

Inter vs. Outer Nexus Multiplexing

SBE, Inc.

02/09/2006

1. Summary

This document explains how MPIO and ERL-2 work in an iSCSI environment and achieve trunking and task migration as well as their advantages and overhead.

2. Terminology

IETF: The Internet Engineering Task Force develops and promotes Internet standards. Their RFC-3720 standard provides guidance on iSCSI implementation.

ERL-2: Error Recovery Level 2 belongs to the connection recovery class specified in IETF's RFC-3720 standard. Upon detection of a broken TCP connection, the iSCSI initiator driver establishes another TCP connection to the target, and the iSCSI initiator driver informs the target that the allegiance of the SCSI command is being changed to another TCP connection. The target can then proceed to continue processing the SCSI command on the new TCP connection. If a new connection cannot be established for the failed connection's network portal address, an existing active iSCSI/TCP connection may be used to reassign/retry commands from the failed connection. The upper level SCSI driver remains unaware that a new TCP connection has been established and that the command has been transferred to the new connection. The iSCSI Session remains active during the period and does not have to be reinstated.

MPIO Software: An application that maintains more than one physical path between networked storage elements to maintain high storage availability. An example would be the Microsoft MPIO driver or Device-Mapper Multipath (DM-MP) tool integrated in RedHat Enterprise Linux 4 update 2 or later.

Trunking: The method to combine multiple network links to achieve aggregated bandwidth in parallel. Link-bonding specified by IEEE's 802.3ad and trunking in iSCSI ERL-2 are different implementations. The former has higher overhead compared to the latter.

Failover: The ability to switch traffic to standby links should the active link fail or terminate abnormally.

Active/active task migration: Different from generic failover, in which standby links remain inactive while active links are functioning normally, this is an ERL-2 feature intended to direct traffic from failed links to other active links within an iSCSI session.

3. MPIO (Outer Nexus Multiplexing)

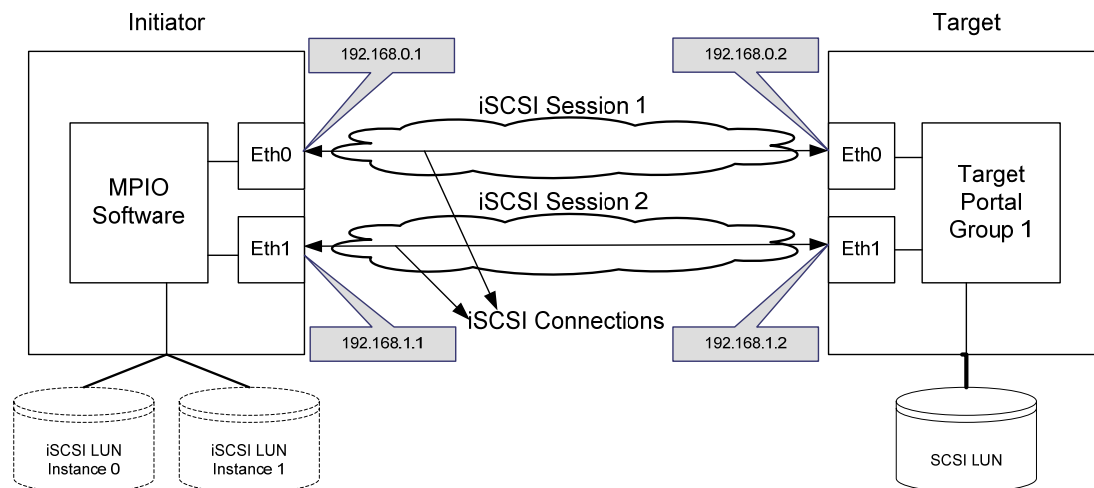


Figure 1: Example Outer Nexus Multiplexing Configuration

Overview: in this configuration, multiple iSCSI sessions to Target Portal Groups (TPG) are created on each physical Ethernet link. LUNs exported by the target are registered multiple times on the initiator. The multi-pathing software run on the initiator identifies identical LUNs by their unique hardware identifiers and manages the commands to access them simultaneously. If functionality permits, it uses all available links for data transfer and balances the load across them; in case of a link failure, it first waits for recovery attempts to time out, closes the iSCSI session on failed link, and then decides whichever link is available to continue the communication.

Advantages:

- This configuration does not require targets that support ERL-2.
- This method can be used independent of storage transport and protocol types, including DAS and Fibre Channel.

Disadvantages:

- Additional MPIO software is required to manage redundant paths.
- The MPIO software is OS dependent.
- MPIO adds extra performance overhead and configuration complexity.
- Extra care has to be taken into account with establishing sessions to multiple TPGs¹.

¹ To access multiple TPGs via multiple sessions, same ISID has to be supplied. Contrarily, to access single TPG via multiple sessions, different ISIDs have to be supplied. Microsoft Initiator chooses different ISIDs automatically.

4. ERL-2 (Inter Nexus Multiplexing)

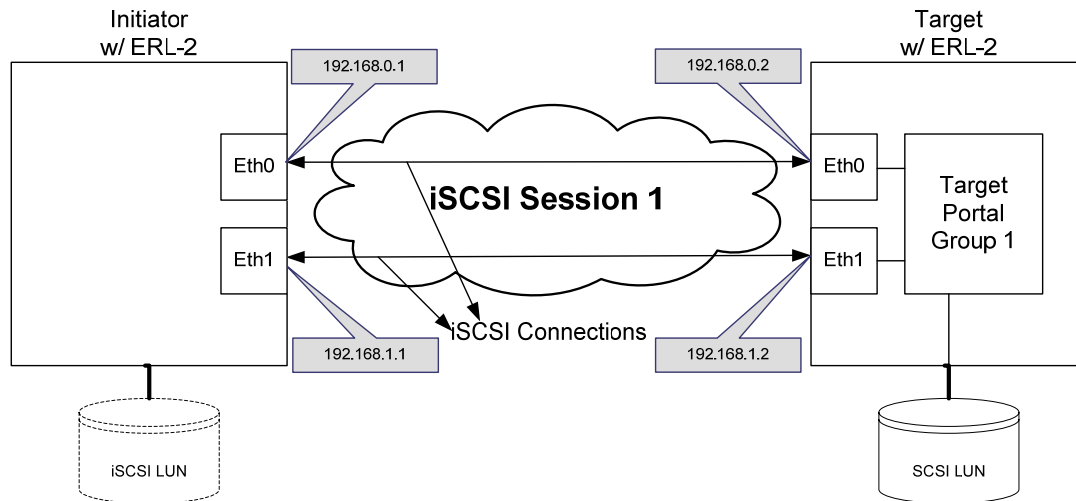


Figure 2: Example Inter Nexus Multiplexing

Overview: in this configuration, a single session containing multiple connections to a single TPG is created across multiple physical links. Initiator registered LUNs exported by target only once. All multi-pathing features, including trunking across subnets and load balancing, are done in the iSCSI layer implementing the optional connection recovery (ERL-2) as specified by IETF's RFC-3720. In case of a link failure, tasks will be re-ordered and transmitted on other active links, without closing a session. Please see ERL-2 under Terminology section.

Advantages:

- The feature is self-contained in the iSCSI stack, requiring no additional administration.
- Trunking and active/active task migration are implemented more efficiently.
- Recovery time is generally less.
- This method can be used in conjunction with MPIO.

Disadvantages:

- Both the target and initiator have to implement ERL-2 functionality fully compliant to IETF's standard
- In order to optimize performance and redundancy in a particular environment, some tuning on a few more parameters may require extra expertise.

5. Conclusion

In summary, MPIO and ERL-2 both achieve the same goal, which is to enable multiple paths between storage endpoints to accomplish aggregated bandwidth and redundancy in case of link failure. For iSCSI targets and initiators that implement ERL-2 properly, this feature is intrinsic in the iSCSI layer, requiring no additional managing software and delivering improved efficiency. In any case, an iSCSI target or initiator, implementing RFC-3720 mandatory feature sets (ERL-0), should work compatibly with the MPIO software.