PS Series Architecture: MPIO with Devices That Have Unequal Link Speeds

Dell Engineering
June 2015

A Dell Technical White Paper

Revisions

Date	Description
June 2015	Initial release

Table of contents

Executive summary

A Best Practice when building a Dell™ PS Series SAN is to use equal network speed on all hardware attached to the SAN. This includes host ports, network switches and PS Series arrays. When this is not possible, the PS Series Multi-Path I/O (MPIO) software included with Windows®, VMware® and Linux® HIT kits attempts to optimize the resources present. This document explains the behavior of the MPIO software in the event that the equipment attached to the SAN is not utilizing the same network speed.

1 Best Practice for network interfaces in a PS Series SANs

A Best Practice when deploying hosts and storage arrays on the same PS Series SAN is to utilize a common network speed throughout the environment. This includes host ports (NIC, offload cards or HBAs), network switching, and PS Series arrays. For example, to obtain the best results install 1GB components in a configuration using a 1Gb array (such as a PS6100) and 10Gb switches and host interfaces when using 10Gb capable arrays (such as the PS6210). When this is not possible, the PS Series MPIO software included in the Windows, VMware and Linux host tools attempts to optimize the environment for performance within the limits imposed by mismatched speeds.

1.1 Distribution of data in a PS Series SAN

Data in a PS Series SAN is automatically distributed across multiple members in a pool. Each of these members contains resources to service the IT workload, including disks, cache and I/O ports. This distribution of resources permits the PS Series SAN to scale up or down easily by simply adding or removing member arrays to the group. When used with non-enhanced MPIO, the PS Series Group automatically forwards the I/O traffic to the appropriate member if the request is received by another member. When using the PS Series MPIO extensions, this forwarding is not needed since there are connections made from the host to each member that holds a portion of the data.

2 Multi-Path I/O

MPIO uses redundant physical (or virtual) connections to deliver high availability to shared storage. Having multiple host connections, switches and SAN interfaces helps to eliminate single points of failure. Using MPIO, servers can send multiple I/O streams to SAN volumes concurrently. MPIO routes I/O over redundant paths that connect servers to storage and manages these paths so that requests can be rerouted through another path in the event of a failure in one of the components along the way. MPIO also provides increased redundancy and can improve performance of application data hosted on the SAN.

MPIO implementations supplied with the Microsoft Windows, VMware or Linux operating systems are necessarily generic in their implementation since they must work reasonably well with a wide range of storage products. These implementations typically require that the administrator manually create individual MPIO connections to the storage and then often uses a blind round-robin algorithm to distribute I/O between these connections. Dell has developed extensions to these generic MPIO implementations for the PS Series arrays that automates the establishment of the MPIO connections, enables the built in multi-path modules of these operations systems to obtain specific information about their PS Series storage and use a Least Queue Depth algorithm to select the port for communicating with the storage. This combination of features results in more efficient and intelligent use and connectivity to the SAN hardware.

2.1 Dell PS Series MPIO extensions

The Dell PS Series Multi-Path I/O extensions are designed to deliver:

- Automatic connection management
- Automatic path failure detection and path failover
- Automatic load balancing across multiple paths
- Support for multiple connections to a single iSCSI target (volume)
- Increased bandwidth
- Reduced network latency
- Easy installation and management
- Automatically throttle down connections when a storage pool reaches 90% of maximum allowable connections

The Dell PS Series MPIO extensions consist of two components:

- A driver component in the I/O Stack that works in conjunction with the native MPIO driver to route I/O to the desired path

  For Microsoft Windows this is referred to as the Device Specific Module (DSM) and for VMware the term used is the Multipathing Extension Module (MEM). Linux has no standard term. This component selects the best path for each I/O by using knowledge about how the volume is laid out on the PS Series Group.

- A service or daemon, (EqualLogic Host Connection Manager – EHCM) that manages connections

The service periodically retrieves the current volume layout from the PS Series Group. It also monitors the iSCSI sessions and SAN topology and reconfigure the iSCSI session mesh when necessary (see Figure 1).

Together, these components allow administrators to easily install and configure MPIO for iSCSI networks. The MPIO extensions, along with a redundant iSCSI hardware configuration, create the infrastructure needed for a completely fault tolerant and efficient MPIO solution.

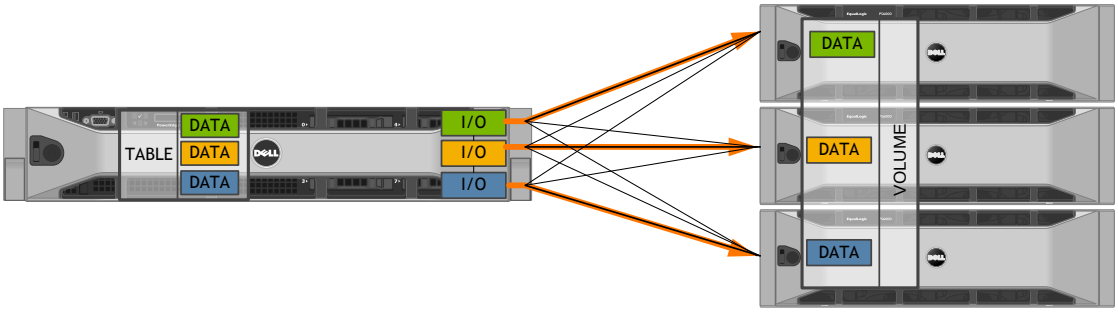

Figure 1 Dell PS Series MPIO extensions automatically manage iSCSI sessions and route I/O directly to the Member servicing it

Multiple I/O streams are concurrently routed through the host iSCSI sessions directly to the PS Series Member servicing them. Using the Dell PS Series MPIO extensions, the best path for each I/O is selected by using knowledge about the volume layout in the PS Series Group. The Dell PS Series MPIO extensions automatically manage these iSCSI sessions.

3 Handling unequal network speeds

The PS Series MPIO extensions attempt to optimize uneven resources when the Ethernet ports on the host and the Ethernet ports on a PS Series array group do not have identical network speeds. In other words, 10Gbps host ports paired with 1Gbps array ports, or 1Gbps host ports paired with 10Gbps array ports. It is assumed that all ports on a single host have identical speeds (all 1Gbps or all 10Gbps) and all array Ethernet ports within a single PS Series array group have identical speeds (all 1Gbps or all 10Gbps).

In all examples, the actual number of connections is subject to the PS Series group pool limits and maximum connections in large configurations. That is, iSCSI connections per volume may be reduced if pool limits or large configuration limits apply. By default, iSCSI connections in each volume are limited by the following factors:

- Maximum connections per member (default: 2)
- Maximum connections per volume overall (default: 6)
- PS Series pool limits: Imposed by the PS Array Group when there are multiple hosts connecting to an PS Series group and the total number of iSCSI connections exceeds 90% of maximum 1024 available connections (approximately 921 connections).
- Large configurations: If the PS Series MPIO software determines that the number of volumes, host ports and PS array ports would result in an extremely large number of iSCSI connections for a single host, the software will trim back the number of connections per volume. For VMware, the default is 512 (configurable to 1024), for Linux it is 1024 and for Windows the number is 248 connections for a single host. Linux and Windows limits are not configurable.

Note: Dell strongly recommends pairing host ports with identical speed PS Series group network interfaces. The recommended practice is to either pair a 1Gbps speed host system with a 1Gbps speed PS Series group, or pair a 10Gbps speed host system with a 10Gbps speed interface PS Series group.

3.1 When host ports have the same speed as PS Series group ports

When the host and PS Series network interface have identical speeds, the host or PS Series group member with the fewest number of ports, limits the overall number of iSCSI Connections. This applies whether 1Gbps host ports to 1Gbps PS Series group ports or 10Gbps host ports to 10Gbps PS Series group ports are used. For example, if the host has two ports for SAN traffic and the PS Series group has four ports per member, the PS Series MPIO software will limit iSCSI connections to two per array member. Remember, all iSCSI connections are further subjected to the max connections per member and max connections per volume limits, in addition to other limiting factors.

3.2 When host ports are faster than PS Series group ports

When the host ports are faster than the PS Series Group ports, the number of slower PS Series ports will limit the number of iSCSI connections. This would apply with 10Gbps hosts and 1Gbps PS Series Group ports. The PS Series MPIO software will make as many host iSCSI connections as possible to the array. With 4 x 1Gbps PS Series Group ports, each 10Gbps host port could use up 4 x 1Gbps ports, resulting in up

to 4Gbps (< 10Gbps) bandwidth. Again, all connections are further subjected to the Max Connections per Member and Max Connections per Volume limits and other limiting factors previously mentioned.

3.3 When host ports are slower than PS Series group ports

When the host ports are slower than the PS Series group ports, the number of slower host ports limits the number of iSCSI connections. This applies with 1Gbps hosts and 10Gbps PS Series group ports. The PS Series MPIO software makes as many PS Series group port iSCSI connections as possible using available ports on the host. Each 10Gbps NIC can make multiple connections to maximize the available host bandwidth. With four 1Gbps host ports, each 10Gbps PS Series group port could use up four 1Gbps ports, providing up to 4Gbps (< 10Gbps) bandwidth. As before, all connections are further subjected to the max connections per member, max connections per volume limits and other factors previously mentioned.

3.4 Connection Examples

Several examples of the number of connections created by the PS Series MPIO software are provided in Table 1. Each example depicts a different size PS Series group size. Pool limits or large configuration limits were not applied in these examples, nor do they depict mixed host interface configurations involving the use of 10Gbps and 1Gbps interfaces on a single host.

The max/mem x max/vol settings information in Table 1 represents the maximum per member and per volume iSCSI connection limits set in the PS Series MPIO software on the host.

Nomenclature:

- 2 x 6 represents 2 max connections / array member / volume, 6 max connections per volume overall
- 4 x 12 represents 4 max connections / array member / volume, 12 max connections per volume overall
- #/Volume: Total number of iSCSI connections per volume
- #/Member: Number of iSCSI connections per member

Table 1

	1 Member		**2 Member**		**3 Member**	
Max/mem x max/vol settings	2 x 6	4 x 12	2 x 6	4 x 12	2 x 6	4 x 12
Equal speeds between host and array 1Gbps to 1Gbps with 2 HBAs / 4 array NICs	2/volume 2/member	2/volume 2/member	4/volume 2/member	4/ volume 2/member	6/ volume 2/member	6/ volume 2/member
Equal speeds between host and array 10 Gbps to 10 Gbps with 2 HBAs / 2 array NICs	2/volume 2/member	2/volume 2/member	4/volume 2/member	4/volume 2/member	6/volume 2/member	6/volume 2/member
Unequal speeds 10G host HBA to 1Gbps array NIC with 2 HBAs / 4 array NICs	2/volume 2/member	4/volume 4/member	4/volume 2/member	8/volume 4/member	6/volume 2/member	12/volume 4/member
Unequal Speeds 1Gbps host HBA to 10 Gbps array NIC with 4 HBAs / 2 array NICs	2/volume 2/member	4/volume 4/member	4/volume 2/member	8/volume 4/member	6/volume 2/member	12/volume 4/member

3.5 A complex example using variable speed host interfaces and limiting total connections per volume

Variable speed host interfaces are interfaces where the bandwidth used can be set to less than the actual link speed of the device. In this case, the lower bandwidth set on the interface determines the number of connections rather than the maximum speed that the interface is capable of supporting. The MPIO software attempts to maximize the bandwidth available. For example:

- A PS Series SAN consisting of 3 members, each with 1 x 10Gbps port enabled
- A host, with 2 x 10Gbps interfaces, with a variable bit rate set to 3Gbps on each, with the Max iSCSI connections per member/volume = 2 (default value) and Max iSCSI connections per Volume = 4 (default = 6),

A configuration such as this results in four iSCSI connections made by the MPIO software from the host to the three members. Two of the iSCSI connections from the host interfaces will be to the same PS Series group member's single interface, as the PS MPIO software will attempt to utilize as much array bandwidth (up to 10Gbps) from the available two 3 Gbps interfaces on the host, subject to pool limits and the host' max iSCSI connection limits.

A MPIO extension controls per platform

The PS Series MPIO extensions default values can be adjusted to provide control over the maximum number of total iSCSI connections per volume and the maximum number of iSCSI connections to each member from a host for a single volume. Refer to the Host Tools documentation for each platform for details on how to set these parameters.

Table 2

Platform	**Manage Maximum iSCSI Connections per Member per Volume**	**Manage Maximum iScsi Connections per Volume**
Windows	EHCM Registry: MaxConnectionsPerMember	EHCM Registry: MaxDevicesPerMpioSession
Linux	rswcli --max-sessions-per-volume-slice	rswcli --max-sessions-per-entire-volume
VMware	Ehcmd.conf file: MemberSessions	Ehcmd.conf file: VolumeSessions

B Additional resources

Support.dell.com is focused on meeting your needs with proven services and support.

DellTechCenter.com is an IT Community where you can connect with Dell customers and Dell employees for the purpose of sharing knowledge, best practices, and information about Dell products and installations.

Referenced or recommended Dell publications:

- *Dell EqualLogic Configuration Guide*
- *Configuring and Installing the EqualLogic Multipathing Extension Module for VMware vSphere and PS Series Arrays*
- *EqualLogic Integration: Installation and Configuration of Host Integration Tools for Linux – HIT/Linux*
- *Deploying and Configuring the Dell EqualLogic Multipath Device Specific Module with Microsoft Windows*

Many other PS Series papers are found on the Dell Tech Center EqualLogic Technical Content page at http://en.community.dell.com/techcenter/storage/w/wiki/2660.equallogic-technical-content.