Technical Report

Automatic load balancing

NetApp SANtricity feature description and implementation

Dean Lang and Joey Parnell, NetApp
March 2022 | TR-4737

Abstract

This document is an overview of automatic load balancing. This feature removes the need for storage administrators to manually manage volume ownership to balance workload across the two I/O controllers in a NetApp® E-Series storage array. This feature enables simple storage management while also providing enterprise-class reliability and best-in-class price/performance. These advantages are made possible by use of a lean I/O path and a data cache coherency algorithm that is only possible with the asymmetric access volume ownership models provided in SCSI by asymmetric logical unit access (ALUA) and in NVMe-oF by asymmetric namespace access (ANA).

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Introduction

Beginning with NetApp SANtricity® OS 11.30 and controller firmware version 08.30.XX.XX, NetApp E-Series storage arrays offer automatic load balancing (ALB), a feature that monitors incoming I/O workload and dynamically makes configuration changes to balance the workload across the array's two I/O controllers. This document is an overview of the behavior of the ALB feature, its key configuration parameters, and its host interoperability enhancements.

Related features

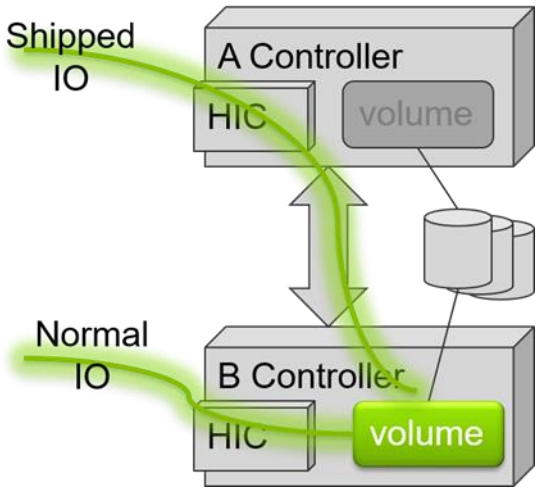
The key concepts described in this section are crucial to understanding the behavior of the ALB feature and are useful for troubleshooting any issues that arise in feature deployment.

E-Series volume ownership model and asymmetric logical unit access

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 concept of volume ownership. This means that I/O for a given volume must be processed by the controller that is assigned as the owner of that volume, forcing all write data for a given volume to be cached on one I/O controller. This design greatly simplifies the data-caching implementation and hardware design, minimizing I/O latency while keeping overall product cost low.

Despite using a design that employs a volume ownership model, E-Series arrays also have multiple target ports on each I/O controller through which hosts can be connected to issue I/O requests. These target ports support an active-active access model in which 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 non-owning controller is processed by the owning controller using intercontroller messaging. For example, Figure 1 shows a case in which 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 and then processed.

Figure 1) I/O shipping example.



This implies that I/O sent to the non-owning controller can result in degraded performance (the I/O shipping penalty), and so the performance observed by a host is not equal across all the controllers on the array. This is referred to as asymmetric logical unit access (ALUA) in SCSI semantics and asymmetric namespace access (ANA) in NVMe-oF semantics, as opposed to a fully symmetric active-active design in which performance is equal across all target ports and controllers.

Target port group support reporting and multipath driver behavior

E-Series arrays support T10-SPC (SCSI) standard methods for reporting groups of controller target ports and their associated performance characteristics. In addition, E-Series arrays support the NVM Express NVMe and NVMe-oF standard methods for reporting asymmetric controller behavior, including the Asymmetric Namespace Access (ANA) log page and associated ANA reporting including performance characteristics.

These methods, referred to in SCSI as target port group support (TPGS), and in NVMe-oF as asymmetric controller behavior, allow reporting of the specific storage controller that offers optimal performance to the host on each individual LUN or volume. The E-Series array controller firmware uses this reporting mechanism to relay information to hosts about the specific controller that owns a given volume. This method enables host multipath drivers to determine how to route I/O to the array controllers so that they can maximize performance under optimal conditions.

Storage administrator control of volume ownership – Manual load balancing

E-Series storage administration tools allow the storage administrator to assign ownership of each volume to one of the two array I/O controllers. This flexibility can be combined with the host multipath driver behavior of routing I/O for a given volume to the owning controller, which gives the storage administrator a method to control host I/O routing to the array. If the storage administrator also understands expected I/O workload to each volume, they can achieve a rudimentary workload balance by carefully distributing volume ownership across the two I/O controllers in the array.

The primary challenge facing the storage administrator in accomplishing this kind of manual workload balance across the two array controllers is that workload on each individual volume is often very dynamic. Therefore, an ownership distribution that accomplishes a satisfactory workload balance at a given point in time often doesn't continue to provide that balance for very long. In addition, upon initial creation of a new storage volume on the array, the storage administrator often doesn't know what the host I/O workload will be on that volume. Therefore, selecting initial controller ownership for a new volume is often difficult.

The ALB feature addresses these shortfalls through the automated volume ownership distribution process described later in this document.

SAN connectivity fault tolerance – Failover and failback

Host multipath drivers identify all physical connection paths to the array I/O controllers and monitor path connectivity in order to provide fault tolerance when SAN connectivity faults occur. When loss of connectivity occurs on all paths to the owning controller, the host multipath driver directs I/O to the remaining connected paths to the non-owning controller. This process results in a temporary I/O performance penalty due to I/O shipping that might last for up to 5 minutes.

This delay is caused by the array controller firmware identifying non-optimal I/O routing and automatically reassigning volume ownership to eliminate I/O shipping if possible. This failover process allows I/O continuity when SAN faults occur and also minimizes the I/O shipping penalty during these fault conditions. This process also has the potential to unbalance workloads across array I/O controllers because of changes to the volume ownership configuration that was carefully created by the storage administrator.

Many SCSI-based host multipath drivers support the concept of a failback process. In such a situation, volume ownership is returned to the preferred distribution assigned by the storage administrator following resolution of a SAN connectivity fault that previously resulted in a failover and volume ownership change. This process is accomplished by the host multipath driver monitoring the connectivity of all paths to each array I/O controller. It then initiates failback after connectivity to the original owning controller for a given volume returns. The failback process is accomplished using a T10-SPC (SCSI) standard command issued by the host multipath driver that instructs the array controller to change volume ownership.

This failback process allows automatic recovery of volume ownership and associated workload balance following resolution of a SAN connectivity fault. However, it is driven by the host multipath driver on each individual host. Therefore, volume ownership thrash can occur in a clustered host solution in which multiple hosts have access to a given volume on the array and those hosts do not agree as to the state of connections in the SAN. As a result, the best practice recommendation is to disable this automatic failback capability in the host multipath driver configuration on hosts that are part of a clustered solution.

Note: This failover and failback feature has long been present in E-Series solutions using array controller firmware capabilities and host-resident multipath drivers specific to E-Series. Therefore, it is not really part of ALB. However, ALB does employ a new method of initiating failback to better solve the ownership thrash issue caused by failback in clustered environments.

Feature benefits

The primary goal of ALB is to remove the need for the storage administrator to manually manage volume ownership and create balanced workloads across the array I/O controllers. This automation allows simple storage management of E-Series arrays. It also provides best-in-class price/performance, made possible by the use of a lean I/O path and simple data cache coherency algorithms associated with the volume ownership model.

In addition, enhancements to the automated failback process implemented by ALB enable automated failback. This is true even in clustered host environments, while improved reporting of SAN connectivity faults allows simpler monitoring and more timely resolution of issues in the SAN.

In essence, the storage administrator does not need to be aware of the concept of volume ownership after ALB is deployed.

Automatic volume ownership distribution

ALB automates the process of balancing workload across the two I/O controllers in the array by periodically evaluating I/O workload to each volume and automatically selecting a volume ownership distribution that maximizes workload balance. The periodic rebalancing process is as follows:

1. I/O workload to each volume is monitored over a period of time, and I/O workload to each volume is predicted over the next time period using the collected workload data.
2. An optimal distribution of volume ownership across the two I/O controllers is determined to balance total workload on each controller in the next time period.
3. Transfer of ownership on a minimum number of volumes is automatically initiated to achieve a workload balance in the upcoming time period.
4. Those volume ownership changes are communicated to the host multipath drivers to cause the hosts to direct I/O to the array controllers according to the new ownership distribution.

ALB periodically initiates this reevaluation and rebalancing process. It can therefore react to the dynamic I/O workload distribution across volumes that is typical in enterprise data centers and automatically adjust to maintain the maximum performance of the array.

Enhanced SAN connectivity reporting

When transferring volume ownership between controllers to rebalance workload, ALB must make sure that the transfer does not cause any performance degradation due to I/O shipping penalties. As a result, ALB cannot transfer volume ownership between controllers unless it is certain that all hosts with access to a given volume being transferred have connectivity to both controllers. It must also determine that the multipath driver on that host has discovered paths to both controllers. To verify this, ALB has implemented a SAN connectivity-tracking module that makes it possible, as a side benefit of ALB, to also provide more robust reporting of SAN connectivity faults to the end user. In particular, the SAN connectivity module reports the following conditions:

- A host appears to be connected at the physical transport-protocol layer to one controller but not the other. Unlike the existing Volume Not on Preferred Path condition reported by the array, a volume ownership failover does not have to occur for this condition to be reported. Because a failover does not occur when connectivity is lost to just the nonpreferred controller, this new alert reports connectivity issues in situations in which the existing reporting would remain silent.
- A host appears to have transport-protocol layer connectivity to a controller, but for some reason the host multipath driver does not appear to have discovered paths to the connected controller. This situation typically indicates that a device rescan is needed on the host or that there is some sort of misconfiguration of the host multipath driver. Note that the host multipath driver device discovery state is inferred by the array controller firmware based on the receipt (or lack thereof) of device-discovery-related SCSI or NVMe-oF commands from that host.

Together, these new error conditions reported by the E-Series array controller firmware as a part of ALB help to determine possible causes of volume ownership failover. In some cases, these error conditions might even allow the storage administrator to correct issues before a failover actually occurs.

Target (implicit) failback

Another side benefit of the SAN connectivity tracking implemented in the array controller firmware for ALB is the concept of implicit or target failback. This form of failback is initiated by the array controller (the target device in SCSI terms) instead of being initiated by an explicit request from the host multipath driver.

When the array controller firmware determines that a failback is possible due to the reconnection of all hosts in the SAN, it initiates a change in volume ownership for failback. It then communicates information about that change to the hosts so that the host multipath driver redirects I/O to the new owning controller. This array-driven volume ownership transfer is called implicit volume transfer because it is initiated by the array itself rather than by an explicit request from one of the hosts to transfer ownership. As a result, the target failback feature is often also called implicit failback.

The array controller can see the connectivity of all hosts in the SAN that have access to the array. Therefore, it can make smarter decisions about when to fail back ownership of a volume to the preferred controller. In clustered host environments, multiple hosts have shared access to the same storage volume. Therefore, the array can avoid failing back ownership on that volume until all hosts with shared access to the volume have reconnected. This configuration avoids the volume ownership thrash problem that can occur when the host multipath driver is initiating the failback process and the two hosts disagree about the state of SAN connections. This opens up the possibility of having automated failback in clustered host deployments in which it was not previously feasible.

An additional benefit of target failback is a potentially significant reduction in the I/O response time spike that can occur during the failback process on some host operating systems. Some host multipath drivers queue and hold all I/O when issuing a command to the array controller to fail back ownership. However, volume ownership transfer implies that the write data cache must be flushed on the array controller that is losing ownership of the volume. This process can take upward of 30 seconds in some cases, although it typically takes less time than that. As a result, a 30-second spike in the I/O response time can occur for the host application generating I/O during the failback process initiated by the host multipath driver.

With target failback, however, the array controller initiates ownership transfer. It first switches the volume to write-through caching mode and flushes unwritten data from cache to media in the background while still handling I/O on that volume. Afterward, the actual ownership change happens in a fraction of the cache flush time. The host multipath driver simply redirects I/O to the new owning controller after it receives notification of the change in ownership. The driver does not queue any I/O during the process because it isn’t even aware that the process is going on until after the transfer completes. The result is a significant reduction in the I/O latency spike during failback because the controller is driving the process.

Configuration options

This section describes configuration settings for enabling or disabling the ALB feature and the settings for target failback and enhanced SAN connectivity reporting, as well as host system effects.

Enabling and disabling ALB

ALB can be enabled or disabled with a user configuration setting available in both SANtricity System Manager and SMcli. Disabling the feature stops the periodic collection of workload data and the evaluation of workload balance across controllers. It also eliminates any volume ownership transfers that were initiated solely for the purpose of balancing the workload. Prior to SANtricity OS 11.40.1 and controller firmware 08.42.XX.XX, disabling ALB also disables enhanced SAN connectivity reporting.

Note: Even with ALB disabled, the target failback feature remains active. Therefore, volume ownership transfers for the purpose of failover and failback still occur when SAN connectivity faults result in complete loss of host access to either array controller.

When should ALB be disabled?

Although overhead for volume ownership transfer has been minimized in development, the feature is not completely free, and there is some amount of added I/O latency for the duration of the transfer. As a result, ALB should be disabled in latency-sensitive environments that cannot tolerate even the slightest fluctuation in I/O response times.

Additionally, ALB should be disabled in any highly tuned deployment in which volume ownership is purposefully assigned to each array controller to achieve a very specific workload balance. Disabling ALB avoids unexpected ownership transfers beyond those intended to handle SAN connectivity faults.

Controller firmware upgrade considerations

The ALB feature is supported in SANtricity OS 11.30 and E-Series controller firmware 08.30.XX.XX and later versions. When an existing array with configured volumes is upgraded from previous firmware versions, the ALB feature is disabled by default at the time of upgrade. This behavior is intentional, because the firmware upgrade process has no method to determine whether the array is deployed in a highly tuned or latency-sensitive environment. You must specifically enable the ALB feature following such an upgrade.

New arrays running SANtricity OS 11.30 and E-Series controller firmware 08.30.XX.XX or later from the factory have ALB enabled by default upon initial startup. When the array controller firmware is upgraded to 08.30.XX.XX or later for the first time, it might determine that there is no existing setting for the ALB feature. It then checks to see whether any user-defined volumes are configured on the array. If not, ALB is enabled by default.

Host operating system effects on ALB behavior

When the array initiates an ownership transfer, there must be cooperation between the host multipath driver and the new owning controller to redirect I/O. Therefore, ALB and target failback work only when specific host OS and multipath driver combinations are in use. ALB and target failback both use a per-host-type setting in NVSRAM to allow array-initiated (implicit) ownership transfers on volumes with LUN mappings only to hosts that have support for ALB or target failback.

Take the example of a heterogeneous, host-type clustered environment, in which one volume is accessible by multiple hosts of different OS types. In this case, any host in that cluster that is not capable of ALB or target-failback precludes ownership transfers of that volume for the purpose of load balancing or failback.

ALB is implicitly disabled if all volumes are mapped to one or more hosts that are not ALB-capable. Consider the case in which a subset of the volumes on the array are mapped to one or more hosts that are not ALB-capable. ALB attempts to create the best workload balance possible by moving those volumes that are mapped only to ALB-capable hosts.

Note: You do not need to disable ALB globally in the array configuration if one or more hosts accessing the array are not running an ALB-capable operating systems and multipath solutions. You should configure the ALB setting based solely on whether the environment in which the array is deployed can tolerate array-initiated volume ownership transfers.

NetApp distributes updates to these NVSRAM settings as new host operating systems and multipath drivers are qualified for use with ALB and target failback.

See the following tables, which outline the ALB and target-failback capability of the most commonly supported host operating systems and multipath drivers.

Table 1) ALB support for E2700, E2800/EF280, E5600, E5700/EF570 series arrays.

Index	Host type	TPGS support	ALUA/ ANA support	Failover mode	Failback (mode + host or target driven)	ALB support	Expected multipath solution
1	Windows	Yes	Yes	Implicit	Implicit/target	Yes	Windows MPIO with NetApp DSM 2.0
2	Solaris (v10 or earlier)	No	No	Explicit	Explicit/host	No	Solaris MPxIO (using legacy failover methods, not TPGS)
6*	Linux MPP/RDAC	No	No	Explicit	Explicit/host	No	Legacy Linux MPP driver
7	Linux DM-MP (kernel 3.9 or earlier)	No	Yes	Implicit	Explicit/host	No	Device mapper multipath (DM-MP) with scsi_dh_rdac device handler
8	Windows clustered	Yes	Yes	Implicit	Implicit/target	Yes	Windows MPIO with NetApp DSM 2.0
10	VMware	Yes	Yes	Implicit	Implicit/target	Yes	VMware NMP/SATP_ALUA
17	Solaris (v11 or later)	Yes	Yes	Explicit	Explicit/host	No	Solaris MPxIO (with TPGS support)
22	Mac OS (ATTO HBA)	Yes	Yes	Explicit	Explicit/host	No	ATTO Multipath Director
23	Windows (ATTO HBA)	Yes	Yes	Explicit	Explicit/host	No	ATTO Multipath Director
24	Linux (ATTO HBA)	Yes	Yes	Explicit	Explicit/host	No	ATTO Multipath Director

Index	Host type	TPGS support	ALUA/ ANA support	Failover mode	Failback (mode + host or target driven)	ALB support	Expected multipath solution
25	Linux (PathManager)	Yes	Yes	Explicit	Explicit/host	No	SGI PathManager
28	Linux DM-MP (kernel 3.10 or later)	Yes	Yes	Implicit	Implicit/target	Yes	SCSI: Device mapper multipath (DM-MP) with scsi_dh_alua device handler NVMeoF: DM-MP or native NVMe multipathing
29**	ATTO clustered (all OS)	Yes	Yes	Implicit	None	No	ATTO multipath

* Present in NVSRAM, but support was dropped in E-Series controller firmware 08.25.XX.XX and later.
** New in SANtricity OS 11.30.2 and controller firmware 08.30.20.00.

Table 2) ALB support for EF300 and EF600 series arrays.

Index	Host type	TPGS support	ALUA/ ANA support	Failover mode	Failback (mode + host or target driven)	ALB support	Expected multipath solution
1	Windows	Yes	Yes	Implicit	Implicit/target	Yes	Windows MPIO with NetApp DSM 2.0
10	VMware	Yes	Yes	Implicit	Implicit/target	Yes	VMware NMP/SATP_ALUA
17	Solaris	Yes	Yes	Explicit	Explicit/host	No	Solaris MPxIO (with TPGS support)
28	Linux	Yes	Yes	Implicit	Implicit/target	Yes	SCSI: Device mapper multipath (DM-MP) with scsi_dh_alua device handler NVMeoF: DM-MP or native NVMe multipathing
29	ATTO clustered (all OS)	Yes	Yes	Implicit	None	No	ATTO multipath

Target failback

By default, Target Failback is enabled in the 08.30.XX.XX or later array firmware for hosts running operating systems and multipath driver combinations that support the capability, based on the settings in the controller NVSRAM as documented in Table 1 and Table 2.

Note: The user Enable/Disable switch for ALB does not affect target failback behavior.

A given volume can be mapped to a host group with hosts running different operating systems or multipath solutions. For example, multiple hosts in the host group can have different host types. In that case, any host in the host group that does not have support for target failback prevents the controller from initiating failback on that volume. Therefore, the control of target failback is a function of selecting the correct host type for each host in the array configuration settings.

Note: With the Linux DM-MP multipath solution, there are two host types in the array configuration settings that control the device handler in use in DM-MP. These host types in turn control whether target failback or host-initiated failback is in use. See section 0, “Deploying ALB with Linux hosts,” for details of the deployment of ALB and target failback with Linux.

Enhanced SAN connectivity reporting

Beginning with SANtricity OS 11.40.2 and controller firmware 08.42.XX.XX on E2800/EF280 and E5700/EF570 arrays and as core technology for all releases of EF300 and EF600, enhanced SAN connectivity reporting can be enabled and disabled independent of ALB through a user configuration setting. This setting can be adjusted with both SANtricity System Manager and SMcli. Disabling the feature stops the periodic evaluation of the SAN connectivity state and prevents any Recovery Guru alerts for SAN connectivity from being posted. This allows the capability to be disabled in deployments for which a lack of redundancy in the SAN is intentional, so that the array does not post a perpetual Needs Attention state.

Making this setting independent of the ALB Enable/Disable setting allows the enhanced SAN connectivity reporting feature to be used even in environments in which ALB-initiated ownership transfers are not desirable.

Existing E2800/EF280s and E5700/EF570s controllers can be upgraded to SANtricity OS 11.40.2 and controller firmware 08.42.XX.XX from previous releases that did not have this independent connectivity reporting setting. With such an upgrade, the initial value for this new configuration parameter defaults to the current setting for ALB. As a result, you must specifically reenable connectivity reporting for arrays on which ALB had previously been disabled.

Deploying ALB with Linux hosts

This section provides additional details on deploying ALB on Linux hosts, including a new device handler plug-in and the considerations for using this plug-in in a host clustered environment.

Support for new DM-MP device handler (scsi_dh_alua)

The Linux Device-Mapper Multipath (DM-MP) multipath/failover solution employs the concept of a device handler. This is a plug-in to the multipath architecture that enables vendor-specific behavior when interacting with a specific type of storage device or array. The device handler selected by DM-MP during device-discovery is based on INQUIRY data reported to the host by the storage array. E-Series arrays vary the INQUIRY data reported based on the host type selected for a given host in the array configuration. Doing so allows the storage administrator to use the host type setup in the array configuration to influence the specific device handler used in DM-MP.

The scsi_dh_rdac device handler for DM-MP is specific to E-Series storage arrays. It uses E-Series vendor-specific commands to report and manage redundant paths to the array. It has long been supported with E-Series firmware and NVSRAM releases.

Beginning with 08.30.XX.XX firmware and NVSRAM, E-Series supports the use of a second device handler (scsi_dh_alua) for DM-MP on Linux kernel version 3.10 or later. This device handler is not specific to E-Series; it is the generic handler for ALUA target devices. Although both device handlers support the same basic failover and failback capabilities for fault tolerance and recovery, only scsi_dh_alua is compatible with the ALB and target failback features.

The Linux DM-MP (kernel 3.10 or later) host type, when selected for a host running kernel version 3.10 or later, results in the use of the scsi_dh_alua device handler in DM-MP and support for ALB and target failback. The Linux DM-MP (kernel 3.9 or earlier) host type results in the use of the scsi_dh_rdac device handler and no support for ALB or target failback.

You might want to upgrade the E-Series firmware to 08.30.XX.XX or later from previous versions to deploy ALB with existing Linux hosts with kernel version 3.10 or later. In that case, the host-type setup in the array configuration must be changed to the new Linux DM-MP (kernel 3.10 or later) host type.

NetApp recommends that all new deployments of Linux kernel version 3.10 or later with E-Series firmware version 08.30.XX.XX or later use the new host type to enable ALB and target failback. This is true even if ALB has been globally disabled on the array. Going forward, scsi_dh_alua is intended to be the supported device handler for DM-MP, although scsi_dh_rdac continues to be supported at this time for the sake of backward compatibility.

Note: Other supported Linux multipath solutions, such as SGI PathManager and the ATTO Multipath Director, are not compatible with ALB or target failback.

See the Linux express configuration overview for further details.

DM-MP configuration in clustered host environments with new host type

You should give special consideration to the DM-MP configuration settings in clustered host environments when switching to the new Linux DM-MP (kernel 3.10 or later) host type. Due to the risk of ownership thrash in a clustered environment, you should place the failback manual option in the E-Series-specific section of `multipath.conf` to avoid explicit host-driven failback.

However, all arrays accessible by a given host might be using target failback; for example, they might have the correct firmware versions and use the new host type. In that case, the E-Series-specific section of `multipath.conf` on that host should contain `failback immediate` despite that being contrary to typical cluster configuration practices. Alternatively, this option can be removed completely from the E-Series-specific section of `multipath.conf` to restore the default option, which is `failback immediate`. Using `failback immediate` forces DM-MP to follow volume ownership changes more aggressively, which is desirable when using target failback.

The host might have access to E-Series arrays that do not have support for target failback or are not using the new host type. Alternatively, the host might be reconnected to such an array at a later time. In that case, you must continue to use `failback manual` to avoid ownership thrash on the arrays that are not using target failback.

Implementation details

This section provides additional details on the implementation of the ALB feature.

Automatic volume ownership distribution

The following information describes the behavior of the ALB load balance optimizer algorithm and how it identifies the optimal volume ownership distribution to achieve a balance of I/O workload across controllers.

Periodic workload balance evaluation

ALB attempts to redistribute volume ownership once per hour to optimize workload balance across the two I/O controllers. It gathers I/O workload statistics in 5-minute intervals throughout the hour and uses that near-term historical data to predict a forward-looking balance of workload based on volume ownership distribution at the time of evaluation. If that workload balance is outside of programmatically determined thresholds, ALB then identifies a new optimized distribution of volume ownership that

achieves a balance between the two I/O controllers within thresholds. The selection process attempts to minimize the total cost of changing volume ownership by minimizing the total number of volumes that need to move.

Under certain fault conditions, such as when one I/O controller has been removed or is in a failed state, ALB pauses periodic evaluations. It does not make sense to attempt to balance workload under such conditions.

Additionally, ALB does not attempt to rebalance workloads unless the workload on both controllers is sufficient to warrant it. This can occur when a very light workload is present on both controllers and there is significant headroom remaining in the capacity of the I/O controllers to handle additional I/O.

In some cases, an ALB cycle might start up and evaluate workload, but it cannot identify a new volume ownership distribution plan that is significantly better. Alternatively, a new distribution might not be enough of an improvement to warrant the overhead of initiating ownership transfers. Sometimes the workload distribution across volumes is such that it isn't possible to produce a new volume ownership distribution that meets the target thresholds. In all of these cases, ALB discards the result of the evaluation and waits for the next hourly cycle to reevaluate conditions.

I/O workload factors

I/O workload statistics gathered by ALB consider several aspects of workload:

- The number of I/O requests being received from all hosts per unit of time (IOPS).
- The total size of I/O requests being received from all hosts per unit of time (bandwidth).
- Utilization of controller write data cache memory. This considers the read/write mix in the incoming workload to make sure that both controllers have a similar level of total-write-cache utilization.
- Total workload (IOPS and bandwidth) sent to the drives per unit of time as a result of the incoming host I/O. This is a normalized workload factor that considers the RAID level and stripe width of the volumes receiving I/O from the hosts. It also considers differences in the cost of read and write operations, the number of full RAID-stripe writes, the cost of write data cache mirroring, and so on.

Workload balance objectives

ALB considers several different (sometimes competing) objectives when determining the optimal volume ownership distribution across the two I/O controllers. These objectives include:

- Achieving a balance of the measured workload factors (noted earlier in this section) across the two I/O controllers.
- Minimizing (and eliminating if possible) any split of ownership between the two controllers for volumes of RAID level 0, 1, 3, 5, or 10 that are using the same HDDs for storage (for example, volumes in the same drive group). Note that volumes that are part of a dynamic disk pool or that are backed by SSD storage are not considered in this objective. The intent of this objective is to minimize head seek time on HDDs when flushing write data cache to media. Excess head seek time can occur when both controllers contain write data in cache destined for the same set of drives.
- Minimizing the total cost of realigning volume ownership for load balancing purposes. This includes the following issues:
  - Selecting a new distribution of volume ownership that achieves a satisfactory workload balance, but that at the same time minimizes the number of volumes that must change ownership.
  - Avoiding transferring ownership of any volumes that have recently transferred ownership, either for workload distribution purposes, for failover or failback, or for any other reason. This minimizes volume ownership thrash caused by load balancing.

Also note that ALB eliminates some volumes as candidates for changing ownership to achieve the workload balance. These volumes include:

- Volumes with LUN mappings to hosts that are not ALB-capable. See section 0, “Host operating system effects on ALB behavior,” for details.
- Volumes that use the SSD cache feature, given that transferring ownership of these volumes purges contents from the SSD read cache and requires a rewarm of the cache.
- Remote mirror secondary volumes. Mirror secondary volumes follow primary volume ownership and are not directly transferable.
- Failed volumes or volumes with drives that are failed or missing.
- Volumes undergoing certain background operations that preclude noncritical ownership transfers.

Implicit failback

This section describes failback functions.

Assessment of periodic failback eligibility

The controller firmware periodically assesses volume ownership conditions, which by default run every 4.5 minutes on each controller, to determine whether target failback is needed. This assessment identifies volumes whose preferred owner is the controller on which the evaluation is running that are actually owned by the alternate controller. In other words, it finds volumes that are not on the preferred owner.

SAN connectivity conditions can indicate that all hosts with access to the given volume have regained connection to the preferred controller. In addition, the multipath driver on those hosts might have discovered the preferred controller through at least one path. If so, then a change of ownership is initiated for that volume to place it back on the preferred controller.

This evaluation is also triggered one minute after any detected change in connection state in the SAN physical transport layer. A new evaluation interrupts the existing evaluation timer so that a scheduled evaluation due to occur within the next minute is deferred for one minute. As a result, hosts in the SAN can complete some level of device discovery to potentially new or reconnected paths before the array attempts a failback.

After a volume is placed back on its preferred owner, a notification is sent to the host multipath driver on all hosts with access to that volume. It can then follow the change and start redirecting I/O to the new controller.

Conditions that delay or preclude failback

To avoid conditions that can result in repeated failover or failback cycles, the following preconditions are enforced before a failback triggers an ownership change back to the preferred controller. These preconditions also help avoid failback for a volume whose ownership has been unstable recently or that is incapable of handling an ownership change:

- A volume must not have failed back within the past 15 minutes.
- A volume must not have changed ownership for any other reason in the past 4.5 minutes. This requirement includes either a failover initiated by a host or an ownership change initiated by the user by using the CLI or UI, and so on.
- A volume must be transferable. Examples of nontransferable volumes include:
  - Remote mirror secondary volumes. Mirror secondary volumes follow the primary volume ownership and are not directly transferable.
  - Volumes with missing drives or failed volumes, because any unwritten data in cache cannot be flushed to media during the transfer.
- Volumes undergoing certain background operations that preclude noncritical ownership transfers.
- Volumes with LUN mappings to hosts that are not capable of handling target failback. See section 0, “Target failback,” for details of how this is determined.

For situations in which the volume is not transferred and remains owned by the nonpreferred controller, the target failback mechanism reevaluates eligibility of that volume for failback in the next 4.5-minute failback cycle.

Suppression of preferred ownership reporting

To facilitate host-initiated failback, the array controller firmware has historically reported information about which controller is the preferred owner of a given volume to the host multipath driver.

When target failback is in use for a given host operating system and multipath driver combination, however, the array controller firmware purposefully suppresses this reporting of preferred volume ownership. This prevents the host multipath solution from attempting to perform failback if the array is already handling the failback process, so that the two failback processes do not conflict.

In addition, users can optionally enable this suppression mechanism for hosts with operating systems and multipath drivers that do not support target failback. This can be useful in clustered host environments with host-initiated failback in which the failback would normally be disabled in the host multipath driver configuration. Instead, this setting offers a way to disable failback from the array configuration. See TR-4604: Clustered Filesystems with E-Series Products – Best Practices Guide for details about how this mechanism is enabled.

Third-party copy handling

E-Series firmware contains support for special third-party copy SCSI commands used by some host operating systems to quickly copy data from one volume to another. This copy operation is most often used in virtualization environments for which guest OS or virtual machine (VM) disks are copied from one volume or file system to another. It is also used when VM disks are cloned as a part of VM replication.

During the copy process, the array controller firmware might change ownership of either the source or destination volume involved in the copy if both are not already owned by the same controller. This ownership change helps optimize the data copy performance, but it has the side effect of potentially moving a volume's ownership off the preferred controller, which is similar to a failover.

The target-failback feature monitors in-flight third-party copy operations on the volume, and it also automatically returns volume ownership to the preferred controller after the copy process has completed. This action makes sure that no user intervention is required to place the volume back on its preferred controller while also making sure that optimal performance can be maintained throughout the copy process.

Enhanced SAN connectivity reporting

This section provides implementation details for SAN connectivity reporting.

Periodic evaluation of SAN connectivity

The ALB feature periodically assesses the SAN connectivity state, which runs by default every 5 minutes on each controller, to determine whether physical connectivity or host multipath driver issues might need to be reported to the user.

This evaluation is also triggered 1 minute after any change in connection state in the SAN physical transport layer or following any volume ownership failover or change in host or LUN mappings. Evaluation interrupts any existing evaluation timer such that a scheduled evaluation that was due to occur within the next minute is deferred until at least 1 minute has passed. As a result, hosts in the SAN can complete some level of device discovery on potentially new or reconnected paths before any connectivity or discovery evaluation is performed.

Before posting any fault conditions to the UI Recovery Guru, the fault must be observed for at least 5 minutes. Depending on when the fault occurs relative to the periodic evaluation cycles, a delay of up to 10

minutes could occur before you are alerted to the condition. This delay is intended to avoid setting the array in a Needs Attention state, and it avoids posting the error when temporary connection changes are being made in the SAN during maintenance activities.

Likewise, when an existing fault in the SAN that was posted to the UI Recovery Guru has been resolved, it can take a few minutes until the next evaluation cycle runs to clear the error. Typically, however, this condition clears within 1 minute because of the quick trigger of the evaluation cycle following observation of a SAN connection state change.

Alert for host redundancy lost

The Recovery Guru alert for Host Redundancy Lost is posted to the UI and the array Needs Attention state is set whenever a host appears to have one or more connections at the physical transport layer to one array controller, but no connections to the other array controller. This condition indicates a loss of redundancy that creates a single point of failure in the SAN. If the single connected array controller fails or is otherwise disconnected from the reported host, all I/O access from that host to the array is lost.

If the lack of redundant connections is intentional in the design of this particular SAN, you should disable connectivity reporting to silence this error. This process is described in section 0, “Enhanced SAN connectivity reporting.”

Alert for host multipath driver incorrect

The Recovery Guru alert for Host Multipath Driver Incorrect is posted to the UI and the array Needs Attention state is set whenever a host appears to have transport-protocol layer connectivity to a controller, but the host multipath driver has not discovered paths to the connected controller. This condition is determined by clearing a Device Discovery indicator flag on a given connection path (I_T Nexus) from the host when the connection is first established at the physical transport layer. You should then set the initiator discovery indicator flag on that connection path when the SCSI or NVMe-oF command sequence used by the multipath driver to discover path states (REPORT TARGET PORT GROUPS for SCSI, Identify Namespace for NVMe-oF) is received from that host over that connection path.

Posting of this error typically indicates that the multipath driver has failed to identify a new connection. Often a device rescan on the host clears this error.

In this condition, a host might have a physical connection to both controllers. However, if the multipath driver has not discovered paths to one of the controllers, a fault in the SAN that results in connection loss to the other controller still results in loss of I/O access. This is because the host multipath driver does not realize it has paths that it can use to reroute I/O.

Implicit ownership changes – Host follow-over failures

Hosts can be notified of ownership changes initiated by the array controller for load balancing or failback purposes. In that case, it is critical for the multipath driver in those hosts to begin directing I/O at the new owning controller to avoid performance penalties of I/O shipping. Typically, the host multipath driver reacts relatively quickly to the ownership change notification sent from the array (an Asymmetric Access state Change Unit Attention notification for SCSI or an ANA Change asynchronous event notice for NVMe-oF). I/O is redirected at the new owning controller within a few seconds, and only a few I/O requests are shipped following the change. However, if the multipath driver does not see the notification or otherwise ignores it, then I/O shipping might continue for an extended period of time.

This is a critical aspect of implicit or array-initiated ownership changes. Therefore, a mechanism has been implemented in the array to detect conditions in which hosts do not follow the change and continue directing I/O at the controller that formerly owned the volume. Such a condition is called a host follow-over failure because the multipath driver failed to follow the ownership change over to the new owning controller.

A follow-over failure is declared on a volume when more than 75% of I/O to the volume is received at the non-owning controller within the first 10 minutes following an array-initiated ownership transfer. Also, there has not been any detected change in the SAN connectivity state that would otherwise explain the host I/O routing decisions. The first check for this condition occurs 2 minutes into the 10-minute window. At that point, ownership is changed back to the controller receiving the majority of the I/O to stop the I/O shipping performance penalty. In addition, a 12-hour moratorium is placed on any further implicit ownership changes for ALB or target failback, given that the host multipath driver is not cooperating with implicit ownership changes.

Conclusion

The ALB feature creates a storage array design that offers an optimized ALUA solution capable of self-managing volume ownership. This allows E-Series storage arrays to simplify storage administration while still employing a design that features a data cache coherency model capable of simultaneously supporting best-in-class price/performance and enterprise-class resiliency.

Additional features bundled with ALB improve reporting of SAN connectivity faults and allow automatic failback of volume ownership even in clustered host solutions. This further simplifies storage management and enhances SAN fault tolerance.

Appendix A - Major event log content

The following major event log (MEL) events are generated by ALB, target failback, and SAN connectivity reporting.

Volume transfer events

These informational events pertain to volumes that have transferred ownership as part of a failover, failback, or ALB workload rebalance.

MEL_EV_LOAD_BALANCING_VD_TRANSFER

Event ID: 0x2044 **Priority**: INFO

Event Description: IO shipping implicit volume transfer

This event is logged when the array controller transfers ownership of a volume because it detects that more than 75% of the I/O to a given volume is being sent to the non-owning controller. This is most typically the failover case for which disruptions in the SAN force one or more hosts to lose all connectivity to the owning controller. This can also happen, however, when ownership is changed, but the host multipath driver does not observe the ownership change notification and then fails to redirect I/O to the new owning controller. See section 0, “Implicit ownership changes – Host follow-over failures” for details.

Note: This MEL event was defined in 07.83.XX.XX firmware and so it predates the implementation of ALB. As a result, the label MEL_EV_LOAD_BALANCING_VD_TRANSFER is a bit misleading, because this actually has nothing to do with ALB.

MEL_EV_IMPLICIT_FAILBACK_VD_TRANSFER

Event ID: 0x2049 **Priority**: INFO

Event Description: Failback implicit volume transfer

This event is logged when the array controller transfers ownership of a volume as part of a target failback operation.

MEL_EV_IMPLICIT_WORKLOAD_VD_TRANSFER

Event ID: 0x204A **Priority**: INFO

Event Description: Auto Load Balancing implicit volume transfer

This event is logged when the array controller transfers ownership of a volume as part of a periodic load balance operation.

Workload balance periodic evaluation cycle execution

These informational events pertain to arrays where a periodic ALB evaluation cycle was executed.

MEL_EV_ALB_OPTIMIZATION_CONSIDERED

Event ID: 0x9104 **Priority**: INFO

Event Description: Analysis of workload balance was performed.

This event is logged when the array controller periodically evaluates workload balance and finds that the conditions on the array merit searching for a better volume ownership distribution that improves workload balance. This event is not logged when the array is under light load or if workload is already balanced such that further evaluation is not necessary. Optional data logged with this event records the conditions on the array used to decide about further evaluation.

MEL_EV_ALB_OPTIMIZATION_PERFORMED

Event ID: 0x9105 **Priority**: INFO

Event Description: Workload was automatically balanced by transferring volumes.

MEL_EV_ALB_OPTIMIZATION_EVALUATED

Event ID: 0x9106 **Priority**: INFO

Event Description: Workload balance was evaluated following load optimization.

This event is logged when an evaluation of the effectiveness of a previous workload rebalance is performed. This evaluation occurs 30 minutes after completion of the workload rebalance. The optional data records workload conditions at the time of the effectiveness evaluation. The intent of this event is simply to capture workload data to assist with long-term engineering evaluations of the effectiveness of the ALB algorithm.

Auto load balancing user enablement

These informational events pertain to arrays for which a change has been made to the global ALB enable/disable setting. See section 0, “Enabling and disabling ALB” for details.

MEL_EV_AUTO_LOAD_BALANCE_ENABLED

Event ID: 0x9100 **Priority**: INFO

Event Description: ALB is enabled.

This event is logged when you set the global ALB enable/disable switch to enable ALB.

MEL_EV_AUTO_LOAD_BALANCE_DISABLED

Event ID: 0x9101 **Priority**: INFO

Event Description: ALB is disabled.

This event is logged when you set the global ALB enable/disable switch to disable ALB.

Connectivity reporting user enablement

These informational events pertain to arrays for which a change has been made to the global SAN connectivity reporting enable/disable setting. See section 0, “Enhanced SAN connectivity reporting,” for details.

MEL_EV_HOST_CONNECTIVITY_REPORTING_DISABLED

Event ID: 0x9107 **Priority**: INFO

Event Description: Host connectivity reporting is disabled.

This event is logged when you set the global connectivity reporting enable/disable switch to disable reporting.

MEL_EV_HOST_CONNECTIVITY_REPORTING_ENABLED

Event ID: 0x9108 **Priority**: INFO

Event Description: Host connectivity reporting is enabled.

This event is logged when you set the global connectivity reporting enable/disable switch to enable reporting.

Connectivity reporting – Alert events

These critical and alertable events pertain to arrays for which the connectivity reporting mechanism identified host connectivity issues or multipath configuration issues that impact I/O path redundancy to the array. See section 0, “Enhanced SAN connectivity reporting,” for details of when these events are generated.

MEL_EV_HOST_REDUNDANCY_LOST

Event ID: 0x9102 **Priority**: CRITICAL / ALERT

Event Description: A loss of host-side connection redundancy is detected.

This event is logged when the Recovery Guru for Host Redundancy Lost is posted to the UI. This alert event generates an SNMP trap (or an email message, if configured to do so).

MEL_EV_MULTIPATH_CONFIG_ERROR

Event ID: 0x9103 **Priority**: CRITICAL / ALERT

Event Description: A host multipath driver configuration error is detected.

This event is logged when the Recovery Guru for the Host Multipath Driver Incorrect is posted to the UI. This alert event generates an SNMP trap (or an email message, if configured to do so).

Where to find additional information

To learn more about the information described in this document, refer to the following documents and/or websites:

- Linux express configuration
  https://docs.netapp.com/us-en/e-series/config-linux/index.html

- VMware express configuration overview
  https://docs.netapp.com/us-en/e-series/config-vmware/index.html
- Windows express configuration overview
  https://docs.netapp.com/us-en/e-series/config-windows/index.html
- Multipath Configuration Power Guides (for SANtricity OS 11.40 and earlier)
  https://mysupport.netapp.com/info/web/ECMLP2522638.html
- Installing and Configuring for Linux: Power Guide for Advanced Users (for SANtricity OS 11.40 and earlier)
  https://mysupport.netapp.com/ecm/ecm_download_file/ECMLP2601371
- TR-4604: Clustered File Systems with E-Series Products – Best Practices Guide:
  https://fieldportal.netapp.com/content/544142

Version history

Version	Date	Document version history
Version 1.0	December 2018	Document initial release.
Version 2.0	June 2020	Updates for NVMe-oF and to reflect new platform support (EF600).
Version 2.1	February 2022	Updates to reflect new platform support (EF300) including host OS support and linking host OS express guides.

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.

TR-4737-0322

NetApp