Technical Report

Clustered File Systems with E-Series Products

Best Practices for Media and Entertainment Customers

Authors: M. K. Jibbe, Dean Lang, and Ahmad Moubarak, NetApp
Reference: M. K. Jibbe, PhD
November 2018 | TR-4604

Abstract

E-Series storage arrays deliver high-bandwidth performance and extreme reliability for file systems that support multiple host types with access to the same back-end LUNs. This document provides a reference architecture, including complex multipath details that must be considered when setting up a StorNext file system using multiple host types and running a media application.

NetApp®

TABLE OF CONTENTS

© 2018 NetApp, Inc. All rights reserved.

LIST OF TABLES

LIST OF FIGURES

 © 2018 NetApp, Inc. All rights reserved.

1 Introduction

NetApp® E-Series storage arrays are often deployed in cluster environments that require high-bandwidth and low-latency performance. However, volume ownership thrashing and other multipath/failover issues have been longstanding challenges in clustered environments, especially when hosts running different OSs access the same storage volumes on the array.

Over the past few years, NetApp has released enhancements in the E-Series firmware that improve interoperability between hosts and E-Series arrays in heterogeneous host OS clustered environments. These changes have culminated in a new failover mode that greatly improves cluster interoperability, and these most recent enhancements are the next step in an ongoing evolution. This includes two new enhancements released in April 2017 that are designed to specifically resist cluster volume ownership thrash.

Options have existed in the past that allow multipath failover to be avoided in deployments where that option made sense, but now these new options exist to build cluster solutions without risk of ownership thrash while still retaining failover and host I/O access redundancy. For some limited configurations, options also exist to provide limited failback in clustered environments. This enables automatic recovery software to move LUN ownership back to preferred paths following the correction of SAN connectivity faults, again without the risk of LUN ownership thrash.

This paper includes an example of an E-Series array deployed using a thrash-resistant, redundant cluster and includes performance test results using a media application from Autodesk to generate the I/O and measure performance. The Autodesk self-certification test suites and stress tests were run on a StorNext cluster and a 96-drive E-Series E5624 storage system. Such workloads mimic the Autodesk Flame application, where the tests simulated multiple workstations playing back media with different frame sizes (HD, 2K, 4K) and combinations of read and write operations simultaneously. The Autodesk self-certification test results and the Flame stress test suites results are covered in this report.

2 E-Series Redundant Host Access/Failover: Background Information

The following information is provided to level-set terminology and to aid in understanding the configuration settings and best practices outlined later in this document. The section presents a significant amount of detailed information, but taking the time to understand the concepts and associated implications of specific multipath driver behavior is useful when following the cluster configuration best practices outlined later in this document. This background can be even more valuable when trying to troubleshoot potential causes of volume ownership thrashing or other multipath issues in an existing clustered environment.

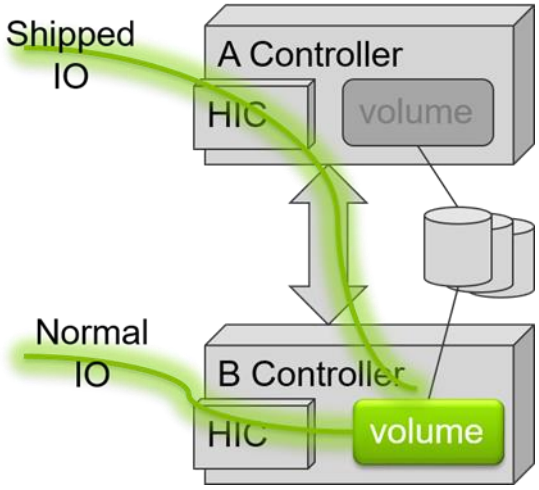
2.1 Volume Ownership Model and ALUA

To achieve consistent, ultralow latency, the E-Series firmware implements a highly optimized write data cache coherency model and a lean I/O path that uses the notion of volume ownership. This means that I/O for a given volume must be processed by the controller that owns that specific volume. This design greatly simplifies the data-caching implementation and hardware design, minimizing I/O latency while keeping the overall product cost low.

Despite using a design that employs an ownership model, E-Series arrays also have multiple target ports on each I/O controller, either on a host interface card (HIC) or on the controller baseboard itself. Both HIC and baseboard ports support an active-active access model. With this model, hosts can access all volumes through any target port on either controller, regardless of volume ownership. This access is accomplished with an “I/O shipping” mechanism, in which I/O received on the nonowning controller is processed by the owning controller using intercontroller messaging. For example, Figure 1 shows the case where a volume is owned by the B controller. I/O directed by the host to the A controller is first shipped through the controller interconnect channels to the B controller, then processed.

© 2018 NetApp, Inc. All rights reserved.

Figure 1) ALUA support implementation and I/O shipping example.



This implies that I/O sent to the nonowning controller can result in degraded performance, and so the performance observed by a host is not equal across all the target ports on the array. This is referred to as Asymmetric Logical Unit Access (ALUA), as opposed to a fully symmetric active-active design in which performance is equal across all target ports.

ALUA was first available with E-Series firmware in late 2011. Prior to that, any I/O directed by a host to ports on the nonowning controller was rejected, and a check condition with a specific sense qualifier indicating the volume is not owned by that controller was returned to the host.

2.2 TPGS Reporting and Legacy Redundant Dual Active Controllers

E-Series arrays also support RDAC (redundant dual active controller) T10-SPC (SCSI) standard methods for reporting groups of controller target ports and their associated performance characteristics or asymmetric access states. This method, typically referred to as target port group support (TPGS), allows reporting of the specific target port group or groups that offer optimal performance to the host.

With TPGS, the host multipath driver can obtain information about the groups of target ports on each array controller using the `REPORT TARGET PORT GROUPS` command. E-Series arrays report two groups of target ports, one group for the ports on each controller.

The scope of the `REPORT TARGET PORT GROUPS` command is volume-specific, meaning each volume can report a different asymmetric access state for each of the two groups of target ports. The target port group on the controller that owns a volume is reported as having an asymmetric access state of active/optimized (AO) because that controller can handle I/O requests with maximum/optimal performance. The target port group for the ports on the controller that does not own the volume is reported as having an asymmetric access state of active/nonoptimized (ANO) because directing I/O requests at the nonowning controller implies a performance penalty. Host multipath drivers use this information to route I/O to the target ports that result in optimal performance (that is, I/O to a given volume is routed to the target ports on the controller that owns that volume). Table 1 summarizes the E-Series volume ownership model. Alternating volume ownership continues as additional volumes are provisioned.

Table 1) ALUA model for volume ownership and access.

Volume Name	Volume Owner	Target Port Group Asymmetric Access States	
		Ports on Controller A	Ports on Controller B
Volume 1	Controller A	Active/optimized	Active/nonoptimized

© 2018 NetApp, Inc. All rights reserved.

Volume Name	Volume Owner	Target Port Group Asymmetric Access States	
		Ports on Controller A	Ports on Controller B
Volume 2	Controller B	Active/nonoptimized	Active/optimized
Volume 3	Controller A	Active/optimized	Active/nonoptimized
Volume 4	Controller B	Active/nonoptimized	Active/optimized

Note: Alternating volume ownership is automatically enabled when using the SANtricity® GUI to create volumes (SANtricity must be able to access both controllers). When using command line scripts, volume ownership is specified in the script and does not automatically alternate controller ownership.

Support for TPGS reporting was introduced initially for a few specific host types in the 07.36.XX.XX firmware release (circa 2008) and prior to support for ALUA. In these earlier firmware releases, the two target port group states were reported as active/optimized and standby instead of active/optimized and active/nonoptimized.

Prior to support for TPGS, all host multipath driver interactions between the host and E-Series arrays utilized a set of vendor-unique inquiry vital product data pages and mode pages defined by LSI/NetApp to report and manage volume ownership. Host multipath drivers that utilize these vendor-unique methods are often referred to as legacy redundant dual active controller (RDAC) drivers.

2.3 Preferred and Current Volume Ownership: Failover and Failback

The controller that owns a given volume for data-caching purposes as described earlier is typically referred to as the current owner of the volume. Each volume is also assigned a preferred owner during the provisioning process that is distinct from the current owner. As noted earlier, the preferred owner is assigned by the user through the CLI and is automatically assigned using the GUI at volume creation time. Ownership can be changed by the user after a volume is initially created.

In optimal conditions, the current and preferred owner for a given volume are the same. However, in some conditions where one or more hosts lose access to the preferred owning controller, a failover might occur, resulting in the current and preferred owner being different. After the SAN connectivity fault that resulted in the failover is corrected, the failback process moves current volume ownership back to the preferred controller (more about this later).

`Preferred owner` is reported to the host multipath driver in the data returned from the array in response to the `REPORT TARGET PORT GROUPS` command. Essentially, one of the two target port groups is marked as the preferred (`PREF=1`) group, as defined in the T10-SPC standard, based on which controller is configured as the preferred owner of the volume. This allows the host multipath driver to monitor both current owner (based on the target port group asymmetric access state) and the preferred owner of a given volume.

2.4 Explicit Versus Implicit ALUA/TPGS and Failover

T10-SPC defines two methods of ALUA support: implicit and explicit. Implicit ALUA means the array controller manages the asymmetric access states of the target port groups (that is, the array manages the volume ownership, which in turn determines the asymmetric access state of each target port group). This is accomplished by monitoring all incoming read/write requests to a given volume on the array and determining which controller is receiving the bulk of the incoming requests. In the absence of any SAN connectivity fault, the preferred owner and current owner of a given volume are the same controller. It tends to stay that way because the host multipath driver directs I/O at the current owning controller (active/optimized target port group) in order to maximize performance. When a fault occurs such that hosts with access to the volume lose connection to the preferred/current controller, I/O is directed at the nonowning/nonpreferred controller, and the array decides to move current ownership to the nonpreferred

© 2018 NetApp, Inc. All rights reserved.

controller to minimize the I/O shipping penalty. By definition, this means more than 75% of the I/O to a given volume over a 5-minute period arrived at the nonowning controller. The result is an implicit failover, and the current and preferred owner are now different.

Note: When an implicit failover occurs, a special notice is sent to all hosts with access to that volume using a `UNIT ATTENTION` check condition, which notifies the host multipath driver that the active/optimized target port group has now changed. A well-behaving host multipath driver simply follows this change and starts directing I/O at the new owning controller (new active/optimized target port group).

Explicit ALUA means that the host multipath driver manages the volume ownership using the `SET TARGET PORT GROUPS` command defined by T10-SPC, which effectively allows the host to change volume ownership by requesting a change to the target port group asymmetric access states. For example, a host could request a change to the controller A target port group asymmetric access state for volume 1 from active/optimized to active/nonoptimized (and vice versa for the controller B target port group), which would tell the array that the volume ownership on volume 1 would have to move from controller A to controller B. An explicit failover would occur when one host loses access to the current owning controller for a given volume. To maintain performance, the host makes an explicit request to the array to change ownership so that the remaining paths retain full performance.

Support for implicit or explicit ALUA (or both) is reported by the array to the host multipath driver using a field in the standard inquiry data reported by the array. If both explicit and implicit modes are reported to the host multipath driver, it is up to the host driver to select which mode is used. Further details about how implicit versus explicit ALUA support is reported to a given host are covered later in this document.

Note: Legacy RDAC drivers using NetApp vendor-unique multipath commands historically operated in a mode equivalent to explicit ALUA.

2.5 Host Versus Implicit/Target (Array-Initiated) Failback

Like the failover process, the failback process is when a volume's current ownership is returned to the preferred controller. This can be initiated either by the host multipath driver (explicit failback) or by the array itself (implicit failback, sometimes referred to as target failback).

With explicit failback, the host issues a `SET TARGET PORT GROUP` request (or in the case of a legacy driver, a vendor-unique command), which results in the current volume ownership being returned to the preferred controller. This typically occurs when the host sees that at least one connection to the preferred controller has been reestablished.

Beginning in 08.30.XX.XX firmware, E-Series added support for implicit failback where the array monitors the connectivity to each controller for all hosts that have access to a given volume. The array initiates an ownership change on that volume to move it back to the preferred controller after all hosts have access again using the preferred controller.

2.6 Mixed-Mode Multipath Drivers

Some multipath drivers rely on the array to manage failover (that is, use implicit failover) but make explicit requests to fail back when connectivity returns to the preferred controller (that is, explicit failback). Although this mixed mode seems odd, it was a very common behavior in many mainstream drivers, including our own Windows MPIO DSM (NetApp E-Series-specific DSM v1.X) and the device handler specific to NetApp E-Series for Linux DM-MP (scsi_dh_rdac). This behavior in the drivers specific to NetApp was essentially an interim phase in the migration from the historically full explicit mode (both explicit failover and failback) to the full implicit mode (both implicit failover and failback) deployed beginning with 08.30.XX.XX firmware and associated multipath drivers.

 © 2018 NetApp, Inc. All rights reserved.

2.7 Explicit Versus Implicit Failover and Failback: Cluster Implications

In a cluster environment where multiple hosts have access to the same volume, explicit ALUA carries significant risk of volume ownership thrash because individual hosts in the cluster might not agree with each other about the appropriate ownership of a given volume, especially if one host loses access to the preferred controller while one or more of the other hosts in the cluster retain access to the preferred controller. The host that must access the volume using the nonpreferred controller changes ownership to retain full performance, but the other hosts in the cluster see that the volume is off the preferred controller even though they still have connectivity to that controller and initiate a failback operation to change ownership back to the preferred controller. This ownership fight can continue until the connectivity problem is resolved. During this time window, volume ownership can oscillate between controllers very rapidly if the host multipath drivers that conflict are aggressive about initiating explicit failover or failback requests. This thrashing can even happen with mixed-mode drivers using implicit ALUA because one host might direct I/O at the nonowning (but preferred) controller even if they don't issue an explicit ownership transfer request. In this case, the ownership thrash typically is not as rapid. The failover takes at least 5 minutes of heavy I/O to the nonowning controller before the array initiates an implicit failover, followed relatively quickly by another host requesting an explicit failback.

Note: This issue was mitigated by having administrators disable failback in the multipath configuration settings on the host to stop at least one side of the ownership fight. Even if a single host in the cluster was misconfigured and did not have failback disabled, ownership thrash could still result in the entire cluster.

In general, any use of explicit failover or failback is toxic in a clustered environment with shared volume access. This means that any driver using explicit failover must be completely avoided because failover cannot be disabled without losing the ability to have redundant access. Drivers using explicit failback in clustered environments are risky but can be used if failback is disabled in the host driver. This means that resolving ownership issues must be manually triggered by the user after correction of SAN connectivity faults.

Use of implicit ALUA for both failover and failback is highly recommended in a clustered environment with shared volume access. The array has visibility to the current connectivity of all hosts in the cluster from both array controllers as well as visibility to incoming I/O patterns and can therefore select the optimal volume ownership. Essentially, the array has the right vantage point to arbitrate shared volume access fairly and determine when ownership transfers are appropriate. With use of implicit ALUA, it is possible to build a cluster with I/O path redundancy and resiliency without risk of ownership thrash.

Beginning in 08.30.XX.XX firmware, full implicit ALUA (both failover and failback) support is available with Windows MPIO, Linux DM-MP, and VMware NMP/SATP_ALUA multipath drivers. This is accomplished by the array reporting only implicit ALUA support to the host multipath driver in the standard inquiry data to hosts running those multipath solutions, essentially denying the host the ability to use explicit ALUA. The array also hides preferred volume ownership information in `REPORT TARGET PORT GROUPS` responses to these specific OS/drivers to make sure the host does not attempt a failback of any kind. Reporting preferred target port group information is optional in the SCSI standard, so it can be withheld by the array. This essentially builds a cluster-safe multipath solution with full redundancy (failover) and autorecovery (failback).

Beginning with controller firmware 08.20.24.00, 08.25.11.00, and 08.30.20.00, support for an ATTO clustered host type is available. This enables implicit ALUA mode for the ATTO multipath solution.

Note: Only implicit failover is supported for the ATTO multipath solution because the current NetApp implicit failback support is not compatible with the ATTO multipath solution behavior.

Historically, the ATTO multipath solution used explicit ALUA for both failover and failback, making ownership thrash a common problem in clustered solutions using the ATTO HBA. The ATTO multipath driver behavior disables failback completely when the array reports implicit ALUA-only support in the standard inquiry data. As a result, there are no compatibility issues with enabling implicit failover for

 © 2018 NetApp, Inc. All rights reserved.

ATTO, just the lack of failback support, which would have been disabled in a clustered environment in the past anyway. There is no need to modify host-side configuration to disable failback anymore.

Note: The ATTO multipath driver has long supported implicit ALUA and just adjusts its behavior based upon the standard inquiry data reported by the array. Therefore, enabling implicit failover support for ATTO simply requires selection of a new host type in the array configuration for hosts with ATTO HBAs.

Having full implicit failover and failback support in the array means the host-side multipath configuration is not as critical in a cluster as long as the host multipath driver adheres to the implicit versus explicit ALUA support reported by the array. This essentially puts the array in charge of all volume ownership and is the direction E-Series engineering is trying to move with all host multipath solutions, given the ever-increasing prevalence of scale-out/clustered solutions.

3 E-Series Storage Partitioning Configuration

During initial device discovery, when the multipath solutions installed on a host are discovering storage devices on the SAN, the standard inquiry data reported by the array controllers and responses to commands such as `REPORT TARGET PORT GROUPS` are key in whether the multipath solution even recognizes the storage, and if it does recognize the storage, what multipath mode (implicit versus explicit ALUA) it selects to use. Proper storage driver operation is critical in making sure of successful interoperability between the array and host.

Note: In some cases, multiple multipath drivers might be installed on a host (that is, in-box OS multipath driver plus vendor-unique drivers installed, and so on). The initial device discovery phase is very critical in making sure the correct multipath solution claims the storage devices associated with E-Series volumes. This primarily depends on the specific inquiry data returned from the array.

Because E-Series supports many different OS/multipath combinations and each of those combinations best operates in different modes requiring different inquiry responses, the host type configuration on the array is critical to make sure of proper responses and successful host/array interoperability.

3.1 Recognizing Host That Initiated a Command (and Associated Host Type)

The E-Series array controller firmware uses the SANtricity storage partitioning configuration information set up by the user (hosts, host ports, host groups) to identify the specific host that sent a command to the array. At the host interface protocol level, the originating worldwide unique port identifier is known for each incoming command, but that port identifier must then be associated with a given host to determine the host type and associated OS/multipath solution. This in turn is used to control some of the host type-specific behavior that must be employed by the array when processing the incoming request. This includes managing access to the specific LUNs/volumes that should be accessible over that host port. Therefore, the mapping of host ports to hosts and the operating system type selected on those hosts in the array configuration are essential configuration parameters that must be correct to make sure of proper responses on incoming device discovery commands received over a given host port.

These configuration parameters are of the utmost importance in a heterogeneous host-type, clustered environment where a single host with a multipath driver operating in explicit ALUA mode could easily trigger volume ownership thrashing in the entire cluster.

Another side effect of an improper setup can be that the installed host multipath drivers cannot claim E-Series volumes during device discovery because the array sends incorrect/unexpected responses to discovery-related commands. This condition is often due to an unknown or misunderstood host operating system type. It can also lead to a condition where host applications have access to the same volume multiple times through multiple device nodes in the OS device stack. By default, the E-Series array exposes each volume in the system on all controller target ports, and the customer's SAN configuration likely exposes each host HBA port to at least one controller target port on each controller in the array. It is the multipath driver that coalesces the multiple paths to the same storage device/volume.

© 2018 NetApp, Inc. All rights reserved.

3.2 Common Host Types and Associated Failover Settings

Table 2 and Table 3 outline a few of the common host types configurable in E-Series arrays and highlights some of their associated multipath-related parameters. The expected multipath solution should claim the storage devices for the E-Series volumes if the associated driver is installed and properly configured on the host.

Note: Any host type listed in this table that uses explicit failover in the multipath solution is essentially not cluster safe.

Note: Any host type that uses explicit failback is not cluster safe without some mechanism of disabling the failback functionality.

The definition or behavior controlled by these specific host types is encoded in the E-Series NVSRAM build and the controller firmware code. Therefore, it is critical that the correct NVSRAM is installed for a given release of E-Series controller firmware. The following tables are based on NetApp product software builds. Table 2 indicates functionality with SANtricity firmware builds 08.20.XX.XX and 08.25.XX.XX.

Table 2) Common host types: multipath settings (08.20.XX.XX–08.25.XX.XX firmware/NVSRAM builds).

Index	Host Type	TPGS	ALUA Support	Failover Mode	Failback (Mode + Host or Target Driven)	Expected Multipath Solution
1	Windows	Y	Y	Implicit	Explicit/host	Windows MPIO with NetApp DSM 2.0
2	Solaris (v10 or earlier)	N	N	Explicit	Explicit/host	Solaris MPxIO (using legacy failover methods, not TPGS)
6*	Linux MPP/RDAC	N	N	Explicit	Explicit/host	Legacy Linux MPP driver
7	Linux DM-MP (kernel 3.9 or earlier)	N	Y	Implicit	Explicit/host	Device mapper multipath (DM-MP) with scsi_dh_rdac device handler
8	Windows clustered	Y	Y	Implicit	Explicit/host	Windows MPIO with NetApp DSM 2.0
10	VMware	Y	Y	Implicit	None	VMware NMP/SATP_ALUA
17	Solaris (v11 or later)	Y	Y	Explicit	Explicit/host	Solaris MPxIO (with TPGS support)
22	Mac OS (ATTO HBA)	Y	Y	Explicit	Explicit/host	ATTO Multipath Director
23	Windows (ATTO HBA)	Y	Y	Explicit	Explicit/host	ATTO Multipath Director
24	Linux (ATTO HBA)	Y	Y	Explicit	Explicit/host	ATTO Multipath Director
25	Linux (PathManager)	Y	Y	Explicit	Explicit/host	SGI PathManager
29**	ATTO clustered (all OS)	Y	Y	Implicit	None	ATTO Multipath Director

© 2018 NetApp, Inc. All rights reserved.

*Present in NVSRAM, but support was dropped in 08.25.XX.XX and later firmware.

**New in 08.20.24.00 and 08.25.11.00.

Table 3 indicates multipath functionality with SANtricity 08.30.XX.XX and later firmware builds. This later E-Series software added the automatic load balancing feature, so the table also indicates if the new feature is supported for the given host type.

Table 3) Common host types: multipath settings (08.30.XX.XX and later firmware/NVSRAM builds).

Index	Host Type	TPGS Support	ALUA Support	Failover Mode	Failback (Mode + Host or Target Driven)	Automatic Load Balancing (ALB) Support	Expected Multipath Solution
1	Windows	Y	Y	Implicit	Implicit/target	Yes	Windows MPIO with NetApp DSM 2.0
2	Solaris (v10 or earlier)	N	N	Explicit	Explicit/host	No	Solaris MPxIO (using legacy failover methods, not TPGS)
6*	Linux MPP/RDAC	N	N	Explicit	Explicit/host	No	Legacy Linux MPP driver
7	Linux DM-MP (kernel 3.9 or earlier)	N	Y	Implicit	Explicit/host	No	Device mapper multipath (DM-MP) with scsi_dh_rdac device handler
8	Windows clustered	Y	Y	Implicit	Implicit/target	Yes	Windows MPIO with NetApp DSM 2.0
10	VMware	Y	Y	Implicit	Implicit/target	Yes	VMware NMP/SATP_ALUA
17	Solaris (v11 or later)	Y	Y	Explicit	Explicit/host	No	Solaris MPxIO (with TPGS support)
22	Mac OS (ATTO HBA)	Y	Y	Explicit	Explicit/host	No	ATTO Multipath Director
23	Windows (ATTO HBA)	Y	Y	Explicit	Explicit/host	No	ATTO Multipath Director
24	Linux (ATTO HBA)	Y	Y	Explicit	Explicit/host	No	ATTO Multipath Director
25	Linux (PathManager)	Y	Y	Explicit	Explicit/host	No	SGI PathManager
28	Linux DM-MP	Y	Y	Implicit	Implicit/target	Yes	Device mapper multipath (DM-MP) with

© 2018 NetApp, Inc. All rights reserved.

Index	Host Type	TPGS Support	ALUA Support	Failover Mode	Failback (Mode + Host or Target Driven)	Automatic Load Balancing (ALB) Support	Expected Multipath Solution
	(kernel 3.10 or later)						scsi_dh_alua device handler
29**	ATTO clustered (all OS)	Y	Y	Implicit	None	No	ATTO Multipath Director

*Present in NVSRAM, but support was dropped in 08.25.XX.XX and later firmware.

**New in 08.30.20.00.

3.3 Default Host Group and Default Host Type

E-Series provides the notion of a default storage partition (default host group). Volumes (LUNs) can be mapped to the default host group, which essentially grants any connected but unconfigured host access to those volumes. When a host is first connected, if the host-port worldwide unique identifiers (HBA ports) from that host are not set up within the storage partitioning configuration in the array, that host by default has visibility to all volumes mapped to the default host group. Therefore, the default host group is essentially a set of volumes with wide-open access permissions.

From an ease-of-use/convenience perspective, the default host group is very attractive because the only requirement is to connect the host cables (or zone the switch), and immediate access to volumes is granted without taking any configuration actions on the array (although this might be of concern to users that are security conscious). However, the default host group does create a potential issue, especially in a heterogeneous host-type cluster, in the fact that the host type cannot be identified for any incoming commands from such unconfigured hosts. The array controller firmware applies a default host-type behavior, which might or might not result in the correct behavior during host device discovery.

Default Host Operating System Type

The array configuration does have the notion of a default host operating system, which is used to determine proper inquiry responses for such unconfigured hosts. But given that multiple OS/driver combinations are in use in a heterogeneous host-type cluster, applying a single host type to all hosts implies that for at least some of the hosts that default host type is incorrect. It is possible, however, to use the default host group without device discovery or multipath issues if all hosts within the group/cluster are of the same operating system type using the same multipath solution and the default host operating system is set accordingly. It is also possible to use the default host group without multipath issues if the single–target port LUN-mapping feature is being used (see Single–Target Port LUN-Mapping Feature for additional details).

Configured Hosts in Default Host Group

E-Series firmware supports the notion of placing configured hosts (each with defined host operating system types) in the default host group. In fact, all hosts are initially placed in the default host group when they are created, so it is possible to use the default host group and still have proper interoperability with hosts in a heterogeneous host-type cluster. In order for the host to remain in the default host group, LUNs cannot be mapped directly to any host defined in the default host group. Hosts in the default group can only have access to volumes mapped to the default group itself. As soon as a volume (LUN) is mapped to a host in the default group, that host leaves the default group, becomes its own storage partition, and now only has access to the volumes specifically mapped to it.

© 2018 NetApp, Inc. All rights reserved.

Factory Default Host Type

The factory default host type (host type 0) is the value set for the default host operating system when a new controller is shipped from manufacturing.

Historically, the settings for this host type mimicked the Windows host type behavior. However, as the behavior for Windows host types changed over time, the factory default host type did not keep pace with the changes. Prior to 08.30.XX.XX firmware, the factory default did not have ALUA or TPGS enabled at all (no support for TPGS in either implicit or explicit ALUA mode), even though ALUA support has been enabled for Windows for some time.

As a result, on 08.25.XX.XX or earlier firmware, use of the factory default host type results in most current generation multipath drivers ignoring E-Series storage devices, which can lead to various issues with host interoperability. If a multipath driver is installed on the host that understands NetApp E-Series legacy vendor-unique multipath commands (vendor-unique mode and inquiry pages), it might claim E-Series storage devices at device discovery time but operates in a legacy mode that is no longer tested by NetApp. Therefore, if the array is running 08.25.XX.XX or earlier firmware, it is critical that the default host operating system is changed to a type that is more representative of the actual OS running on hosts with access to the default host group.

Beginning with 08.30.XX.XX and later firmware, the factory default was updated to enable TPGS reporting of implicit ALUA support, which makes it much more compatible with current generation multipath solutions and would likely result in better interoperability with most hosts.

Note: It is still recommended that the user set the default host operating system type to the most common host type with access to the default host group.

To avoid all issues and limitations associated with the default host group, creating a new E-Series host group specifically used to support cluster access to a given set of volumes is a preferred method for setting up E-Series arrays in any clustered host environment.

3.4 Specific Host Operating Systems and Associated Multipath Solutions

The following information about specific host operating systems and associated multipath solutions is provided for background information and can be used in determining the appropriate host type to use for a given deployment.

Microsoft Windows: NetApp E-Series Specific MPIO DSM

NetApp develops an E-Series device-specific module (DSM) for the Windows MPIO multipath architecture. The current DSM revision 2.0 utilizes T10-SPC standard TPGS commands. The E-Series DSM prior to SANtricity firmware version 8.10.XXX.XX used the legacy vendor-unique multipath mode and inquiry pages used by prior versions.

As with all drivers and host plug-ins, it is essential that the version installed on the host was certified with the particular firmware versions in use on the attached arrays. Table 4 lists changes in the associated host type settings on the Windows host types over recent E-Series firmware releases:

Table 4) Windows MPIO: NetApp DSM revisions.

Firmware/NVSRAM Version	Expected DSM Version	Failover Method	Failback Method
Pre-07.83.XX.XX	1.X	Explicit failover/legacy (not TPGS) No ALUA support (active/standby)	Explicit/host-driven

© 2018 NetApp, Inc. All rights reserved.

Firmware/NVSRAM Version	Expected DSM Version	Failover Method	Failback Method
07.83.XX.XX–07.86.XX.XX	1.X	Implicit ALUA/legacy (not TPGS)	Explicit/host-driven
08.10.XX.XX–08.25.XX.XX	2.X	Implicit ALUA/TPGS (uses standard T10-SPC command set, not legacy mode pages)	Explicit/host-driven
08.30.XX.XX and later	2.X	Implicit ALUA/TPGS	Implicit/Array-driven (target failback)

Support for legacy failover methods has been retained in both array firmware and the DSM to facilitate backward/forward compatibility of the driver and firmware, so current firmware most likely works with an older driver and vice versa. Nonetheless, the NetApp Interoperability Certification test matrix should be consulted to determine proper versions for fully certified support because not all combinations are guaranteed to work.

Note that Microsoft also has a generic T10-SPC compliant DSM that might appear to work in most cases to manage multipath support of E-Series storage devices because both that DSM and E-Series firmware are T10-SPC compliant. However, unlike the NetApp specific DSM, the Microsoft DSM is not tuned to interoperate specifically with E-Series and so does not handle all SAN fault scenarios optimally. Therefore, NetApp does not test with the generic/Microsoft DSM, and it is not officially supported.

There are two Windows host types defined in E-Series firmware/NVSRAM, one for clustered solutions and one for nonclustered solutions. However, the settings for these two host types are effectively identical, and the reason for two separate host types is only historic at this point. Prior to 07.83.XX.XX firmware, the settings were different, and we had to keep both types because of existing configured hosts set to one of the two types. As a result, these two host types are essentially interchangeable with current generation firmware.

Linux Multipath Solutions

MPP/RDAC

Historically, NetApp developed an E-Series-specific multipath driver for Linux using a vendor-unique architecture and legacy/vendor-unique mode and inquiry pages to interact with the E-Series array. This is the Linux MPP driver that is sometimes called RDAC. Support for this driver was officially end of life with 08.20.XX.XX firmware (that is, support was dropped in 08.25.XX.XX and later), although the host type remains in NVSRAM/firmware for the sake of facilitating existing deployments upgrading to later firmware.

Note: After SANtricity is upgraded to 8.25.XX.XX or a later version, a new host type must be selected to use the new multipath solution.

Device Mapper Multipath (DM-MP)

Linux device mapper multipath (DM-MP) uses a plug-in architecture where a device handler understands the specific attributes of a particular storage device. There are two device handlers available in box in most Linux distributions that understand how to manage multipath for E-Series storage: scsi_dh_rdac and scsi_dh_alua. The scsi_dh_rdac device handler uses the E-Series vendor-unique mode and inquiry pages to manage failover and failback, whereas scsi_dh_alua uses T10-SPC TPGS standard methods and is not vendor-specific at all. The host type set in the array configuration determines which device handler claims E-Series storage devices. Therefore, the exact same Linux distribution might select scsi_dh_rdac as the DM-MP device handler if the Linux DM-MP (kernel version 3.9 and earlier) host type

© 2018 NetApp, Inc. All rights reserved.

is selected in SANtricity during the array configuration, or it might select scsi_dh_alua as the DM-MP device handler if the Linux DM-MP (kernel version 3.10 and later) host type is selected.

Support for scsi_dh_alua and the associated Linux DM-MP (kernel version 3.10 and later) host type was released in SANtricity 08.30.XX.XX firmware/NVSRAM. As the name suggests, Linux distributions based on kernel versions 3.10 and later have the appropriate version of the scsi_dh_alua device handler to properly interoperate with E-Series. Earlier versions of Linux must use the scsi_dh_rdac device handler with DM-MP for proper interoperability, so the Linux DM-MP (kernel version 3.9 and earlier) host type should be selected for hosts running Linux distributions based on kernel version 3.9 or earlier.

Note that sometimes one of the DM-MP device handlers distributed in box on Linux claims E-Series storage devices if the host type is set to another host type that happens to cause the array to return the expected answers for TPGS/ALUA support. For example, if the host type were set to Linux (ATTO) but the ATTO multipath solution was not installed on the host, it is entirely possible that the Linux in-box DM-MP/scsi_dh_alua multipath solution would claim the E-Series volumes. Therefore, care must be taken to understand exactly what is installed on the host, what is in box with the Linux distribution in use, and what the host type is set to in the array configuration to fully understand which multipath solution (or lack of any potentially) is in effect.

Cluster Thrash Risk with Linux (DM-MP/scsi_dh_rdac and MPP/RDAC)

It should be noted that the scsi_dh_rdac device handler on DM-MP and legacy MPP/RDAC driver on Linux use explicit host-driven failback. As noted earlier, when using these drivers in a cluster environment, the multipath configuration on the host must be set to disable failback to avoid ownership thrash.

Note that beginning with 08.20.24.00, 08.25.11.00, 08.30.20.00, and later firmware builds, support for a special failback disable mechanism within the array firmware can be applied for DM-MP/scsi_dh_rdac to assist with avoiding ownership thrash when DM-MP/scsi_dh_rdac must be used due to specific Linux versions in use. See the Preventing Failback and Cluster Thrash with Non-ATTO Host Types section later for details.

Because the legacy MPP multipath solution uses explicit failover and explicit failback and does not have ALUA support at all, it is not cluster-friendly and should be avoided in clustered solutions.

3.5 Host Context Agent (HCA)

Given both the complexity and importance of properly configuring hosts and host ports, selecting the correct host type when configuring the array is the first step to a successful multipath implementation. The SANtricity host context agent software was developed to assist and automate much of the host setup process. This agent, packaged with SANtricity Storage Manager, detects when new storage arrays are attached to a host. The agent gathers information from the OS to determine which drivers are installed, and it gathers information from the newly attached array to confirm the supported host types and the port worldwide unique identifiers that connect the host to the new array. This information is then used to automatically configure the host and host ports in the array as well as set the host type correctly.

Note that the HCA does not update an existing configuration, so if the host port worldwide unique identifier has already been configured/registered in the array configuration, it is not reregistered if the host is removed and then reattached. This is intentional to preserve any configuration changes the user applied manually after the initial configuration. Similarly, if the host port worldwide unique identifier and host had previously been manually configured by the user, the HCA does not modify that manual configuration.

In some cases, it might not be desirable or feasible to install SANtricity software packages on the I/O attach host to enable HCA support, but if it is possible, the HCA might provide one alternative to simplify the initial setup.

 © 2018 NetApp, Inc. All rights reserved.

3.6 Volume Not on Preferred Path Needs Attention Condition

Whenever volume ownership is transferred away from the preferred controller, the E-Series firmware posts a volume not on preferred path (VNOPP) alert to the user. This is intended to indicate that some sort of SAN connectivity fault resulted in one or more hosts in the cluster directing I/O at the nonowning controller, which in turn resulted in a failover and volume ownership changed from the preferred E-Series controller to the nonpreferred controller.

However, in cluster configurations where the multipath driver or the host interoperability with the array is suboptimal in some way (such as use of an explicit ALUA driver or lack of any multipath driver at all on one or more hosts in the cluster), the VNOPP alert can be triggered if hosts begin directing I/O at the nonowning controller or an explicit ALUA driver specifically requests an ownership change on a volume. This can occur without any connectivity fault in the SAN. The result is a phantom alert from the user's viewpoint because there appears to be no reason for the issue and no issue that can be resolved to make the alert go away.

The key to eliminating these phantom VNOPP alerts is to make sure of proper array/host interoperability, and that starts with building a cluster that avoids using any explicit ALUA drivers (or at least employs implicit failover and disables any explicit failback). The mechanisms and best practices in this document are intended to reduce phantom VNOPP alerts such that any alert should be a true indicator of SAN connectivity issues that need to be addressed.

Note that because failback is disabled in most cluster environments, when a real SAN connectivity fault causes a legitimate VNOPP alert, the alert is not automatically cleared after the connectivity issue is corrected unless the user takes manual action to realign volume ownership back with the preferred controller. Long term, as solutions are built that can allow implicit failback within a cluster environment, the need to manually clear VNOPP alerts might eventually be eliminated.

4 Media/StorNext-Specific Features and Enhancements

The following E-Series features and enhancements were implemented with the specific needs of the media and entertainment user, targeting use cases where storage configurations are more dynamic than would usually be the case in a typical enterprise data center. In these environments, host access redundancy is often less of a requirement.

4.1 Single–Target Port LUN-Mapping Feature

In some cases, avoiding any multipath/failover handling at all can be desirable. In specific use cases with StorNext or other clustered file systems where host access redundancy is not a requirement, E-Series firmware implemented a new feature beginning with version 08.20.11.00 firmware, which supports exposing a given volume on only one controller target port on the array. This feature only applies to E-Series arrays with Fibre Channel host ports and is available by product variation request (PVR) only. You must contact your account team to file a PVR.

Note: Activating this feature requires a system reboot and interrupts all existing host connectivity.

The single–target port LUN-mapping feature, also known as port to LUN, restricts access to a given volume using a specific target port on one controller based on the logical unit number used to map the volume to a host or host group. Essentially, a static range of logical unit numbers is assigned to each controller target port (0–15 for target port 1 on controller A, 16–31 for target port 2 on controller A, and so on). Table 5, Table 6, Table 7, and Table 8 provide the LUN ID range static mappings for E-Series arrays with FC host interfaces.

© 2018 NetApp, Inc. All rights reserved.

Table 5) E-Series port-to-LUN ID mapping: 2-port FC base on E2800 and 2-port FC HIC on E2700, E5500, and EF550 controllers.

Controller ID	Port 1	Port 2	Unavailable
Controller A	LUN 0 to LUN 15	LUN 16 to LUN 31	LUN 32 to LUN 127
Controller B	LUN 128 to LUN 143	LUN 144 to LUN 159	LUN 160 to LUN 255

Table 6) E-Series port-to-LUN ID mapping: 4-port FC base on E5400 and EF540 and 4-port FC HIC on E2600, E2700, E5500, E5600, EF550, and EF560 controllers.

Controller ID	Port 1	Port 2	Port 3	Port 4	Unavailable
Controller A	LUNs 0 to 15	LUNs 16 to 31	LUNs 32 to 47	LUNs 48 to 63	LUNs 64 to 127
Controller B	LUNs 128 to 143	LUNs 144 to 159	LUNs 160 to 175	LUNs 176 to 191	LUNs 192 to 255

Table 7) E-Series port-to-LUN ID mapping: 2-port FC base with 4-port FC HIC on E2800, EF280, E5700 and EF570 controllers.

Controller ID	Port 1	Port 2	Port 3	Port 4	Port 5	Port 6	Unavailable
Controller A	LUNs 0 to 15	LUNs 16 to 31	LUNs 32 to 47	LUNs 48 to 63	LUNs 64 to 79	LUNs 80 to 95	LUNs 96 to 127
Controller B	LUNs 128 to 143	LUNs 144 to 159	LUNs 160 to 175	LUNs 176 to 191	LUNs 192 to 207	LUNs 208 to 223	LUNs 224 to 255

Table 8) E-Series port-to-LUN ID mapping: 4-port FC base and 4-port FC HIC on E5400 and EF540 controllers.

Controller ID	Port 1	Port 2	Port 3	Port 4	Port 5	Port 6	Port 7	Port 8
Controller A	LUNs 0 to 15	LUNs 16 to 31	LUNs 32 to 47	LUNs 48 to 63	LUNs 64 to 79	LUNs 80 to 95	LUNs 96 to 111	LUNs 112 to 127
Controller B	LUNs 128 to 143	LUNs 144 to 159	LUNs 160 to 175	LUNs 176 to 191	LUNs 192 to 207	LUNs 208 to 223	LUNs 224 to 239	LUNs 240 to 255

Because of the unusual behavior of this feature, it can only be activated by applying a specialized feature pack key that customers must obtain from NetApp. After the PVR process is completed, the customer receives instructions to generate the key using an online tool.

Note: The Premium Feature Activation website includes a link to premium feature activation instructions. Do not attempt to use those instructions for this procedure. Instead, follow the procedure that is provided as part of the PVR process.

There is no additional cost to obtain the key, but after being applied, the key activates the behavior for all LUN mappings on the array.

4.2 ATTO Host Types

E-Series customers in the media and entertainment space are the most common users of the HBAs produced by ATTO and the corresponding ATTO Multipath Director multipath solution. E-Series firmware and NVSRAM have support for several host types that use ATTO hardware.

© 2018 NetApp, Inc. All rights reserved.

ATTO Host Types: Explicit Mode ALUA

There are three existing (long supported) SANtricity host types defined in E-Series NVSRAM for ATTO HBAs and the ATTO Multipath Director failover solution:

- Mac OS (ATTO)
- Linux (ATTO)
- Windows (ATTO)

The array behavior settings for each of these three host types are effectively the same because all three host types imply both explicit failover and failback behavior, and all use the same ATTO Multipath Director failover solution. Only the underlying OS is different.

Because these host types imply use of explicit mode ALUA (host-controlled failover) and because the ATTO Multipath Director driver is very aggressive at failing over upon loss of connectivity and failing back after connection recovery, ownership thrash in a cluster environment with any of these host types in use is very likely. Great care must be taken to disable failback behavior in the ATTO multipath configuration on the host if that host is in a cluster with shared access to E-Series storage volumes. Even a single host in the cluster attempting to fail back can start an ownership thrash storm in conditions where connectivity to the preferred controller is lost on a subset of the hosts in the cluster.

ATTO Cluster (All OS Types)

Beginning with NVSRAM builds associated with firmware 08.20.24.00, 08.25.11.00, and 08.30.20.00, a new ATTO cluster (all supported OSs) host type has been added. This new host type enables implicit failover and effectively disables failback behavior in the ATTO Multipath Director failover solution. This new host type just enables using implicit mode in the driver by changing the standard inquiry responses from the array to report implicit mode ALUA support only.

Note: With all E-Series or EF-Series implementations, you must use the NetApp Interoperability Matrix (IMT) to confirm that your specific combination of hardware and software is supported and that any qualifying notes in the IMT for your configuration have been carefully considered before the implementation begins.

This change in the ATTO driver behavior significantly reduces the risk of ownership thrash in a cluster of hosts with ATTO HBAs by disabling the aggressive and dangerous explicit ALUA behavior. This also does not rely on a host-side multipath configuration to disable failback as with the earlier ATTO host types. The new host type significantly reduces the risk of ownership thrash if a new host being added to the configuration does not have proper failback settings.

Because the settings were the same for all three existing ATTO host types, a single new common ATTO cluster host type was created for all three OS platforms. This simplifies the host type choice in SANtricity when configuring a clustered host using ATTO hardware. The ATTO cluster (all OS) host type should be used in favor of any of the three explicit mode ALUA ATTO host types when deploying E-Series arrays in a clustered configuration.

4.3 Preventing Failback and Cluster Thrash with Non-ATTO Host Types

Suppressing reporting of preferred path information from the E-Series array is required when using host types that utilize explicit failover and explicit failback in a cluster configuration. To enable the suppression, an SMcli command is used to set a parameter on a per–host type basis. This is intended to be used primarily if the cluster has Linux hosts with kernel version 3.9 or earlier and DM-MP is being used as the failover solution (utilizing the scsi_dh_rdac device handler, which uses explicit failback), or if the cluster has Windows hosts and pre-08.30.XX.XX firmware on the array, which implies the Windows MPIO DSM is operating in explicit failback mode.

Note: This method is not intended for use on any host type/driver mode that already supports implicit failback (or no failback, as with the new ATTO cluster all OS host type).

 © 2018 NetApp, Inc. All rights reserved.

Use of this mechanism is only supported under a NetApp PVR. This is to make sure the specific host OS/drivers and HBAs in use on the customer configuration have been tested and are known to react favorably to the firmware behavior of not reporting any preferred ownership information.

To enable this for the Windows host types, issue the following commands to the E-Series array using SMcli:

```
set controller[a] HostNVSRAMbyte[1, 61]=1,1;
set controller[a] HostNVSRAMbyte[8, 61]=1,1;
set controller[b] HostNVSRAMbyte[1, 61]=1,1;
set controller[b] HostNVSRAMbyte[8, 61]=1,1;
```

To enable this for the Linux DM-MP (kernel 3.9 and earlier) host type, issue the following commands to the E-Series array using SMcli:

```
set controller[a] HostNVSRAMbyte[7, 61]=1,1;
set controller[b] HostNVSRAMbyte[7, 61]=1,1;
```

Note: All hosts must be forced to rescan/rediscover storage volumes mapped from the array after executing these CLI commands for the changes to be picked up by the host multipath drivers.

The E-Series controller firmware persists this new setting enabled using the CLI in such a way as to apply it to all hosts of the same type, so it only should be enabled once for each of the unique host types that have a need to use it. This does disable all failback behavior for the associated hosts of that type, so the user must expect that failback following recovery of a SAN fault does not work after this option is enabled. Given that in a cluster, the user should have been disabling failback from the host settings anyway, this should be expected behavior. This setting also persists through subsequent updates of NVSRAM or controller firmware.

Important note for Linux DM-MP: If all arrays accessible by a given host have this option enabled to suppress preferred ownership on the Linux OS with kernel version 3.9 or earlier host type or in cases where the Linux kernel version 3.10 or later host type is being used (which suppresses preferred ownership information reporting to the host by default), then the DM-MP configuration on that host should be set to enable failback `immediate`, even in a cluster environment. This forces DM-MP to follow target port group asymmetric access state changes more aggressively, which is desirable if the host is not provided with preferred ownership information and the path routing decision is based solely on target port group asymmetric access states. However, if the host has access to any array that does not have this new option set to suppress preferred ownership information reporting (or is potentially reconnected at a later time to such an array), then failback should be set to `manual` in the DM-MP configuration on all cluster nodes to avoid ownership thrash.

5 Recommended Configuration Best Practices

The following procedures are intended to assist in building a cluster that is resistant to volume ownership thrashing. Although many of these concepts are applicable to any clustered solution where multiple hosts have shared access to the same storage volumes, this is specifically targeted at deployments utilizing heterogeneous host-type clustered file systems such as StorNext or CXFS.

5.1 Determine Requirement for Host Access Redundancy and Solution Design

First, the requirement for host access redundancy (failover) must be understood. If loss of access due to SAN connectivity faults or array controller faults/failures cannot be tolerated in the customer environment, then some sort of multipath solution must be deployed on the hosts in the cluster, and the array must be configured accordingly. Note that with the addition of support to E-Series firmware for more implicit mode ALUA multipath solutions such as the ATTO clustered host type and new mechanisms available for disabling failback using the array configuration, it is very possible to support this redundancy requirement

© 2018 NetApp, Inc. All rights reserved.

with minimal risk of ownership thrash. These new solution options should be considered when discussing the redundancy requirements during the design work for a new deployment.

Explicit Setup of Host Group/Host/Host Ports: Best Practice

Regardless of whether a multipath solution is deployed or not, it is a best practice to inventory all hosts in the cluster accessing the clustered file system and set up a host group. Next, individually define the hosts with their associated host type based on the OS and multipath solution in use on each and assign the host port worldwide unique identifiers to each host. This allows the array to properly interoperate with the OS/multipath driver on each host in the cluster.

Using Default Host Group

In a highly dynamic environment where hosts are connected to and disconnected from the cluster frequently, it is possible to use the default host group for this purpose and not strictly configure each host in the array, taking advantage of the fact that hosts physically connected without assigned host ports (let's call them "unconfigured hosts") are automatically granted access to volumes mapped to the default host group. This assumes, of course, that granting such default access is not a security concern for the customer.

Note: The default host type for the default host group must match the OS and multipath solution used by these unconfigured cluster nodes. This must be manually checked, and if required, the setting must be manually altered using SANtricity Storage Manager or SMcli.

For example, in a cluster composed predominantly of hosts with ATTO HBAs that are frequently added or removed and also a few Windows or Linux hosts with non-ATTO HBAs that are relatively static components of the configuration, the default host type could be set to ATTO clustered, and then only the non-ATTO Windows and Linux hosts would have to be strictly configured in the array with assigned host ports and associated host type. The ATTO hosts would assume the default host type (ATTO clustered), which would be correct, while the remaining hosts would also have the correct host type because they are specifically configured with the appropriate type for the OS and multipath solution being deployed on those hosts. This would provide for minimal configuration effort for the hosts being frequently connected/disconnected while still providing proper interoperability with all hosts in the cluster because the array would know the actual host OS and multipath solution on each. The key factors here are (1) making sure the hosts that are not using the HBA/OS/multipath solution implied by the default host type are strictly configured with the correct type and (2) the default host type is properly set. However, using the default host group also assumes that all hosts in the cluster need access to the exact same set of volumes. If the relatively static hosts are file system metadata servers that need access to additional volumes for storing metadata to which the other hosts must not have visibility, then this solution does not work because mapping additional volumes to those hosts results in the specifically configured hosts being ejected from the default host group (see notes about default host group behavior in the Default Host Group and Default Host Type section, earlier).

Important: Whenever the default host group is used, it is imperative that the default host type is set properly when using one of the new host type–dependent features to avoid volume ownership thrash. For example, if the ATTO clustered host type is being used to enable implicit ALUA with the ATTO Multipath Director, the default host type must be set to ATTO clustered. Otherwise, new ATTO hosts that are improperly configured (that is, with failback enabled) and attached to the array without any specific host configuration might still cause issues.

Note: It is especially important that the factory default host type is not used on 08.25.XX.XX or earlier firmware.

Additionally, if the new E-Series multipath mechanisms are used to suppress preferred ownership information for Linux DM-MP or Windows hosts, those mechanisms do not work if a Linux or Windows host is attached and unconfigured such that the default host type applies unless the default host type is set to the associated Linux DM-MP or Windows host type.

 © 2018 NetApp, Inc. All rights reserved.

Also note that any host using the default host group but not specifically configured to be part of that group (that is, hosts that are just attached and take advantage of the default LUN mappings) is not able to take advantage of features such as implicit/target failback or automatic load balancing because those features require strict host and host group setup. See the Multipath-Related Features in 08.30.XX.XX Firmware section later for details.

Design Option 1: Single–Target Port LUN Mapping

If temporary loss of access to a LUN or set of LUNs due to link faults can be tolerated in the customer environment, then the next consideration must be whether to utilize the single target port LUN-mapping feature. There are a few restrictions implied by that feature that must be taken into consideration:

- There is no direct UI support for this feature, and the target port through which a given volume is exposed is implied by the logical unit number to which that volume is mapped. This requires careful setup to achieve the desired results. Therefore, use of this feature requires skilled storage administration and data center network knowledge.
- Although the controller limits access to a given volume using a single target port, a host might still see multiple paths to the volume depending on how the switches in the fabric are configured (that is, if multiple HBA ports on the host have access to the given controller target port through which a specific volume is exposed). In this case, a multipath driver is still necessary on the host to avoid the file system/application seeing multiple instances of the volume, but the single-port LUN-mapping feature is likely not the best choice for that specific deployment.
- Because a given volume is only accessible using a single target port, bandwidth to a single volume is limited to the bandwidth of a single path from the host to the array. The anticipated workload on each volume should be considered to make sure this does not create an unacceptable performance limitation. Worst-case workloads on each controller target port (aggregate maximum anticipated load on all volumes mapped to LUNs implying the same target port) should also be considered.
- It might not be possible to evenly distribute load across all controller target ports, depending on how many volumes are on the array and the relative workload on each. In situations where data is striped across all volumes, then the workload in theory should be distributed across all target ports if the number of volumes is a multiple of the total target port count on the array. This might even need to be considered when sizing the number of drives in the configuration because the number of volumes is likely determined by the number of drives and selected RAID level, hot-spare coverage, and so on, assuming one volume per drive group as is typical in high-bandwidth configurations. One advantage of using a multipath driver and not using single-port LUN mapping is more even workload distribution across all target ports on a controller by using a round-robin path selection policy for host I/O distribution.
- One advantage of using single-port LUN mapping is less sensitivity to multipath driver behavior, allowing less stringent requirement for strict configuration of the array in terms of host and host type setup because there is less risk of volume ownership thrash with only one path to each volume. Nonetheless, best practice is to make certain all host groups/hosts/host ports are strictly configured in the array according to actual SAN topology and host OS/multipath installation of all hosts in the cluster.

Note that use of the single–target port LUN-mapping feature doesn't completely preclude use of a multipath driver in the host because in-box or installed multipath drivers on the host might still claim E-Series storage devices based upon the data reported in standard inquiry responses from the array. However, the multipath driver typically only has one path to manage and so typically is not able to cause any ownership thrashing (but is also not able to provide any redundant access). When used for this purpose, the single–target port LUN-mapping feature is essentially a big hammer approach to the ownership thrash problem. NetApp engineering recommends using newly supported implicit failover/failback methods if possible in a cluster configuration as the preferred method of avoiding ownership thrash.

 © 2018 NetApp, Inc. All rights reserved.

A more common use of the single–target port LUN-mapping feature is to manage conditions where it is known that no multipath solution is installed on a host at all and no in-box driver claims E-Series storage devices. If a configuration has been certified by NetApp Interoperability Testing for a given OS platform, then single–target port LUN mapping is a useful tool for preventing the applications accessing E-Series storage devices from being presented multiple instances of the same volume by avoiding visibility of redundant paths.

If the configuration design decision is to use the single–target port LUN-mapping feature, then follow the associated feature documentation provided with the feature pack key.

Design Option 2: Ownership Thrash-Resistant Multipath Cluster

If a multipath solution is selected for the configuration design to meet the solution redundancy requirement or if there is not a strict redundancy requirement but the single–target port LUN-mapping feature is not deemed appropriate for the configuration design, then the proper cluster configuration must be set up in order to create a thrash-resistant configuration. This setup simply involves configuring the hosts in the array to have the appropriate host type and host group association. On the hosts, it's a matter of making sure the appropriate NetApp certified multipath solution is installed (if not already in box on the OS) and configured on each host. In some cases, there might not be any host-side installation or setup at all if the driver is in box. You might just have to validate that a certified OS version is in use and the appropriate host type in the array configuration was selected.

Assuming the hosts combined in the cluster are using one of the cluster-compatible multipath solutions listed in Table 9, risk of ownership thrash is greatly reduced. Clusters configured this way with properly installed (or in-box) multipath solutions should be able to fail over when fault conditions occur and provide redundant host access to the E-Series storage.

Table 9) Cluster-compatible multipath solutions.

OS	HBA	Multipath Solution	Recommended Host Type
Linux[1]	ATTO	ATTO Multipath Director	ATTO clustered (all OS)[4]
Windows[1]	ATTO	ATTO Multipath Director	ATTO clustered (all OS)[4]
Mac[1]	ATTO	ATTO Multipath Director	ATTO clustered (all OS)[4]
Linux kernel version 3.9 or earlier[2]	NetApp certified non-ATTO HBA	In-box, DM-MP (scsi_dh_rdac)	Linux (kernel 3.9 or earlier)
Linux kernel version 3.10 or later[2,3]	NetApp certified non-ATTO HBA	In-box, DM-MP (scsi_dh_alua)	Linux (kernel 3.10 or later)
Windows Server 2012/2016 cluster (v2.0 or later NetApp E-Series DSM)	NetApp certified non-ATTO HBA	E-Series MPIO DSM v2.0 or later	Windows clustered

1. Limited to OS revisions supported by ATTO and certified by NetApp Interoperability Testing.
2. Limited to revisions of the OS and HBAs certified by NetApp Interoperability Testing.
3. Supported with 08.30.XX.XX and later firmware only.
4. Requires 08.20.24.00, 08.25.11.00 or 08.30.20.00 firmware and compatible NVSRAM or later.

In cases where 08.30.XX.XX and later firmware is in use on the E-Series array and hosts in the cluster are only using non-ATTO HBAs and running either Linux with kernel versions 3.10 and later or Windows with version 2.X of the NetApp MPIO DSM, it is also possible to provide autofailback functionality without risk of ownership thrash. Any hosts with ATTO HBAs in the cluster or hosts running older versions of

© 2018 NetApp, Inc. All rights reserved.

Linux or older versions of the Windows MPIO DSM preclude support for this autofailback capability but still retain the redundant failover capability.

Section 8 describes the configuration of a StorNext cluster that is ownership thrash resistant yet still provides failover capability. Test results from media test applications using this environment are also provided for demonstration of the capability the configuration design delivers.

6 Multipath-Related Features in 08.30.XX.XX Firmware

The following information is provided as additional background on some of the multipath-related features in the SANtricity 08.30.XX.XX firmware code. Apart from using the implicit or target failback feature for a few specific host types, most of the following feature descriptions do not apply directly to media customers. However, understanding the details of how these features operate (and how they are disabled if necessary) can be useful for anyone assisting in deploying E-Series arrays running 08.30.XX.XX or later firmware into a media and entertainment environment.

6.1 Automatic Load Balancing

The automatic load balancing (ALB) feature takes advantage of the behavior of current generation host multipath drivers and implicit ALUA to rebalance dynamic workloads across the two array controllers automatically. The incoming I/O workload to each volume is monitored, and periodic rebalances occur with the array controller triggering volume ownership transfers as needed to shift workload from one controller to the other. This uses the fact that the host multipath drivers are notified of the implicit target port group asymmetric access state change when an ownership transfer occurs, and drivers that operate in implicit ALUA mode then follow that change and start directing I/O at the new active/optimized target port group (new owning controller for a given volume). This is especially useful in environments where the workload distribution across volumes is highly dynamic and a simple fixed balance of ownership between the two controllers does not achieve an acceptable workload balance.

In the media and entertainment use cases where a clustered file system is striped across a small number of volumes on the array, ALB likely does not provide much added value. The real benefit of this feature in the media space is the shift to support for implicit ALUA within E-Series, which was a necessary move to support the array-initiated ownership transfers for rebalancing workloads. Without implicit ALUA support, there would be no opportunity to provide thrash-resistant failover capabilities within a clustered environment.

ALB is sensitive to the fact that some environments cannot tolerate regular volume ownership transfers, and the design encompasses several aspects to deal with that:

- ALB has a global switch that allows the user to enable or disable the functionality. ALB is not enabled by default when upgrading firmware to 08.30.XX.XX or later from a prior release that did not support ALB; the user must explicitly enable the feature through the UI or CLI.
- ALB does not transfer ownership of a volume if any host with access to that volume is not running a multipath solution that can tolerate implicit ownership transfers. See the ALB column in Table 3 for details of which host types/multipath solutions support ALB.
  - For example, any cluster with hosts using ATTO HBAs suppresses any ALB-initiated ownership transfers on the volumes mapped to that cluster.
  - This also means that environments using explicit ALUA that would be sensitive to ownership thrash are automatically disqualified for ALB functionality.
- ALB has a built-in thrash avoidance mechanism to avoid the same volume from being transferred too often for the purpose of rebalancing load. There is also a mechanism that monitors host-side behavior and disables ALB transfers if a host multipath driver does not follow an implicit ownership transfer for some reason.

© 2018 NetApp, Inc. All rights reserved.

- The ownership transfer process used during ALB was highly optimized to minimize any I/O latency spikes when the process is activated.

Implicit/Target Failback

Support for implicit failback (also known as target failback or array-initiated failback) was added with ALB in 08.30.XX.XX firmware but is not enabled/disabled with ALB. Like ALB, the functionality is only applied to volumes accessible to hosts that are running a multipath solution that can tolerate implicit ownership transfers. See Table 3 for the specific host types and associated multipath solutions that can support implicit or target failback.

Target failback support provides failback functionality for hosts with multipath drivers that do not support a failback process at all or drivers that fail back by simply redirecting I/O to the nonowning (but preferred) controller upon recovery of a SAN connectivity fault. This implies a performance penalty for I/O shipping for some period of time.

The implicit failback feature is a side benefit from the ALB feature. In order for ALB to succeed in rebalancing workloads between peer controllers or get hosts to redirect I/O to the desired controller, it must be certain that all hosts with access to that volume are connected to the controller that takes ownership of the volume. Additionally, the multipath driver must have discovered paths to the new owning controller as well. That means ALB has implemented tracking within the controller firmware of all host-to-controller connectivity and multipath discovery. This combines transport protocol–level connectivity status information for all connections between the hosts and controllers and command-level protocol monitoring that watches for incoming commands that are indicative of host multipath driver discovery of a given path. The net result is that the array controller is now aware of connectivity faults within the SAN and when those faults recover such that all hosts with access to a given volume have regained access to the preferred controller. The controller uses this tracking data to initiate a failback ownership transfer.

Perhaps the most important aspect of this is the fact that in a clustered environment, the array controller has visibility to all hosts in the cluster and their associated connectivity status, which allows it to determine when an appropriate time is to initiate a failback without creating any ownership thrash (no “fights” for setting ownership between the hosts in the cluster).

Given that the E-Series host type controls this mechanism as noted in Table 3, there is no need to explicitly disable the target failback capability in a media environment as long as host types are properly configured.

6.2 Connectivity Reporting

Another benefit of the connectivity tracking added for ALB support in 08.30.XX.XX and later firmware is new reporting mechanisms to augment the VNOPP alert.

Rather than waiting for a volume to be transferred off the preferred controller before raising an alert, the array uses the internal SAN connectivity tracking logic to detect and report loss of redundant host connectivity as soon as it happens. This means that an alert is raised even if connectivity is lost to the nonpreferred path, which previously would not have raised the VNOPP alert. The new alert for host redundancy loss is also more specific in indicating the host that lost connectivity and to which array controller. This was not certain previously in cluster environments when multiple hosts were sharing access to the volume that reported VNOPP because the alert was only reported on a volume and not a specific host or hosts.

Note: This new connectivity reporting is only enabled if ALB is enabled. It was added primarily because ALB cannot function properly in environments where SAN connectivity faults have occurred.

6.3 Host Setup Requirement and Implicit Failback/ALB/Connectivity Reporting

It is worth noting that there is a restriction in 08.30.XX.XX and later firmware such that many of the new multipath-related features such as connectivity reporting, target failback, and ALB do not fully function in

 © 2018 NetApp, Inc. All rights reserved.

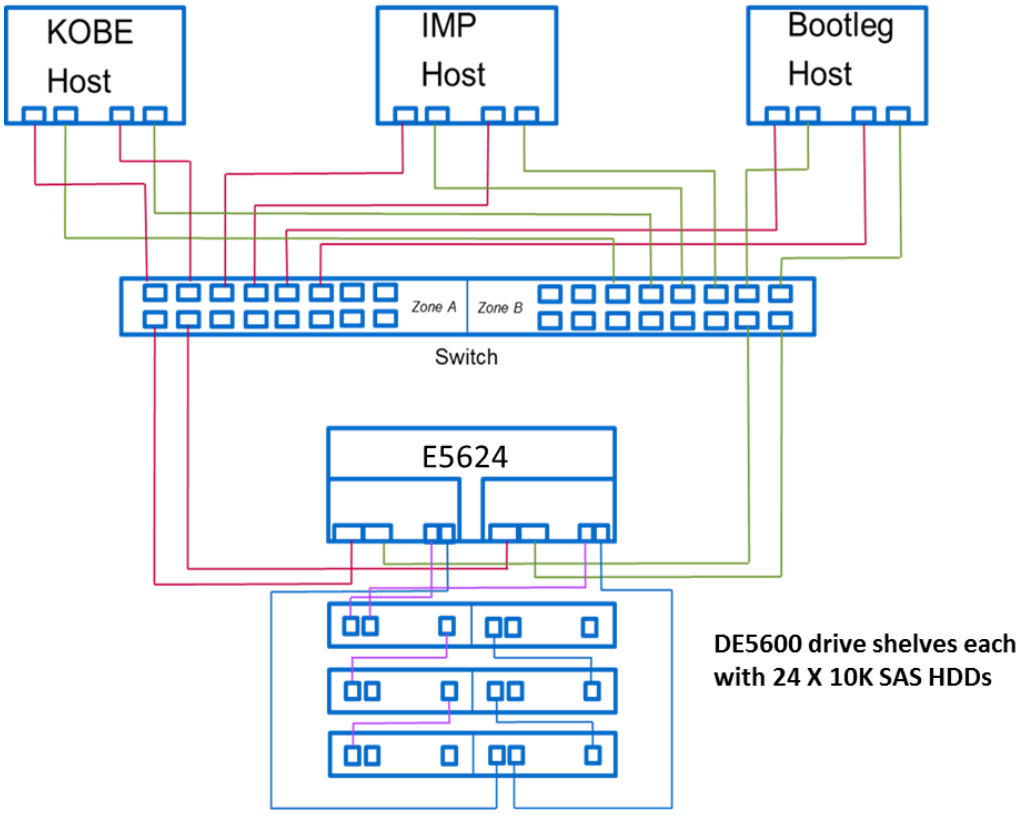
environments using the default host group. These features must understand the current connectivity of a given host in the SAN topology in order to initiate ownership transfers, so hosts must be set up in the storage partition configuration (dedicated host group) in the array. All their associated host-port worldwide identifiers and host types must be set properly. Without this, the features cannot determine whether a given host has the necessary redundant access before taking actions that would potentially affect that host's volume access.

When hosts are connected to the array and given access to the volumes mapped to the default host group and no explicit hosts or host ports are set up, all connected hosts appear as if they have single-port, nonredundant access from the controller firmware viewpoint. It is not possible for the firmware to automatically determine that one host-port worldwide identifier connected to controller A is also associated with the same host as another host-port worldwide identifier connected to controller B.

7 Reference Configuration

For this report, we used an E5600 array controller in a DE5600 shelf (RBOD) with 24 x 10k SAS drives (model HUC101818CS4205) loaded with drive firmware version NE02. This RBOD was connected to 3 DE5600 expansion shelves (EBODs), each with 24 x 10k SAS drives (model HUC101818CS4205) and loaded with drive firmware version NE02. The four shelves were connected as shown in Figure 2.

Figure 2) Drive tray configuration: StorNext cluster with E-Series.



7.1 Storage Configuration Steps

The storage array system was configured by using the following steps:

1. Established IP addresses for both controllers in the RBOD and specified the storage array name (fn1014fp-flame-5624_03 was used).

© 2018 NetApp, Inc. All rights reserved.

2. Connect the Fibre Channel (FC) HIC of the controllers in the RBOD to the two FC switches using fiber cables. The switches were connected to three hosts, namely, IMP, Bootleg, and Kobe, using FC cables. Each host had two FC ATTO HBAs installed, and the hosts had the following characteristics:
   a. HP z840 40cpu 3.1ghz
   b. 64G RAM
   c. M6000 12G (used K6000) server video card
3. Installed SANtricity Storage Manager (version 11.25.0G00.0024) on a local management server.
4. Used SANtricity Storage Manager to download and upgrade to the following software package on the storage array system:
   a. Current package version: 08.25.06.00
   b. Current NVSRAM version: N5600-825834-D03
5. Checked that all the ESMs had firmware version 039E.
6. Created eight volume groups with one volume per volume group using 8+2 RAID 6 (total capacity: 104.450TB usable, 104.450TB used).

Name	Status	Usable Capacity	Used Capacity	Free Capacity	RAID Level	Drive/Media Type	Volumes	Secure Capable	DA Capable
fbc_5624_03_vg00	Optimal	13.056 TB	13.056 TB	0.000 MB	6	Serial Attached SCSI (SAS), Hard Disk Drive	1	Yes (Non Secure)	Yes
fbc_5624_03_vg01	Optimal	13.056 TB	13.056 TB	0.000 MB	6	Serial Attached SCSI (SAS), Hard Disk Drive	1	Yes (Non Secure)	Yes
fbc_5624_03_vg02	Optimal	13.056 TB	13.056 TB	0.000 MB	6	Serial Attached SCSI (SAS), Hard Disk Drive	1	Yes (Non Secure)	Yes
fbc_5624_03_vg03	Optimal	13.056 TB	13.056 TB	0.000 MB	6	Serial Attached SCSI (SAS), Hard Disk Drive	1	Yes (Non Secure)	Yes
fbc_5624_03_vg04	Optimal	13.056 TB	13.056 TB	0.000 MB	6	Serial Attached SCSI (SAS), Hard Disk Drive	1	Yes (Non Secure)	Yes
fbc_5624_03_vg05	Optimal	13.056 TB	13.056 TB	0.000 MB	6	Serial Attached SCSI (SAS), Hard Disk Drive	1	Yes (Non Secure)	Yes
fbc_5624_03_vg06	Optimal	13.056 TB	13.056 TB	0.000 MB	6	Serial Attached SCSI (SAS), Hard Disk Drive	1	Yes (Non Secure)	Yes
fbc_5624_03_vg07	Optimal	13.056 TB	13.056 TB	0.000 MB	6	Serial Attached SCSI (SAS), Hard Disk Drive	1	Yes (Non Secure)	Yes

7. The eight volumes were confirmed to have the following characteristics:
   a. Tray loss protection: No
   b. Data assurance (DA) capable: Yes
   c. DA-enabled volume present: No
   d. Read cache: Enabled
   e. Write cache: Enabled
   f. Write cache without batteries: Disabled
   g. Write cache with mirroring: Enabled
   h. Flush write cache after (in seconds): 10.00
   i. Dynamic cache read prefetch: Enabled
   j. Enable background media scan: Enabled
   k. Media scan with redundancy check: Disabled
   l. Preread redundancy check: Disabled
8. Following is an example of a 10-drive distribution for a RAID 6 volume group and, in this case, its associated single volume between trays (shelves) 99, 0, 1, and 2. All eight volume groups were striped across the 4 shelves in a similar manner.

© 2018 NetApp, Inc. All rights reserved.

Tray	Slot
1	3
2	3
99	4
0	4
1	4
2	4
99	5
0	5
1	5
2	5

9. Defined the zoning on switch A and switch B. The zones are defined such that any HBA in each host can access one target port from one array controller from each switch.

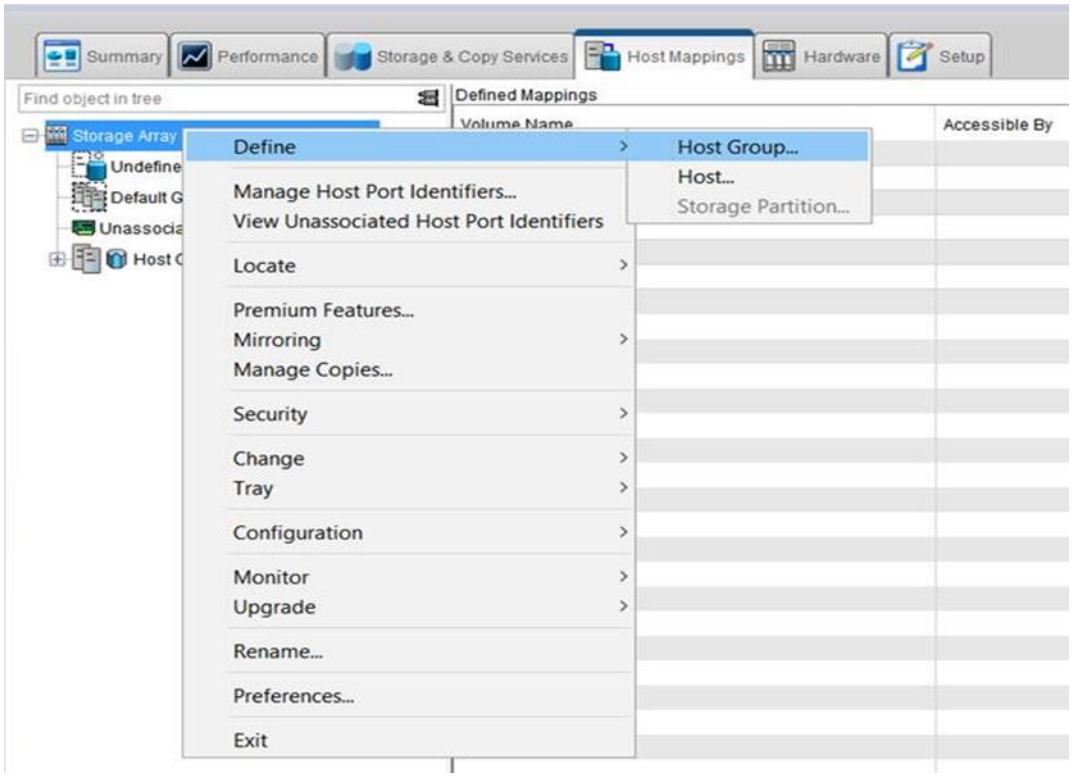
© 2018 NetApp, Inc. All rights reserved.

Switch A	Switch B
Defined configuration: cfg: StorNext Bootleg_E5600A0; IMP_E5600A1; Kobe_E5600B1; MDC1_E5600; MDC2_E5600 zone: Bootleg_E5600A0 E5600A0; Bootleg zone: IMP_E5600A1 E5600A1; IMP zone: Kobe_E5600B1 E5600B1; Kobe zone: MDC1_E5600 E5600A0; E5600A1; E5600B0; E5600B1; MDC1 zone: MDC2_E5600 E5600A0; E5600A1; E5600B0; E5600B1; MDC2 alias: Bootleg 1,13 alias: E5600A0 1,0 alias: E5600A1 1,4 alias: E5600B0 1,1 alias: E5600B1 1,5 alias: IMP 1,14 alias: Kobe 1,16 alias: MDC1 1,3 alias: MDC2 1,7 Effective configuration: cfg: StorNext zone: Bootleg_E5600A0 1,0	**Defined configuration:** cfg: StorNext Bootleg_E5600B2; IMP_E5600B3; Kobe_E5600A3; MDC1_E5600; MDC2_E5600 zone: Bootleg_E5600B2 E5600B2; Bootleg zone: IMP_E5600B3 E5600B3; IMP zone: Kobe_E5600A3 E5600A3; Kobe zone: MDC1_E5600 E5600A2; E5600A3; E5600B2; E5600B3; MDC1 zone: MDC2_E5600 E5600A2; E5600A3; E5600B2; E5600B3; MDC2 alias: Bootleg 1,13 alias: E5600A2 1,0 alias: E5600A3 1,4 alias: E5600B2 1,1 alias: E5600B3 1,5 alias: IMP 1,14 alias: Kobe 1,16 alias: MDC1 1,3 alias: MDC2 1,7 Effective configuration: cfg: StorNext zone: Bootleg_E5600B2 1,1 1,13 zone: IMP_E5600B3 1,5

© 2018 NetApp, Inc. All rights reserved.

Switch A	Switch B
1,13 zone: IMP_E5600A1 1,4 1,14 zone: Kobe_E5600B1 1,5 1,16 zone: MDC1_E5600 1,0 1,4 1,1 1,5 1,3 zone: MDC2_E5600 1,0 1,4 1,1 1,5 1,7	1,14 zone: Kobe_E5600A3 1,4 1,16 zone: MDC1_E5600 1,0 1,4 1,1 1,5 1,3 zone: MDC2_E5600 1,0 1,4 1,1 1,5 1,7

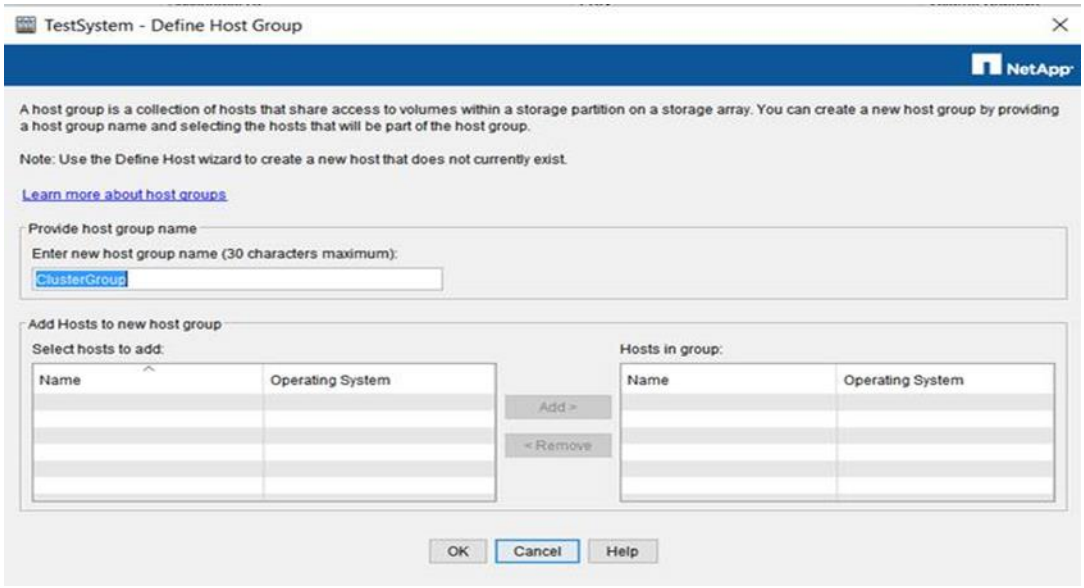
10. Defined host groups using the Array Management window:
    a. Select Host Mappings tab in SANtricity AMW.
    b. Right-click the storage array name and select Define -> Host Group.



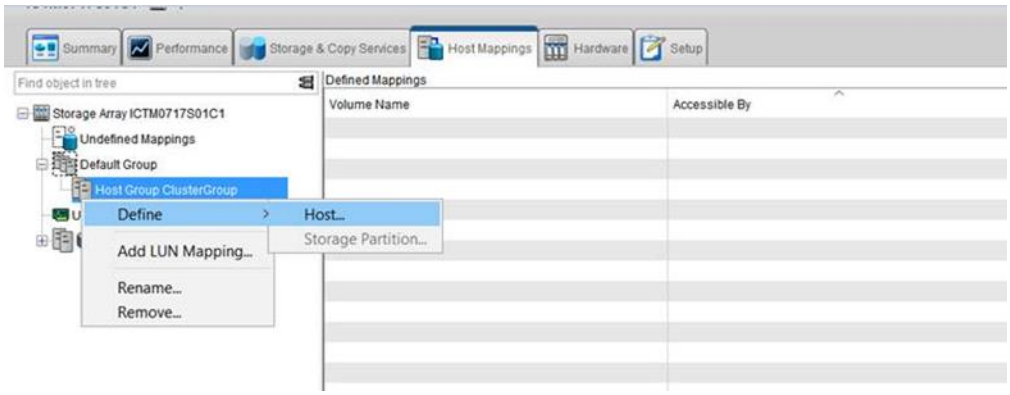
    c. Enter the host group name (for example, ClusterGroup).

© 2018 NetApp, Inc. All rights reserved.

d. Select OK.

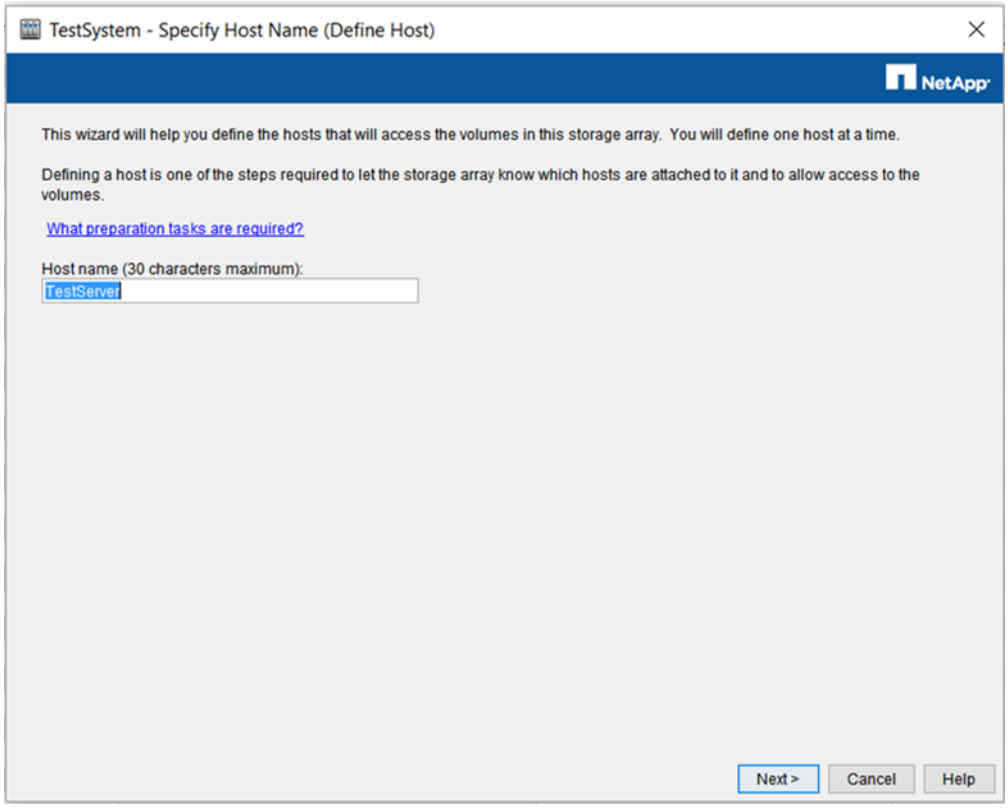


11. Defined host or client under the host group:
    a. Select Host Mappings tab in SANtricity.
    b. Right-click the host group and select Define -> Host.



    c. Input the host or client name.
    d. Select Next.

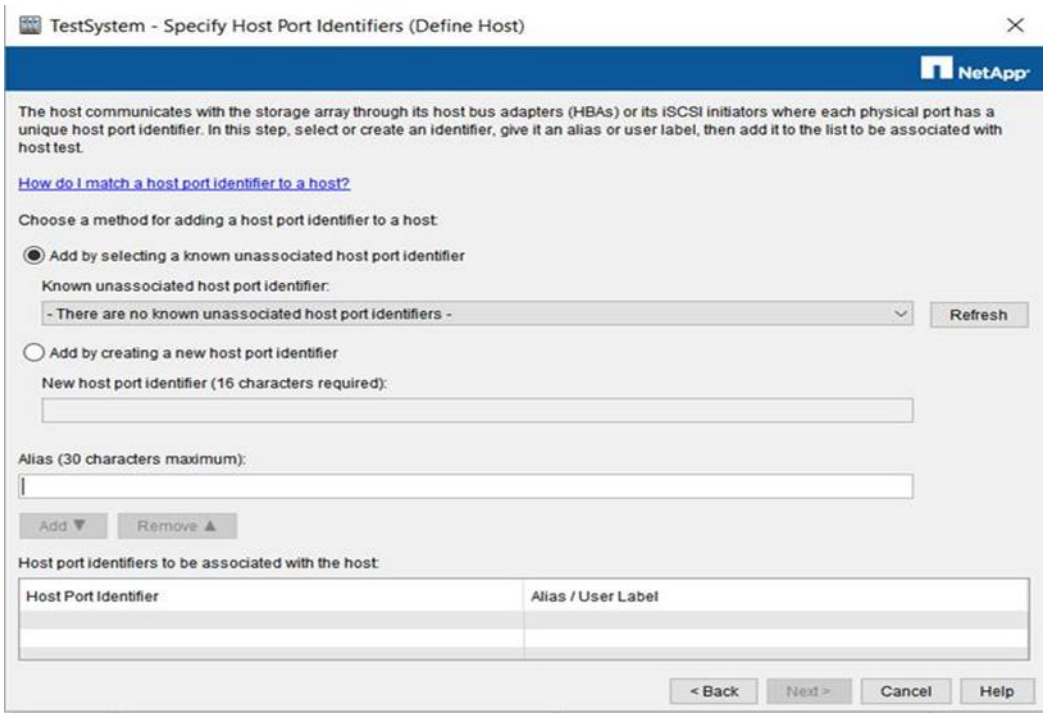
© 2018 NetApp, Inc. All rights reserved.



e. Specify the HBA port identifiers that belong to the host being defined:
   i. Add the first HBA port by using the “Add by selecting a known unassociated host port identifier” or by manually inputting HBA port identifier under the “Add by creating a new host port identifier” field.
   ii. Provide an alias to the HBA port.
   iii. Select Add.
   iv. Repeat steps i through iii for the remaining HBA ports.

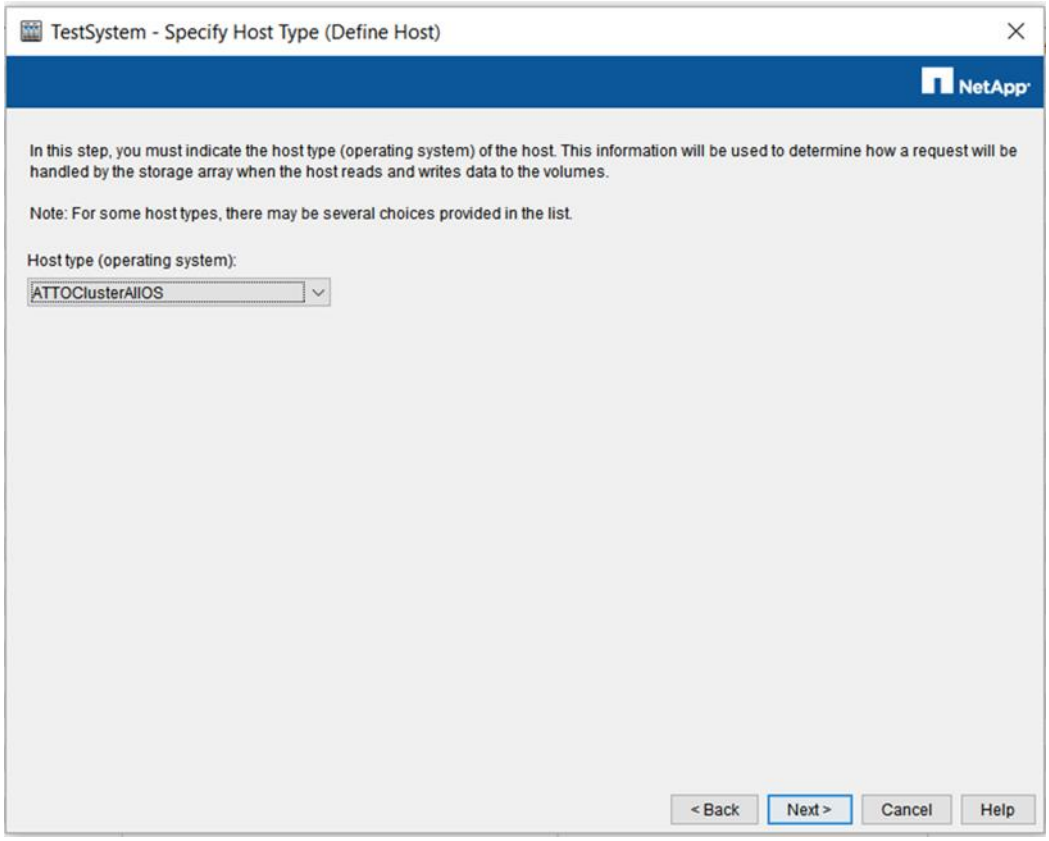
f. Select Next.

© 2018 NetApp, Inc. All rights reserved.

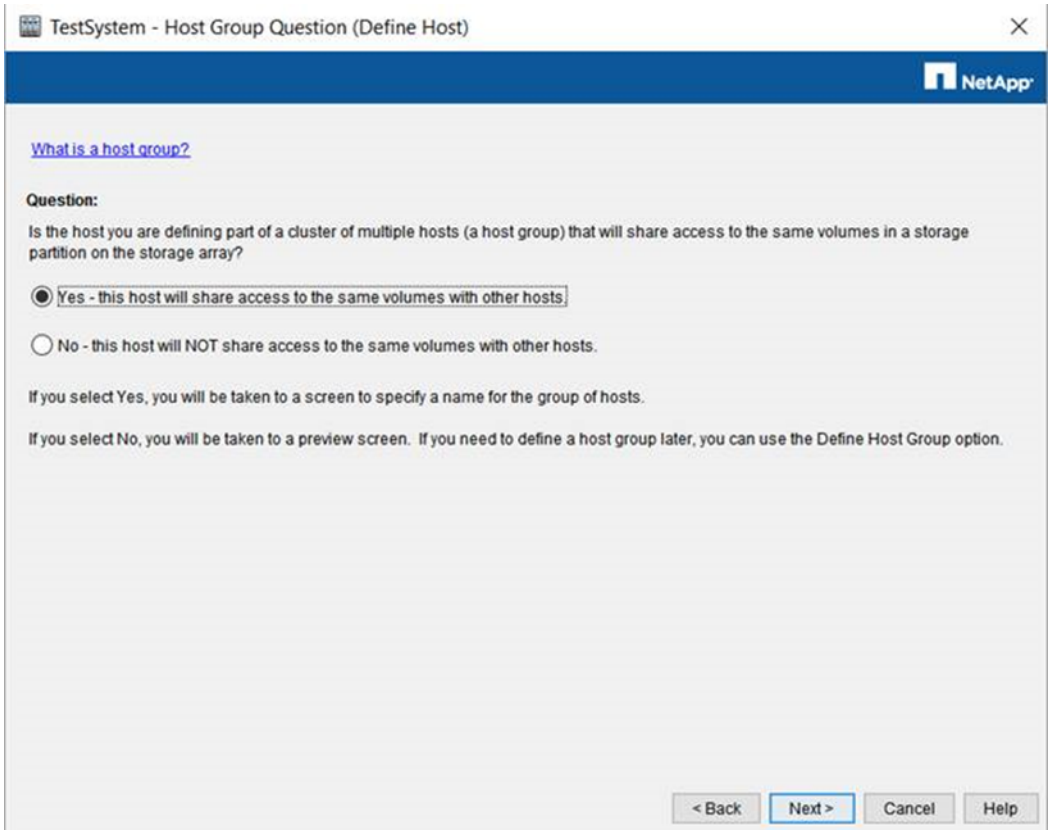


g. Select the host type (operating system) from the drop-down list.

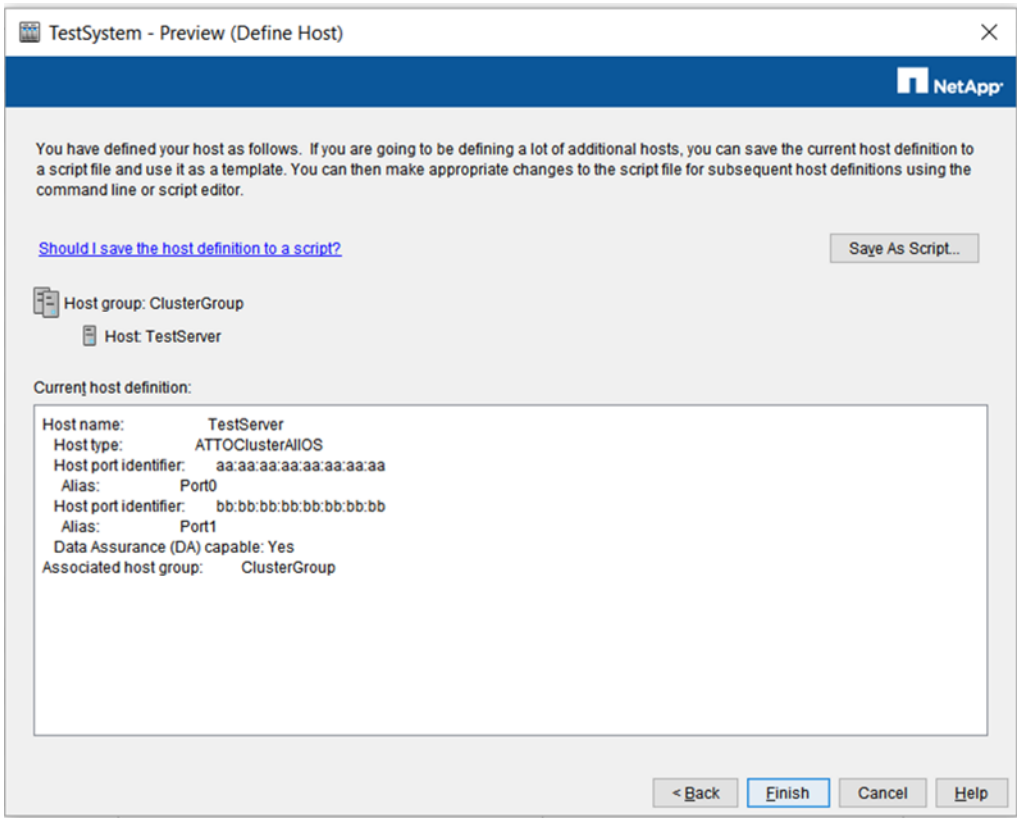
Note: Select ATTO cluster all OS host type for all hosts or clients configured with ATTO HBA and ATTO Multipath Director.



© 2018 NetApp, Inc. All rights reserved.



h. Verify the current host definition, then select Finish to complete the host definition process.



© 2018 NetApp, Inc. All rights reserved.

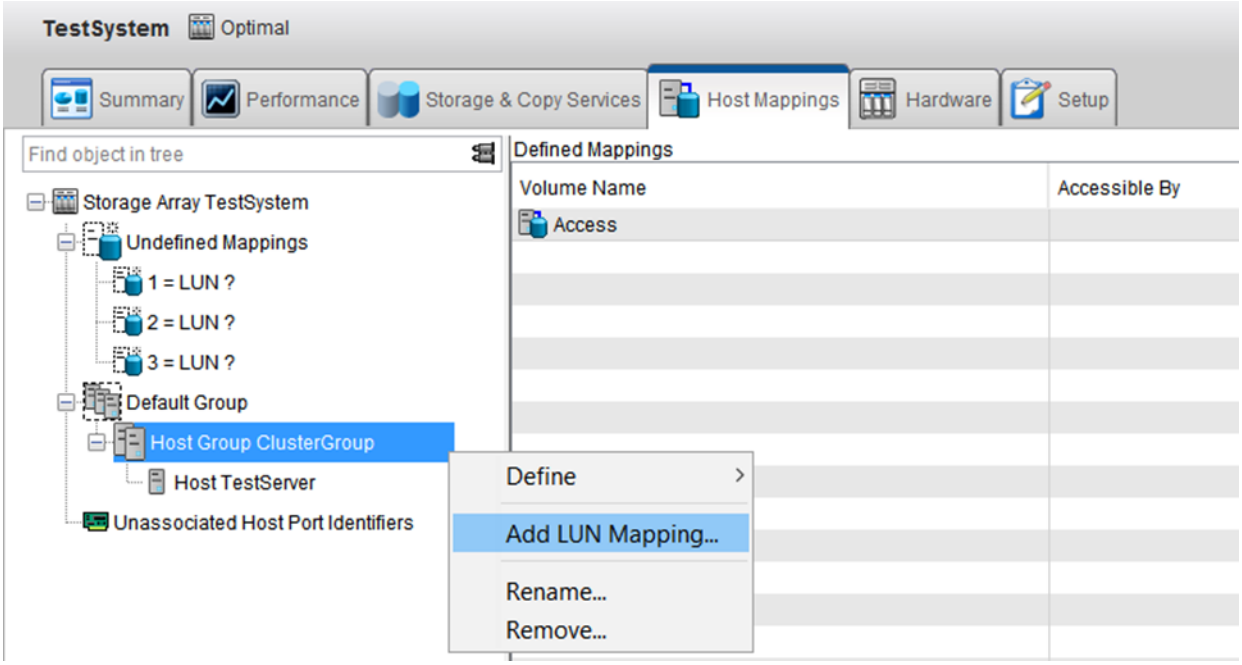
i. The new defined host is now listed under the host group (cluster).

j. Repeat steps b through h to define the remaining hosts or clients.

12. Assign volumes to the host group (cluster):

a. Select Host Mappings tab in SANtricity.

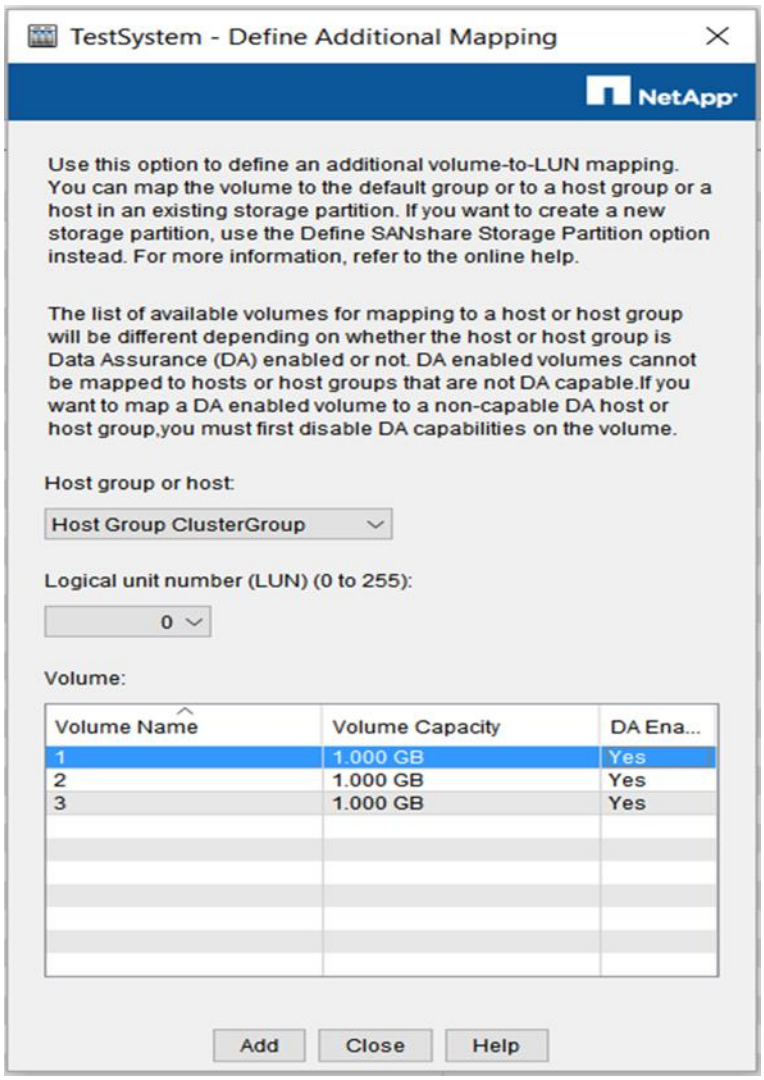
b. Right-click the host group (cluster) and select Add LUN Mapping.



c. Select a volume from the volume list.

d. Select Add.

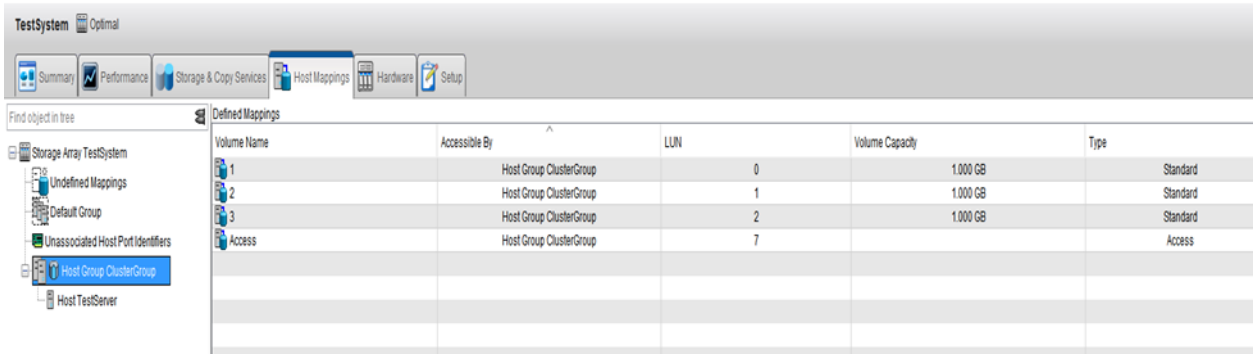
© 2018 NetApp, Inc. All rights reserved.



e. Repeat steps c and d for the remaining volumes.

f. After all volumes are defined, select Close.

g. Select the host group from the left window to see all the volumes defined.



© 2018 NetApp, Inc. All rights reserved.

7.2 Cluster Configuration

Configure the StorNext MDCs HA cluster by following the StorNext installation guide. The following are the major critical steps in setting up the StorNext environment:

1. Install and configure the StorNext software on the MDCs and clients.
2. Complete the steps in the StorNext configuration wizard in the StorNext GUI
   a. Create the StorNext shared file system using the StorNext GUI.

   Note: Create two separate stripe groups: one for data and one for metadata.

   b. Convert the MDCs to HA
3. Create the file systems for the E-Series LUNs and set up affinities for each client if needed.

For more information, see the following reference documents:

- StorNext 5 and StorNext FX 5 Installation Guide
- StorNext 5 Documentation

8 Self-Certification with Autodesk Test Suites

On our StorNext on E-Series environment, we ran the certification test suites from Autodesk. The tests mimic the Flame application. The certification tests suites and associated test results are summarized in Figures 3 through 5.

Figure 3) Play all frames playback test results.

Self Certification Program Results

Results below achieved with the flowing Workstation configuration

Hpz840 40cpu 3.1ghz

64 G Ram

M6000 1 2G (USED K6 000)

Play all frames Playback

23,98 FPS project expected Results

Codec	State	resolution	bit depth	FPS	Expected Result		Your Result		Comments
DPX	Linked	1920x1080	10	23,98	118 sec	39.9 FPS	57.58 sec	77.5 fps	Array BAUD rate= 2.8 GBPS
DPX	Linked	4096x2160	10	23,98	65.8 sec	49.7 FPS	35.88 sec	84.95 fps	Array BAUD rate= 2.5 GBPS
RED	Linked	4094x2104	12bit u	24p	193.9 sec	42.3 FPS	197.3 sec	41.6 fps	Array BAUD rate= 44.7 MBPS
ARRI	Linked	2880x1620	12bit u	23,98	106.7 sec	44 FPS	153.7 sec	30.6 fps	Array BAUD rate= 400 MBPS
ARRI	Linked	3414x2202	12bit u	24p	90 sec	57.9 FPS	149.2 sec	33.4 fps	Array BAUD rate= 440 MBPS
Open EXR u	Linked	2048x1556	16fp	23,98	118.8 sec	37.9 FPS	85.49 sec	51.2 fps	Array BAUD rate= 2.3 GBPS
Open EXR zip scanline	Linked	2048x1556	16fp	23,98	121.8 sec	34.8 FPS	134 sec	33 fps	Array BAUD rate= 320 MBPS
Open EXR u	Linked	4096x2160	16fp	24p	106.7 sec	45.5 FPS	110.8 sec	44 fps	Array BAUD rate= 2.5 GBPS
XAVC	Linked	3840x2160	10bit	23,98	77.31 sec	90.8 FPS	89.72 sec	77.9 fps	Array BAUD rate= 284 MBPS
Prores 422HQ	Linked	1920x1080	10bit	23,98	74 sec	61.5 FPS	79.61 sec	58.9 fps	Array BAUD rate= 330 MBPS
Prores 422HQ	Linked	4096x2160	10bit	23,98	97.85 sec	43.2 FPS	91.92 sec	49.2 fps	Array BAUD rate= 320 MBPS
DNX HD	Linked	1920x1080	10bit	23,98	38.31 sec	114.5 FPS	38.18 sec	116.2 fps	Array BAUD rate= 410 MBPS
DNX-HR	Linked	4096x2160	10bit	23,98	154 sec	29.3 FPS	168.3 sec	26.4 fps	Array BAUD rate= 380 MBPS

59p FPS project expected results

Codec		resolution	bit depth	FPS	Expected Result		Your Result		
DPX	Rendered DPX	4096x2160	10bit	59p	125 sec	71.2 FPS	117 sec	149.4 fps	Array BAUD rate= 2.3 GBPS
Prores 422 HQ	Linked	4096x2160	10bit	59p	62.7 sec	155 FPS	85.65	235.1	Array BAUD rate= 650 MBPS,9 621 Frames
XAVC	Linked	4096x2160	10bit	59p	108 sec	97 FPS	124.9 sec	166.1	Array BAUD rate= 400 MBPS,9 621 Frames

© 2018 NetApp, Inc. All rights reserved.

Figure 4) Real-time playback test results.

Realtime Playback

23,98 FPS project expected Results

Codec	State	resolution	bit depth	FPS	Expected Result	Your Result	Comments
DPX	Linked	1920x1080	10	23,98	4 streams	4 streams	Staggered
DPX	Linked	4096x2160	10	23,98	1 stream	1 streams	
RED	Linked	4094x2104	12bit u	24p	1 stream	1 streams	
ARRI	Linked	2880x1620	12bit u	23,98	2 Streams	2 streams	
ARRI	Linked	3414x2202	12bit u	24p	1 stream	1 streams	
Open EXR u	Linked	2048x1556	16fp	23,98	2 Streams	2 streams	
Open EXR zip scanline	Linked	2048x1556	16fp	23,98	2 Streams	2 streams	
Open EXR u	Linked	4096x2160	16fp	24p	1 stream	1 stream	
XAVC	Linked	3840x2160	10bit	23,98	1 stream	1 stream	
Prores 422HQ	Linked	1920x1080	10bit	23,98	10 Streams	10 streams	Staggered
Prores 422HQ	Linked	4096x2160	10bit	23,98	4 streams	4 streams	Staggered
DNX HD	Linked	1920x1080	10bit	23,98	4 Streams	4 streams	Staggered
DNX-HR	Linked	4096x2160	10bit	23,98	4 streams	4 streams	Staggered

59p FPS project expected results

Codec		resolution	bit depth	FPS	Expected Result	Your Result	Comments
DPX	Rendered DPX	4096x2160	10bit	59p	1 stream	1 stream	
Prores 422 HQ	Linked	4096x2160	10bit	59p	1 stream	1 stream	frame-rate limited by graphics card
XAVC	Linked	4096x2160	10bit	59p	1 stream	1 stream	frame-rate limited by graphics card

Figure 5) Rendering project test results.

Rendering Project expected results

TEST		Epected Result	Your Result	Comments
Heavy EXR batch setup	Rendered DPX	97 sec	91.63	Run test 5 times
BFX timeline	Rendered DPX	162.5 sec		Did not run the test because It is graphic card senstive and K6000 is not appropriate for this test
Heavy EXR batch setup	Rendered Prores 422 HQ	105 sec	89.1	Run test 5 times
BFX timeline	Rendered Prores 422 HQ	164 sec		Did not run the test because It is graphic card senstive and K6000 is not appropriate for this test
Heavy EXR batch setup	Rendered DNXHD HQ	91 sec	89.1	Run test 5 times
BFX timeline	Rendered DNXHD HQ	167 sec		Did not run the test because It is graphic card senstive and K6000 is not appropriate for this test

Referring to Figure 3, Figure 4, and Figure 5, the reader can observe the following:

- The E-Series array delivered better than expected results, even while using a less powerful graphics card (K6000) compared to what was used for the benchmark expected results (M6000).
- The four entries in the “Your Result” column that were less than expected were caused by an incorrect setting in the analysis tool.

9 Performance Stress Testing of StorNext Cluster with Video Applications from Autodesk

On our StorNext cluster with E-Series SAN, we ran different video applications (similar to the Flame application) from the three hosts. We captured the following excellent responses with the E-Series products under test with zero dropped frames:

© 2018 NetApp, Inc. All rights reserved.

9.1 Test 1

Test Specifications

- 1 client running 4 streams of unthrottled 4K
- 1 client running 2 streams of HD
- 1 client running 1 stream of unthrottled HD
- Pixel/depth of 8 bits

Test Results

- E-Series:
  - Average IOPS of 12k
  - Average throughput of 6.2GBps

9.2 Test 2

Test Specifications

- Unthrottled 4K: 1 stream
- 3 clients running write/read I/O
- Each client had affinities set to 2 volumes on the storage array; 6 out of 8 volumes were used during this test
- Pixel/depth of 8 bits
- Duration of test is 2 minutes

Test Results

- E-Series:
  - Average IOPS: 10,156
  - Maximum IOPS: 14,400
  - Average latency: 6.5ms
  - Maximum throughput: 7.0GBps
  - Average throughput: 5.0GBps
- Client results:
  - Maximum latency: 162/195/200ms

9.3 Test 3

Test Specifications

- Throttled 4K at 30fps: 2 streams
- 3 clients running write/read I/O
- Each client had affinities set to 2 volumes on the storage array; 6 out of 8 volumes were used during this test
- Pixel/depth of 8 bits
- Duration of test is 2 minutes

Test Results

- E-Series:
  - Average IOPS: 9,502

© 2018 NetApp, Inc. All rights reserved.

  - Maximum IOPS: 14,220
  - Average latency: 7ms
  - Maximum throughput: 6.8GBps
  - Average throughput: 4.6GBps
- Client results:
  - Maximum latency: 340/370/490ms

9.4 Test 4

Test Specifications

- Unthrottled 4K: 2 streams
- 3 clients running write/read I/O
- Each client had affinities set to 2 volumes on the storage array; 6 out of 8 volumes were used during this test
- Pixel/depth of 8 bits
- Duration of test is 2 minutes

Test Results

- E-Series:
  - Average IOPS: 10,760
  - Maximum IOPS: 14,441
  - Average latency: 6.7ms
  - Maximum throughput: 7.0GBps
  - Average throughput: 5.3GBps
- Client results:
  - Maximum latency: 415/520/350ms

9.5 Test 5

Test Specifications

- Unthrottled 4K: 3 streams
- 3 clients running write/read I/O
- Each client had affinities set to 2 volumes on the storage array; 6 out of 8 volumes were used during this test
- Pixel/depth of 8 bits
- Duration of test is 2 minutes

Test Results

- E-Series:
  - Average IOPS: 11,280
  - Maximum IOPS: 14,161
  - Average latency: 7ms
  - Maximum throughput: 6.9GBps
  - Average throughput: 5.5GBps
- Client results:

© 2018 NetApp, Inc. All rights reserved.

- Maximum latency: 600/650/600ms

9.6 Test 6

Test Specifications

- Unthrottled 4K: 1 stream
- Pixel/depth of 16 bits
- 3 clients running write/read I/O
- Each client had affinities set to 2 volumes on the storage array; 6 out of 8 volumes were being used during this test
- Duration of test is 2 minutes

Test Results

- E-Series:
  - Average IOPS: 11,950
  - Maximum IOPS: 14,730
  - Average latency: 6.5ms
  - Maximum throughput: 7.4GBps
  - Average throughput: 6.0GBps
- Client results:
  - Maximum latency: 350/310/290ms

9.7 Test 7

Test Specifications

- Frame test: 2,000 frames; 4K; writes only
- Total of 3 clients had affinities set to 2 volumes on the storage array; 6 out of 8 volumes were used during this test
- Duration of test is 1 minute

Test Results

- E-Series:
  - Average IOPS: 9,300
  - Maximum IOPS: 10,200
  - Average latency: 5ms
  - Maximum throughput: 5.0GBps
  - Average throughput: 4.6GBps
- Client results:
  - Slowest frame: 170/165/204ms
  - Fastest frame: 18/17/18ms

9.8 Test 8

Test Specifications

- Frame test: 2,000 frames; 4K; reads only

© 2018 NetApp, Inc. All rights reserved.

- Total of 3 clients had affinities set to 2 volumes on the storage array; 6 out of 8 volumes were used during this test
- Duration of test is 1 minute

Test Results

- E-Series:
  - Average IOPS: 9,200
  - Maximum IOPS: 11,000
  - Average latency: 6ms
  - Maximum throughput: 5.5GBps
  - Average throughput: 3.2GBps
- Client results:
  - Frame rate: 27.34/27.61/61.06fps

9.9 Test 9

Test Specifications

- Frame test: 2,000 frames; 4K; reads only
- 2 clients were reading from 2 volumes on the storage array (1 StorNext stripe group)
- Duration of test is 1 minute

Test Results

- E-Series:
  - Average IOPS: 11,890
  - Maximum IOPS: 12,000
  - Average latency: 4.5ms
  - Maximum throughput: 6.0GBps
  - Average throughput: 6.0GBps
- Client results:
  - Frame rate: 60.51/60.27fps

9.10 Test 10

Test Specifications

- Frame test: 2,000 frames; 4K; reads only
- All 3 clients were reading from 2 volumes on the storage array (1 StorNext stripe group)
- Duration of test is 1 minute

Test Results

- E-Series:
  - Average IOPS: 17,875
  - Maximum IOPS: 18,104
  - Average latency: 4.5ms
  - Maximum throughput: 9.0GBps
  - Average throughput: 8.8GBps

© 2018 NetApp, Inc. All rights reserved.

- Client results:
  - Frame rate: 60/61.39/60.58fps

9.11 Test 11

Test Specifications

- Frame test: 2,000 frames; 4K; write/read test
- 1 client was writing to 1 stripe group (2 volumes), while the other 2 clients were reading from the same stripe group
- Duration of test is 1 minute

Test Results

- E-Series:
  - Average IOPS: 8,640
  - Maximum IOPS: 15,000
  - Average latency: 7ms
  - Maximum throughput: 7.5GBps
  - Average throughput: 4.3GBps
- Client results:
  - Frame rate: 28.93/40.33/29.29fps

9.12 Test 12

Test Specifications

- Frame test: 2,000 frames; 4K; write/read test (same as test 11)
- 1 client was writing to 1 stripe group (2 volumes), while the other 2 clients were reading from the same stripe group
- Duration of test is 1 minute

Test Results

- E-Series:
  - Average IOPS: 7,273
  - Maximum IOPS: 11,000
  - Average latency: 6ms
  - Maximum throughput: 5.4GBps
  - Average throughput: 2.8GBps
- Client results:
  - Frame rate: 26.39/39.29/25.94fps

9.13 Test 13

Test Specifications

- Frame test: 2,000 frames; 4K; write/read test (same as test 11)
- 1 client was writing to 1 stripe group (2 volumes), while the other 2 clients were reading from the same stripe group (different stripe group/volumes than in previous test)
- Duration of test is 1 minute

© 2018 NetApp, Inc. All rights reserved.

Test Results

- E-Series:
  - Average IOPS: 6,000
  - Maximum IOPS: 11,500
  - Average latency: 6.5ms
  - Maximum throughput: 5.8GBps
  - Average throughput: 4GBps
- Client results:
  - Frame rate: 26.69/26.1/38.21fps

9.14 Test 14

Test Specifications

- We ran 4 tests with Frame: 2,000 frames
- Write/read test like in tests 10 and 11

Test Results

The test results were lower latency and slightly higher throughput.

10 Conclusion

NetApp engineering continues to improve host/array interoperability and compatibility with host multipath solutions to better support clustered host configurations with shared access to the same set of storage volumes.

In particular, the controller firmware and NVSRAM enhancements released in April 27, 2017, directly benefit users of E-Series storage in the media and entertainment space deploying ATTO HBAs and other clustered file systems by minimizing risk of volume ownership thrash while retaining failover/redundancy capabilities. This addresses issues that have historically been very problematic in clustered configurations.

In addition, testing of media-specific configurations has demonstrated real-world performance aspects of solutions using these new capabilities. E-Series products with the following controller firmware releases (released April 27, 2017) passed the Autodesk self-certification suites with much better than expected stream results. The Quantum-Autodesk stress tests mimicking the Flame application also delivered superior performance throughput (sustained 9.1GBps end to end) without any frame drop:

- 08.20.24.00 (also known as Kingston maintenance “M9”)
- 08.25.11.00 (also known as Lancaster maintenance “M4”)
- 08.30.20.00 (also known as Lehigh maintenance “M2”)

E-Series design also continues to evolve in a direction that is more cluster-friendly, emphasizing implicit ALUA to support failover in shared volume access environments today and identifying potential opportunities to continue to expand these capabilities going forward.

Where to Find Additional Information

To learn more about the information that is described in this document, review the following documents and/or websites:

- Multipath Configuration Power Guides: https://mysupport.netapp.com/info/web/ECMLP2522638.html

© 2018 NetApp, Inc. All rights reserved.

- Installing and Configuring for Linux: Power Guide for Advanced Users
- Installing and Configuring for Windows: Power Guide for Advanced Users
- E-Series and SANtricity 11 Resources page
  https://mysupport.netapp.com/info/web/ECMP1658252.html
- E-Series and SANtricity 11 Documentation Center
  https://docs.netapp.com/ess-11/index.jsp
- NetApp Product Documentation
  https://www.netapp.com/us/documentation/index.aspx

Version History

Version	Date	Document Version History
Version 1.0	June 2017	Document initial release.
Version 1.1	November 2018	Minor updates to reflect new array models and to clarify command lines.

© 2018 NetApp, Inc. All rights reserved.

Refer to the Interoperability Matrix Tool (IMT) on the NetApp Support site to validate that the exact product and feature versions described in this document are supported for your specific environment. The NetApp IMT defines the product components and versions that can be used to construct configurations that are supported by NetApp. Specific results depend on each customer's installation in accordance with published specifications.

Copyright Information

Copyright © 2018 NetApp, Inc. All Rights Reserved. Printed in the U.S. No part of this document covered by copyright may be reproduced in any form or by any means—graphic, electronic, or mechanical, including photocopying, recording, taping, or storage in an electronic retrieval system—without prior written permission of the copyright owner.

Software derived from copyrighted NetApp material is subject to the following license and disclaimer:

THIS SOFTWARE IS PROVIDED BY NETAPP "AS IS" AND WITHOUT ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE, WHICH ARE HEREBY DISCLAIMED. IN NO EVENT SHALL NETAPP BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

NetApp reserves the right to change any products described herein at any time, and without notice. NetApp assumes no responsibility or liability arising from the use of products described herein, except as expressly agreed to in writing by NetApp. The use or purchase of this product does not convey a license under any patent rights, trademark rights, or any other intellectual property rights of NetApp.

The product described in this manual may be protected by one or more U.S. patents, foreign patents, or pending applications.

Data contained herein pertains to a commercial item (as defined in FAR 2.101) and is proprietary to NetApp, Inc. The U.S. Government has a non-exclusive, non-transferrable, non-sublicensable, worldwide, limited irrevocable license to use the Data only in connection with and in support of the U.S. Government contract under which the Data was delivered. Except as provided herein, the Data may not be used, disclosed, reproduced, modified, performed, or displayed without the prior written approval of NetApp, Inc. United States Government license rights for the Department of Defense are limited to those rights identified in DFARS clause 252.227-7015(b).

Trademark Information

NETAPP, the NETAPP logo, and the marks listed at http://www.netapp.com/TM are trademarks of NetApp, Inc. Other company and product names may be trademarks of their respective owners.

TR-4604-1118