White Paper

EMC CLARiiON
Asymmetric Active/Active Feature

A Detailed Review

Abstract

This white paper provides an overview of the EMC® CLARiiON® Asymmetric Active/Active feature. It highlights the configuration, best practices, implementation details of EMC PowerPath®, and native multipathing software when the host initiator is configured in Asymmetric Active/Active mode.

October 2010

Part Number H2890.4

Table of Contents

Executive summary

In FLARE® release 26, EMC introduced the Asymmetric Active/Active feature for CLARiiON® storage systems. Asymmetric Active/Active provides a powerful new way for CLARiiON storage systems to present logical unit numbers (LUNs) to hosts, and eliminates the need for hosts to use the LUN ownership model. This changes the way in which a host manages multiple communication paths to LUNs on the array (commonly referred to as *path management*) by permitting input/output (I/O) to flow to either or both storage processors. This white paper discusses the benefits and implementation details of this feature.

Prior to release 26, all CLARiiON storage systems used the standard active/passive feature in which one storage processor (SP) owned the LUN, and all I/O for that LUN was sent to that SP. If all paths from a host to the SP failed, host-based path management software regulated the I/O path by issuing a **trespass** command. This caused the storage system to change the ownership of the LUN to the peer SP, and I/O was then sent to the peer SP.

CLARiiON Asymmetric Active/Active is a feature that introduces a new initiator Failover Mode (Failover Mode 4). When configured as Failover Mode 4, initiators can send I/O to a LUN regardless of which SP owns the LUN. While this feature allows a trespass-like command, explicit trespass behavior is not required.

Introduction

This white paper introduces the CLARiiON Asymmetric Active/Active feature and its configuration. The white paper describes the implementation of the feature on CLARiiON and compares *explicit* and *implicit* trespass commands. The paper provides details about the various configurations for ALUA and CLARiiON and describes how EMC PowerPath®, native Multipath I/O (MPIO) software, and Veritas DMP work with ALUA.

This white paper discusses the performance considerations of routing data through optimal and non-optimal paths. It also lists some best practices to optimize performance. Then, the paper moves onto a brief discussion about the benefits of the CLARiiON Asymmetric Active/Active feature and the impact of the new redirector for FLARE release 26. Finally, the paper makes a detailed comparison between the various path management software that are used with CLARiiON to manage multiple paths to storage devices.

Audience

This white paper is intended for customers, partners, and EMC field personnel who want a better understanding about the implementation, benefits, and configuration of the CLARiiON Asymmetric Active/Active feature and its potential impact on their storage environment.

Terminology

- **ALUA (Asymmetric Logic Unit Access) —** A SCSI standard that allows multiple controllers to route I/O to a given logical unit.
- **CLARiiON LUNs —** Logical subdivisions of RAID groups in a CLARiiON storage system. These are volumes that are presented to hosts.
- **CLARiiON Messaging Interface (CMI) —** Redundant Peripheral Component Interconnect Express connections that provide communication between the two SPs.
- **Failover Mode —** Determines how the array responds to I/O requests that are directed to LUNs on the non-owning SP.
- **Non-optimal path —** A path that is available to transport I/O, but that may not yield the best performance.
- **Optimal path —** A path that is ready to do I/O and will yield the best performance.

- **Preferred bit —**This bit represents that a given SP is the default owner of a LUN
- **SP (storage processor) —** CLARiiON controller.
- **Target port group —** A set of target ports that are in either primary SP ports or secondary SP ports.
- **Trespass —** A command that allows an SP or its peer to take ownership of the LUN.

Overview of the Asymmetric Active/Active feature

CLARiiON Asymmetric Active/Active is based on the Asymmetric Logical Unit Access (ALUA) standard. ALUA uses SCSI 3 primary commands that are part of the standard SCSI SPC-3 specification (not a CLARiiON-specific implementation) to determine I/O paths. In dual-SP systems, such as a CLARiiON, I/O can be routed through either SP. For example, if I/O for a LUN is sent to an SP that does not own the LUN, that SP redirects the I/O to the SP that *does* own the LUN. This redirection is done through internal communication within the storage system. It is transparent to the host, and the host is not aware that the other SP processed the I/O. Hence, a trespass is not required when I/O is sent to the non-owning (or non-optimal) SP.

Dual-SP storage systems that support ALUA define a set of target port groups for each LUN. One target port group is defined for the SP that currently owns the LUN, and the other target port group is defined for the SP that does not own the LUN. Standard ALUA commands enable the host failover software to determine the state of a LUN's path.

The `REPORT TARGET PORT GROUP` command provides a report about the following three attributes of a port group:

- Preferred — Indicates whether this port group is the default (preferred) port group.
- Asymmetric access state — Indicates the state of the port group. Port group states include Active/Optimal, Active/Non-optimal, Standby, and Unavailable.
- Attribute — Indicates whether the current asymmetric access state was explicitly set by a `SET TARGET PORT GROUP` command or was implicitly set or changed by the storage system.

The SET TARGET PORT GROUP command allows the access state (Active/Optimal, Active/Non-optimal, Standby, and Unavailable) of each port group to be set or changed. Access states are:

- Active/Optimal — Best performing path, does not require upper level redirection to complete I/O
- Active/Non-optimal — Requires upper level redirection to complete I/O
- Standby — This state is not supported by CLARiiON
- Unavailable — Returned for port groups on a SP that is down. The user cannot set it by using set target port groups.

Target port group commands are implemented in the ALUA layer of the storage system. However, host-based path management software executes the commands and manages the paths. The explicit ALUA approach is similar to the traditional path management mechanisms except that ALUA has standardized the previously vendor-specific mechanisms.

Implementation

FLARE 26 or later includes a redirector driver that improves CLARiiON availability; this driver consists of an upper and lower redirector for each SP. The upper redirector is placed closer to the host connection, while the lower redirector is placed closer to the CLARiiON back end. The layered drivers such as EMC Navisphere® Quality of Service Manager (NQM), SnapView™, and so forth sit between the upper and lower redirectors as shown in Figure 1.

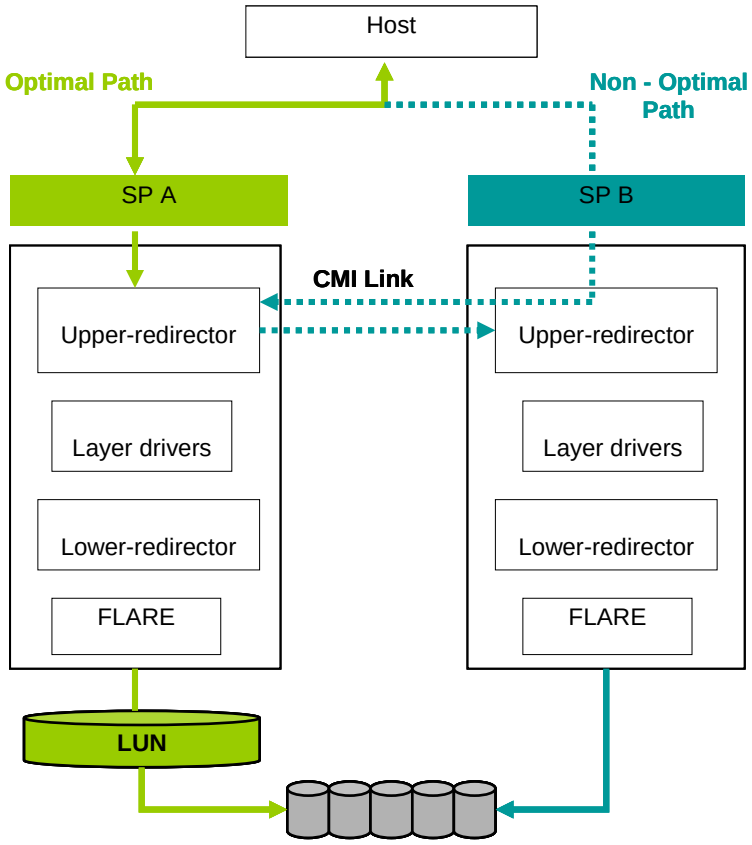

Figure 1. CLARiiON Asymmetric Active/Active feature

The CLARiiON Asymmetric Active/Active feature supports two target port groups:

- Optimal — An optimal target port group represents the ports that belong to the current owner of the LUN
- Non-optimal — A non-optimal port group, implemented in the redirector driver, represents the ports owned by an SP that is not the current owner of the LUN

I/O is accepted on all ports. In the event of a front-end path failure, an SP that does not own the LUN may receive I/O. The upper redirector of the SP that does not own the LUN routes the I/O to the owning SP through the internal CLARiiON Messaging Interface (CMI) channel. The SP that owns the LUN services the I/O request. An I/O request that is received by an SP has to be acknowledged over the same path by that SP.

Load balancing across ports on one SP works just like it does in PowerPath multipathing software. Load balancing across optimal and non-optimal paths is not recommended, and is only supported in failover situations.

CLARiiON's Asymmetric Active/Active feature supports two target port groups: optimal and non-optimal. An optimal target port group represents the ports that belong to the current owner of the LUN. A non-optimal port group represents the ports owned by an SP that is not the current owner of the LUN. This is implemented in the redirector driver.

I/O is accepted on all ports. In event of front-end path failure I/O may be received by an SP that does not own the LUN. The upper redirector of the SP that does not own the LUN routes the I/O to the owning SP through the internal CMI channel. The I/O request is serviced by the SP that owns the LUN. Note that an I/O that is received by an SP has to be *acknowledged* over the same path by that SP.

Load balancing across ports on one SP works just like it does in PowerPath multipathing software. Load balancing across optimal and non-optimal paths is not recommended.

Layered drivers impact

The behavior of the layered drivers (MetaLUNs, LUN Migration, SnapView, MirrorView™ and SAN Copy™) does not change with Asymmetric Active/Active. For example, if a LUN is trespassed during a LUN Migration, the array still trespasses the destination LUN and the synchronization continues from a checkpoint.

With SnapView, the source and replica (snap or clone) are always owned by the same SP. Hence, if the source or SnapView replica trespasses, the source or the SnapView replica will still follow its replica or source, respectively. However, depending on the failover software deployed on the host, the *likelihood* of trespassing the source or SnapView replica is reduced in the ALUA mode. See the "Avoids LUN ownership thrashing situations that may occur in SnapView and cluster configurations" bullet for more information.

For MirrorView/Synchronous (MirrorView/S), if a primary LUN is trespassed, the secondary trespasses immediately. For MirrorView/Asynchronous (MirrorView/A), if a MirrorView primary LUN is trespassed, the secondary MirrorView LUN trespasses during an update or at the start of an update.

For local and remote SAN Copy sessions, if the source or destination LUN trespasses, the session must be restarted on the peer SP because the SAN Copy initiators do not use ALUA mode. As a result, MirrorView and SAN Copy will behave the same as before.

Explicit and implicit trespass

An explicit trespass is the result of an external command from a user or the failover software. When an SP receives this command (from the failover software or a user issuing the LUN trespass in Navisphere), LUN ownership is transferred to that SP. PowerPath path management software can issue this trespass using CLARiiON proprietary commands, or other ALUA-compatible path management software can issue this trespass using the SET TARGET PORT GROUP command.

An implicit trespass is the result of software controls within the storage system. For example, an implicit trespass occurs when the amount of I/O transferred across the non-optimal path exceeds the optimal path I/O by a certain amount (threshold). The software uses counters to keep track of optimal and non-optimal path I/O. When it detects that the non-optimal path has received 128,000 more I/Os than the optimal path, it initiates a trespass.

Configuring for ALUA

The options for failover software on the host are PowerPath; Dynamic Multipathing (DMP) failover within Veritas Storage Foundation Suite; and native operating system (OS) failover software, such as MPIO, which are compliant with ALUA. The following sections discuss these options.

To ensure that application data is highly available, the host must be configured to withstand a single point of failure, including a failure in the host bus adapter (HBA), fibre cable, or failover software. The *EMC CLARiiON Open Systems Configuration Guide* on EMC Powerlink® outlines the attached-storage methodologies that CLARiiON supports.

Configuring the host

Table 1 on page 8 illustrates that many hosts, using their native failover software, can take advantage of ALUA mode on CLARiiON. Redundancy for all points of a configuration is essential for optimal high availability (including maximum data accessibility and server uptime). This means that each server must have at least two HBAs (or network interface cards in an iSCSI environment), and be attached to both SPs.

Two switches provide independent, discrete paths to the storage system from the server. If more than two switches are employed, redundant switches should be connected via interswitch links for fabric failover purposes.

Each HBA can be configured to see an SP port (if direct connected) or zoned to see both SPs. This protects the user's configuration against the loss of an HBA, cable, switch, or SP, and takes advantage of the failover capabilities and load-balancing features of the user's path management software.

Having multiple active paths to a LUN available from a server ensures that the user can use the load-balancing algorithms in path management software to avoid becoming path bound (restricted to a single port).

Using PowerPath with ALUA

PowerPath version 5.1 is the first ALUA compliant release. Ensure that the PowerPath version is 5.x or later, and consult the EMC E-Lab™ Navigator on Powerlink for various host considerations.

PowerPath load balances across optimal paths. If PowerPath detects that all optimal paths have failed, PowerPath initiates a trespass to change the LUN's ownership.

Table 1. PowerPath and native MPIO software support

	Native with Active/Passive	Native with ALUA	PowerPath with Active/Passive	PowerPath with ALUA
W2K8	No	Yes	Yes	Yes
W2K3	No	No	Yes	Yes
Win2K	No	No	Yes	Yes
HP-UX 11i v1 and v2	Yes	No	Yes	Yes
HP-UX 11iv3	No	Yes (11.31.*0709* or later)	No	Yes
Solaris 9/10	Yes	Yes	Yes	Yes
Linux (RH and SuSE)	Yes	Yes (RH 5.1, 4.6 or later) (SLES 10.1 or later)	Yes	Yes
AIX	Yes	Yes (AIX 5.3, AIX 6.1)	Yes	Yes
VMware	Yes	Yes (ESX 4.x or later)	Yes	Yes (ESX 4.x or later)

Definitions for columns:

- Native with Active/Passive: Standard Active/Passive failover (*not* ALUA) is provided when using the noted operating system (OS) alone. In this case, the OS issues trespass commands to enable alternate paths.
- Native with ALUA: ALUA features are provided when using the noted OS alone (PowerPath is not required).
- PowerPath with Active/Passive: Standard Active/Passive failover (not ALUA) is provided by PowerPath with the noted OS; PowerPath issues trespass commands to enable alternate paths.

- PowerPath with ALUA: ALUA features are provided when the specified PowerPath release is used with the noted OS.

Using native failover with ALUA

MPIO and other native host-based failover applications can work with ALUA if they are ALUA-compliant. Native failover software for all operating systems only sends I/O down optimal paths. If the optimal paths are not available for some reason, I/O is be sent to the non-optimal paths, and these non-optimal paths eventually become optimal.

Please note the following:

- For HPUX 11i v3.1, patch 11.31.0709 or later is required for native failover support with ALUA.
- For Solaris 9, MPIO requires StorEdge SAN Foundation Software 4.4.12 or later.
- Solaris 10 update 3 or later is required for native failover support with ALUA.
- For Sun Cluster support with ALUA, Solaris MPxIO requires FLARE release 28 version 04.28.000.5.704 or later.
- For VMware ESX 4.x, MPIO requires FLARE release 28 version 04.28.000.5.704 or later to support ALUA.
- For AIX MPIO, FLARE 29 or later is required; see E-Lab Navigator for more information about specific Technology Levels of AIX 5.3 and 6.1 that support (implicit) ALUA.

Using Veritas DMP with ALUA

Symantec Storage Foundation Suite includes explicit ALUA support including autorestore for the following operating system versions at the specific minimum revisions listed below:

Table 2. Veritas DMP ALUA support

Operating system	Storage Foundation Suite minimum version with ALUA	Notes
W2K3, W2K8	5.1	
HP-UX 11i v3	5.0.1	
Solaris	5 MP3 RP1	
Linux (RHEL & SLES)	5 MP3 RP2 HF3	
AIX	5.1	FLARE 29 or later

It is important to consult E-Lab Navigator on Powerlink for up-to-date host considerations to support failover software on various operating systems. The *Host Connectivity Guides* on E-Lab Navigator provide more information on various operating systems.

Configuring the CLARiiON

To configure Asymmetric Active/Active, use Navisphere Manager or the CLI to set the Failover Mode for host initiators to 4. For more information about Failover Mode see the "CLARiiON Failover Modes" section.

As shown in Figure 2, the **Failover Mode** pull-down menu (in the **Group Edit Initiators** dialog box) includes **4** as an option. Since Failover mode is an *initiator* option (rather than *storage group* option) both ALUA hosts and non-ALUA hosts (hosts not configured with a Failover Mode 4) can be attached to the same LUN.

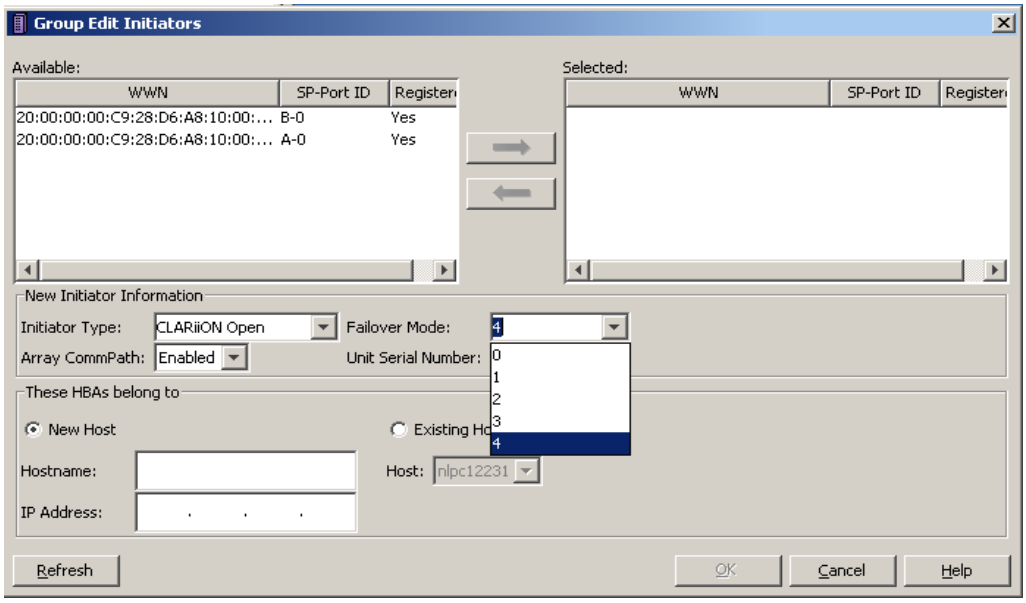

Figure 2. Group Edit Initiators dialog box in Navisphere displaying new Failover Mode for Asymmetric Active/Active

To set Failover Mode to 4 (Asymmetric Active/Active) on the host, type:

```
naviseccli -h <SP_IP_Address> -user a -password a -scope 0 storagegroup -
sethost -ip < IP Address> -failovermode 4
```

To set Failover Mode to 4 (Active/Active) for each HBA initiator record, type:

```
naviseccli -h <SP_IP_Address> -user a -password a -scope 0 storagegroup -
setpath -hbauid xxxxx -sp a -spport xxxxx  -failovermode 4
```

To display the default Failover Mode value, type:

```
naviseccli -h <SP IP Address> -user a -password a -scope 0 port -list -
failovermode
```

Note: After configuring devices for ALUA (Failover Mode 4), reboot the host.

CLARiiON Failover Modes

A CLARiiON LUN trespass may be initiated by the storage system, or by the path management software residing on the server. The Failover Mode of the initiator specifies how the CLARiiON should respond to a trespass condition. CLARiiON supports five distinct Failover Modes depending on the operating system type attached:

- Failover Mode= 0 — Auto trespass mode; any media access to the non-owning SP is rejected
- Failover Mode= 1 — Passive Not Ready; a command failure when I/O is sent to a non-owning SP
- Failover Mode= 2 — (DMP mode); Quiet Trespass on I/O to non-owning SP
- Failover Mode= 3 — Passive Always Ready; some commands (for example, Test Unit Ready) return Passive Always Ready status
- Failover Mode= 4 — Asymmetric Active/Active

Getting the ALUA state from a CLARiiON

The Navisphere CLI **getlun** command displays additional information about the LUN connected to an Asymmetric Active/Active host. Example output of this command is follows:

```
# naviseccli –h [SPipaddress] getlun 0

    Read Requests SPA:        236
    Read Requests SPB:        1480
    Write Requests SPA:       426
    Write Requests SPB:       627
    LUN Busy Ticks SPA:        273
    LUN Busy Ticks SPB:        297
    LUN Idle Ticks SPA:        0
    LUN Idle Ticks SPB:        0
    Number of arrivals with non-zero queue SPA:     398
    Number of arrivals with non-zero queue SPB:     398
    Sum queue lengths by arrivals SPA:              398
    Sum queue lengths by arrivals SPB:              398
    Explicit Trespasses:  5800
    Explicit Trespasses SPA:  2346
    Explicit Trespasses SPB:  3454
    Implicit Trespasses:  320
    Implicit Trespasses SPA:  214
    Implicit Trespasses SPB:  106
```

The meaning of the output for this command has not changed; however, with Failover Mode 4, a LUN reports statistics for both SPA and SPB. In addition, the output displays explicit and implicit trespasses.

The **LUN Properties** dialog box as shown in Figure 3 has been enhanced to display statistics for LUNs that are connected to initiators with Failover Mode 4. These statistics are available when **Statistics Logging** is enabled on the storage system. The statistics include the number of reads and writes routing through the optimal and non-optimal paths. The Redirector Reassignment value records the number of implicit trespasses issued by Asymmetric Active/Active since **Statistics Logging** was enabled.

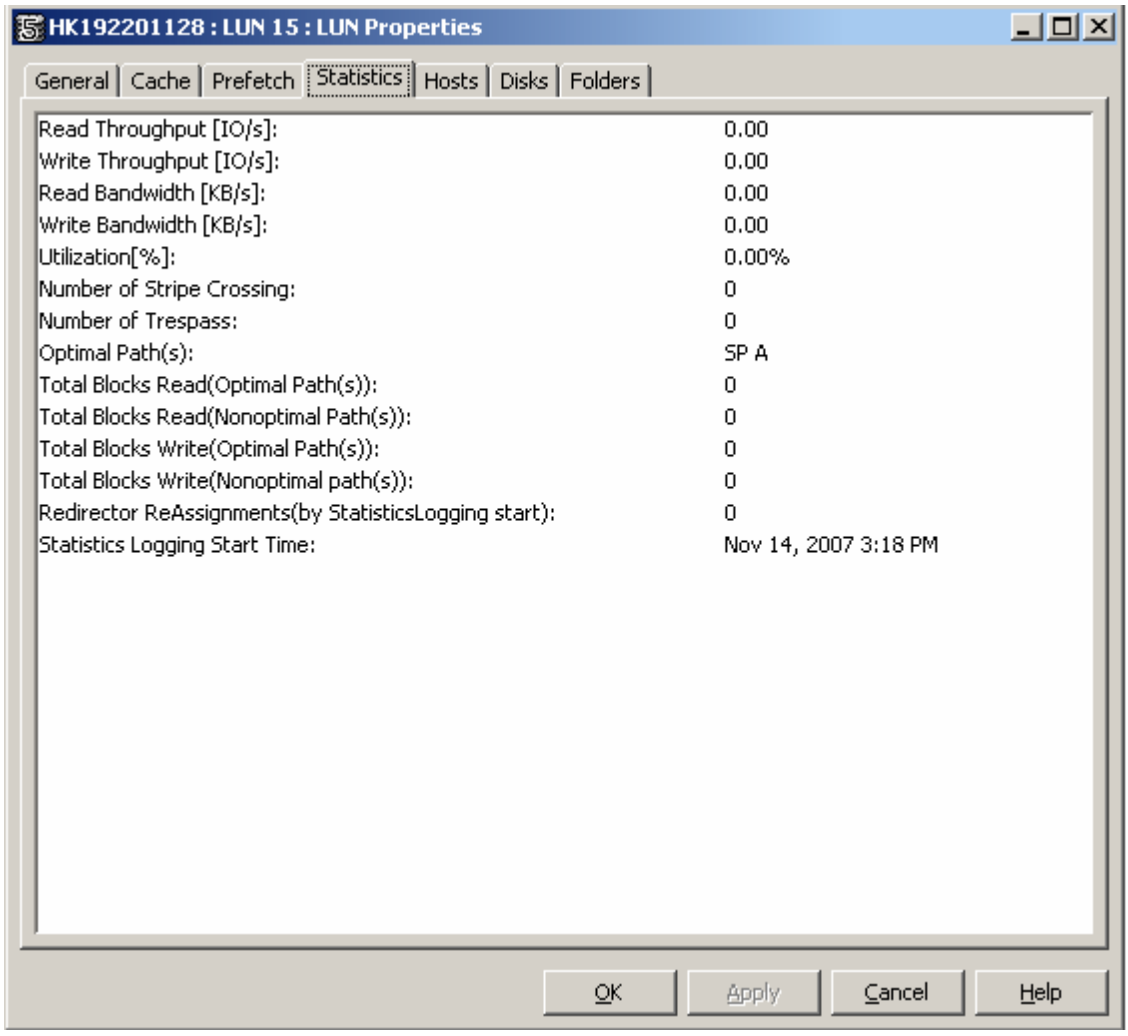

Figure 3. LUN Properties dialog box statistics

Storage system support

CX4 systems running FLARE release 28 or later support ALUA. The CX3 and CX (700/500/300) storage systems running FLARE release 26 are ALUA compliant. The ALUA functionality is available for both Fibre Channel (FC) and iSCSI attach hosts if they are connected using Failover Mode 4 and have ALUA-aware failover software that supports FC, iSCSI, or both connections. AX systems do not support ALUA.

Performance considerations

There will be some performance impact if I/O is routed to the non-owning SP (through a non-optimal path). Non-optimal paths are slower and therefore are not the preferred method for normal access to the storage. New statistics are available on the array to help you determine if I/O is flowing through the optimal or non-optimal paths. The *EMC CLARiiON Storage System Fundamentals for Performance and Availability* white paper provides more details on the performance impact when routing I/O through non-optimal paths.

Best practices

To avoid performance impact when I/O is routed to the non-owning SP of a LUN, EMC recommends the following best practices:

- Balance LUN ownership between the two SPs.

- Configure failover software so that it only load balances across the active-optimal path for a given LUN. PowerPath does this by default.
- After SAN changes (component failures, replacements) that may cause I/O paths to change, ensure that hosts are still using optimal paths to their LUNs.
- After an NDU operation, ensure that all LUNs are returned to the default owner. PowerPath does this automatically.
- In case of failure or a performance issue, ensure I/O is routing through the optimal path.

Impact of the Asymmetric Active/Active feature

Asymmetric Active/Active is a request-forwarding implementation that honors the LUN ownership feature of the storage system (only one controller owns a given LUN). However, it allows I/O to route through either controller. The controller that is not the current owner of the LUN redirects the I/O to the controller that owns the LUN by using internal communication paths within the storage system.

There is no benefit of ALUA in the event of an SP failure because the redirector driver is not accessible to redirect the I/O to a non-owning SP. As a result, in the event of SP failure ALUA works as PNR (Failover Mode 1) mode.

Benefits of CLARiiON Asymmetric Active/Active

Improved customer experience

The CLARiiON Asymmetric Active/Active feature:

- **Avoids unavailability of boot from SAN during path failure**
  ALUA avoids issues when the BIOS attempts to boot from non-optimal path and there is no failover software available (because the system is booting).

  In the past, when the boot server could not get to an SP (for example, all paths to that SP had failed before the operating system boot), the user had to manually trespass the LUN to the other SP for the server to boot successfully. With the request forwarding method, users do not need to explicitly trespass the LUN to the other SP. If the HBA boot BIOS is configured so that it can issue an I/O to the surviving SP, I/O will route through the upper redirector to the owning SP and boot the operating system successfully.

  The *Host Connectivity Guides* on E-Lab Navigator provide information on how to configure the HBA boot BIOS.

- **Reduces data unavailable situations due to misconfiguration of the host**
  Occasionally a user might misconfigure a host in such a way that an application sends an I/O to a non-optimal path (meaning the SP that does not own the LUN). In this case, depending on the failover software installed on the host, the CLARiiON storage system does not return an I/O error condition. Instead, due to the request forwarding feature, the I/O is routed to the SP that owns the LUN.

  Furthermore, when the CLARiiON addresses a change in the access of storage it automatically adjusts the optimal path setting for a LUN. (This "implicit trespass" is discussed in the "Explicit and implicit trespass" section.) This automatic adjustment is extremely beneficial in larger environments where the chances of misconfiguration are higher, and in environments where access of a LUN may vary over time.

- **Avoids LUN ownership thrashing situations that may occur in SnapView and cluster configurations**
  For Active/Active cluster configurations, if the LUNs are shared and written to by multiple hosts, ALUA avoids the trespass of LUNs between the two SPs in a one-path-per-SP host configuration.

  For SnapView, the same SP that owns the source LUN must own the clone or snapshot. Therefore, in a case where a production server is writing to the source LUN while the snapshot/clone is mounted and written to by a backup server, a one-path-per-SP configuration on both the production and backup hosts (where each server has a path to the other SP than its peer) can cause path thrashing with servers in the non-ALUA mode.

  For both cluster and SnapView configurations, with the introduction of the ALUA standard, the LUN will not trespass back and forth between the two SPs, but will be owned by the SP through which maximum I/O requests for that LUN are received by a given host in the ALUA mode.

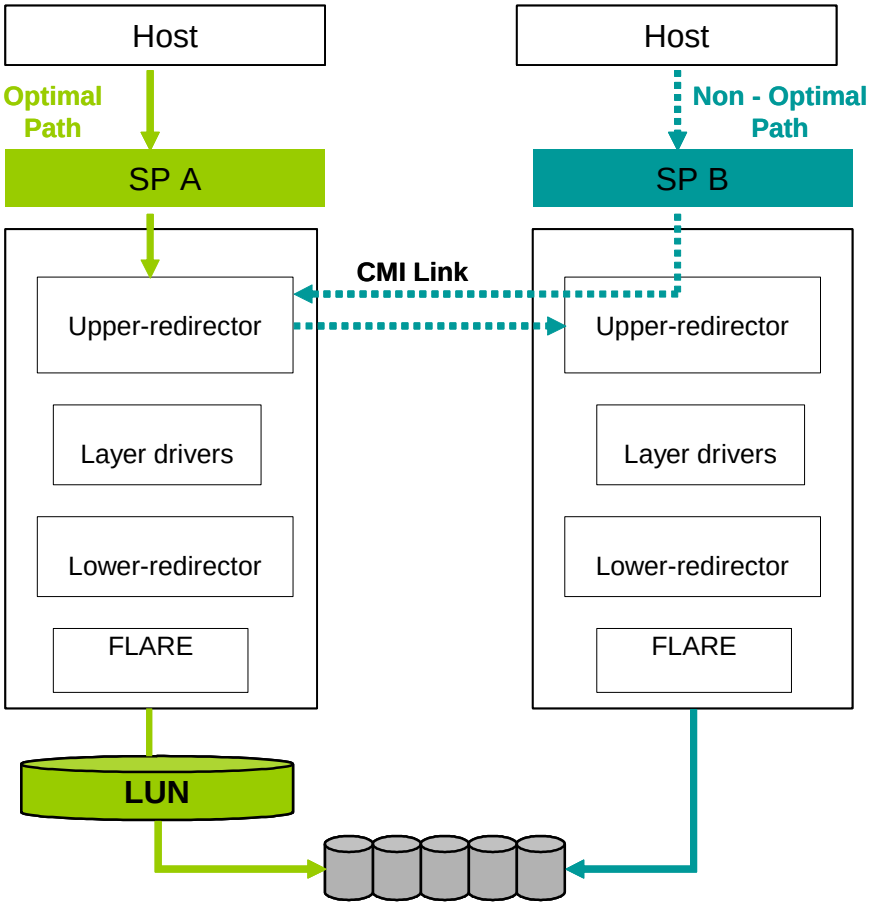

Figure 4. Impact of ALUA with SnapView and cluster configurations

- **Supports standard SCSI multipath interfaces**
  ALUA standardizes the implementation of OS vendors' native multipathing and other failover software. Host failover software does not need to contend with CLARiiON-specific trespass commands because CLARiiON implements the ALUA SCSI standard.

Impact of the redirector

In addition to front-end pathing advantages, the new FLARE release 26 redirector provides benefits for all attach (traditional PNR as well as ALUA-based) types by providing redirection services for the "back end" as well as the front end.

Back-end failure masking

In the case of a back-end failure (such as an LCC failure) on the SP that is the current owner of the LUN, by using the request forwarding feature in ALUA, I/O is routed through the lower redirector to the peer SP with the stable back end. The I/O acknowledgement is sent through the SP that owns the LUN, by using the lower redirector, as shown in Figure 5. No intervention of failover software is required on the host, thus masking certain CLARiiON back-end failures.

Note that in the following example, the LUN is trespassed by FLARE to the SP that can access the LUN at the back end. The failover software configured on the host will send I/O to that LUN through the peer SP. After the back-end error is corrected, the LUN is trespassed back to the previous SP, and any I/O for that LUN is not be redirected to the peer SP. As a result, the host sees a minimum delay in I/O during the trespass operations.

This benefits CLARiiON layered applications such as SnapView, MirrorView, and SAN Copy since it isolates those layers from having any knowledge that redirection is taking place. The back-end fault masking feature is provided to all Failover Mode types.

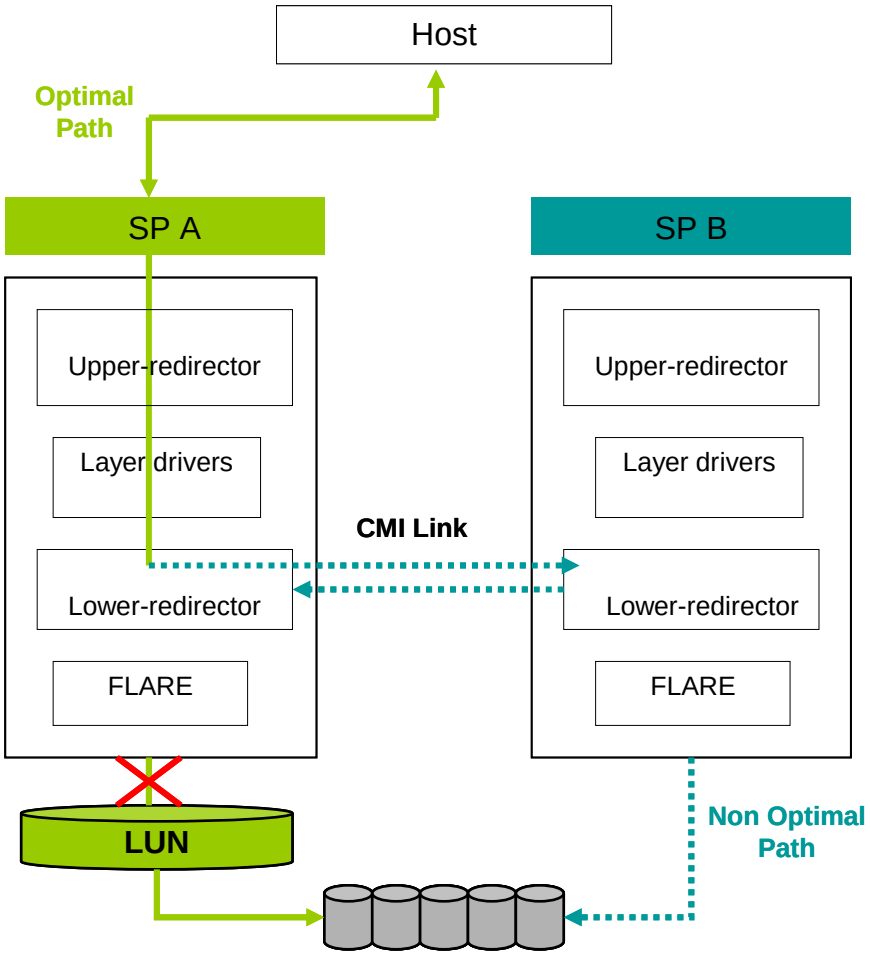

Figure 5. Masking back-end failures without failover software

Comparison of failover methodologies

Path management software is generally a server-based application that interacts with the storage system to automate and manage multiple paths to data devices or LUNs. Several options of path management software are supported by the CLARiiON storage system. As described in the "CLARiiON Failover Modes" section, CLARiiON can be configured to run in several different Failover Modes with different path management software.

EMC PowerPath is host-based software that provides path management. PowerPath works with several storage systems, on several operating systems, with both Fibre Channel and iSCSI data channels. Implementation details and various features of PowerPath and respective native failover software (MPIO) are discussed next.

Explicit and implicit failover software

Explicit and implicit failover *software* should not be confused with the explicit and implicit trespass *commands* mentioned earlier.

Failover software that supports *explicit* ALUA commands monitors and adjusts the active paths between optimal and non-optimal using the `REPORT TARGET PORT GROUPS` and `SET TARGET PORT GROUPS` commands, respectively. The `SET TARGET PORT GROUPS` command, when issued on one controller, trespasses the LUN to the other controller. PowerPath has the same net effect as an explicit trespass, although it uses traditional CLARiiON proprietary commands to initiate a trespass. Windows 2008 MPIO supports explicit ALUA commands.

It is possible for host multipathing software to support *implicit* ALUA functionality. The host could redirect traffic to the non-optimal path for some reason that is not apparent to the storage system. In this case, the storage system must implicitly trespass once it detects enough I/O on the non-optimal paths. To get the optimal path information, the failover software uses the `REPORT TARGET PORT GROUP` command. AIX MPIO only supports implicit ALUA commands.

Both *explicit* and *implicit* failover software may or may not provide autorestore capability. Without the autorestore capability, after an NDU, all LUNs could end up on the same SP. Cluster software could also make use of this functionality.

PowerPath

PowerPath works with the storage system to intelligently manage I/O paths. Under normal conditions, PowerPath only issues I/O to LUNs only on optimal paths (paths to the owning SP). If optimal paths have failed, PowerPath issues a trespass command; changes the LUN ownership to the other SP; and redefines the non-optimal path as the optimal path.

A unique feature of PowerPath is that it uses a vendor-unique query to the peer SP to determine if the failed path is due to a failed SP. In that case, PowerPath issues a trespass immediately without checking path by path. This is an important advantage of PowerPath over MPIO.

PowerPath is an explicit failover software because it load balances only across the optimal paths and has the ability to issue a trespass operation. If all optimal paths fail, PowerPath issues a trespass and makes the non-optimal paths optimal, and load balances across them. PowerPath also supports the Asymmetric Active/Active feature for operating systems that do not have native failover software that supports ALUA. For example, a Windows 2003 host can be configured in ALUA mode with the 5.1 version of PowerPath that supports ALUA.

PowerPath:

- Load-balances all I/O only across the optimal paths, as opposed to load balancing I/O across optimal and non-optimal paths, which can lead to lower performance. PowerPath supports various policies for load balancing.
- Has an autorestore capability to restore LUNs to their default SPs after an SP returns to health following a failure/NDU/cable or switch failure. This ensures even workload balancing of LUNs across SPs and more predictable performance.
- Supports device prioritization and proactive path testing.
- Version 5.1 and later support both ALUA and non-ALUA mode LUNs on the same host attached to multiple arrays.
- Has ALUA-specific I/O retry optimization to improve failover time in the event of SP failures and non-disruptive upgrades (NDUs).
- Displays for each LUN, the modes (ALUA versus non-ALUA) along with CLARiiON nice-names
- Supports boot from SAN.

A PowerPath license (full functionality) is embedded in the PowerPath package for the AX4 series. PowerPath SE is bundled free with CX4-120 and higher models of CLARiiON. PowerPath SE provides single HBA support and limited load-balancing options. To get full functionality PowerPath support with CX4-120 and higher models, full PowerPath software is required and can be purchased. Full PowerPath has more flexible options for attach (multiple HBA) and load-balancing policies, together with unique features as mentioned previously.

Native failover software (MPIO and others)

To many people, native failover software means "MPIO." MPIO is a commonly used name for the host-side interface that handles multipathed LUNs. However, MPIO implementations on different vendor operating systems do not have the same API. Some operating systems require storage vendors to provide libraries to implement their MPIO framework, while some operating systems include libraries (such as Solaris MPIO).

MPIO framework can be multi-mode. Some vendors support Active/Active, Active/Passive, and ALUA. HP-UX 11iv3 and MS Windows Server 2008 limit support to ALUA or Active/Active.

As mentioned in Table 1 on page 8, CLARiiON supports various native MPIO software.

Native failover systems (ALUA and non-ALUA) often do not offer all the features that PowerPath offers. These should be noted before a commitment to native failover software is made. Some of the features not offered in failover systems include:

- HP-UX non-ALUA-based PVLinks (before 11iv2) has no load balancing.
- HP-UX 11iv3.1 MPIO offers load balancing but not autorestore.
- LINUX non-ALUA-based MPIO for RH4 and SUSE 9 do not support boot from SAN.
- LINUX ALUA-based MPIO, for both RHEL and SuSE, supports autorestore and load balancing; however, it cannot detect a non-responding (hung) SP. There are no timers in Linux. PowerPath can detect a hung SP as PowerPath times and checks the peer SP.
- Solaris non-ALUA-based MPxIO does not support RAID 6.
- Solaris ALUA-based MPxIO does not support autorestore. EMC PowerPath provides autorestore capability. This ensures even workload balancing of LUNs across SPs and more predictable performance.
- AIX MPIO supports both the "Round Robin" and "Failover" polices with *implicit* ALUA commands; however, it does not support autorestore.
- VMware MPIO for ESX 4.x supports both the FIXED and Round Robin policies with ALUA. The Round Robin policy offers load balancing but no autorestore. The FIXED policy supports autorestore but no load balancing.

Nondisruptive upgrade (NDU)

Nondisruptive upgrade (NDU) requires a reboot of each SP in turn. For the most part, ALUA does not change how the storage system works in an NDU. Using PowerPath with Active/Passive mode (Failover Mode 1), PowerPath trespasses the LUN to the other SP while the owner SP is rebooting. After the NDU operation completes, the PowerPath auto restores the LUN to the SP that is the default owner of the LUN.

In Asymmetric Active/Active mode, during the SP reboot the failover software detects the failed paths and redirects I/O to the SP that is up and running. When the NDU completes, CLARiiON relies on the failover software to interrogate the preferred bit returned by the REPORT PORT GROUPS command to move LUNs back to the default owner – thus restoring the original path structure.

PowerPath trespasses a LUN to the peer SP during an NDU operation. After an NDU completes, PowerPath auto restores the LUN to its default owner. This is true whether the initiators are configured using Failover Mode 1 (Active/Passive) or 4 (ALUA). Note that PowerPath does not use the preferred bit to perform the autorestore but uses CLARiiON proprietary commands to issue a trespass.

Manual trespass

In the case of Active/Passive mode (Failover Mode 1), when a manual trespass is issued (using Navisphere Manager or the CLI), subsequent I/O for that LUN is rejected over the SP on which the manual trespass was issued. This would result in a unit attention condition for the host; failover software detects that error and re-routes the I/O to the SP that owns the LUN.

Any trespass operation, automatic or manual, causes the ownership of the LUN to change and a **unit attention** to be sent to connected hosts. These hosts, being ALUA aware, would interpret the unit attention and query the current status using appropriate Target Port Group commands.

PowerPath, DMP, and most native MPIO software that provide Explicit ALUA support immediately act on the unit attention by routing all I/O to the optimal path so that I/O does not go through the non-optimal path.

Therefore, if a user manually trespasses a LUN, PowerPath, DMP and all MPIO software will continue to use optimal paths. As per the CLARiiON ALUA implicit trespass mechanism, changes in I/O balance may cause the LUN to trespass implicitly (as described in the “Explicit and implicit trespass” section on page 7).

Path, HBA, and switch failure

If a host is configured with Failover Mode 1 and all the paths to the SP that owns a LUN fail, the LUN is trespassed to the other SP by the host’s failover software.

With Failover Mode 4, in the case of a path, HBA, or switch failure, when I/O routes to the non-owning SP, the LUN may not trespass immediately (depending on the failover software on the host). If the LUN is not trespassed, FLARE will trespass the LUN to the SP that receives the most I/O requests to that LUN based on the implicit trespass mechanism.

SP failure

In case of an SP failure for a host configured as Failover Mode 1, the failover software trespasses the LUN to the surviving SP.

With Failover Mode 4, if an I/O arrives from an ALUA initiator on the SP that does not own the LUN (non-optimal), failover software or FLARE initiates an internal trespass operation. This operation changes the ownership of the target LUN to the surviving SP, because its peer SP is dead. The host (failover software) must have access to the secondary SP so that it can issue an I/O under these circumstances.

Conclusion

Significant benefits can be achieved through the Asymmetric Active/Active feature available with FLARE release 26 or later. This feature increases configuration flexibility and investment protection for a changing environment, and provides increased availability for systems.

Asymmetric Active/Active is a SCSI standard. It supports host failover methods that adhere to this specification. Hosts can now avoid unwanted trespassing in certain scenarios, as I/O can be redirected to the SP that owns the LUN

The back-end fault-masking feature is provided to all Failover Mode types. In the case of a back-end failure on the SP that is the current owner of the LUN, I/O is routed via the lower redirector of the peer SP with the stable back end. This avoids trespassing, which can often impede replication operations.

All ports can be used to access the same LU simultaneously. The Asymmetric Active/Active multipathing feature is a software enhancement to the current base software package. It will be upward compatible with the existing Failover Modes. It presents an Asymmetric Active/Active model, enabling host I/O to a LUN over all ports based on the optimal and non-optimal path as reported by REPORT TARGET PORT GROUP.

Asymmetric Active/Active is part of the SPC-3 SCSI standard and is a new selectable Failover Mode just like the current modes of 0 (Auto-trespass), 1 (Passive Not Ready), 2 (DMP), and 3 (Passive Always Ready). Note that the Asymmetric Active/Active model can be applicable for any given attach to the array but is not the default system behavior. A PNR initiator and an Asymmetric Active/Active initiator can connect through the same physical SP port and have a different failover behavior where PNR behaves the same as in legacy systems.

References

Name: *EMC CLARiiON Storage System Fundamentals for Performance and Availability — Applied Best Practices*
Type: White paper
URL: http://powerlink.emc.com
Audience: Customer, Employees, Partners
Technical Depth: High

Name: *EMC CLARiiON Integration with VMware ESX — Applied Technology*
Type: White paper
URL: http://powerlink.emc.com
Audience: Customer, Employees, Partners
Technical Depth: High

Name: *EMC PowerPath Product Guide*
Type: Product documentation
URL: http://powerlink.emc.com
Audience: Customer, Employees, Partners
Technical Depth: High