Dell PowerStore: Clustering and High Availability

October 2022

H18157.7

White Paper

Abstract

This white paper provides an overview of PowerStore clustering technology along with the highly available features of the platform for the PowerStoreOS releases.

Dell Technologies

Contents

Executive summary

Overview
This white paper describes the fully redundant hardware and high-availability features that are available on Dell PowerStore. These features are designated to withstand component failures in the system and in the environment, such as network or power failures. If an individual component fails, the storage system can remain online and continue to serve data. The system can also withstand multiple failures if they occur in separate component sets. After the administrator is alerted about the failure, they can order and replace the failed component without any impact.

Audience
This document is intended for IT administrators, storage architects, partners, and Dell Technologies employees. This audience also includes any individuals who evaluate, acquire, manage, operate, or design a Dell networked storage environment using PowerStore systems.

Revisions

Date	Description
April 2020	Initial release: PowerStoreOS 1.0.0
September 2020	Updates to high availability section; other minor updates
April 2021	PowerStoreOS 2.0.0 updates
January 2022	PowerStoreOS 2.1.0 updates; template update
April 2022	PowerStoreOS 2.1.1 updates
July 2022	PowerStoreOS 3.0 updates
August 2022	Minor correction in the 'Distributed sparing' section
October 2022	PowerStoreOS 3.2 updates

We value your feedback
Dell Technologies and the authors of this document welcome your feedback on this document. Contact the Dell Technologies team by email.

Author: Jonathan Tang, Ryan Meyer

Note: For links to other documentation for this topic, see the PowerStore Info Hub.

Introduction

Document purpose
Having constant access to data is a critical component in any modern business. If data becomes inaccessible, business operations might be affected and can potentially cause revenue loss. Because of this, IT administrators are tasked with ensuring that every component in the data center has no single point of failure. This white paper details how to cluster PowerStore systems and describes the high availability features that are available on the PowerStore system. PowerStore is designed for 99.9999% availability (based on the Dell Technologies specification for PowerStore, April 2020; actual system availability might vary).

PowerStore product overview

PowerStore achieves new levels of value, flexibility, and simplicity in storage. It uses a container-based microservices architecture, advanced storage technologies, and integrated machine learning to unlock the power of your data. It is a versatile platform with a performance-centric design that delivers scale-up and scale-out capabilities, always-on data reduction, and support for next-generation media.

PowerStore brings the simplicity of the public cloud to on-premises infrastructure, streamlining operations with an integrated machine learning engine and seamless automation. It offers predictive analytics to easily monitor, analyze, and troubleshoot the environment. PowerStore is highly adaptable, providing the flexibility to host specialized workloads directly on the appliance and modernize infrastructure without disruption. It also offers investment protection through flexible payment solutions.

Terminology

The following table provides definitions for some of the terms that are used in this document.

Table 1.   Terminology

Term	Definition
Appliance	Solution containing the base enclosure and any attached expansion enclosures.
Base enclosure	Used to reference the enclosure containing both Nodes (Node A and Node B) and 25x NVMe drive slots.
Cluster	One or more appliances in a single grouping and management interface. Clusters are expandable by adding more appliances to the existing cluster, up to the allowed amount for a cluster.
Expansion enclosure	Enclosures that can be attached to a base enclosure to provide additional storage.
Fibre Channel (FC) protocol	A protocol used to perform Internet Protocol (IP) and Small Computer Systems Interface (SCSI) commands over a Fibre Channel network.
File system	A storage resource that can be accessed through file-sharing protocols such as SMB or NFS.
Internet SCSI (iSCSI)	Provides a mechanism for accessing block-level data storage over network connections.
Intra-cluster Management (ICM) Network	An internal management network that provides continuous management connectivity between appliances in the PowerStore cluster.
Intra-cluster Data (ICD) Network	An internal network that provides continuous storage connectivity between appliances in the PowerStore cluster.
Network-attached storage (NAS) server	A virtualized network-attached storage server that uses the SMB, NFS, FTP, and SFTP protocols to catalog, organize, and transfer files within file system shares and exports. A NAS server, the basis for multitenancy, must be created before you can create file-level storage resources. NAS servers are responsible for the configuration parameters on the set of file systems that it serves.

Term	Definition
Network File System (NFS)	An access protocol that allows data access from Linux/UNIX hosts on a network.
Node	Component within an appliance that contains processors and memory. Each appliance consists of two nodes.
NVMe over Fibre Channel (NVMe/FC)	Protocol used to perform Non-Volatile Memory Express (NVMe) commands over a Fibre Channel network.
NVMe over TCP (NVMe/TCP)	Protocol used to perform Non-Volatile Memory Express (NVMe) commands over an Ethernet network.
PowerStore Command Line Interface (PSTCLI)	An interface that allows a user to perform tasks on the storage system by typing commands instead of using the UI.
PowerStore Manager	An HTML5 user interface used to manage PowerStore systems.
PowerStore T model	Container-based storage system that is running on purpose-built hardware. This storage system supports unified (block and file) workloads, or block-optimized workloads.
PowerStore X model	Container-based storage system that is running inside a virtual machine that is deployed on a VMware hypervisor. In addition to the block-optimized workloads that this storage system offers, it also allows users to deploy applications to be deployed directly on the array.
Representational State Transfer (REST) API	A set of resources (objects), operations, and attributes that provide interactive, scripted, and programmatic management control of the PowerStore cluster.
Server Message Block (SMB)	A network file sharing protocol, sometimes referred to as CIFS, used by Microsoft Windows environments. SMB is used to provide access to files and folders from Windows hosts on a network.
Snapshot	A point-in-time view of data stored on a storage resource. A user can recover files from a snapshot, restore a storage resource from a snapshot, or clone the snapshot to provide access to a host.
Storage Policy Based Management (SPBM)	Using policies to control storage-related capabilities for a VM and ensure compliance throughout its life cycle.
Volume	A block-level storage device that can be shared out using a protocol such as iSCSI or Fibre Channel.
vSphere Virtual Volumes (vVols)	A VMware storage framework that allows VM data to be stored on individual Virtual Volumes. This ability allows for data services to be applied at a VM-level of granularity and according to SPBM. Virtual Volumes can also refer to the individual storage objects that are used to enable this functionality.
vSphere API for Array Integration (VAAI)	A VMware API that improves ESXi host utilization by offloading storage-related tasks to the storage system.
vSphere API for Storage Awareness (VASA)	A VMware vendor-neutral API that enables vSphere to determine the capabilities of a storage system. This feature requires a VASA provider on the storage system for communication.

Clustering

Overview

Every PowerStore appliance is deployed into a PowerStore cluster. There is a minimum of one PowerStore appliance and a maximum of four PowerStore appliances that you can configure in the cluster. You can create a cluster made up of all PowerStore T model appliances or all PowerStore X appliances, but you cannot mix appliances with both models in the same cluster. When a multi-appliance cluster is deployed, you can perform this task during the initial configuration process or add appliances to an existing cluster. You can scale down PowerStore clusters by removing appliances from an existing cluster. As of PowerStoreOS 3.2, multi-appliance clustering for PowerStore X appliances is no longer supported.

Clustering PowerStore appliances can provide many benefits:

- Easy scale-out to increase CPU, memory, storage capacity, and front-end connectivity

- Independent scaling of storage and compute resources

- Centralized management for multi-appliance cluster

- Automated orchestration for host connectivity

- Increased resiliency and fault tolerance

Consider the example shown in the following figure. Users can scale out appliances with different model numbers to create a four-appliance cluster. Besides the scale-out benefit, each appliance can scale up with different numbers of expansion enclosures. For example, each appliance that is shown has a different number of expansion enclosures attached. Finally, each appliance in the cluster can have different media types. For example, the PowerStore 1000T model could have NVMe SCM drives, while the PowerStore 3000T, 5000T, and 7000T models could have NVMe SSD drives. This flexible scale-out and scale-up deployment gives customers the ability to grow their clusters with no dependence on the model number, drive count, or even drive type.
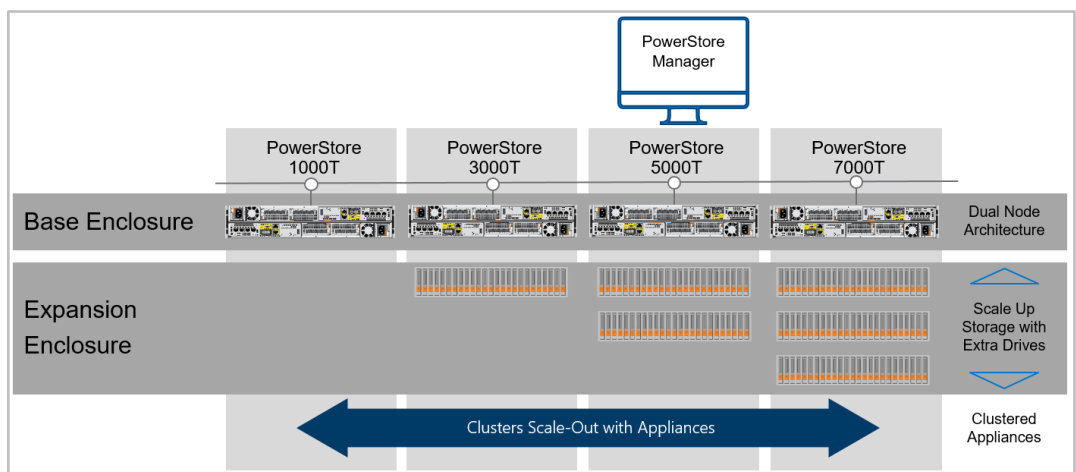

Figure 1.   Multi-appliance cluster

Requirements

Ensure that the PowerStore systems are cabled correctly by referencing the *PowerStore Quick Start Guide* on dell.com/powerstoredocs. Each node in every appliance that is a part of a PowerStore cluster must communicate to other nodes through the bonded ports (see the following figure). The network that allows the nodes to communicate to each other is an internal network named the Intra-cluster Management (ICM) and Intra-cluster Data (ICD) networks. In multi-appliance PowerStore clusters, the ICM and ICD networks communicate through the top-of-rack switch network with untagged VLAN network packets that have auto-generated IPv6 addresses. All appliances in the cluster should be in the same rack or multiple racks in the same data center. If PowerStore appliances span multiple switches, ensure that the untagged network (or native VLAN) is configured on the switch ports and are shared across the switches. Deploying a cluster in different buildings but on the same campus, also known as stretched clusters, is not supported. For single-appliance clusters, starting in PowerStoreOS 1.0.2, the ICM network communicates through the backplane within the appliance instead of through the top-of-rack switch. In PowerStoreOS 3.0 single-appliance clusters, both the ICM and ICD networks communicate through the backplane of the appliance. By moving these networks away from the top-of-rack switch, single-appliance PowerStore T systems can be deployed with minimal top-of-rack switch configuration.
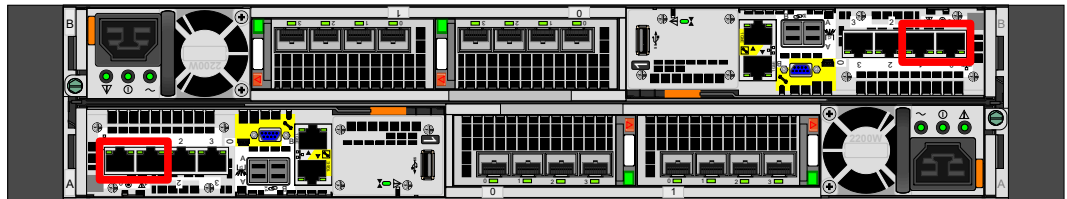

Figure 2.   Back view of PowerStore appliance

Primary appliance

Each cluster has a primary appliance which is the first appliance that is selected during the initial configuration. For PowerStore T model appliances, the primary appliance can be configured in unified mode or block-optimized mode during the initial configuration process. Starting in PowerStoreOS 2.0, the primary appliance is automatically selected by the operating system for Block-optimized clusters.

The following figure shows an example of an appliance with unified mode selected. File services only run on the primary appliance of a unified cluster. If an appliance uses file services, ensure that the **Storage Configuration** field has **Unified** mode selected. If an appliance only uses block workloads such as iSCSI, FC, NVMe/TCP, or NVMe/FC, ensure that the **Storage Configuration** has **Block Optimized** mode selected.

Note: You cannot change the configuration type (Unified or Block Optimized) of an appliance after the cluster is configured.

Figure 3.   Unified storage configuration

Initial configuration

When completing the Initial Configuration Wizard, the PowerStore system scans for other appliances on the network. If there are multiple types of PowerStore T model or PowerStore X model appliances on the network, the wizard only displays the PowerStore model that is initially selected. For example, if a PowerStore T model is selected as the primary appliance, only PowerStore T model appliances will be displayed. To add extra appliances to the cluster, click the **Select Additional Appliances** button. The resulting view shows all available appliances that are eligible to be a part of the cluster.

The following figure shows an example of selecting a multi-appliance cluster. There are no PowerStore X model appliances available for selection because a PowerStore T model appliance was selected as the primary appliance. For more information about the discovery process, see the document PowerStore: Introduction to the Platform.
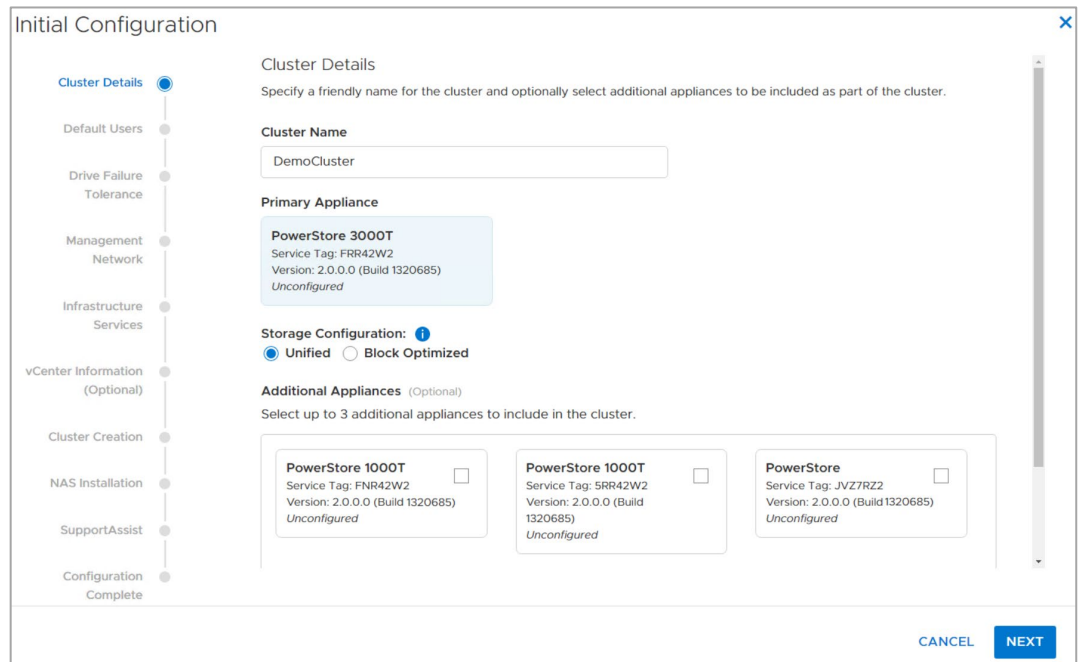
Figure 4.   Multi-appliance selection

PowerStore T cluster

Below are some general guidelines to consider when creating a multi-appliance PowerStore T cluster for the first time. Unified mode and block-optimized mode are detailed in the following sections of the white paper.

- The final cluster size cannot exceed four appliances.

- You can only select healthy and uninitialized appliances.

- Each appliance is configured synchronously starting with the primary appliance.

- Only the primary appliance can be in unified mode if selected.

- All additional appliances are automatically rebooted into a block-optimized configuration when added into a cluster.

- The appliances must be able to communicate with each other with untagged VLAN and IPv6 addresses on the bonded interfaces of the PowerStore system.

- The system bonds of all appliances should be on the same native VLAN.

Unified cluster

When setting up a cluster for the first time, determine if the cluster will use file services. If a decision has not been made, we recommend deploying a unified cluster to provide the most flexibility in the future. As mentioned in Primary appliance, services only run on the primary appliance of a unified cluster. If other appliances are configured in the cluster, those appliances are deployed in a block-optimized configuration.

The following figure shows an example of creating a multi-appliance cluster. The primary appliance is set to unified, and the other appliances are rebooted into a block-optimized configuration.
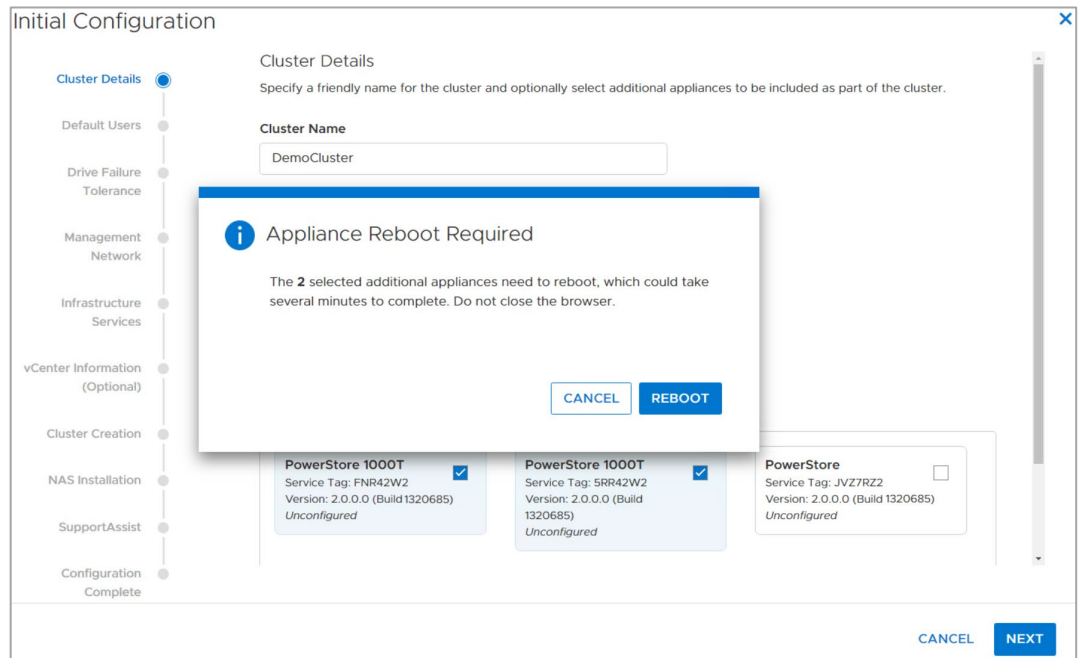
Figure 5.   Reboot non-primary appliances to block optimized configuration

If file services are not used, you can configure a cluster to be block optimized to deliver the highest possible block performance. The following figure shows an example of a multi-appliance cluster that will be configured as block optimized. The primary appliance is set to block optimized configuration. This means that all the appliances that are selected must be automatically rebooted into a block-optimized configuration before the cluster can be configured.

Note: You cannot change the configuration type (unified or block optimized) of the cluster after the cluster is configured.
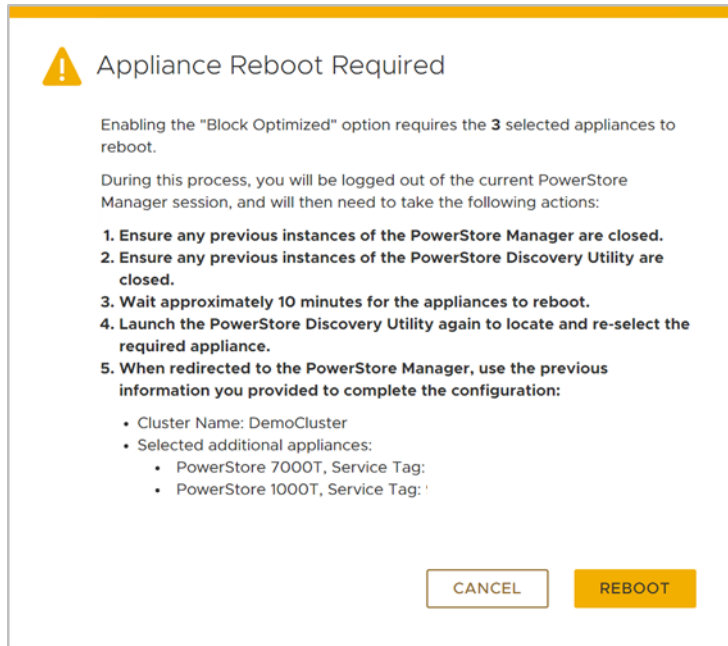
Figure 6.   Reboot all appliances in block-optimized cluster

PowerStore X cluster

Below are some general guidelines to consider when creating a multi-appliance PowerStore X[1] cluster for the first time.

- The final cluster size cannot exceed four appliances.

- The user can only select healthy and uninitialized appliances.

- Each appliance is configured synchronously starting with the primary appliance.

- The appliances must be able to communicate with each other with untagged VLAN and IPv6 addresses on the bonded interfaces of the PowerStore system.

- The first two ports of the 4-port card of all appliances should be on the same native VLAN.

Add appliance   PowerStore allows scaling a cluster, scaling up and scaling out independently. To add more capacity, scale up by adding expansion enclosures. To add compute, memory, and connectivity, scale out by adding appliances to an existing cluster. You can add an appliance using the PowerStore Manager, REST API, and PSTCLI.

Consider these general guidelines when adding appliances to an existing cluster:

- The final cluster size cannot exceed four appliances.

- You can only add uninitialized appliances.

- You can only add one appliance at a time.

- The new appliance and the existing cluster must be in a healthy state.

---

[1] As of PowerStoreOS 3.2, multi-appliance clustering for PowerStore X appliances is no longer supported.

- All appliances must be able to communicate with each other using untagged VLAN and IPv6 addresses on the system bond of the PowerStore system.

- The system bonds of all appliances should be on the same native VLAN.

Selecting the appliance

PowerStore Manager simplifies adding an appliance into an existing cluster. The following figure shows an example of adding an appliance into an existing cluster. To add the appliance, go to the **Hardware** page and click the **Add** button. This action presents the available unconfigured appliances that can be added. If you are adding a PowerStore T model to the cluster, the appliance automatically reboots into the block-optimized configuration. For PowerStore clusters running PowerStoreOS 3.0 and later, any unconfigured appliances being added to that cluster automatically sync to the same version of PowerStoreOS that is running on the cluster. Users are prompted to click the "Synchronize" button after selecting the unconfigured appliance. When the synchronization is complete, users can then continue with the Add Appliance wizard to add unconfigured appliances. When adding an appliance, the unconfigured appliance must be within one major release (N-1 or N+1) in order for the software to sync.
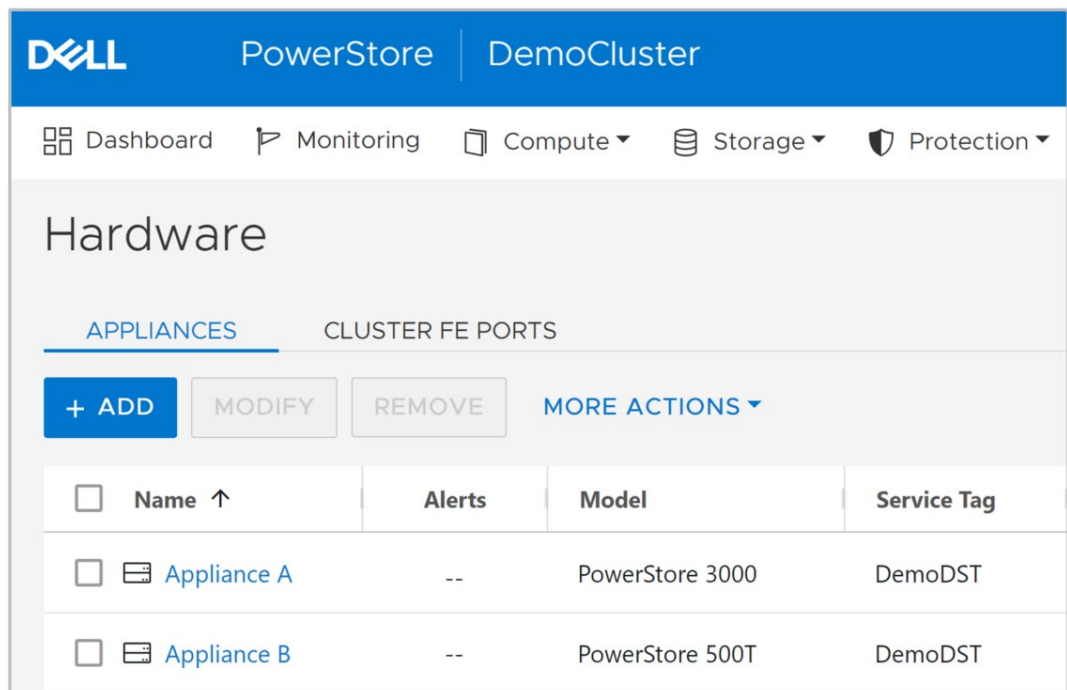

Figure 7.   How to add an appliance

Configuring IP addresses

After you select an appliance, extra IP addresses must be added to the cluster if none are available. There are two methods for adding additional IP addresses. For example, Figure 8 and Figure 9 show how to add IP addresses during the **Add appliance** workflow. In this example, a single PowerStore T model appliance is being added to the cluster. For cluster expansion, you must add three more management IP addresses and two storage IP addresses (not pictured below) to expand the cluster. Click **Add Network IPs** to supply new IP addresses for their management or storage networks.
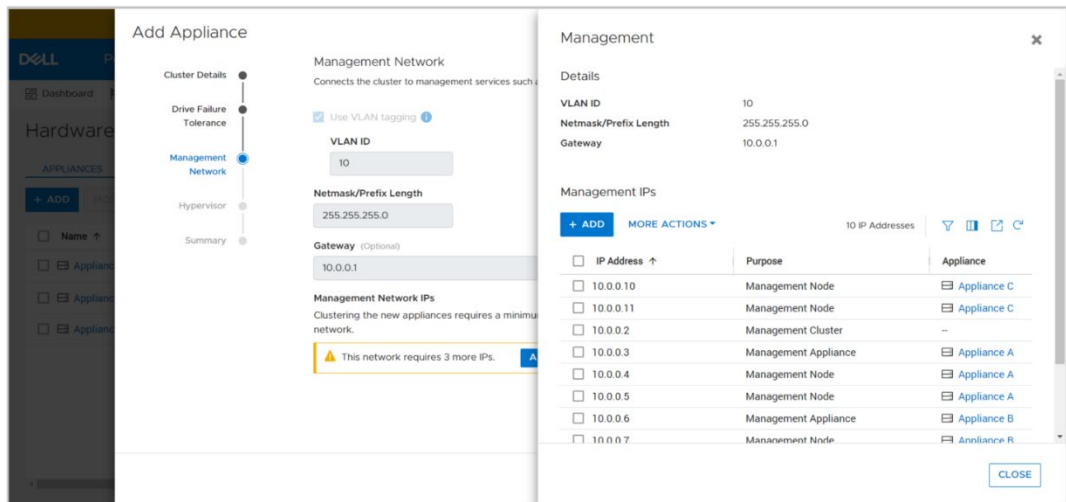
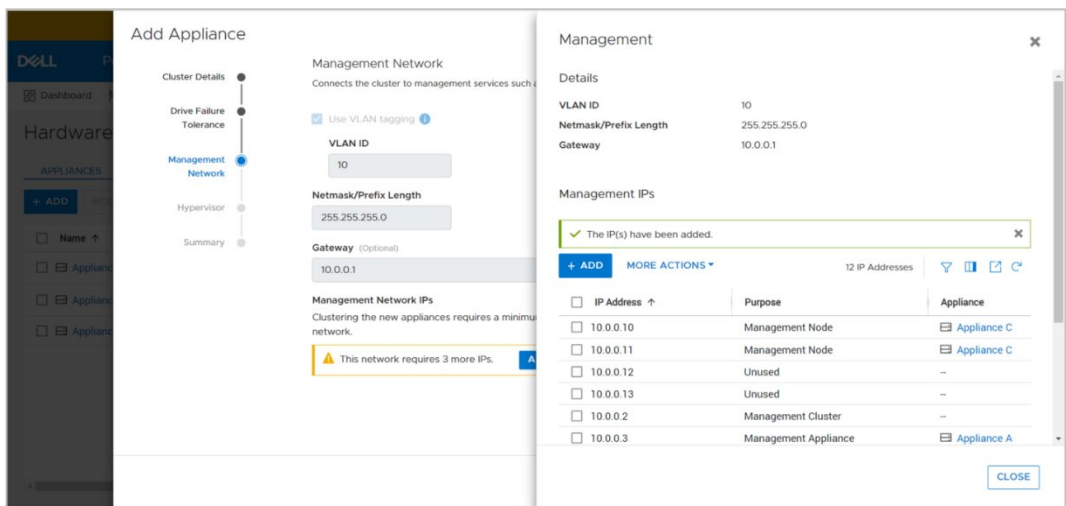**Figure 8.   How to add IPs from the Add Appliance Wizard**

Figure 9.   Adding management IPs

The alternative method for adding IP addresses is to go to **Settings** > **Networking** > **Network IPs**. The following figure shows an example of this page. A user can reserve a range of extra IP addresses where the PowerStore system might initially use only a small set of those addresses.
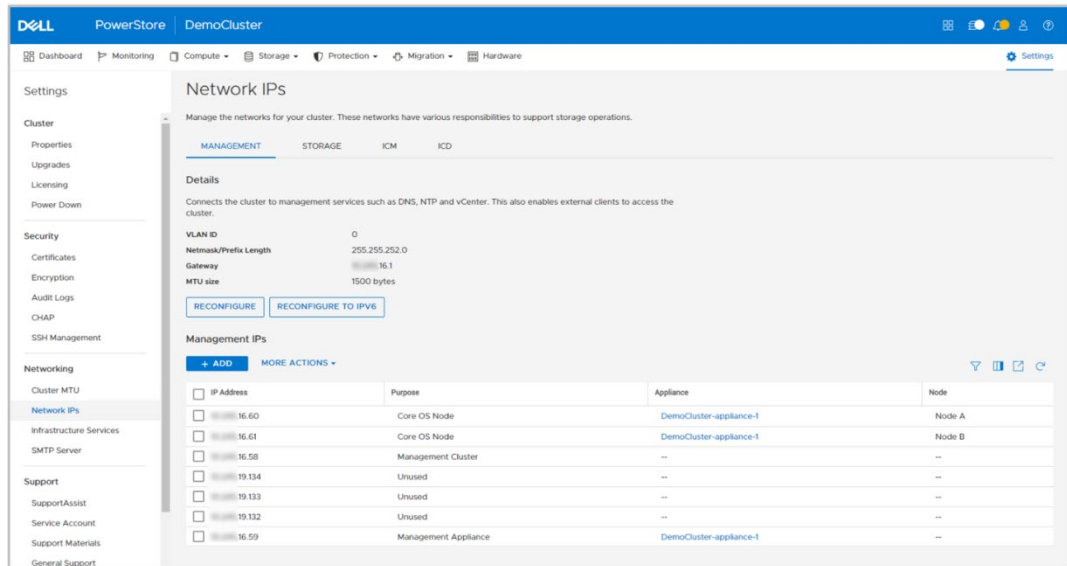
Figure 10.   How to add IPs from the Settings page

In the **Add Appliance** > **Summary** page, perform a network validation check before starting the Add Appliance job. Warnings or errors might be displayed after the network validation check is performed. We recommend addressing warnings before beginning a cluster creation, but you can bypass the warnings. However, if there are any errors, you must correct them before starting the Add Appliance job. If the errors are not addressed, users are blocked from starting a cluster-creation job. Figure 11 shows an example of the Summary page where you can validate the network.

Figure 11.   Add appliance network validation
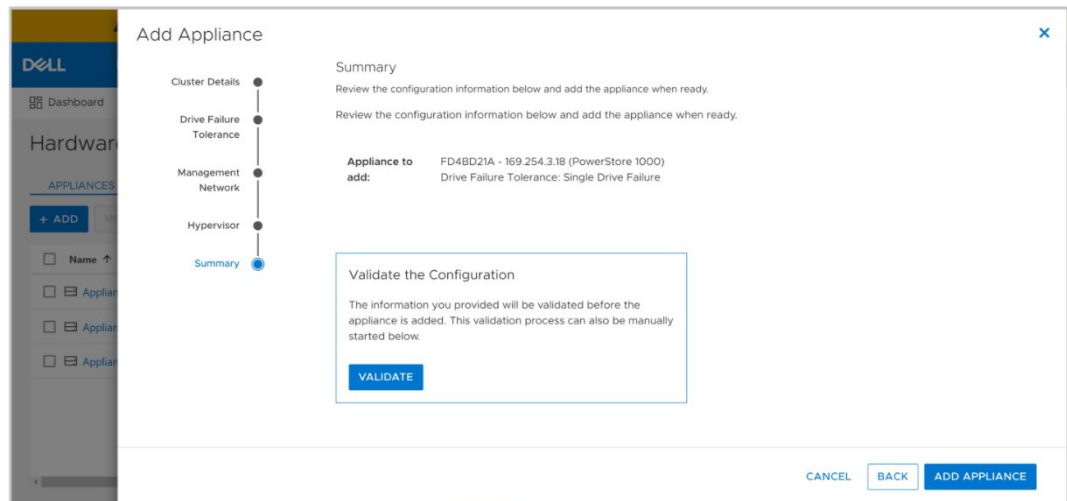
Review the job

After you complete the Add Appliance wizard, a new job is created. To see more details about the job, click the **Jobs** icon in the top-right area of PowerStore Manager. Figure 12 shows an example of an Add Appliance job that is running in the background. This job allows you to perform other tasks in PowerStore Manager while the Add Appliance job runs.
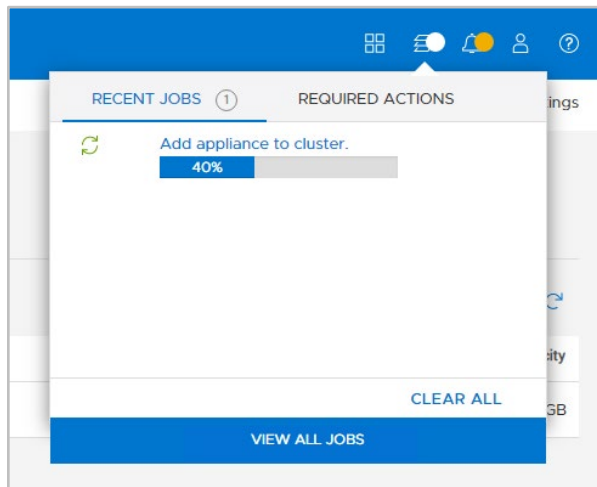
Figure 12.   Add appliance job

Remove appliance

You can scale down both PowerStore T and PowerStore X model clusters by removing individual appliances from the cluster. Removing an appliance from a cluster requires careful planning and consideration. Before starting this process, you can automatically or manually migrate storage resources from the appliance. You can manually delete any remaining storage resources before removing the appliance from the cluster. After you remove the appliance from the cluster, the appliance is reverted to the original factory settings.

When removing an appliance from the cluster, there are different steps that users need to be aware of depending on the model of the PowerStore appliance. Table 2 shows the high-level steps for removing a PowerStore T model appliance from the cluster. However, for the PowerStore X model appliance, users must migrate storage resources and virtual machines (VMs) to other appliances in the cluster as shown in Table 3. For detailed steps regarding removing a PowerStore X model appliance, see the PowerStore Manager Online Help.

Table 2.   High-level steps to remove a PowerStore T model appliance

	Entity	Operation
1	PowerStore Manager	Migrate storage resources to other appliances in the cluster
2	PowerStore Manager	Disable support notifications
3	PowerStore Manager	Remove the appliance from the cluster

Table 3.   High-level steps to remove a PowerStore X model appliance

	Entity	Operation
1	PowerStore Manager	Migrate storage resources to other appliances in the cluster
2	PowerStore Manager	Disable support notifications
3	PowerStore Manager	Identify the non-primary node
4	vSphere Web Client	vMotion VMs located on the non-primary node to other appliances on the cluster

	Entity	Operation
5	vSphere Web Client	Enable maintenance mode of the non-primary node
6	vSphere Web Client	Remove the non-primary node from the ESXi cluster
7	vSphere Web Client	Exit maintenance mode of the non-primary node
8	vSphere Web Client	Power on the Controller VM
9	vSphere Web Client & PowerStore Manager	Repeat the steps above for the primary node
10	vSphere Web Client & PowerStore Manager	Remove the appliance from the cluster

PowerStore Manager

Starting in PowerStoreOS 2.0, the recommended method to remove an appliance from a cluster is now available from PowerStore Manager. You can perform this operation with a two-step process. The first step of the procedure is to migrate storage resources from the appliance that will be removed. You can easily perform this procedure from PowerStore Manager. Starting from the **Hardware** page, select the appliance to be removed, and select **More Actions** > **Migrate**. You are guided through a wizard-based workflow to automatically move storage resources to other appliances in the cluster. More information regarding appliance space evacuation is discussed later in this document.

The second step is to remove the appliance from the cluster. As shown in the following figure, from the **Hardware** page, select the appliance to be removed, and select **Remove**.
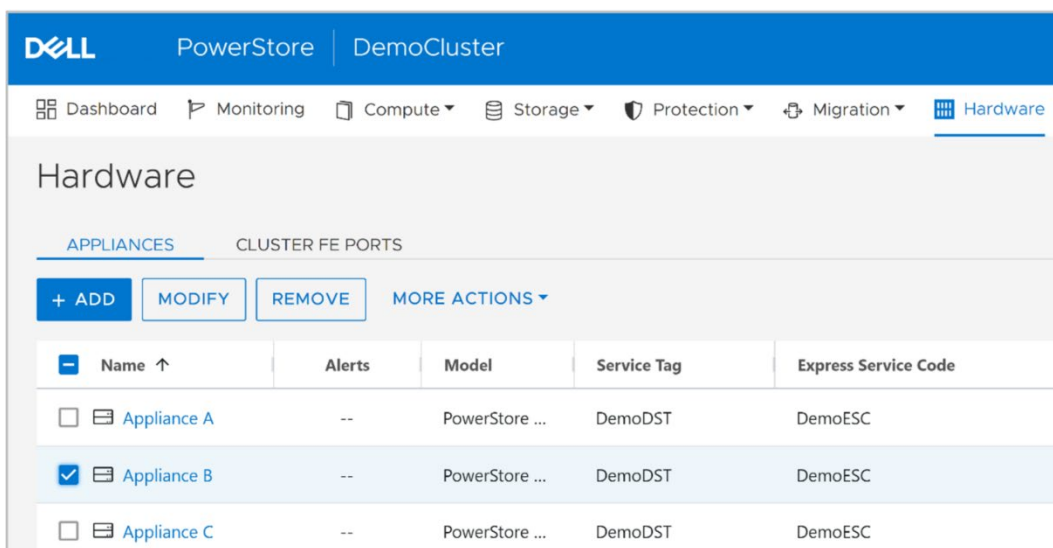

Figure 13. Remove appliance

Service script

For users that are not running PowerStoreOS 2.0 and later, see the following guidelines and workflow for running the service script to remove an appliance from a cluster:

- Notify all users to avoid creating storage resources or virtual machines while this operation is in progress.

- Run the **svc_remove_appliance** script and note the storage resources and workloads on the appliance.

- Using the prompts from the script, review and stop the scheduler and automated storage placement, allowing all active jobs to complete.

- Migrate storage resources or virtual machines to another appliance in the cluster.

- Cycle through the prompts in the script to review and ensure that there are no more active workloads on the appliance.

- Follow the prompts and confirm to start removal of the appliance.

Note: The **svc_remove_appliance** service script does not automatically migrate storage resources off the appliance that is being removed.

VMware integration

PowerStore T models and PowerStore X models offer deep integration with VMware vSphere. These integrations include as VAAI and VASA support, event notifications, snapshot management, storage containers for virtual volumes (vVols), and virtual machine discovery and monitoring in PowerStore Manager. By default, when a PowerStore T model or PowerStore X model is initialized, a storage container is automatically created on the appliance. Storage containers are used to present vVol storage from PowerStore to vSphere. vSphere mounts the storage container as a vVol datastore and makes it available for VM storage. This vVol datastore is then accessible by internal ESXi nodes or external ESXi hosts. If an administrator has a multi-appliance cluster, the total capacity of a storage container spans the aggregation of all appliances in the cluster. For PowerStore X model appliances, the cluster Distributed Resource Scheduler (DRS) is set to partially automated mode. Because the appliance is optimized for and expects this configuration, changing these settings is not supported. For more information about VMware Integration with PowerStore, see the white paper PowerStore: Virtualization Integration.

Resource balancer

In the modern data center, administrators must be quick and agile to support mission-critical applications. PowerStore offers an intelligent analytical engine that is built into the PowerStore operating system. This engine offers many benefits such as helping administrators make decisions that are based on initial placement of data, and automatically balancing node performance within an appliance. It also assists with volume migrations or storage expansions that are based on analytics and capacity forecasting.

Initial placement of data

When storage administrators create storage resources, the resource balancer can provide many benefits. For example, when a multi-appliance cluster is deployed, the resource balancer intelligently and automatically places newly created volumes on different appliances in the cluster. This placement is based on which appliance has the most unused capacity. If you choose to override the decision of automatic placement, you can manually select the appliance on which to place the volume.

The following figure shows an example of a volume that will be placed automatically on a specific appliance. Note that when volumes or volume groups are created, they are only placed on a single appliance and do not span multiple appliances in the cluster. After these storage resources are created, they do not move until they are migrated manually to another appliance in the cluster. Migration of storage resources within a cluster is discussed in the following sections of this paper.

Figure 14.   Automatic placement of volumes or manual assignment

Volume groups

In PowerStore, when you create volume groups for the first time, all corresponding members in the volume group are placed on the same appliance. If you have existing volumes, there might be situations in which these volumes could have been automatically placed on different appliances in the cluster. In these situations, you must manually migrate the storage resources to a single appliance before placing them into a volume group.

Migration overview

After you create a volume or volume group on an appliance, capacity and demand might change over time. PowerStore supports moving a volume, volume group, or vVol to another appliance within the cluster. You can perform this operation with a Manual Migration, Assisted Migration, or Appliance Space Evacuation in a PowerStore cluster by using PowerStore Manager, REST API, and PSTCLI. These methods are detailed in the next sections of this paper. For more information regarding vVol migrations, see the document PowerStore: Virtualization Integration.

Before the start of a migration job, the administrator must verify the following steps:

- If applicable, verify or setup zoning on FC switches between the host, source appliance, and destination appliance.

- Ensure that the host has connectivity to the source and destination appliances with the iSCSI, FC, NVMe/TCP, or NVMe/FC protocol.

- Ensure that the host object in PowerStore Manager shows initiators going to the source and destination appliances.

- From the host, perform a rescan of the storage object that will be migrated.

After you verify the above steps, you can start a migration job which is non-disruptive and transparent to the host. Although it is not visible to the end user, the storage resource is transferred to the new appliance within the cluster by using asynchronous replication technology. When the storage resource is being moved to the destination, all associated

storage objects, such as snapshots and clones that are tied to the storage resource, are moved to the destination also. After the data transfer is complete, the host switches paths automatically to the destination appliance and the migration job is complete.

The following figure shows an example of a multi-appliance cluster that is made up of four appliances. This example involves moving a storage resource from the PowerStore 1000T model over to a PowerStore 5000T model. The host may have connectivity to the two appliances by using iSCSI, FC, NVMe/TCP, or NVMe/FC protocols. However, the PowerStore 3000T and 7000T models do not have host connectivity. This means that the storage resource would not be able to migrate over to those appliances until host connectivity is established.
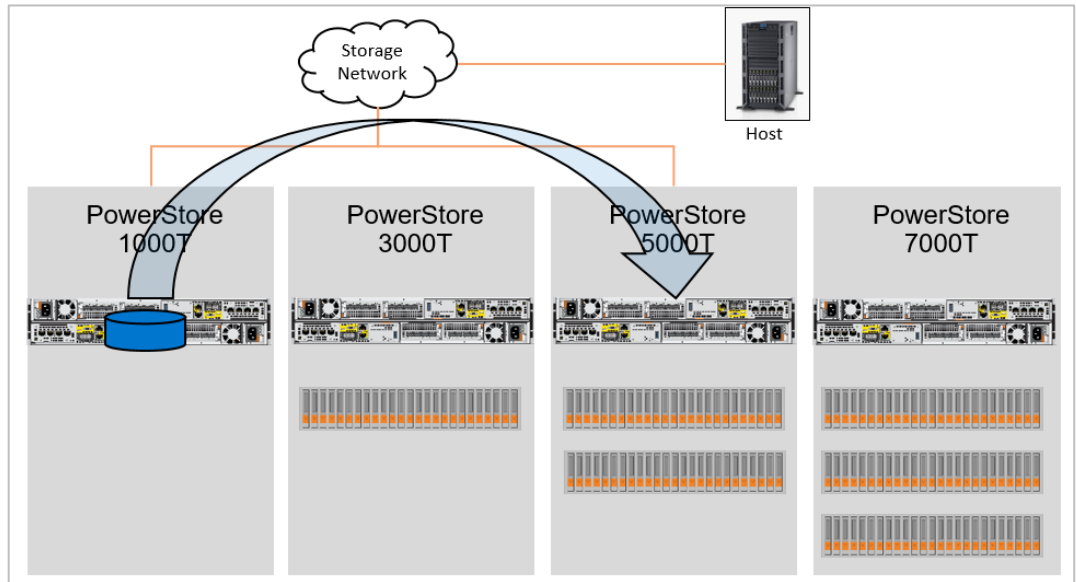

Figure 15.   Volume migration example

Manual migration

PowerStore supports moving storage resources such as a volume, volume group, or vVol to another appliance within the cluster. Before moving the storage resource, ensure that paths are seen from the PowerStore Manager and a host rescan has been performed. After the host initiator paths are verified, you can perform a manual migration on the storage resource.

Consider the following points when performing a manual migration in PowerStore for a volume, volume group, or vVol depends on the storage resource:

- Volume: Select **Storage** > **Volumes** to view a list of all available volumes in the cluster. Select a volume to migrate and click **More Actions** > **Migrate.**

- Volume group: Select **Storage** > **Volume Group** to view a list of all available volumes in the cluster. Select a volume group to migrate and click **More Actions** > **Migrate.**

- vVol: Select **Storage** > **Storage Containers** and select the storage container where the virtual machine is deployed. Select the **Virtual Volumes** tab to view all associated vVols in the cluster, select a vVol to migrate, and click **Migrate.**

The example shown in the following figure depicts a volume that is named Test-Volume-001 and is created on a two-appliance cluster. In this example, the volume was created on appliance 2 and the wizard is presenting a list of available appliances in the cluster for where the volume can be migrated. When the Migrate Volume wizard appears, the other appliance in the two-appliance cluster is appliance 1, and the administrator can click **Start Migration** to initiate the migration job.
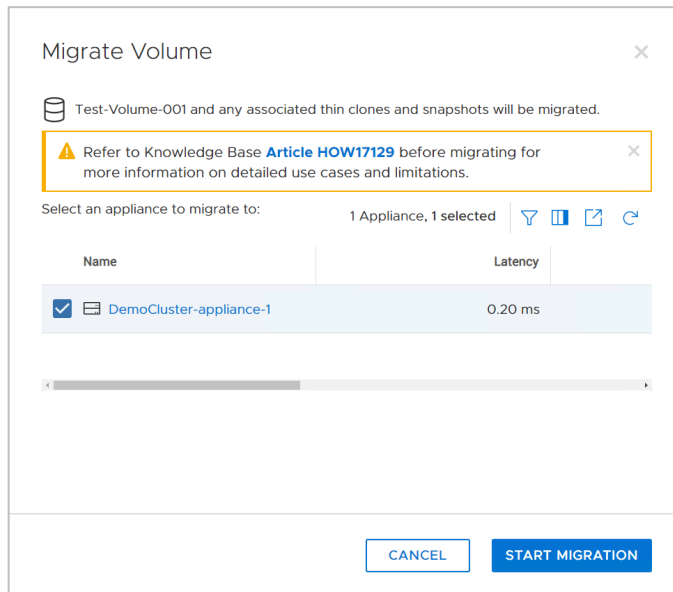

Figure 16.   Start volume migration job

Assisted migration

While the system is running, it periodically monitors storage resource utilization across all appliances within the cluster. Over time, an appliance might approach the maximum usable capacity. In this scenario, the system generates migration recommendations as shown in the following figure. These recommendations are generated based on factors such as drive wear, appliance capacity, and health. An administrator can view assisted-migration recommendations from the alerts or from the **Migration** > **Migration Actions** page in PowerStore Manager. If the administrator accepts a migration recommendation, a migration session is automatically created. Any snapshots or clones that are associated with the original storage resources are migrated over to the new appliance, as with a manual volume migration.
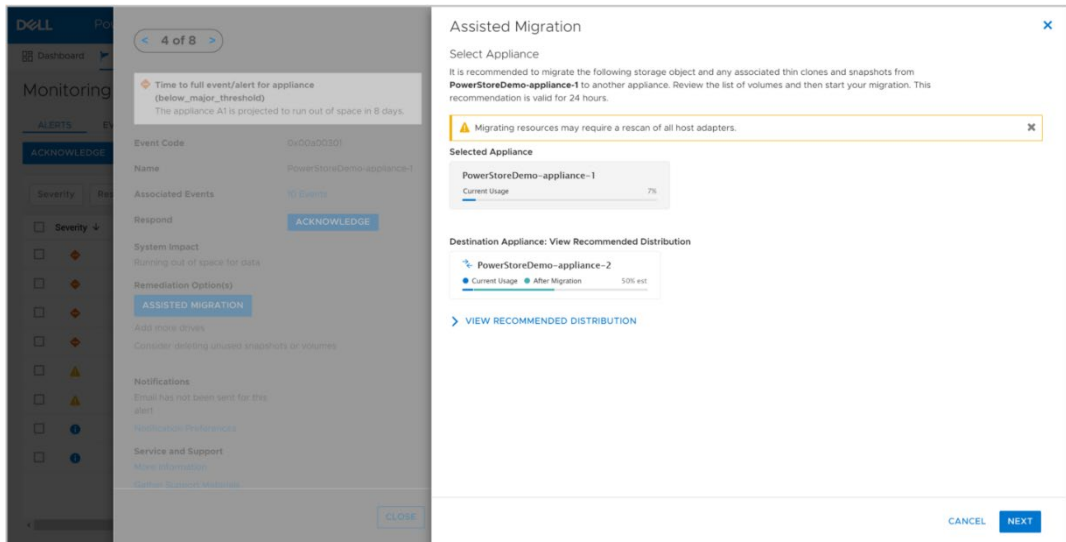
Figure 17.   Assisted migration recommendations

Appliance space evacuation

PowerStoreOS 2.0 introduces the ability to automatically move storage resources such as Volumes, Volume Groups, and vVols from an appliance, all within PowerStore Manager. This easy-to-use, wizard-based workflow can help you prepare an appliance to be powered off or removed from a cluster, or migrate storage resources for an appliance that is full.

To perform this task, go to the **Hardware** page, select an appliance, and click **More Actions** > **Migrate**. This action launches a new workflow as shown in the following figure. The wizard presents a list of storage resources that include Volumes, Volume Groups, and vVols. You can optionally select all storage resources that are shown or select individual storage resources manually. All storage resources might not be available when you run the wizard. Storage resources that are in an active import session or in an active internal migration session are marked as Not Eligible for Migration. After the import sessions or migration sessions are complete, you can rerun the wizard to move the storage resources to other appliances in the wizard.
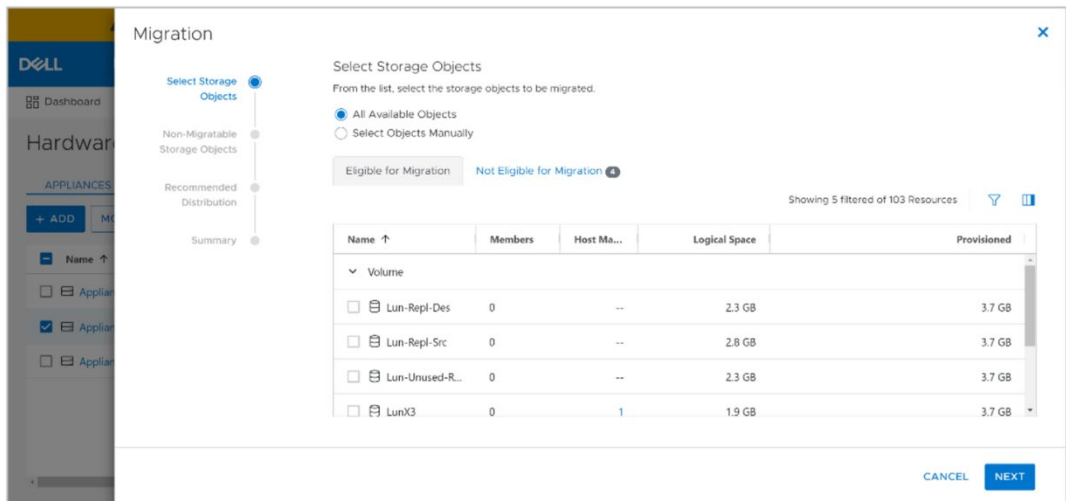

Figure 18.   Migration wizard

When you select storage resources, PowerStore intelligently calculates which appliance is best for the various storage resources. As shown in the following figure, PowerStore recommends that LunX3 be moved to the destination Appliance C. PowerStore makes these recommendations automatically, but there might be scenarios when you would like to place the storage resource on a specific appliance. As shown in the following figure, you can override the destination by clicking the pencil icon next to Appliance C.
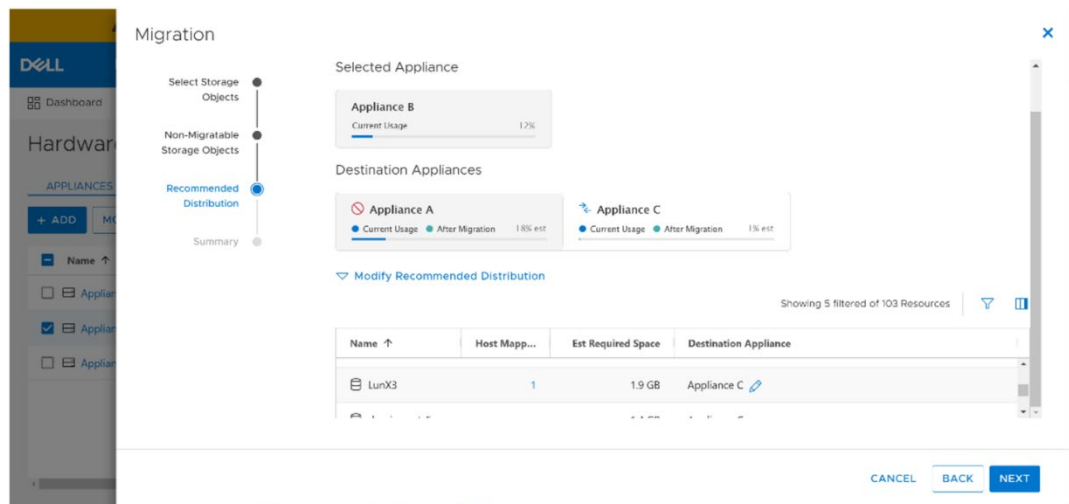

Figure 19.   Storage resource appliance recommendation

At the end of the wizard, PowerStore might provide a list of associated hosts that require a rescan. After the rescan is completed, PowerStore automatically generates the migration sessions and starts the migration jobs.

Migration states

After migration for a storage resource is generated, an administrator can view the various sessions that are active within the cluster. Administrators can pause, resume, or delete any active migration sessions through PowerStore Manager, REST API, or PSTCLI. The following table shows a list of all the possible states for a migration session.

Table 4.   Migration Session States

State	Definition
Initializing	Migration session starts and stays in this state until the session initialization completes
Initialized	Migration session transitions to this state when the session initialization completes
Synchronizing	Background copy is in progress
Idle	Migration session transitions to this state when the initial background copy completes
Cutting_Over	A final portion of the copy is performed in this state, and the ownership of the storage resource is transferred to the new appliance.
Deleting	Migration session is being deleted
Completed	Migration session is completed, and it is safe to delete the session

State	Definition
Pausing	Migration session transitions to this state when the pause command is issued
Paused	Migration session is Paused, and user intervention is required to resume the session
System_Paused	Migration session transitions to this state if it encounters any error, and user may resume or delete the migration after resolving the error
Resuming	Migration session background copy is resumed
Failed	Migration session encountered an error

Capacity forecasting

PowerStore provides details on when a cluster might run out of space. These details help when planning to scale a cluster based on business needs. To inform on these decisions, PowerStore maintains historical statistics about the consumed storage on an appliance to intelligently predict how this storage will be used over time.

To provide the most accurate information, the system collects statistics over 15 days. After 15 days, PowerStore Manager displays a forecast of when a system might run out of space. The following figure shows an example of viewing capacity forecasting in PowerStore Manager.
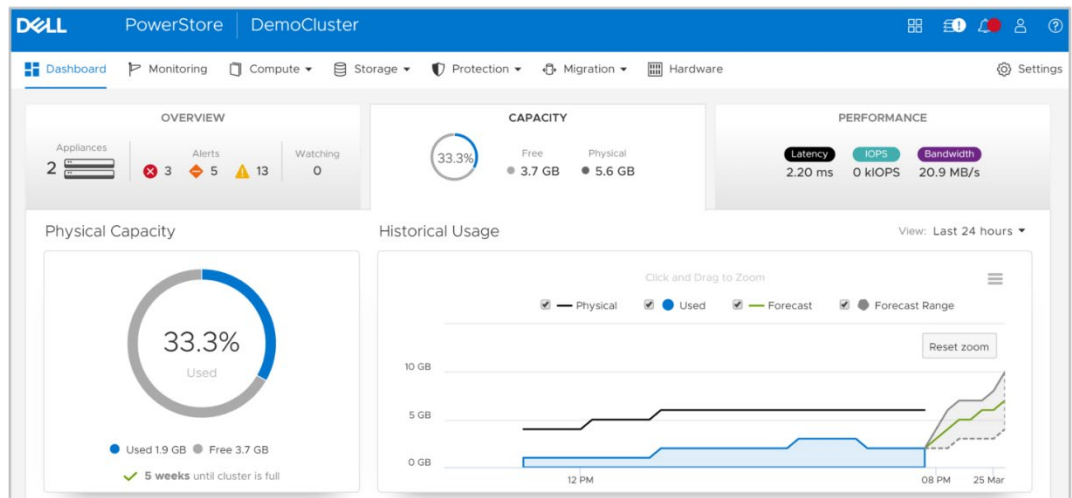

Figure 20.   Capacity forecasting

Dynamic Node Affinity

When mapping volumes to hosts, PowerStore dynamically sets a node for preferred host access, also known as Node Affinity, as shown in the following figure. As more storage resources are created and used over time, there might be situations where one PowerStore node is more heavily used than the peer node. In this scenario, PowerStore might automatically change the Node Affinity of one or more storage resources if certain performance metrics checks are met.

Figure 21. Node Affinity

The first performance check is to compare the read and write IOPs of the source and destination nodes. The second performance check is to ensure that at least one of the nodes has greater than 50% CPU utilization. The last performance check is to see if the system is exceeding a latency check as reported by the data path engine. PowerStore checks these performance metrics every 30 minutes. If all three of these performance checks are met throughout the time duration of 30 minutes, PowerStore dynamically changes the node affinity of the volume. This change is transparent and non-disruptive to the host.

High availability

Introduction

Although PowerStore appliances support multi-appliance configurations to increase redundancy and fault tolerance throughout the cluster, each PowerStore appliance also has two nodes that make up an HA pair. PowerStore features fully redundant hardware and includes several high availability features. These features are designed to withstand component failures within the system itself and in the environment. Examples of these include network or power failures. If an individual component fails, the storage system can remain online and continue to serve data. If the failures occur in separate component sets, the system can also withstand multiple failures. After a failure alert is issued, the failed component can be ordered and replaced without impact. This section reviews PowerStore redundancy and fault tolerance within the platform and the cluster.

Hardware redundancy

PowerStore has a dual-node architecture which includes two identical nodes for redundancy. It features an active/active controller configuration in which both nodes service I/O simultaneously. This feature increases hardware efficiency since there are no requirements for idle-standby hardware. The base enclosure includes these nodes and up to twenty-five 2.5-inch drives.

Figure 22 and Figure 23 show the internal components of PowerStore 1000 to 9000 models, including:

- Power supply
- Battery backup unit (BBU)
- Intel CPU
- Intel QAT Hardware Offload Engine

- Motherboard

- Cooling fans

- 2 x M.2 solid state drives
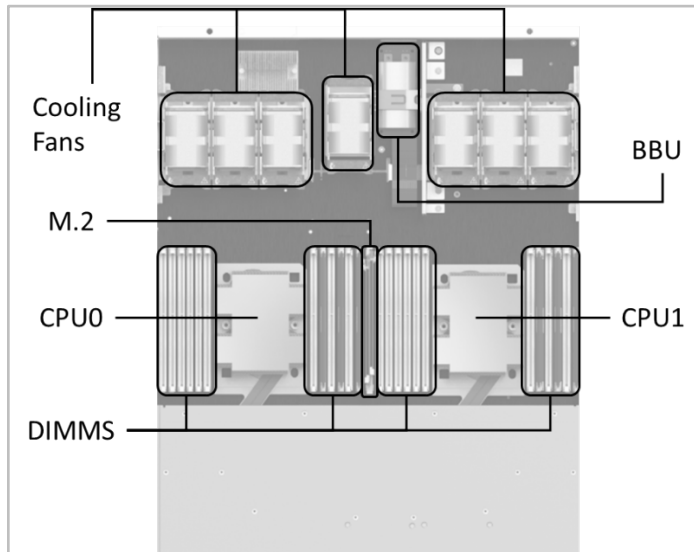
- Memory DIMMs

- I/O modules (optional)

Figure 22.   Internal components of PowerStore 1000 to 9000 nodes (top view)

Figure 23.   Back view of PowerStore 1000 to 9000 nodes

For details about the components of PowerStore 1000 to 9000, PowerStore 1200 to 9200 (available since PowerStoreOS 3.0), and PowerStore 500 appliances, see these documents:

- *PowerStore Hardware Information Guide* on dell.com/powerstoredocs

- *PowerStore Hardware Information Guide for 500T* on dell.com/powerstoredocs

See also the white paper PowerStore: Introduction to the Platform on the PowerStore Info Hub.

Management software

The management software runs PowerStore Manager, the management interface (cluster IP), and other services such as REST API. For PowerStore systems running on versions earlier than PowerStoreOS 3.0, all the management software runs on one node in the appliance at a time. In these previous versions of PowerStoreOS, the node that runs the management software is designated as the primary node in PowerStore Manager. Figure

24 shows an example of a primary node from the hardware properties page, which is highlighted by going to **Hardware** > **Rear View**.

In PowerStoreOS 3.0 and later, the management resources have been balanced and split between both nodes of the appliance to promote better CPU utilization on each of the appliance nodes. In PowerStoreOS 3.0 and later, PowerStore Manager runs on the secondary node while the other management services run on the primary node.
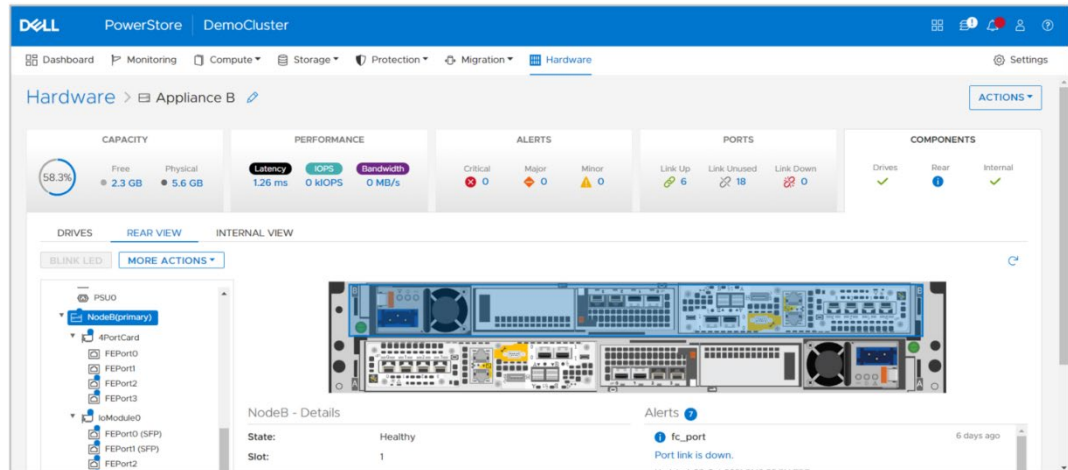

Figure 24. Primary node

If a node reboots, crashes, or the management connection goes down, the services running on that node automatically fail over to the peer node. After a failover, it might take several minutes for all services to start completely. Users that are logged in to PowerStore Manager during the failover might see a message indicating that the connection has been lost. When the failover process completes, access to PowerStore Manager is automatically restored, however users can also manually refresh the browser to regain access. Host access to storage resources is prioritized and available before PowerStore Manager is accessible.

After the failover, the services temporarily run on the new node. When the peer node has recovered from the failover event, the services automatically rebalance after five minutes. During the rebalance, PowerStore Manager might lose connection for up to three minutes while services are brought back up on the recovered node. Note that after failover events, the primary node might change to the peer node and will remain this way until the system is rebooted or failed over again.

System bond

PowerStore systems ensure that there is no single point of failure throughout the system. For a PowerStore T model appliance, we recommend cabling the first two ports of the 4-port card to a top-of-rack switch as shown in the following figure. For the PowerStore T model, these ports are internally referred to as the **system bond**. This interface is created to ensure that there is no single point of failure and data services and production data are always available. The figure shows a list of different types of traffic that flows through the system bond if the data services are leveraged on the PowerStore system:

- Management traffic (for PowerStore X models)
- Cluster communication

- iSCSI or NVMe/TCP host traffic (optional)

- Replication traffic (optional)

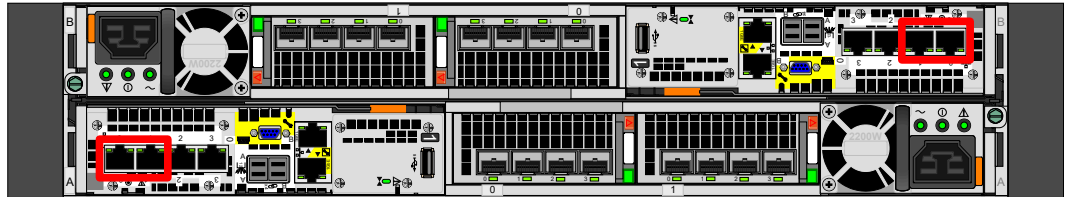- NAS traffic (for PowerStore T models) (optional)

Figure 25.   First two ports of the 4-port card (system bond)

For PowerStore T model appliances, the individual ports within the system bond could be running in active/active mode or active/passive mode. This state depends on whether Link Aggregation Control Protocol (LACP) is configured on the network switches. To ensure best resiliency and network performance, we recommend enabling LACP on the network. To learn more about LACP with PowerStore, see the section Link Aggregation Control Protocol (LACP) in this document. For more information about configuring LACP for a PowerStore T model appliance, see the document *PowerStore Guide for PowerStore T Models* on the dell.com/powerstoredocs.

For PowerStore X model appliances, there is no system bond, and no LACP configuration is required because the link aggregation and failover are handled by the virtual switches in the VMware hypervisor.

PowerStore enables adding more ports for extra bandwidth or increasing fault tolerance. Administrators can extend storage traffic by mapping the extra physical ports, or virtual ports (if applicable) associated with the appliances in the cluster. For a PowerStore T model appliance, administrators can also untag and remove replication traffic from the system bond ports if replication traffic is already tagged to other ports on the 4-port card or I/O modules. For more information about how to add additional ports for host connectivity, see the section Ethernet configuration. For information about how to enable and scale replication traffic, see the white paper PowerStore: Replication Technologies.

Note: NAS interfaces are permanently fixed to the system bond interfaces and cannot be moved or assigned to any other ports on the PowerStore system.

Dynamic Resiliency Engine (DRE)

Enterprise-class storage systems require high levels of reliability and protection from data loss and latent drive failures. Traditional data protection schemes are based on RAID groups of a fixed layout that protect a volumes data. The bandwidth and rebuild speed in this traditional design is limited by the number of drives participating in the group. Also, the speed of the rebuild is limited by the number of tolerated drive failures of that RAID level. For example, a 6+2 could achieve the read speed of 6 drives but only sustain 2 drives worth of rebuild speed in the case of a dual drive failure.

The reliability of the data being protected depends heavily on the bit error rate (BER) of the drive, the amount of data that must be rebuilt, and the number of drives. As capacity and drive counts increase, it becomes more difficult to maintain reliability when traditional RAID protection schemes are used.

Furthermore, economics and reliability are key buying decisions. The cost of a storage solution is driven by the cost of the drives and ultimately, it is price/performance and effective capacity ($/IOP and $/Effective Capacity). As storage and drive capacity grows, the following occurs:

- Performance scales

- Relative cost of the controller diminishes

- The probability of encountering drive failures increases

- Protection and system metadata overhead increases, impacting effective storage capacity

- Higher rebuild speeds are needed to maintain reliability

DRE overview

The PowerStore Dynamic Resiliency Engine (DRE) is a 100% software-based approach to redundancy that is more distributed, automated, and efficient than traditional RAID. It meets RAID 6 and/or RAID 5 parity requirements with superior resiliency and at a lower cost. PowerStore implements proprietary algorithms where every drive is partitioned into multiple virtual chunks and redundancy extents are created by using the chunks across several drives.

It automatically consumes the drives within an appliance and creates appropriate redundancy using all the drives. This process improves overall performance and allows performance to scale as more drives are added to the appliance. Data written to a volume can be spread across any number of drives within an appliance. As new drives are added, the data is automatically rebalanced.
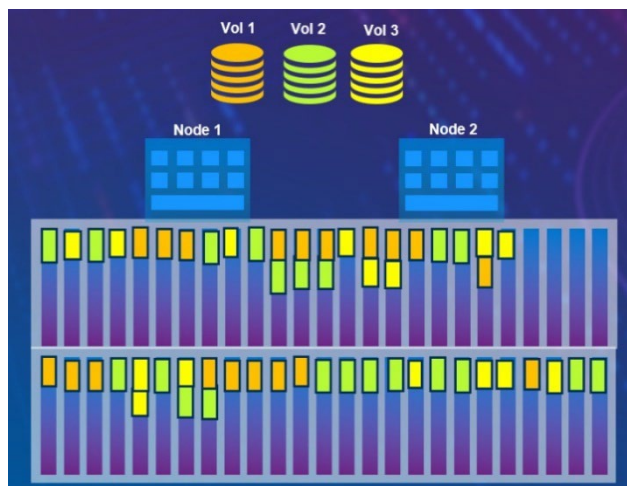

Figure 26.   Data placement in PowerStore

Distributed sparing

Unlike traditional RAID protection strategies, PowerStore does not require dedicated spare drives. Spare space is distributed across the entire appliance, and a small chunk of space is reserved from each drive used for sparing if a drive fails. A single drive's worth of spare space is reserved for every resiliency set in an appliance. Resiliency sets are explained later in this section.

When a drive fails, only the portion of the drive which has data written will be rebuilt. By doing so, the spare capacity is efficiently managed by consuming only the required space. This feature also shortens rebuild time because only data that has written to the drive must be rebuilt.
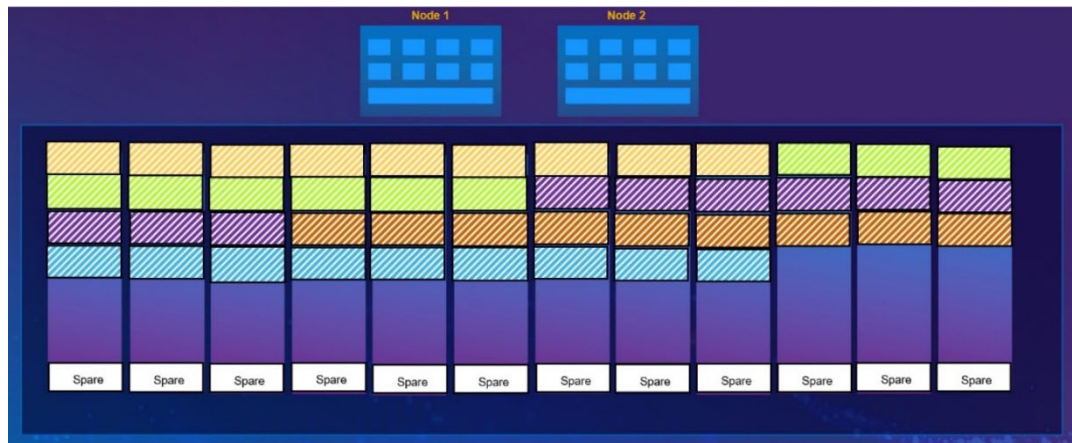

Figure 27.   Distributed sparing

Resiliency sets

PowerStore implements resiliency sets to improve reliability while minimizing spare overhead. Having multiple failure domains (resiliency sets) increases the reliability of the system since it allows the appliance to tolerate a drive failure in each of the sets if the failure occurs simultaneously.
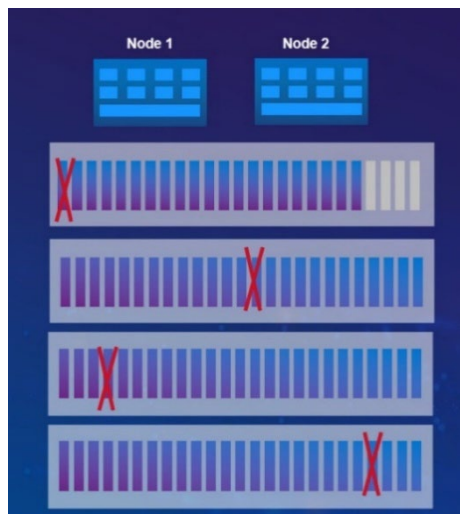

Figure 28.   Tolerance for simultaneous drive failure in multiple resiliency sets

The appliance can tolerate multiple drive failures even within the same resiliency set, if the failure occurs at different instances (a second drive fails after the rebuild on first drive failed drive is complete). Starting with PowerStoreOS 2.0, during the initial configuration of an appliance, you can select a single drive failure tolerance for a 25-drive resiliency set or double drive failure tolerance for a 50-drive resiliency set. If deploying a multi-appliance cluster, you can mix different drive failure tolerance on the appliances. As shown in the following figure, Appliance A could be set to single drive failure tolerance while Appliance B is set to double drive failure tolerance.

Figure 29.   Appliance fault tolerance level

As capacity and drives are added to the system over time, resiliency sets dynamically increase. For example, you might have an appliance that has a single-drive-fault-tolerance set. This means that the system is configured with a 25-drive resiliency set. If you add a 26th drive to the system, the resiliency set dynamically splits into two sets. Furthermore, resiliency sets can span across physical enclosures based on the number of drives in the appliance and have mixed drive sizes.

Resiliency sets offer the following key benefits:

- Enterprise class availability

  - Faster rebuild times with distributed sparing

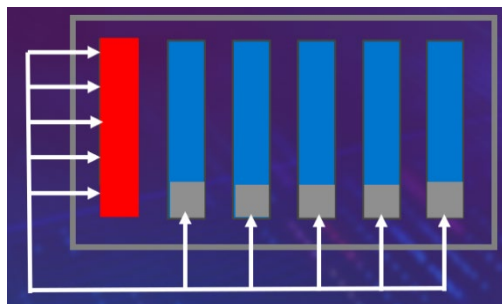 - Rebuild smaller chunks of the drive simultaneously to multiple drives in the appliance

Figure 30.   Parallel rebuild of single drive to distributed spare space

- Intelligent infrastructure

  - Automatically allocate unused user space to replenish spare space to handle multiple failures

    DRE dynamically transfers unused user capacity to replenish spare capacity if there is sufficient unused capacity available on the appliance.

  - Intelligently vary the rebuild speed based on incoming IO traffic while maintaining availability

    PowerStore uses machine-learning algorithm and automatically adjusts the rebuild rate to prioritize host IO when there is a drive failure to optimize performance, while maintaining reliability.

Figure 31. Rebuild data chunks to spare space after drive failure and replenish spare space with unused user space

- Flexible configurations
  - Lower TCO with ability to expand storage by adding single drives

    An appliance can have a minimum of six drives and can scale up to 96 drives. Capacity can be added non-disruptively, giving customer the flexibility to expand their storage by adding one or more drives based on their need.

  - Flexible options to add different drive sizes based on storage need

    PowerStore implements proprietary algorithms to manage drives with different sizes by optimizing the distribution of redundancy extents across multiple drives.

Figure 32. Single drive expansion

Block storage   The PowerStore architecture implements a fully active/active thinly provisioned mapping subsystem that enables IOs from any path to be committed and fully consistent without the need for redirection. The primary advantage of this implementation is to reduce the time to service I/O requests if some paths were to become unavailable. This approach contrasts with competing architectures that use redirection and single node locking that can result in long-trespass times which increases the chances for data unavailability during failover.

PowerStore implements a dual-node, shared-storage architecture that uses Asymmetric Logical Unit Access (ALUA) for SCSI-based host access and Asymmetric Namespace Access (ANA) for NVMe-oF host access. I/O requests received on any path (active/optimized or active/non-optimized) are committed locally and are fully consistent

with its peer node. I/O is normally sent down an active/optimized path. However, in the rare event that all active/optimized paths become unavailable, I/O requests are processed on the local node through the active/non-optimized paths without the need to send any data to the peer node. Then, the I/O is written to the shared NVRAM write cache where the I/O is deduplicated and compressed before being written to the drives. For more information about how the I/O is written to the drives, see the document PowerStore: Data Efficiencies.

ALUA and ANA multipathing ensures high availability. Also, the underlying fully symmetric thin provisioning architecture eliminates the complexity and overhead that is associated with making the volumes available on the surviving path which impacts time to service I/O. To use ALUA and ANA, you must install multipathing software, such as Dell PowerPath, on the host. You should configure multipathing software to use the optimized paths first and only use the non-optimized paths if there are no optimized paths available. If possible, use two separate network interface cards (NICs) or Fibre Channel host bus adapters (HBAs) on the host. This use avoids a single point of failure on the card and the server card slot.

Because the physical ports must always match on both Nodes, the same port numbers are always used for host access in the event of a failover. For example, if 4-port card port 3 on node A is currently used for host access, the same port would be used on node B in the event of a failure. Because of this, connect the same port on both Nodes to the multiple switches for host for multipathing and redundancy purposes.

Ethernet configuration

PowerStoreOS 2.1 introduced support for NVMe/TCP on PowerStore T model appliances, which allows users to configure Ethernet interfaces for iSCSI or NVMe/TCP host connectivity. For PowerStore X model appliances, NVMe/TCP host connectivity is supported with PowerStoreOS 2.1.1 and later. You can deploy Ethernet interfaces in mirrored pairs to both PowerStore nodes since these interfaces do not fail over. This configuration ensures that the host has continuous access to block-level storage resources if one node becomes unavailable. For PowerStore T model appliances, storage networks are configured after the cluster is created. For a robust HA environment, create additional interfaces on other ports after the cluster has been created.

For PowerStore X model appliances, iSCSI or NVMe/TCP is enabled during ICW. Enter the IP address information about the **Storage Network** page as shown in Figure 33. These IP addresses are assigned to the first two ports on the 4-port card on node A and node B.
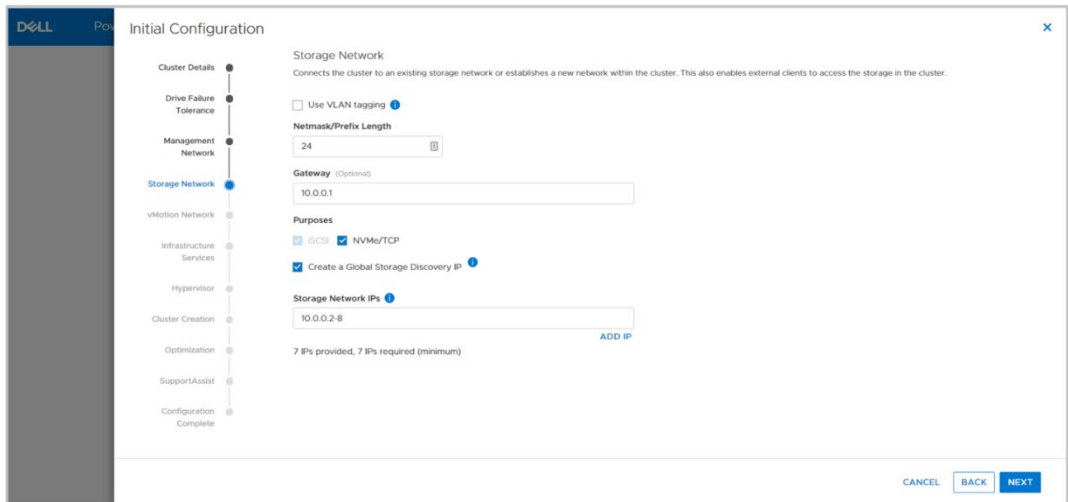
Figure 33. Storage network for PowerStore X model appliances

To remove a single point of failure at the host and switch level, verify that the PowerStore system has the first two ports of the 4-port card cabled to switches, as shown in the following figure.
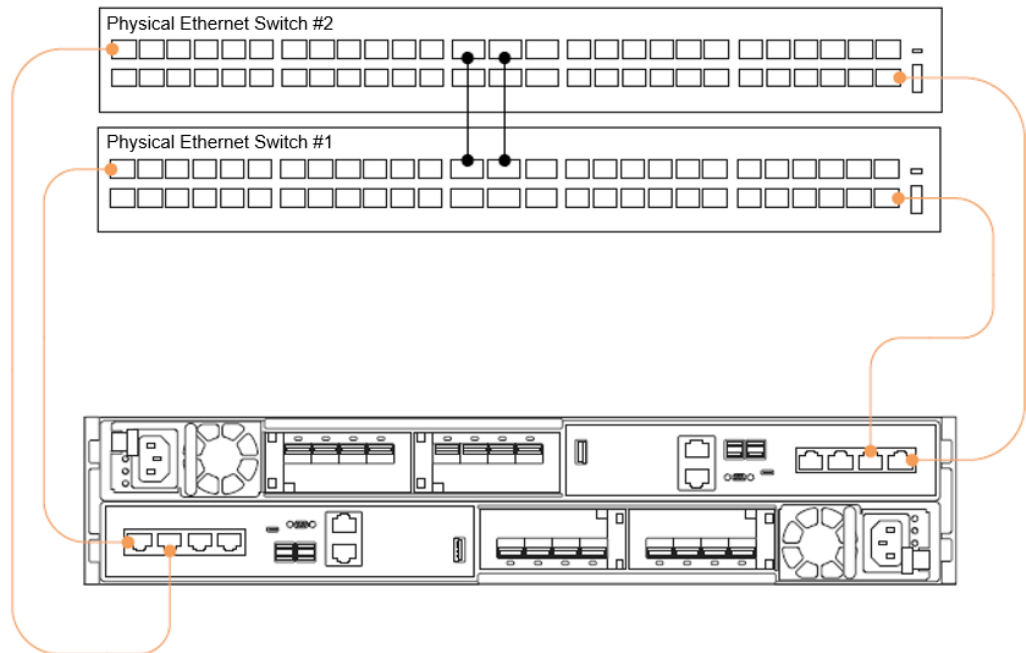

Figure 34. First two ports of the 4-port card cabled to top-of-rack switches

After you configure an appliance, enable other ports for extra bandwidth, throughput, dedicated host connectivity, or dedicated replication traffic. For these ports, cable other ports that are available on the 4-port card or on the I/O modules if available. Because the physical ports match on both nodes, ensure that the ports are cabled to multiple switches for host multipathing and redundancy purposes. For more information about how to enable additional interfaces for replication traffic, see the document PowerStore: Replication Technologies.

After completing cabling, add IP addresses in PowerStore Manager. Figure 35 shows the first step of making sure that there are unused IP addresses configured on the storage network. If there are no other IPs available, click the **Add** button to supply more IP addresses. The following figure shows the next step on the Ports page for mapping extra ports for host connectivity. The Ports page varies depending on the model type and is accessed as follows.

- PowerStore T models: **Hardware** > **Appliance Details** > **Ports**

- PowerStore X models: **Hardware** > **Appliance Details** > **Ports** > **Virtual**

These ports obtain the next-available unused IP address from the storage network and assign them to the newly configured port. As seen in this example, newly configured ports that are assigned for host traffic are enabled in mirrored pairs. If only one port is selected, there is a notification indicating that the associated port on the peer node is also enabled for host traffic. For more information about the virtual ports for a PowerStore X model appliance, see the document PowerStore: Virtualization Integration.

Network IPs

Manage the networks for the cluster. These networks have various responsibilities to support storage operations.

| MANAGEMENT | STORAGE | VMOTION | ICM | ICD |

Details

Connects the cluster to management services such as DNS, NTP and vCenter. This also enables external clients to access the cluster.

VLAN ID	10
Netmask/Prefix Length	255.255.255.0
Gateway	10.0.0.1
Network MTU Size	1500 bytes

RECONFIGURE   RECONFIGURE TO IPV6

Management IPs

+ ADD   MORE ACTIONS ▾

	IP Address ↑	Purpose	Appliance	Node
☐	10.0.0.10	Management Node	Appliance C	Node A
☐	10.0.0.11	Management Node	Appliance C	Node B
☐	10.0.0.12	Unused	--	--
☐	10.0.0.13	Unused	--	--
☐	10.0.0.14	Unused	--	--
☐	10.0.0.15	Unused	--	--
☐	10.0.0.16	Unused	--	--

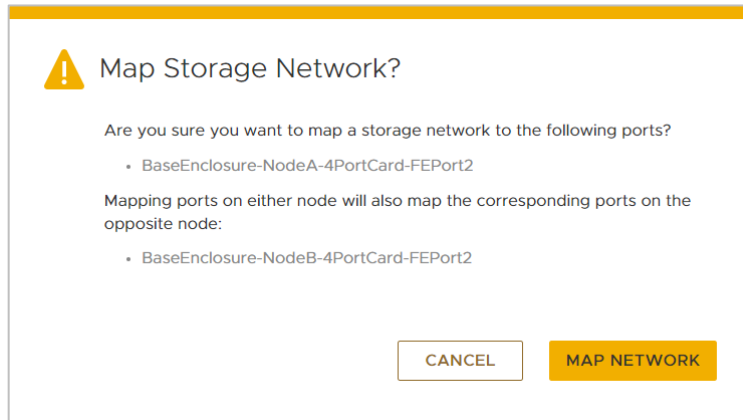**Figure 35.   Verify unused storage network IP addresses**

Figure 36.   Map Storage Network

Fibre Channel configuration

To achieve high availability with Fibre Channel (FC) and NVMe/FC, configure at least one connection to each node. This practice enables hosts to have continuous access to block-level storage resources if one node becomes unavailable.

With FC and NVMe/FC, you must configure zoning on the switch to allow communication between the host and the PowerStore appliance. If there are multiple appliances in the cluster that are using FC, ensure that each PowerStore appliance has zones that are configured to enable host communication. Create a zone for each host HBA port to each node FC port on each appliance in the cluster. For a robust HA environment, you can zone more FC ports to provide more paths to the PowerStore appliance.

There are two locations where you can locate the port World Wide Names (WWNs) in PowerStore Manager. The first location is in the **Hardware** > **Cluster FE Ports** page, as shown in the following figure. The second location is in the **Components** > **Rear View** page, as shown in Figure 38. For more details about configuring hosts, see the document *PowerStore Host Configuration Guide* on the dell.com/powerstoredocs.
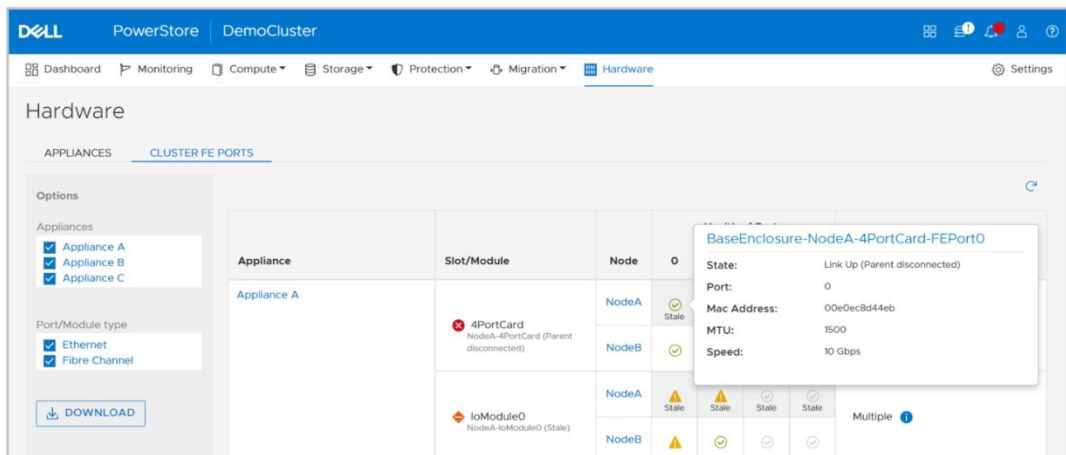

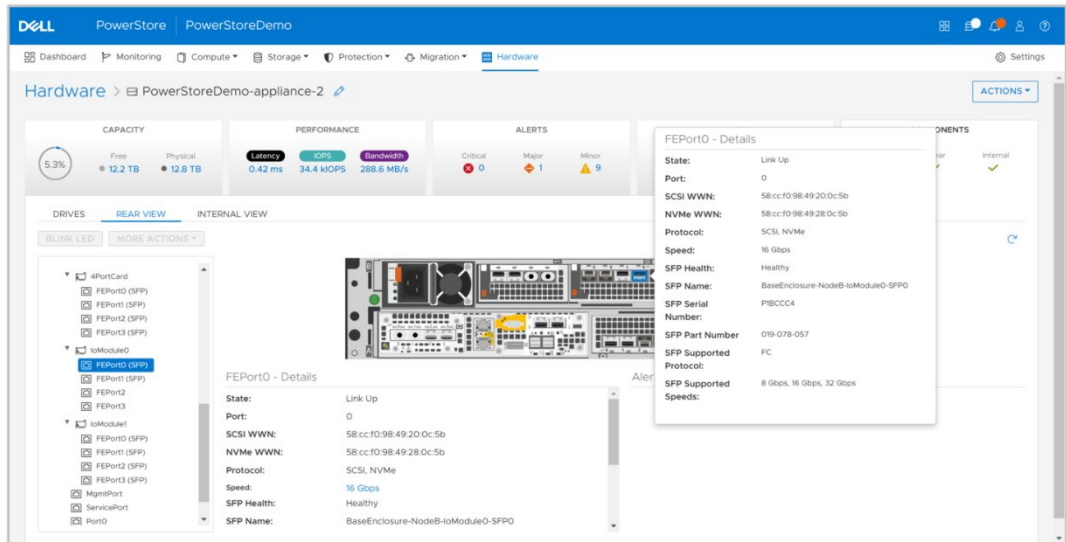**Figure 37.   PowerStore Manager hardware cluster front-end ports page**

Figure 38.   PowerStore Manager components rear view

Block example

When designing a highly available infrastructure, components that connect to the storage system must also be redundant. This design includes removing single points of failure at the host and switch level to avoid data unavailability due to connectivity issues. The following figure shows an example of a highly available configuration for a PowerStore T model system, which has no single point of failure. See the documents *PowerStore Network Planning Guide* and *PowerStore Network Configuration for PowerSwitch Series Guide* on the dell.com/powerstoredocs for detailed information about cabling and configuration of the network infrastructure.
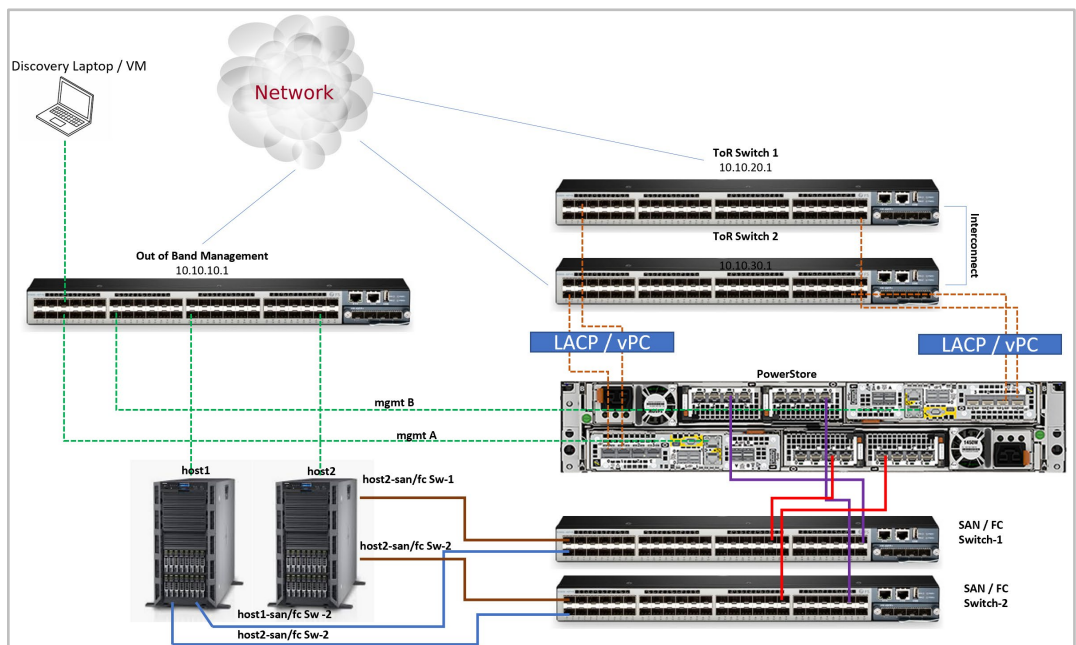

Figure 39.   Highly available block configuration

File storage

Before file-level resources are shared from a PowerStore T model system, NAS server interfaces must be configured using either the default system bond on the first two ports of the 4-port card or a user-defined Link Aggregation Group. User-defined link aggregation Groups are available since PowerStoreOS 3.0. For more information about link aggregation, see the section Link Aggregation Control Protocol (LACP).

An administrator can create a NAS server that holds the configuration information for SMB, NFS, FTP, or SFTP access to the file systems. New NAS servers are automatically assigned on a round-robin basis across the available nodes. The preferred node acts as a marker to indicate the node on which the NAS server should be running, based on this algorithm. After it is provisioned, the preferred node for a NAS server never changes. The current node indicates the node on which the NAS server is running. Changing the current node moves the NAS server to a different node, which can be used for loading balancing purposes. When you move a NAS server to a new node, all file systems on the NAS server move along with it.

The following figure shows the current and preferred node columns in PowerStore Manager.

Figure 40.   Current Node and Preferred Node

During a PowerStore node failure, the NAS servers automatically fail over to the surviving node. This process generally completes within 30 seconds to avoid host timeouts. After the failed node is recovered, a manual process is required to fail back the NAS servers and return to a balanced configuration.

NAS servers are automatically moved to the peer node and back during the upgrade process. After the upgrade is completed, the NAS servers return to the node that they were assigned to at the beginning of the upgrade. For more information about NAS Servers, see the document PowerStore: File Capabilities.

Link Aggregation Control Protocol (LACP)

Link loss can be caused by many environmental factors, such as cable or switch port failure. Configure high availability on the ports to protect against these types of failure scenarios.

Link aggregation combines multiple network connections into one logical link. This provides increased throughput by distributing traffic across multiple connections, and provides redundancy in case one connection fails. If connection loss is detected, the link is immediately disabled, and traffic is automatically moved to the surviving links in the aggregation to avoid disruption. The switch should be properly configured to add the ports back to the aggregation when the connection is restored. Although link aggregations provide more overall bandwidth, each individual client still runs through a single port. Dell

PowerStore systems use the Link Aggregation Control Protocol (LACP) IEEE 802.3ad standard.

NAS Servers include one or more network interfaces that are created on the Ethernet ports for client access. On PowerStoreOS versions prior to 3.0, LACP is only available through the System Bond ports, which are ports 0 and 1 on the 4-port card of the embedded module. In PowerStoreOS 3.0 and later, link aggregations can be configured on PowerStore T systems with two to four ports from different I/O Modules and also between I/O Modules and Ethernet ports on the embedded module. All ports within the aggregation must have the same speed, duplex settings, and MTU size. Link aggregations are created from the Hardware > Appliance > PORTS page in PowerStore Manager (Figure 41).
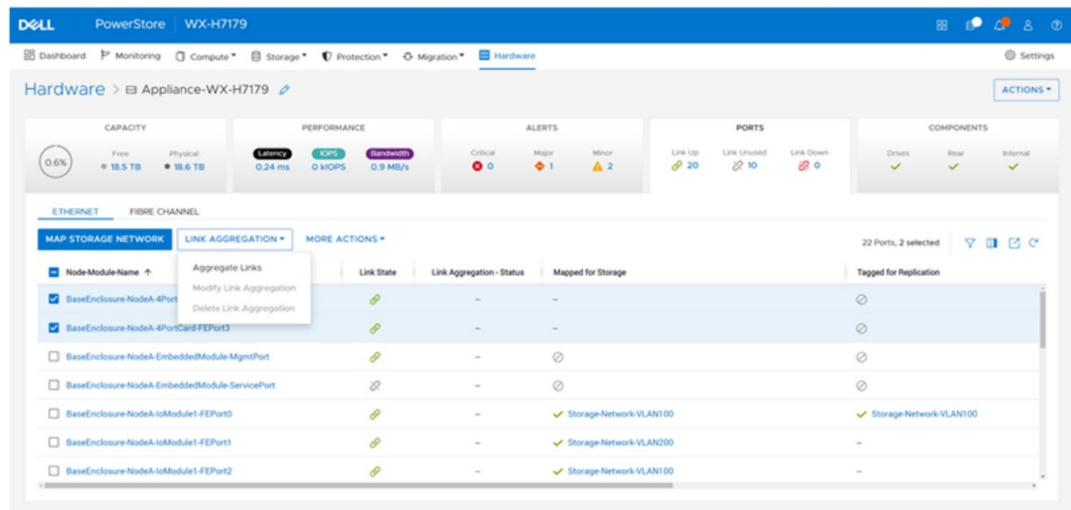

Figure 41.   Creating Link Aggregation

Link aggregation can be used for NAS Servers and File Import interfaces. PowerStore supports the sharing of protocols on the same interfaces, so it is possible to have Link aggregation across multiple ports while also using the same individual ports for iSCSI, Replication, and Import traffic. Note that you cannot map iSCSI traffic to a link aggregation directly because multipathing is used for block access. As a best practice, it is recommended to keep link aggregation traffic on separate ports from block traffic, where possible, to maintain optimal bandwidth and performance.

When configuring link aggregation, ensure that the same ports are cabled on both nodes of the system. This is necessary because in case of failover, the peer node uses the same ports. Also, ensure that the appropriate switch ports connected to the nodes are configured for link aggregation. If the switch is not properly configured or if the cabling does not match, communication issues might occur.

A single link aggregation can be used for PowerStore ports connecting to the same switch or different switches. For switches that are stacked, link aggregation provides multiple paths to multiple switches for redundancy purposes. Figure 42 shows a cabling diagram for the default System Bond LACP configuration on the first two ports of the system (4-port card ports 0 and 1). Each port on node A connects to a different Dell switch, and the configuration is mirrored on node B. For switches that are stacked, VLT or an equivalent technology for a different switch vendor must be configured. A link aggregation group is

then configured on the blue paths in the example below, and separately for the green paths.
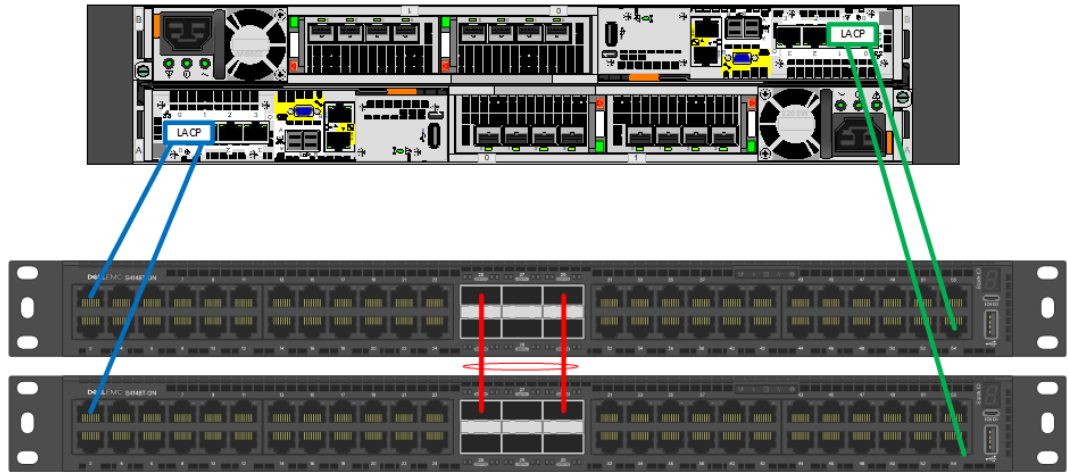

Figure 42. Link aggregation example 1

For networks that contain switches that are not interconnected, a single link aggregation can be used to provide redundancy to one network switch, while a second aggregation can be used for redundancy to another. In Figure 43, the first link aggregation is created on ports 0 and 1 on the 4-port card, shown in blue, and connects to the top switch. A second link aggregation is created on ports 2 and 3 of the 4-port card, shown in green, and connects to the bottom switch. A link aggregation group then needs to be configured for each of the four groups of cables. On node A, a link aggregation group would be configured on the blue cables to the top switch, and separately on the green paths to the bottom switch. This is then repeated for the ports on node B.
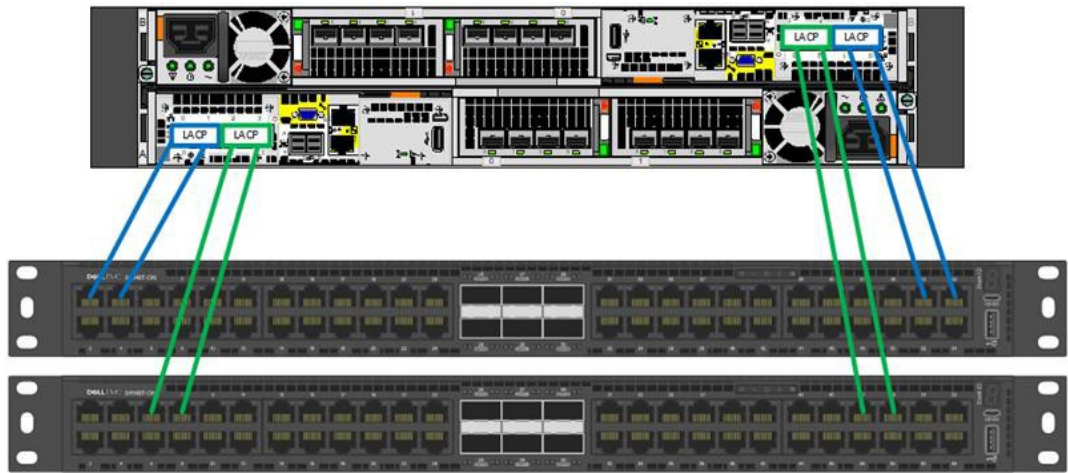

Figure 43. Link aggregation example 2

When replicating from one system to another, it is recommended to configure link aggregations the same way on both systems. If a link aggregation with the same name is not found on the destination system, the interfaces on the destination NAS Server are created without a port assignment and the user must manually assign a port in this case. Alternatively, if a nonmatching configuration is wanted, you can override the interfaces on the destination NAS Server to assign them to a valid port. Otherwise, data access becomes unavailable in the event of a failover.

Link aggregation should also be configured at the host level/client level to provide resiliency against port or cable failures. Depending on the vendor, this might also be referred to as trunking, bonding, or NIC teaming. See the vendor's documentation for more information.

VMware integration

The following subsections describe some HA capabilities with the PowerStore X model. For more details about PowerStore integration with VMware, see the document [PowerStore: Virtualization Integration](#).

Controller VMs

The PowerStore X model appliance is fully redundant with nodes that are in the form of controller VMs. Each node has a controller VM that provides data services. Each of the controller VMs is deployed into private datastores that represent the physical M.2 device that is inside the node. Because these datastores are not shared, if a node fails, the controller VM does not restart on any other host in the cluster. However, since PowerStore is fully redundant, the peer node that is running the other controller VM continues to service I/O to the appliance. The two controller VMs act as the HA pair for the appliance. The controller VM should not be moved and should not have any settings changed.

vSphere HA

When administrators are using virtual machines on either the PowerStore T or PowerStore X model appliances, we recommend enabling vSphere HA on the cluster. If a node or host in the cluster loses power, virtual machines restart on a surviving host or node in the cluster.

vCenter connection

Both PowerStore T and PowerStore X model appliances enable integrating with VMware environments by being managed and monitored from a vCenter. For PowerStore T model appliances, a vCenter connection is optional after the cluster has been set up. For PowerStore X model appliances, a vCenter connection is required during the initial configuration wizard. After a cluster has been deployed, there could be certain situations where a vCenter connection is lost to the PowerStore appliance. In this scenario, management tasks through vCenter might be temporarily unavailable. Note that I/O continues to be processed for all storage objects in the PowerStore appliance.

Replication

To protect against outages at a system or data-center level, we recommend using replication to a remote site. This use case includes planned maintenance events, unplanned power outages, or natural disasters. PowerStore supports replication to simplify disaster recovery for block resources. Starting in PowerStoreOS 3.0, file resources are also supported for replication. In PowerStoreOS 3.0, PowerStore appliances support native metro volume replication, which provides synchronous replication of spanned block storage volumes across two PowerStore clusters in metro distance for VMFS datastores. For details on how to set up and use the replication features on PowerStore, see the white papers [Dell PowerStore: Replication Technologies](#) and [Dell PowerStore: Metro Volume](#) on the [Dell Technologies Info Hub](#).

Platform high availability

On PowerStore, each storage resource is assigned to either node A or node B for load-balancing and redundancy purposes. Besides storage-resource assignments, each node has various containers running on them that make up the PowerStore operating system. If one node becomes unavailable, its resources (storage and containers) automatically fail over to the surviving node.

The time that it takes for the failover process to complete depends on several factors such as system utilization and the number of resources. The peer node assumes ownership of the resources and continues servicing I/O to avoid an extended outage. Failovers occur if the following occurs on a node:

- Node reboot: The system or a user rebooted the node.

- Hardware or software fault: The node has failed and must be replaced.

- Service mode: The system or a user placed the node into service mode. This occurs automatically when the node is unable to boot due to a hardware or software issue.

- Powered off: A user powered off the node.

Note: Manually putting a node into service mode is only available through a service script. It is not available from the PowerStore Manager, REST API, or PSTCLI.

While the node is unavailable, all the resources of the node are serviced by the peer. After the node is brought back online or the fault is corrected, block-storage resources automatically fail back to the proper node owner. File-storage resources must be failed back manually.

During a code upgrade, both nodes reboot in a coordinated manner. All resources on the rebooting node are failed over to the peer node. When the peer comes back online, the resources are failed back to their original owner. This process repeats for the second node. Users can run a pre-upgrade health check before starting a code upgrade to ensure a smooth upgrade process.

Cluster high availability

Every PowerStore appliance uses the pacemaker stack that is used for cluster resource management and making sure that the cluster has a quorum. The pacemaker is the cluster brain; it processes and reacts to cluster events such as appliances being added or removed from the cluster, or resource events that are caused by failures. During an appliance failure, management services are still serviced if there is a quorum.

Cluster quorum

A quorum is defined as N/2+1 appliances being in active communication. If there is no quorum, management operations are temporarily lost, but data continues to be serviced if available. The following figure shows an example of a two-appliance cluster that is servicing I/O to a host with Fibre Channel connectivity but has temporarily lost management access. In this example, there is no quorum between Appliance 1 and Appliance 2. Because the appliances are servicing I/O to the host with a Fibre Channel connectivity, there is no impact to the host. However, since quorum is lost, some management services are temporarily suspended until quorum is restored:

- Access to PowerStore Manager

- Running scheduled snapshots
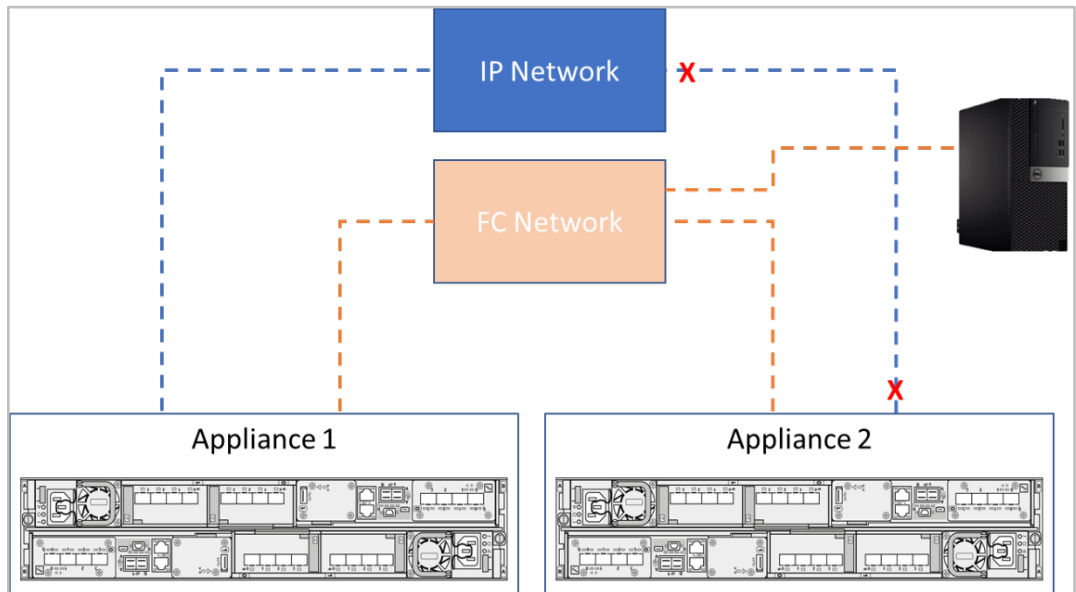
- Syncing replication sessions

Figure 44.   Two-appliance cluster with no quorum

The following figure shows a three-appliance cluster where Appliance 3 crashes. In this scenario, quorum is met because two appliances are still in active communication. I/O remains accessible from Appliance 1 and Appliance 2. However, since Appliance 3 failed, I/O is inaccessible for that appliance.
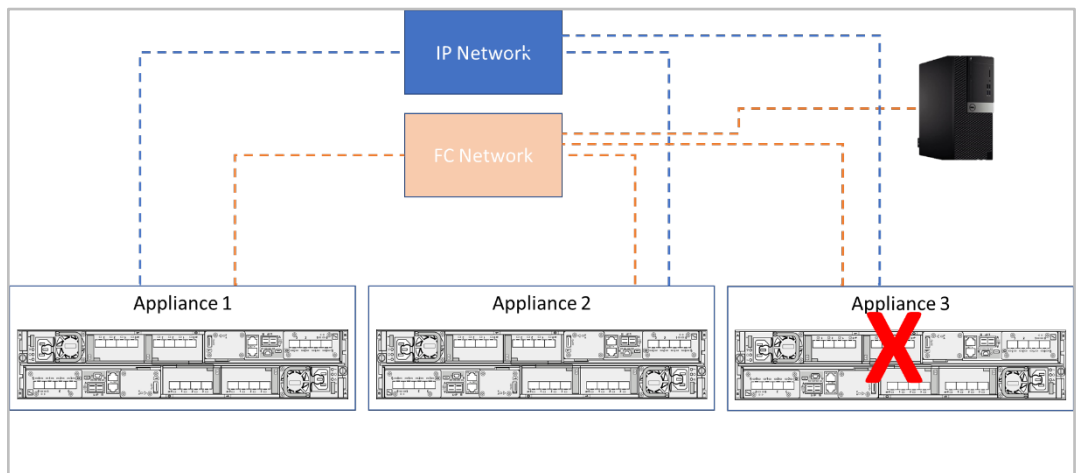

Figure 45.   Three-appliance cluster with quorum

Cluster IP

The cluster IP address is a required field that you can set during the initial configuration of the cluster. This cluster is a highly available IP address that is used to access PowerStore Manager. The cluster IP address is typically on the primary node of the primary appliance. In the rare event that the primary appliance experiences a dual-node failure, the cluster IP address fails over to another appliance in the cluster if there is quorum. When this cluster

IP address fails over, it might take several minutes to regain access to the PowerStore Manager.

Global storage IP

The Global Storage IP (GSIP) address is an optional field that can be set during or after the initial configuration of the cluster. This IP address is a global, floating storage-discovery IP. iSCSI or NVMe/TCP hosts only require one GSIP and can discover all the storage paths for all the appliances in the cluster. Otherwise, iSCSI or NVMe/TCP hosts require a list of storage IPs so that if one IP is down, the host can try the next IP.

Conclusion

Designing an infrastructure with high levels of availability ensures continuous access to business-critical data. If data becomes unavailable, day-to-day operations are impacted, which could lead to loss of productivity and revenue. PowerStore systems are designed with full redundancy across all components at both the hardware and software level. These features enable the system to be designed for 99.9999% availability.[2] By using the clustering and high availability features in PowerStore, organizations can minimize the risk of data unavailability.

References

The [Dell Technologies Info Hub](#) > [Storage](#) site provides expertise that helps to ensure customer success with Dell storage platforms.

[Dell.com/powerstoredocs](#) provides detailed documentation about how to install, configure, and manage PowerStore systems.

---

[2] Based on the Dell Technologies specification for PowerStore, April 2020; actual system availability might vary.