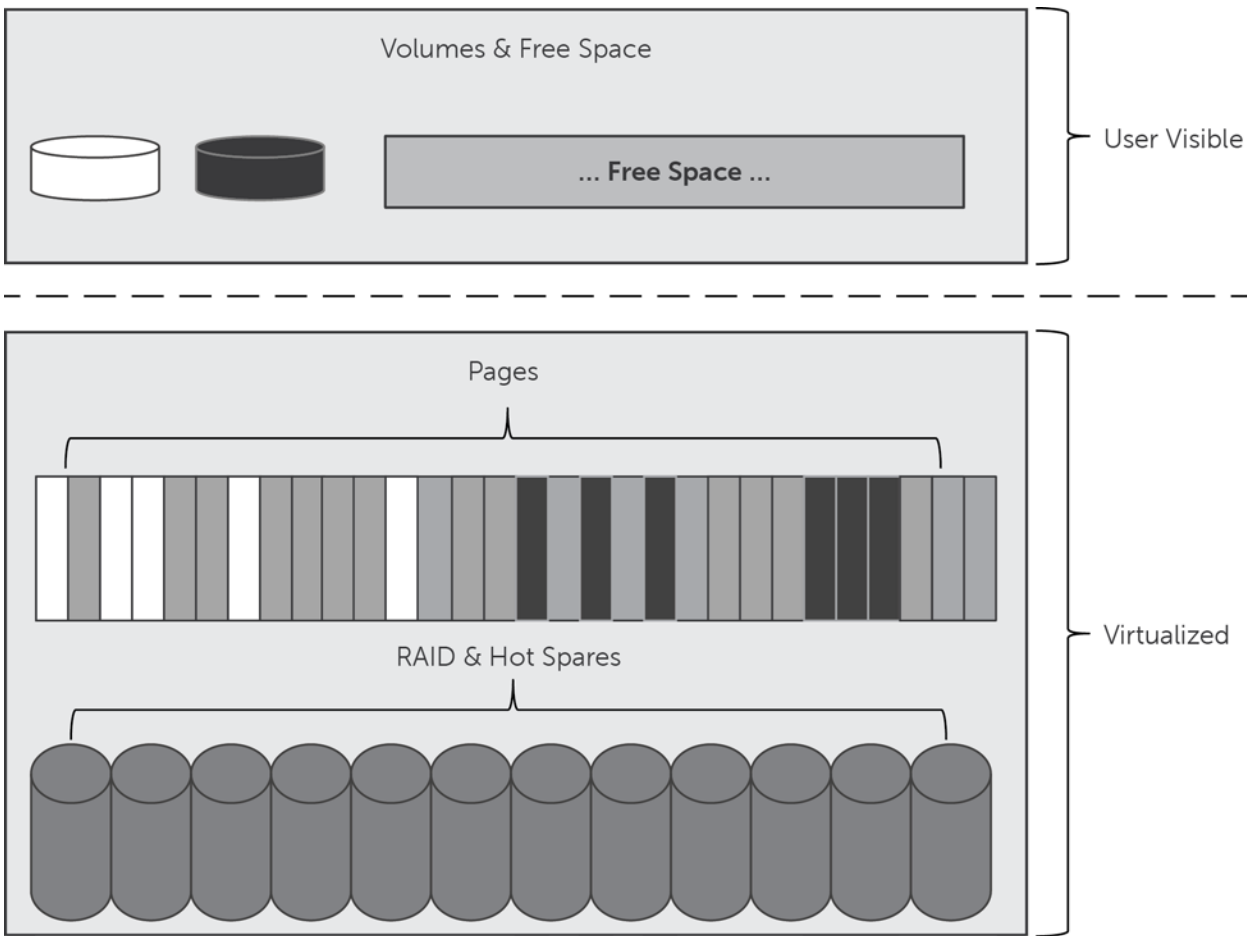

Figure 5 - How EqualLogic storage is virtualized

While many storage administrators may be accustomed to determining exact placement of volumes on physical disks (through the use of RAID groups or sets), the virtualized architecture of EqualLogic PS Series arrays can offer advantages over other designs. In the case where a pool contains multiple RAID policies, automatic tiering of data will occur over time, placing more frequently accessed data on faster disks and/or RAID types.   When an additional array is added to a pool, volumes may automatically redistribute across multiple members, which can lead to higher performance. Administrators can also easily migrate volumes from one pool to another without downtime.

From a host perspective, a group of EqualLogic arrays appears as a single storage system. A built-in network load balancer transparently redirects iSCSI requests to the proper member as needed. However, a host-side software component is also available to automatically manage and optimize iSCSI sessions. The EqualLogic Host Integration Tools (HIT) kit for Windows or Linux contains a host-side service or process that allows it to quickly reference which member contains the portion of data or “slice” of the volume that is needed for an I/O operation. This reduces the need for iSCSI redirections by array members and improves overall performance. For VMware, the same functionality is provided in the Multipathing Extension Module (MEM) which is a Path Selection Plugin (PSP) for the VMware Pluggable Storage Architecture (PSA). The HIT kit (or MEM) also automatically creates the optimal number of connections to each volume once the initial connection is made.

7 Security concept comparison

Security within the concepts of data storage provides benefits including:

- Ensuring that hostile hosts do not gain access to stored sensitive data
- Limiting access to storage devices to only the devices that are designated and allocated to each host (i.e., LUN masking and mapping)
- Because Ethernet switches are common to both LAN and iSCSI SAN connectivity, it is tempting to put them all together on the same fabric. However, when SANs are physically isolated from the LAN or WAN, they are inherently more secure. Of course the other aspect of physical security simply involves securing the SAN (such as in a data center) where unauthorized personnel do not have physical access to it and this is no different for FC or iSCSI SANs.

Beyond the basic physical security that is necessary for most every SAN deployment, there are a few other features that should be considered. Zoning on an FC switch restricts access to a device by World Wide Name (WWN) or physical port. While zoning is configured on the switch, some hosts and storage arrays can also use WWN masking to prevent unauthorized or accidental access to a device.

Ethernet devices also contain a unique Media Access Control (MAC) address which is unique to each device, and usually burned into an EEPROM. A MAC address is the Ethernet equivalent to the FC WWN and identifies each unique network port on a switch, NIC, or array port. However, a MAC address is not always part of the Access Control List (ACL) in iSCSI environments and that is true for EqualLogic PS Series storage as well.

For iSCSI targets and initiators, the iSCSI Qualified Name (IQN) can be assigned as a unique value to identify initiators and targets. EqualLogic PS Series arrays assign unique IQN names to every volume and can use these values to limit access to volumes by specific initiators which will also have unique IQN values assigned. The iSCSI protocol also provides for Common Handshake Authentication Protocol (CHAP). EqualLogic PS Series arrays support CHAP and it can be used exclusively, or in conjunction with other ACL methods (such as IQN or IP address filtering) to control access to volumes.

Ethernet switches are not aware of IQN names or CHAP logins and therefore do not provide any filtering via the switch hardware like is done with FC zoning. Ethernet does contain a similar feature however in the Virtual Local Area Network (VLAN). Besides isolating certain types of traffic on the switch, a VLAN can also be used to prevent access to unauthorized or accidental connections to a switch.

8 Storage management

Many traditional enterprise storage systems utilize licensing tiers or add-ons to enable advanced features such as snapshots, clones, replication, and advanced monitoring or systems management. When one of these features is required, a license must be purchased to enable that additional functionality. Sometimes even the addition of storage capacity requires a license or feature upgrade.

EqualLogic PS Series storage provides many features that are found in other high-end enterprise systems, but instead of offering these at additional cost, they are included as part of its all-inclusive feature set. Group Manager allows the storage admin to configure, maintain, and monitor the arrays as well as provision volumes to hosts. Group Manager also provides capabilities for e-mail notification and configuring SNMP monitoring.

Beyond the day to day administration capabilities of the Group Manager application, Dell also offers the EqualLogic SAN Head Quarters (SAN HQ) tool – again at no additional charge. SAN HQ can be used to monitor storage capacity and performance over time. It gathers statistics and can offer real-time as well as historical information about entire SANs, groups of arrays, pools, individual arrays, or even individual volumes.

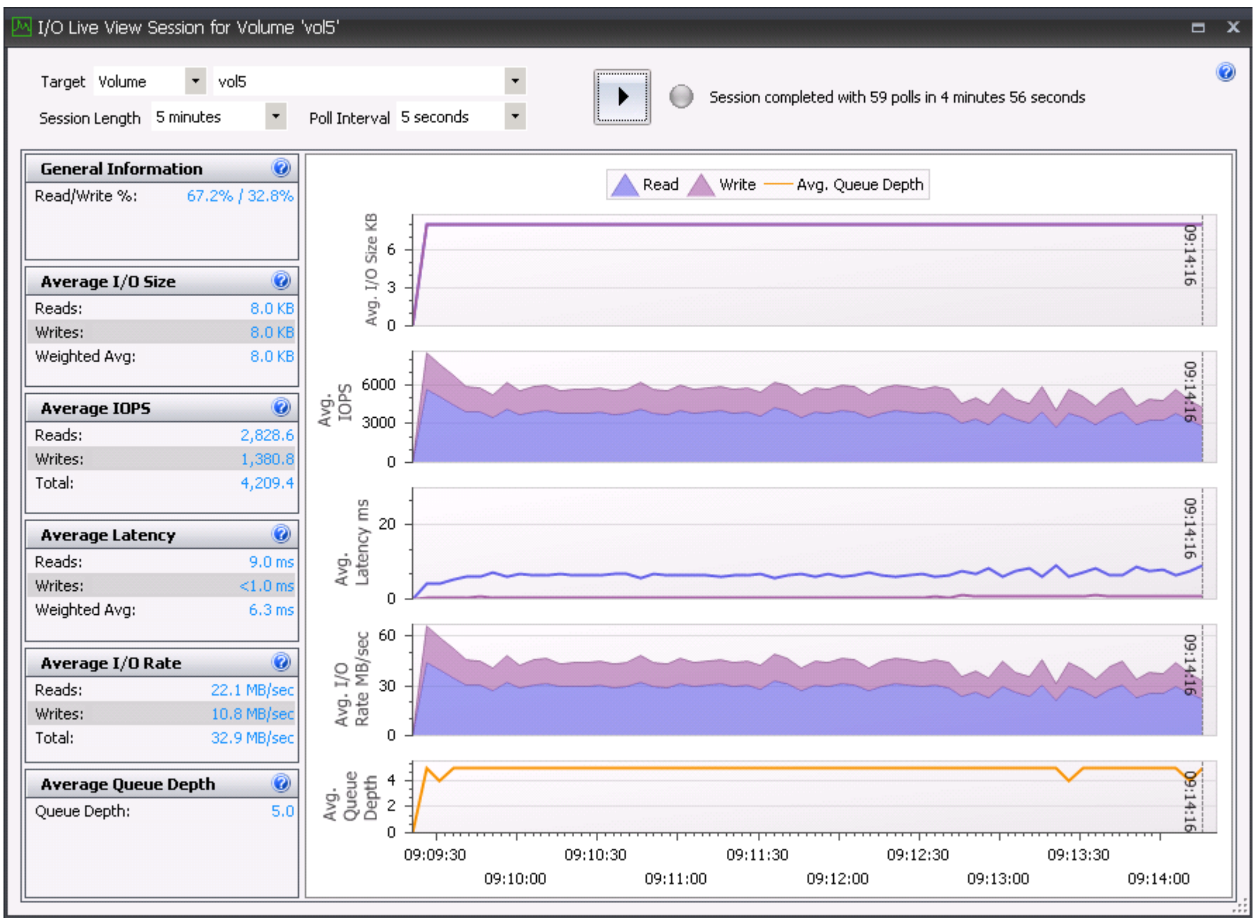

Figure 6 - SAN HQ "Live View" Session

9 Conclusion

Both FC and iSCSI Storage Area Networks require proper planning and deployment to ensure optimal functionality and performance. There are various hardware choices available for iSCSI connectivity, just as there is with FC SANs. Proper configuration and monitoring of the entire SAN is necessary to ensure deterministic performance for both FC and iSCSI environments.

The differences between EqualLogic iSCSI and the various FC storage systems are sometimes subtle, and in fact there are many similar concepts. Having an understanding of these similarities and differences can help to avoid mistakes that could affect planning, deployment, or even administration of a production storage environment. The *EqualLogic Configuration Guide* is an excellent reference for planning and deploying of EqualLogic iSCSI SANs (see Appendix B).

Although most standard Ethernet devices will support the iSCSI protocol, in order to ensure proper iSCSI SAN performance, Dell recommends selecting switches and initiators that are proven and validated for use with the EqualLogic array. Refer to the list of EQL approved Ethernet switches in the *Validated Component List for EqualLogic PS Series SANs* (see Appendix B) for confirmation of certified components.

Note: Refer to the *EqualLogic Configuration Guide* for the latest information and recommended settings.

Appendix A Common terminology

The following terms are used throughout this paper:

DCB: Data Center Bridging – a set of enhancements to Ethernet Protocols to prioritize specific types of traffic in data center environments.

E-port: Expansion-Port. A port on a FC switch that connects to another FC switch fabric.

Fabric: The switching hardware in a FC or iSCSI SAN

F-port: A port on a FC switch that connects to a node or end device.

HBA: Host Bus Adapter

ISL: Inter-Switch Link. Usually one or more links connecting switches. Similar to LAG.

LAG: Link Aggregation Group or a way to aggregate the bandwidth of two or more links and appear as a single link. Similar to ISL.

LUN: Logical Unit Number – a method for addressing and identifying a unique device presented by a SCSI target.

NIC: Network Interface Card

N-port: A FC port on a node – usually a host bus adapter or storage device.

Pool: A group of EqualLogic PS Series arrays in which a volume can be created. A Storage Group can have up to 4 pools.

SAN: Storage Area Network – a network that is usually dedicated to providing connectivity between storage and hosts.

Slice: The portion of a volume that is located on a single array, even though the volume is distributed across multiple members.

Spanning Tree Protocol: A protocol that prevents loops in Ethernet networks.

Storage Group: A group of EqualLogic PS Series arrays that can be managed as a single entity.

Appendix B Related publications

The following Dell publications are referenced in this document and are recommended sources for additional information.

- *Dell EqualLogic PS Series SANs Validated Component List*:

  http://en.community.dell.com/techcenter/storage/w/wiki/equallogic-validated-components.aspx

- *Dell EqualLogic Configuration Guide*:

  http://en.community.dell.com/techcenter/storage/w/wiki/equallogic-configuration-guide.aspx

- PS Online help

  http://psonlinehelp.equallogic.com/V5.1/groupmanager.htm