Technical White Paper

Dell EMC PowerStore: Clustering and High Availability

Abstract

This white paper provides an overview of Dell EMC™ PowerStore™ clustering technology along with the highly available features of the platform.

September 2020

H18157

Revisions

Date	Description
April 2020	Initial release: PowerStoreOS 1.0
September 2020	PowerStoreOS 1.0.2, updates to high availability section, and other minor updates

Acknowledgments

Author: Jonathan Tang

Table of contents

Executive summary

This white paper discusses the redundant hardware and high-availability features that are available on Dell EMC™ PowerStore™. 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, 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 may evaluate, acquire, manage, operate, or design a Dell EMC networked storage environment using PowerStore systems.

1 Introduction

Having constant access to data is a critical component in any modern business. If data becomes inaccessible, business operations may be impacted 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, which is designed for 99.9999% availability.[1]

1.1 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. PowerStore 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 also 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.

1.2 Terminology

The following terms that are used with PowerStore:

Appliance: Term used for solution containing a base enclosure and any attached expansion enclosures. The size of an appliance could be only the base enclosure or the base enclosure plus expansion enclosures.

Base enclosure: Used to reference the enclosure containing both Nodes (Node A and Node B) and 25x NVMe drive slots.

Cluster: Multiple appliances in a single grouping. Clusters can consist of one appliance or more. Only PowerStore T model appliances are expandable by adding more appliances (up to 4 in total).

Expansion enclosure: Enclosures that can be attached to a base enclosure to provide additional storage in the form of SAS drives.

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.

Intracluster Management (ICM) Network: An internal management network that provides continuous management connectivity between appliances in the PowerStore cluster.

Intracluster Data (ICD) Network: An internal network that provides continuous storage connectivity between appliances in the PowerStore cluster.

Network-attached storage (NAS) server: A file-level storage server used to host file systems. A NAS server is required in order to create file systems that use SMB or NFS shares, and VMware NFS datastores and VMware® vSphere® Virtual Volumes™ (vVols) (file).

Network File System (NFS): An access protocol that allows data access from Linux®/UNIX® hosts on a network.

Node: A storage node that provides the processing resources for performing storage operations and servicing I/O between storage and hosts.

PowerStore Manager: The web-based user interface (UI) for storage management.

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: 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.

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.

Virtual Volumes (vVols): A VMware storage framework which allows VM data to be stored on individual Virtual Volumes. This allows for data services to be applied at a VM-granularity level while using Storage Policy Based Management (SPBM).

vSphere API for Array Integration (VAAI): A VMware API that allows storage-related tasks to be offloaded to the storage system.

vSphere API for Storage Awareness (VASA): A VMware API that provides additional insight about the storage capabilities in vSphere.

Volume: A block-level storage device that can be shared out using a protocol such as iSCSI or Fibre Channel.

2 Clustering

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 can be configured into the cluster. When a multi-appliance cluster is deployed, this task can be performed during the initial configuration process or appliances can be added to an existing cluster. PowerStore clusters can be scaled down by removing appliances from an existing cluster.

Note: Clustering is only supported on PowerStore T models.

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 in Figure 1, users can scale out appliances with different model numbers to create a four-appliance cluster. In addition to the scale-out benefit, each appliance can scale-up with different numbers of expansion enclosures. For example, each appliance that is shown in Figure 1 has 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 give customers the ability to grow their clusters with no dependence on the model number, drive count, or even drive type.

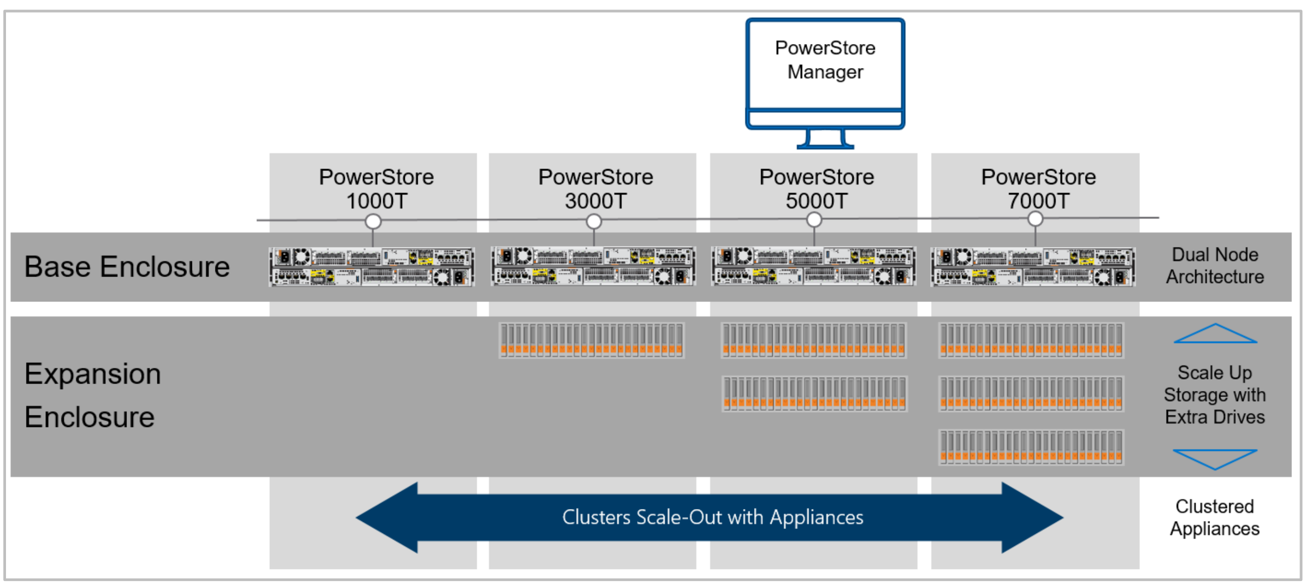

Figure 1 Multi-appliance cluster

2.1 Requirements

Ensure that the PowerStore systems are cabled correctly by referencing the *Dell EMC PowerStore Quick Start Guide* on the PowerStore Info Hub. Each node in every appliance that is a part of a PowerStore cluster must communicate to other nodes through the bonded ports (see Figure 2). The network that allows the nodes to communicate to each other is an internal network named the Intracluster Management (ICM) and Intracluster Data (ICD) networks. In PowerStore, the ICM and ICD networks communicates through the 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.

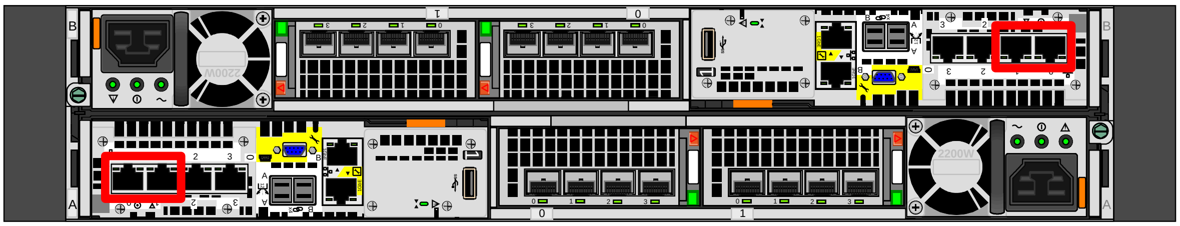

Figure 2 Back view of PowerStore appliance

2.2 Primary appliance

Each cluster has a primary appliance which is the first appliance that is selected during the initial configuration. A primary appliance can be configured in unified mode or block-optimized mode during the initial configuration process.

Figure 3 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 or FC, ensure that the **Storage Configuration** has **Block Optimized** mode selected.

Note: An appliance cannot change its configuration type (Unified or Block Optimized) after the cluster has been configured.

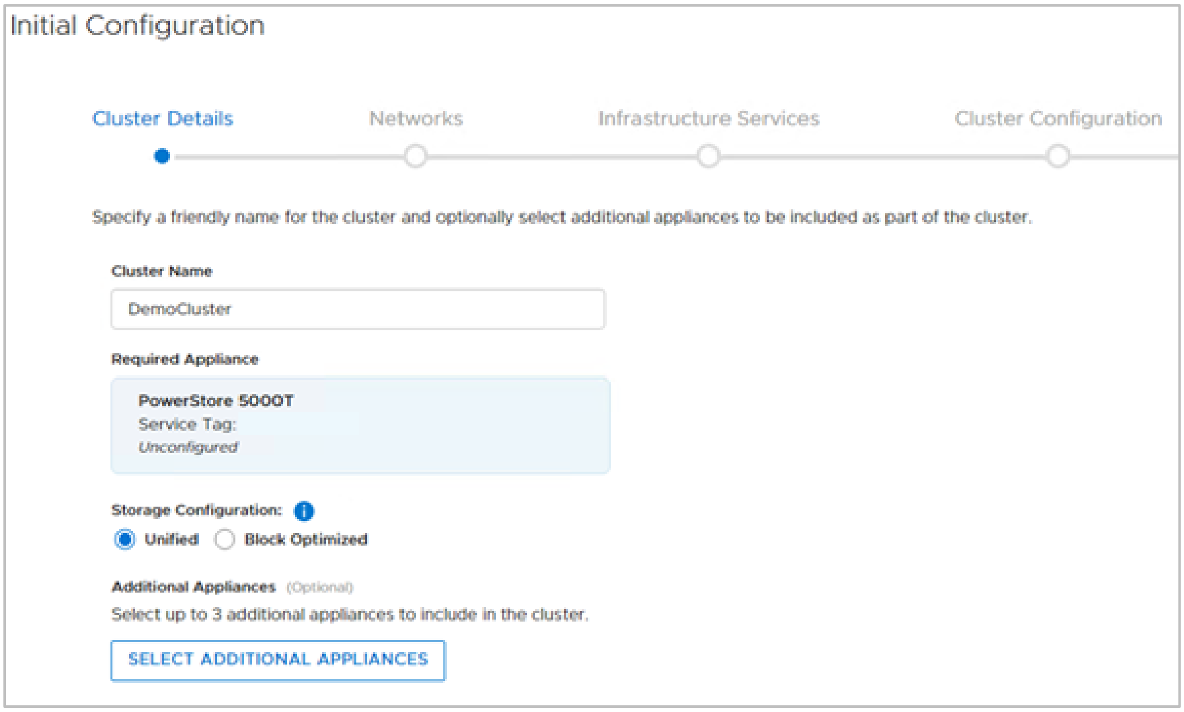

Figure 3 Unified storage configuration

2.3 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 T model appliances. 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.

Figure 4 shows an example of a selecting a multi-appliance cluster. There are no PowerStore X model appliances available for selection. For more information about the discovery process, see the document *Dell EMC PowerStore: Introduction to the Platform* on Dell.com/StorageResources.

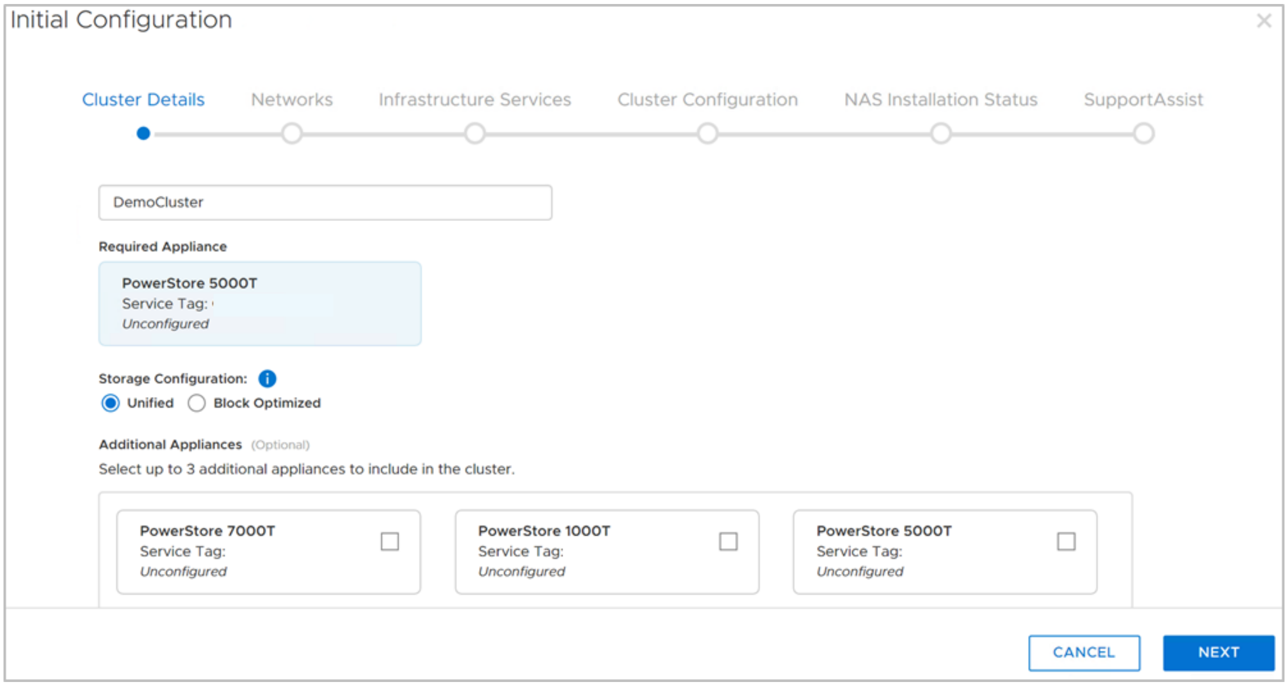

Figure 4 Multi-appliance selection

Below are some general guidelines to consider when creating a multi-appliance 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.
- 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.

2.3.1 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, it is recommended to deploy a unified cluster to provide the most flexibility in the future. As mentioned in section 2.2, file 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.

Figure 5 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.

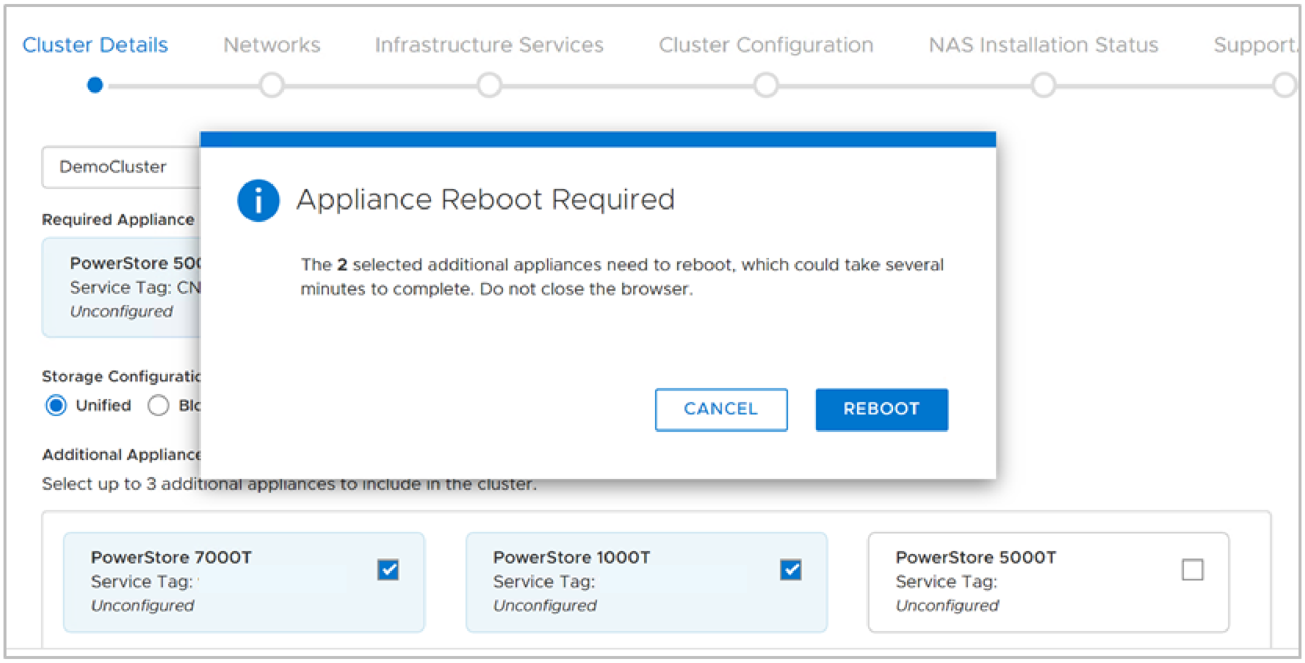

Figure 5 Reboot non-primary appliances to block optimized configuration

2.3.2 Block-optimized cluster

If file services are not used, a cluster can be configured to be block optimized to deliver the highest possible block performance. Figure 6 shows an example of a multi-appliance cluster that will be configured as block optimized. The primary appliance is now 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: A cluster cannot change configuration type (unified or block optimized) after the cluster has been configured.

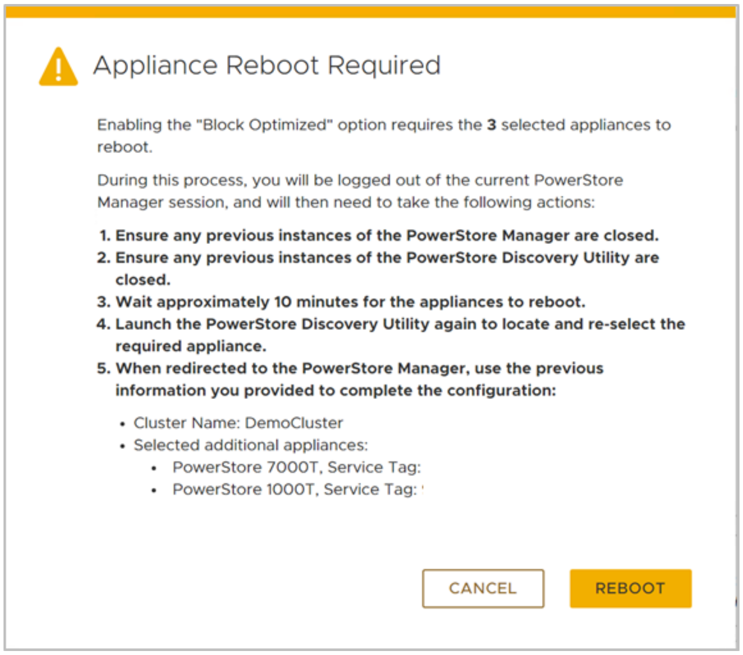

Figure 6 Reboot all appliances in a block-optimized cluster

2.4 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. Adding an appliance is only supported using the PowerStore Manager. There is no REST API or PSTCLI support at this time.

General guidelines to consider when adding appliances to an existing cluster:

- Final cluster size cannot exceed four appliances
- Can only add uninitialized appliances
- Can only add one at a time
- New appliance as well as the existing cluster must be in a healthy state
- Extra appliances are automatically rebooted into a block-optimized configuration when added to a cluster
- 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

2.4.1 Selecting the appliance

PowerStore Manager simplifies adding an appliance into an existing cluster. Figure 7 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. Once an appliance is selected, the appliance reboots into the block-optimized configuration.

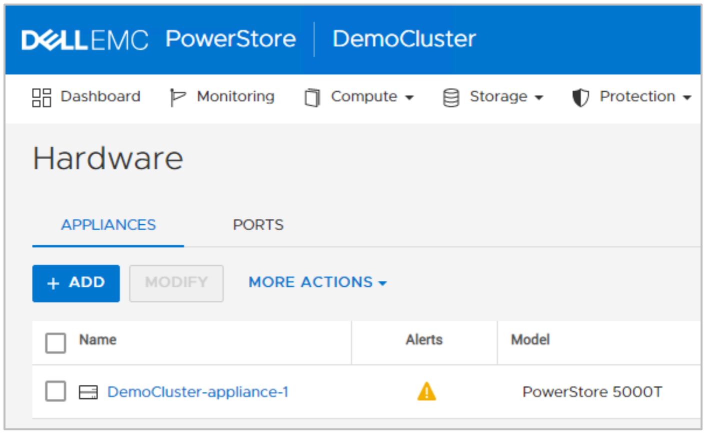

Figure 7 How to add an appliance

2.4.2 Configuring IP addresses

After an appliance has been selected, 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, three additional management IP addresses and two storage IP addresses (not pictured below) must be added in order to expand the cluster. Users can simply click **Add Network IPs** to supply new IP addresses for their management or storage networks.

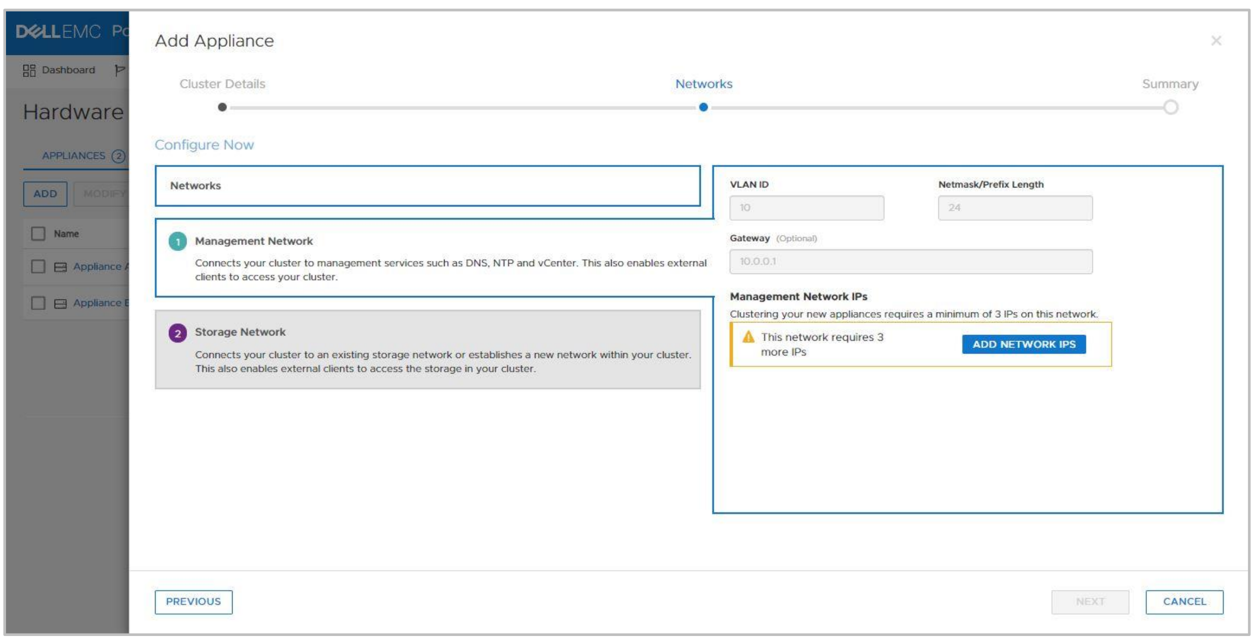

Figure 8 How to add IPs from the Add Appliance Wizard

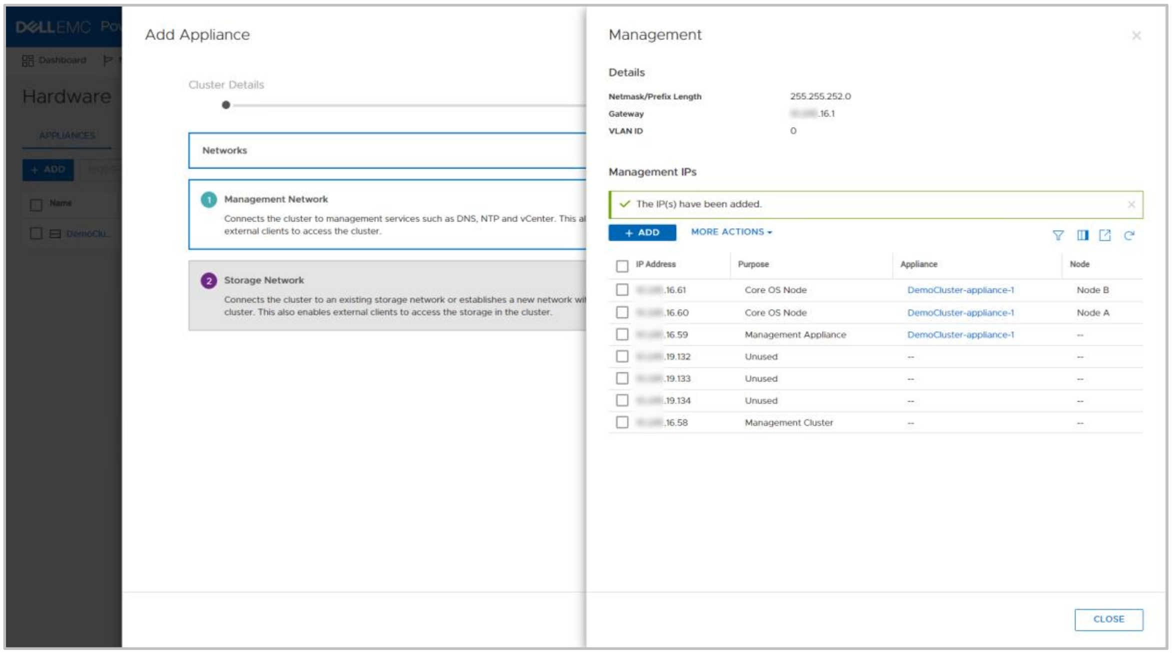

Figure 9 Adding Management IPs

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

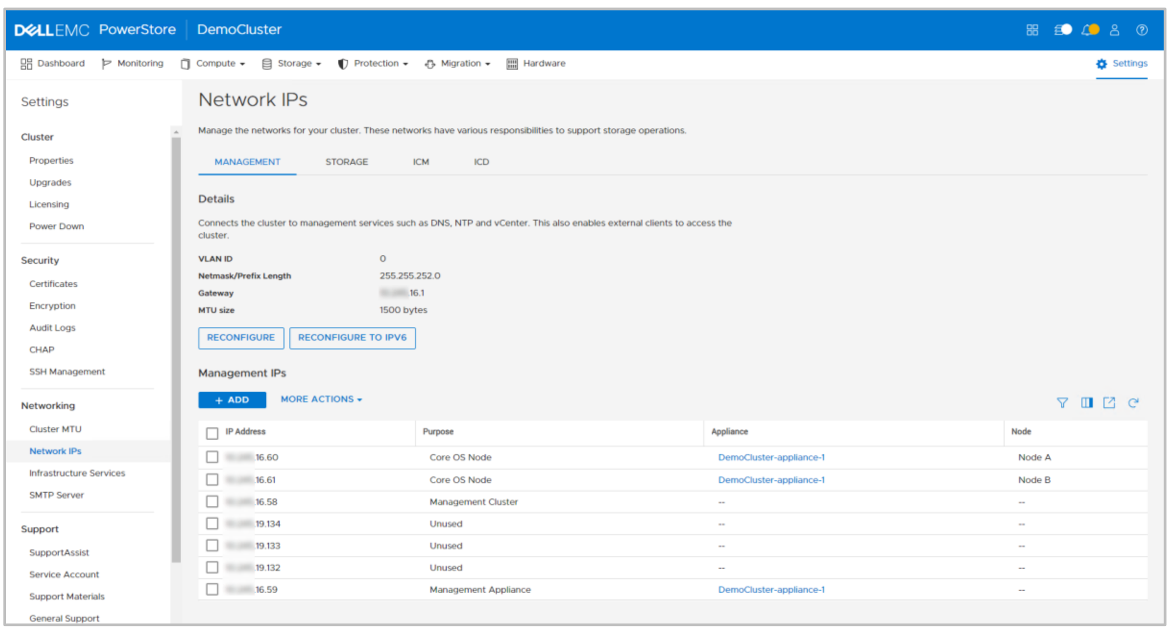

Figure 10 How to add IPs from the Settings Page

2.4.2.1 Validate network configuration

In the **Add Appliance** > **Summary** page, perform a network validation check before starting the Add Appliance job. Warnings or errors may display after the network validation check is performed. It is recommended to address warnings before beginning a cluster creation, but the warnings can be bypassed. However, if there are any errors, it is required to 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 and the network can be validated.

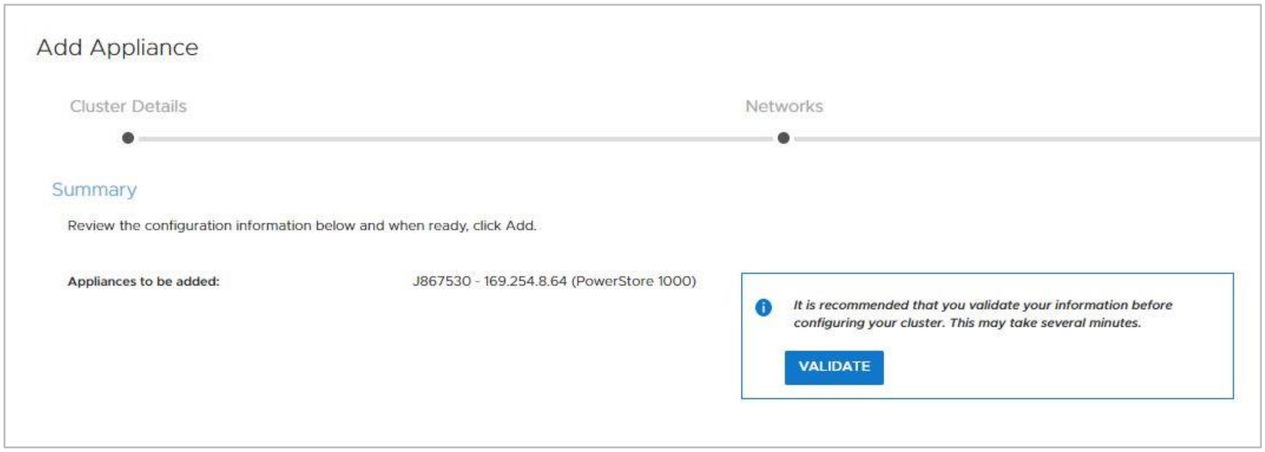

Figure 11 Add appliance network validation

2.4.3 Review the job

Once the Add Appliance wizard is completed, a new job is created. To see more details about the job, click **Jobs** icon in the top-right area of PowerStore Manager. Figure 12 shows an example of an Add Appliance job that is being run in the background. This job allows other tasks to be performed in PowerStore Manager while the Add Appliance job is running.

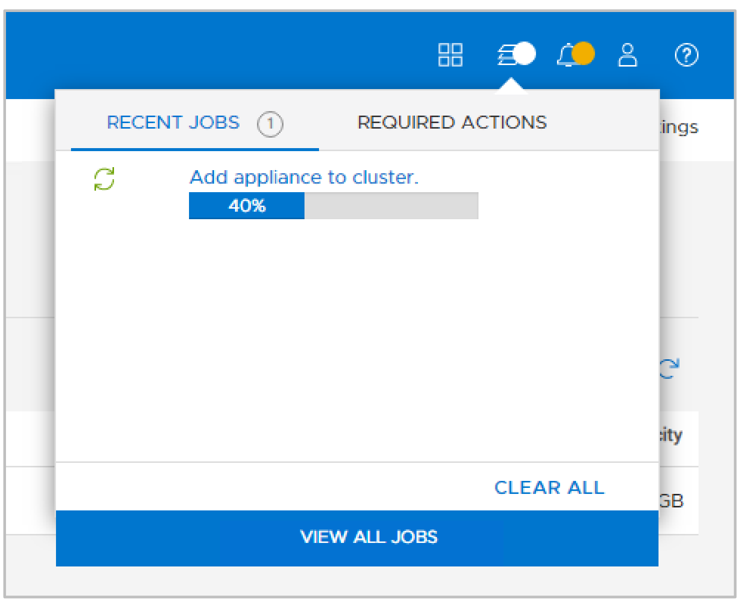

Figure 12 Add appliance job

2.5 Remove appliance

Clusters can be scaled down by removing individual appliances from the cluster. Removing an appliance from a cluster requires careful planning and consideration. Users must move all storage resources off the appliance before starting this process. After an appliance is removed from the cluster, the appliance is placed in the original factory settings. Removing an appliance from the cluster is only available with the service script **svc_remove_appliance**.

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.
- Continue with 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.

2.6 VMware integration

PowerStore T models and PowerStore X models offer deep integration with VMware vSphere such 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 then mounts the storage container as a vVol datastore and makes it available for VM storage. This vVol datastore is then able to be 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 within the cluster. It is important to note for PowerStore X model appliances, the cluster Distributed Resource Scheduler (DRS) is set to partially automated mode. Since 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 document *Dell EMC PowerStore: Virtualization Integration* on Dell.com/StorageResources.

2.7 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. It also assists with volume migrations or storage expansion that are based on analytics and capacity forecasting.

2.7.1 Initial placement of data

When storage administrators are creating 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 choosing to override the decision of automatic placement, there is an option to manually select which appliance to place the volume.

Figure 13 shows an example of a volume that will be placed automatically on a specific appliance. It is important to 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 migrated manually to another appliance in the cluster. Migration of storage resources within a cluster is discussed in the following sections of this paper.

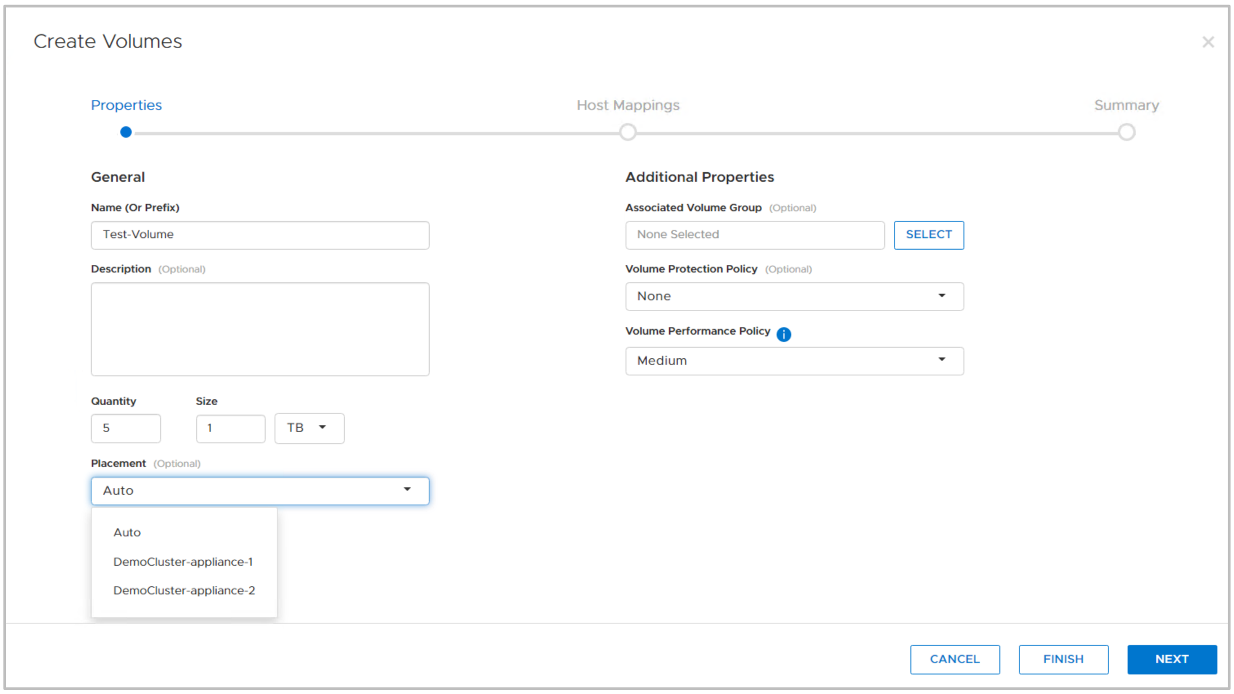

Figure 13 Automatic Placement of volumes or manual assignment

2.7.2 Volume groups

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

2.7.3 Migration overview

After a volume or volume group has been created on an appliance, capacity and demand may change over time. PowerStore supports moving a volume, volume group, or vVol to another appliance within the cluster. For volume and volume groups, this is non-disruptive. However, for vVol, users must power off the vVol based virtual machine before migrating the vVol to another appliance. This operation can be performed with a Manual Migration or an Assisted Migration within a PowerStore cluster by using PowerStore Manager, REST API, and PSTCLI. Both methods will be detailed in the next sections of this paper.

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 iSCSI or FC protocol.
- 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.

Once the above steps have been verified, a migration job can be started which is non-disruptive and transparent to the host. Although not visible to the end user, the storage resource is transferred to the new appliance within the cluster by leveraging asynchronous replication technology. It is important to note that when the storage resource is being moved over to the destination, all the associated storage objects such as snapshots and clones that are tied to the storage resource will be moved over to the destination as well. After the data transfer is complete, the host will switch paths automatically to the destination appliance and the migration job is complete.

Figure 14 shows an example of a multi-appliance cluster that is made up of four appliances. In this example, a user would like to move a storage resource off the PowerStore 1000T model over to a PowerStore 5000T model. In this example, the host may have connectivity to the two appliances by using iSCSI or FC protocols. Notice however that 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.

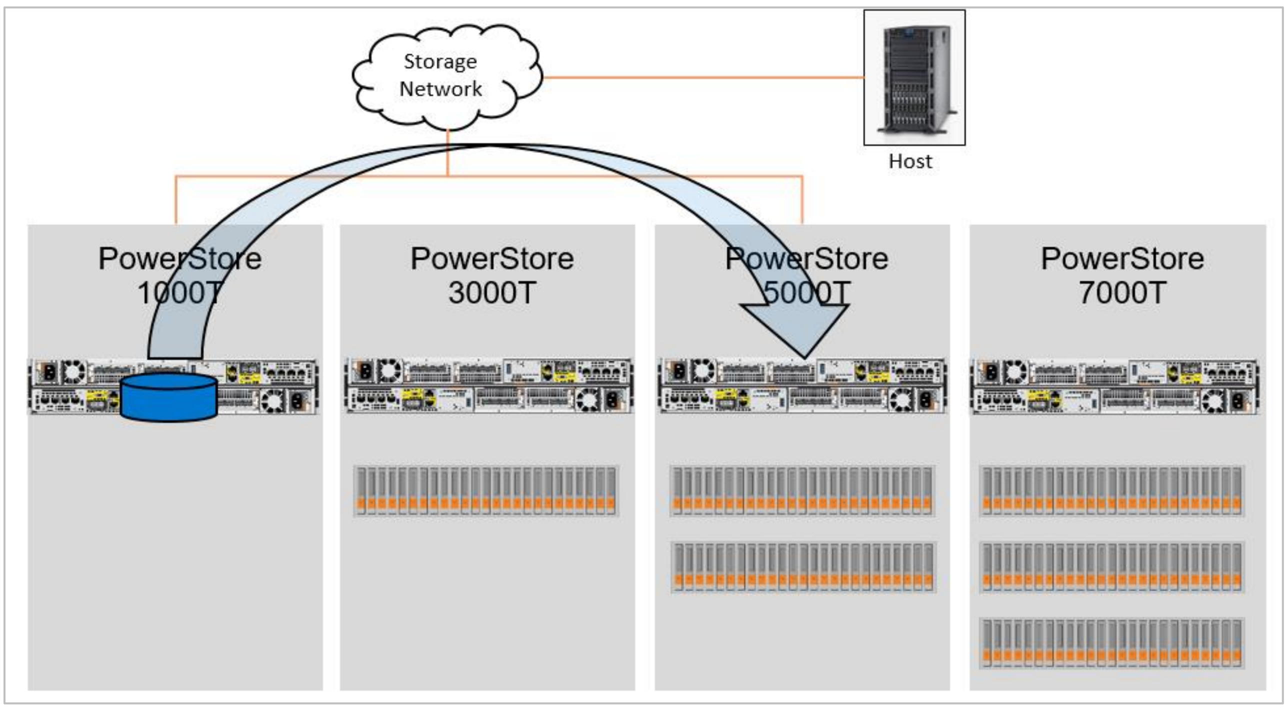

Figure 14 Volume migration example

2.7.4 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 have been verified, a manual migration can be performed on the storage resource.

Performing a manual migration in PowerStore for a volume, volume group, or vVol depends on the storage resource:

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

Take the example shown in Figure 15 where there is a volume that is named Test-Volume-001 that 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.

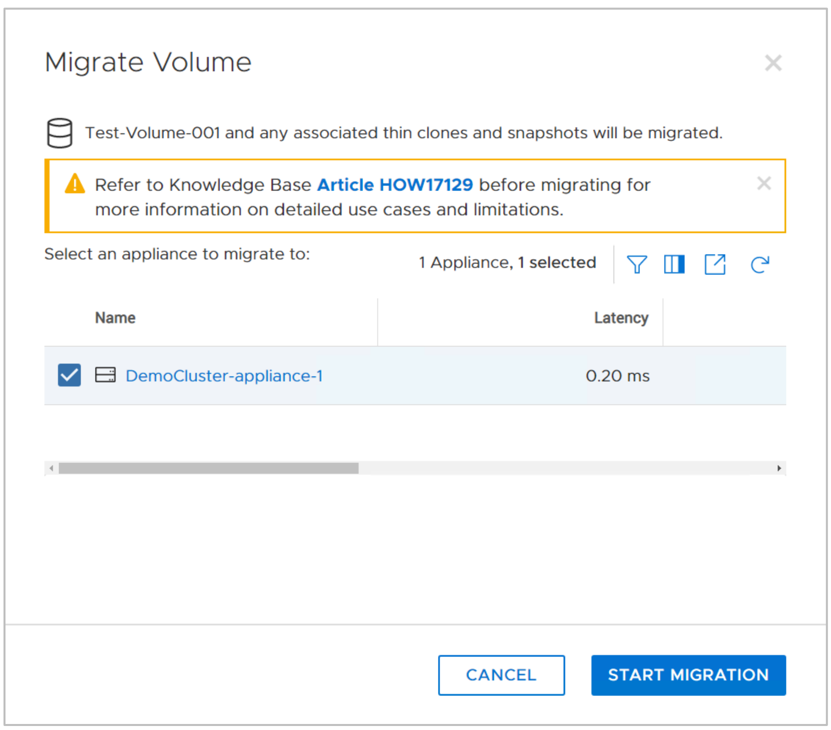

Figure 15 Start volume migration job

2.7.5 Assisted migration

While the system is running, it periodically monitors storage resource utilization across all appliances within the cluster. Over time, an appliance may approach the maximum usable capacity. In this situation, the system generates migration recommendations as shown in Figure 16. These recommendations are generated based on factors such as drive wear, appliance capacity, and health. An administrator can view assisted-migration recommendations from either the alerts or going to the **Migration** > **Migration Actions** page in PowerStore Manager. If the administrator accepts a migration recommendation, a migration session is automatically created. Like a manual volume migration, any snapshots or clones that are associated with the original storage resources are migrated over to the new appliance.

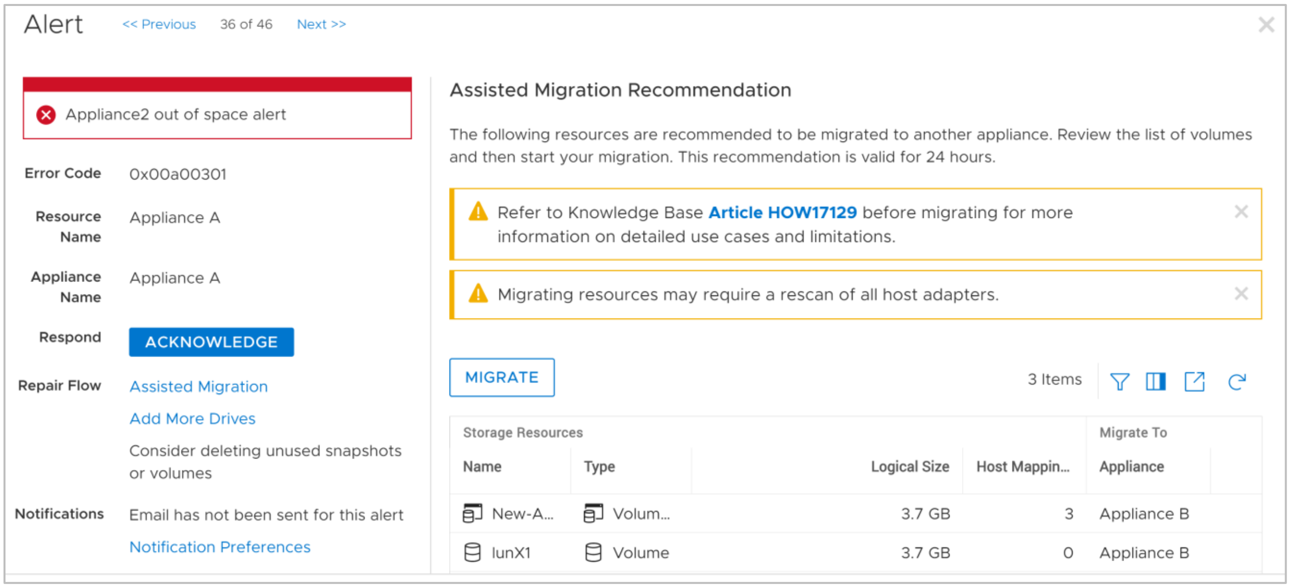

Figure 16 Assisted Migration Recommendations

2.7.6 Migration states

Once 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. Table 1 shows a list of all the possible states for a migration session.

Table 1 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 as soon as session initialization completes
Synchronizing	Represents that a background copy is in progress
Idle	Migration session transitions to this state as soon as 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	Represents a migration session being deleted
Completed	Migration session is completed and it is safe to delete the session
Pausing	Migration session transitions to this state as soon as pause command is issued
Paused	Migration session is Paused, user intervention is required to resume the session
System_Paused	Migration session transitions to this state if it encounters any error, user may resume or delete the migration after resolving the error

State	Definition
Resuming	Migration session background copy is resumed
Failed	Migration session encountered an error

2.7.7 Capacity forecasting

PowerStore provides details on when a cluster may run out of space. These details help when planning to scale a cluster based on business needs. To inform 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 may run out of space. Figure 17 shows an example of viewing capacity forecasting in PowerStore Manager.

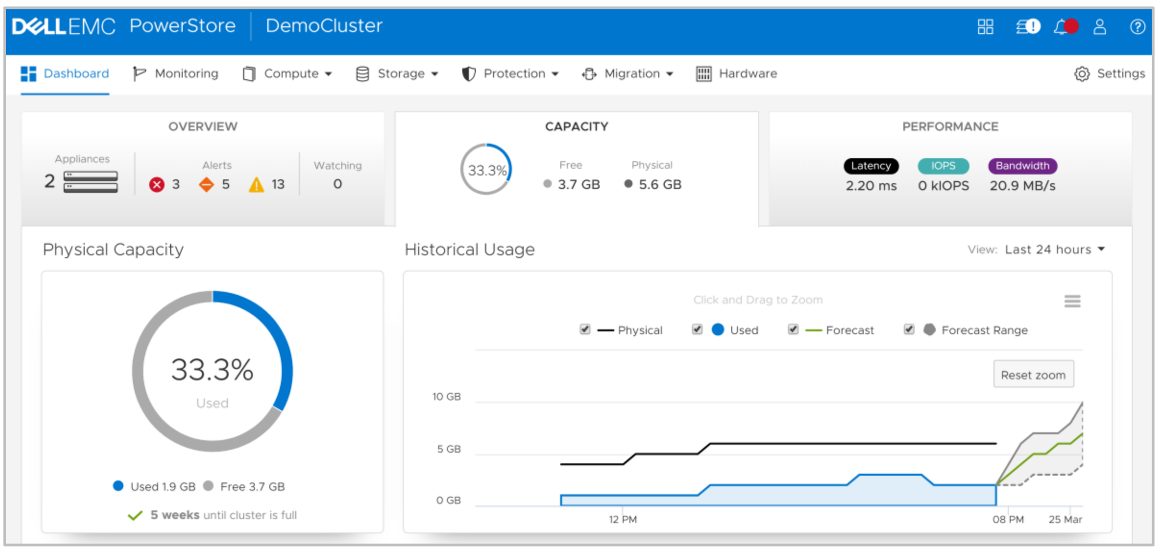

Figure 17 Capacity forecasting

3 High availability

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, such as 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.

3.1 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.

Some of the following components are listed in Figure 18 and Figure 19. For details about the components of a PowerStore appliance, see the document *Dell EMC PowerStore Hardware Information Guide* on the PowerStore Info Hub. Also, see the document *Dell EMC PowerStore: Introduction to the Platform* on Dell.com/StorageResources.

- Power supply
- Battery backup unit (BBU)
- Intel CPU
- Intel® QAT Hardware Offload Engine
- Motherboard
- Cooling fans
- 2x M.2 solid-state drives
- Memory DIMMs
- I/O modules (optional)

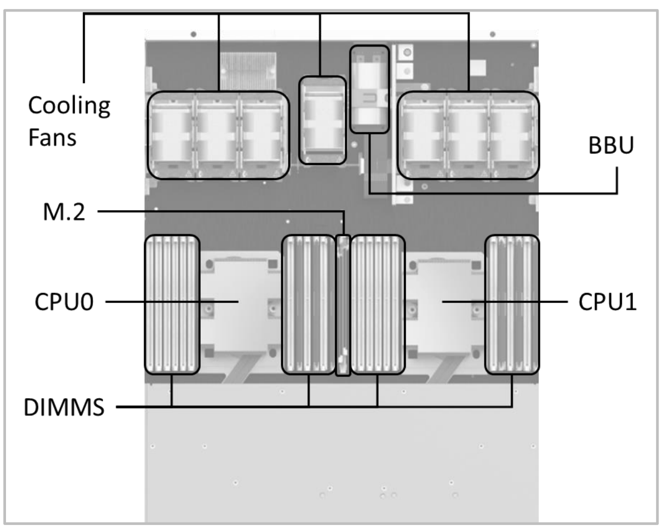

Figure 18 Internal components of node (top view)

1. 4-Port Card
2. Management Ports
3. Service Ports
4. I/O Module

Figure 19 Back view of node

3.1.1 Management software

The management software runs PowerStore Manager, the management interface (cluster IP), and other services. This software runs on one node within a cluster at a time. The node that is running the management software is designated as the primary node in PowerStore Manager which can be found on the hardware properties page. Figure 20 shows an example primary node highlighted by going to the **Hardware** > **Rear View**.

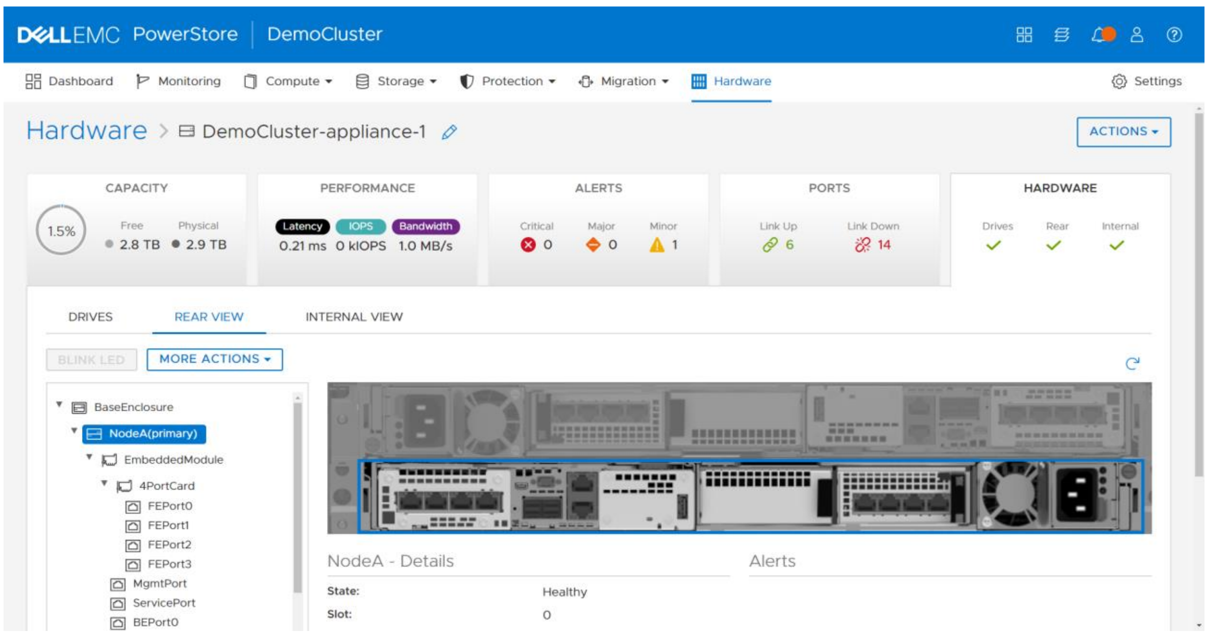

Figure 20 Primary node

If the primary node reboots, panics, or the management connection goes down, the management software automatically fails over to the peer node. After a failover, it may take several minutes for all services to fully start. Users that are logged in to PowerStore Manager during the failover may see a message indicating that the connection has been lost. When the failover process completes, restore access to PowerStore by refreshing the browser. Host access to storage resources are prioritized and are available before PowerStore Manager is accessible. After the failover, the management software continues to run on the new node, even if the other node returns to a healthy state. The node continues to run as the primary node, until it is rebooted or failed over.

3.1.2 System bond

PowerStore systems ensure that there is no single point of failure throughout the system. For a PowerStore T model appliance, it is recommended to cable the first two ports of the 4-port card to a top-of-rack switch as shown in Figure 21. 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. Below 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 (for PowerStore T models)
- iSCSI host traffic
- Replication traffic
- NAS traffic (for PowerStore T models)

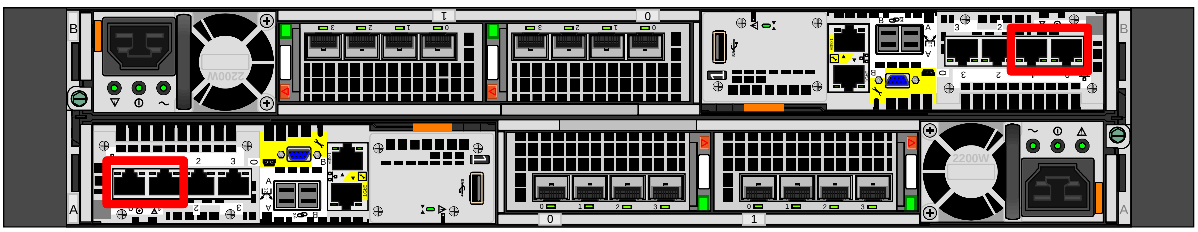

Figure 21 First two ports of the 4-port card (system bond)

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) has been configured on the network switches. To ensure best resiliency and network performance, it is recommended to enable LACP on the network. For more information about configuring LACP for a PowerStore T model appliance, see the documents *PowerStore Network Planning Guide* and *PowerStore Network Configuration for Dell PowerSwitch Series Guide* on the PowerStore Info Hub. For PowerStore X model appliances, there is no system bond and no LACP configuration is needed because the link aggregation and failover is handled by the virtual switches within the VMware hypervisor.

PowerStore also supports the ability to add additional ports for additional bandwidth or increasing fault tolerance. Administrators can extend iSCSI traffic by mapping the additional physical or, if applicable, virtual ports 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 as long as replication traffic has already been 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, reference section 3.3.1 in this white paper. For information about how to enable and scale replication traffic, see the document *Dell EMC PowerStore: Replication Technologies* on Dell.com/StorageResources.

When installing file services, starting from PowerStoreOS version 1.0.2 or later, Unified PowerStore T model appliances will route Intracluster Management (ICM) traffic within the appliance rather than the Top-of-Rack (ToR) switch.

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.

3.2 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 and 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's price/performance and effective capacity ($/IOP and $/Effective Capacity). As storage and drive capacity grows:

- 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

3.2.1 Overview

PowerStore implements proprietary algorithms where every drive is partitioned into multiple virtual chunks and redundancy extents are created by utilizing the chunks across several drives.

It automatically consumes the drives within an appliance and creates appropriate redundancy using all the drives. This 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 re-balanced.

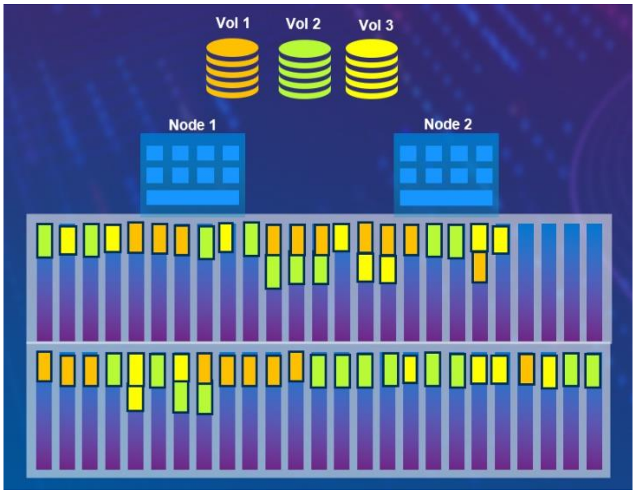

Figure 22 Data placement in PowerStore

3.2.2 Distributed Sparing

Unlike traditional RAID protection strategies, PowerStore does not require dedicated spare drives. Spare space is distributed across the entire appliance, a small chunk of space is reserved from each of the drive to be used for sparing in the event of a drive failure. A single drive worth of spare space is reserved for every 25 drives in an appliance.

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 also shortens rebuild time as only data that has written to the drive needs to be rebuilt.

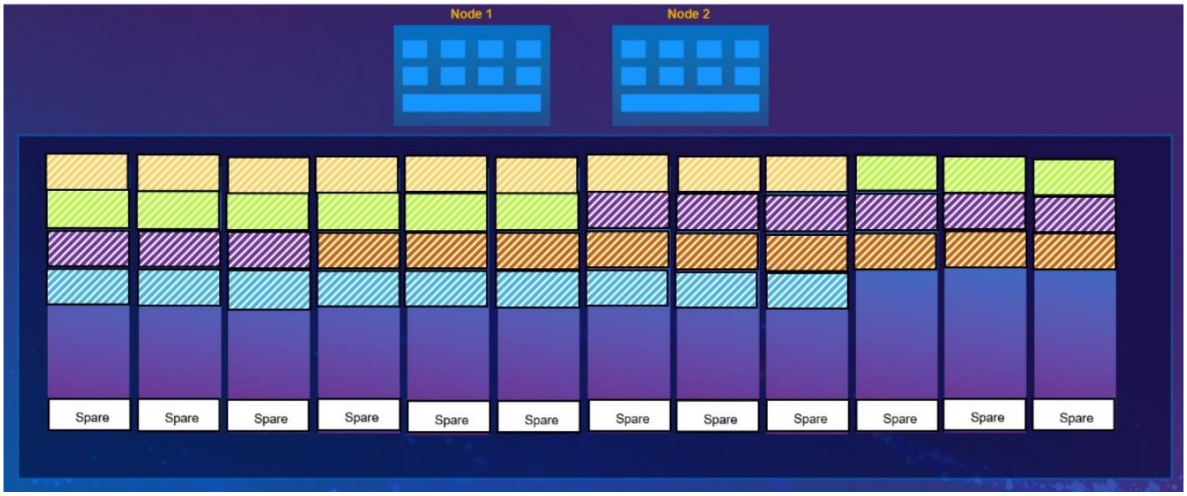

Figure 23 Distributed Sparing

3.2.3 Resiliency Sets

PowerStore implements resiliency sets to improve the reliability while minimizing spare overhead. Having multiple failure domains aka resiliency sets increases the reliability of the system since it allows the appliance to tolerate a drive failure within each of these resiliency sets if the failure occurs at the same time.

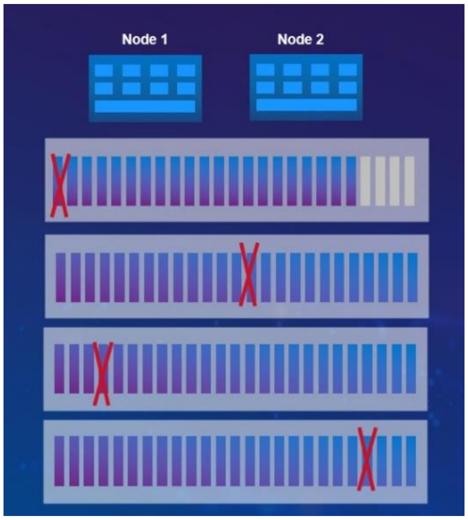

Figure 24 Tolerance for drive failure within multiple resiliency sets simultaneously

The appliance can tolerate multiple drive failures even within the same resiliency set, if the failure occurs at different instances (second drive fails after the rebuild on first failed drive is complete)

Resiliency sets can have up to 25 drives and the number of resiliency sets dynamically increases as more drives are added. For example, if a 26th drive is added, the resiliency set dynamically splits into 2.

Resiliency sets can span across physical enclosures based on the number of drives in the appliance and can have mixed drive sizes

Key Benefits:

Enterprise Class Availability

- Faster rebuild times with distributed sparing
- Rebuild smaller chunks of the drive simultaneously to multiple drives in the appliance

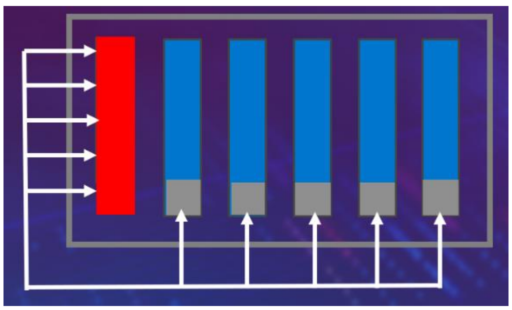

Figure 25 *Parallel rebuild of single drive to distributed spare space*

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 utilizes 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.

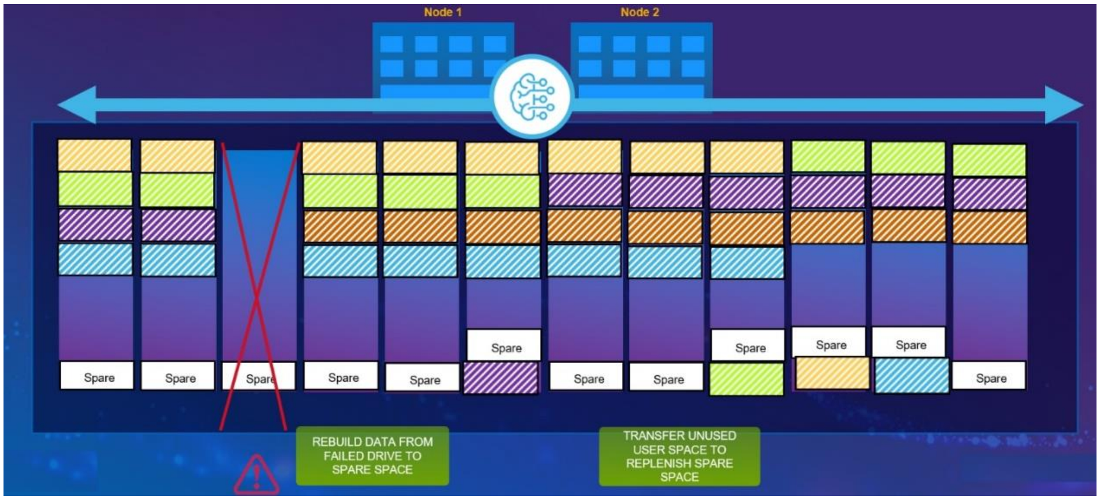

Figure 26 *Rebuild data chunks to spare space after drive failure and replenish spare space with unused user space*

Flexible Configurations

- Lower TCO with ability to expand storage by adding single drives
  - An appliance can have a minimum of 6 drives and can scale up to 96 drives. Capacity can be added non disruptively, customers have 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.

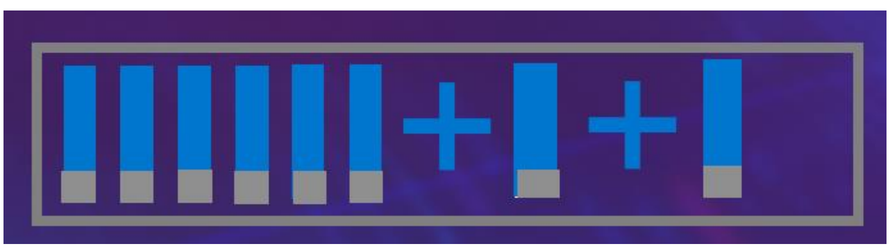

Figure 27 *Single Drive Expansion*

3.3 Block storage

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 is in contrast to competing architectures that leverage redirection and single node locking that can result in long trespass times which increase the chances for data unavailability during failover.

PowerStore implements a dual node, shared storage architecture that leverages Asymmetric Logical Unit Access (ALUA) for host access. I/O requests received on any path (active/optimized or active/non-optimized) are committed locally and 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 were to become unavailable, I/O requests will be processed on the local node through the active/non-optimized paths without the need to send any data to the peer node. From this point, the I/O will be written to the shared NVRAM write cache where the I/O will be deduplicated and compressed before being written to the drives. For more information on how the I/O is written to the drives, reference the *Dell EMC PowerStore: Data Efficiencies* white paper on Dell.com/StorageResources.

ALUA multi-pathing ensures high availability and 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, multipathing software, such as Dell EMC PowerPath™, must be installed on the host. Multipathing software should be configured 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.

Since 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 multi-pathing and redundancy purposes.

3.3.1 iSCSI configuration

iSCSI interface creation is deployed 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. iSCSI interfaces can be created during the Initial Configuration Wizard (ICW), or in post-cluster creation. For a robust HA environment, create additional interfaces on other ports after the cluster has been created.

If iSCSI is enabled during ICW, enter the iSCSI information on the **Storage Network** page as shown in Figure 28. These IP addresses are assigned to a virtual bonded interface that is assigned to the first two ports on the 4-port card on node A and node B. 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 Figure 29.

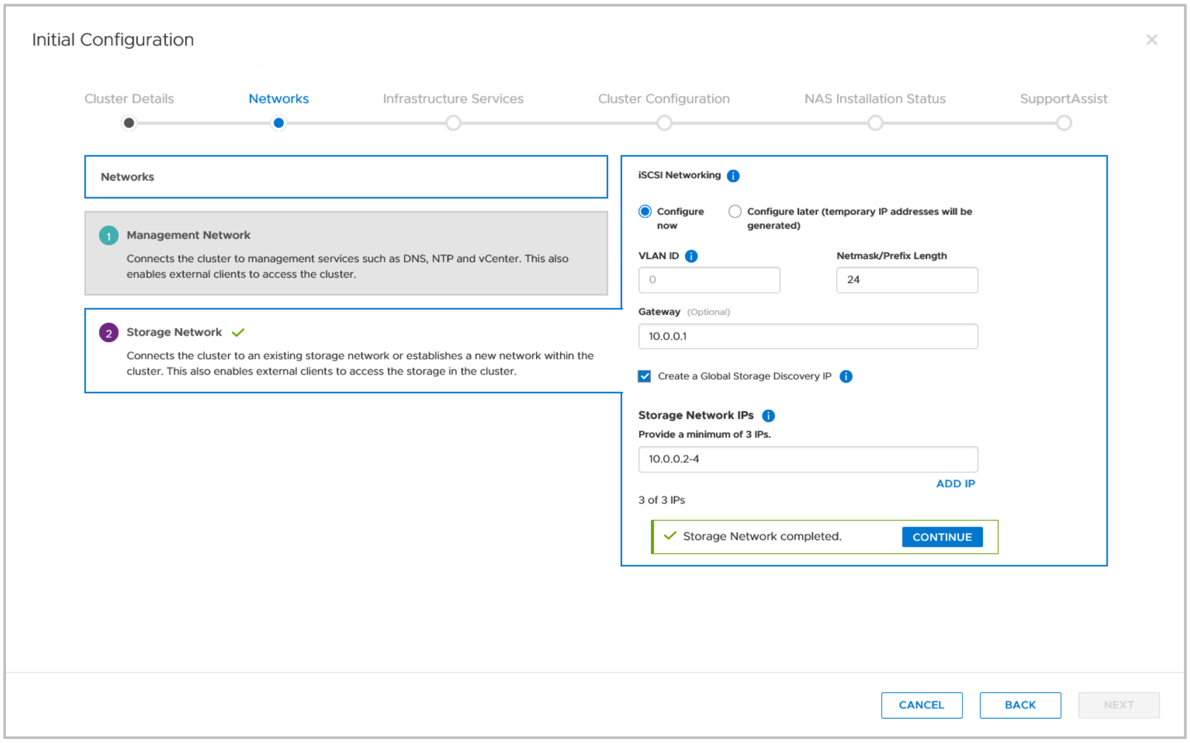

Figure 28 Storage network

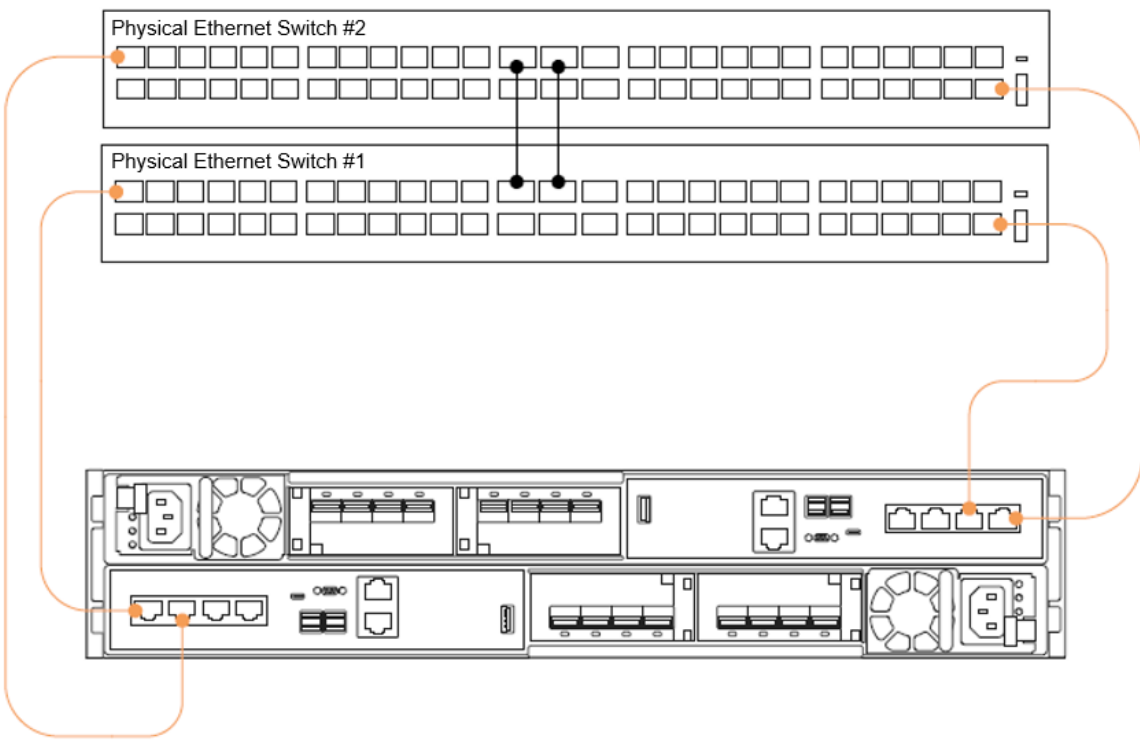

Figure 29 First two ports of the 4-port card cabled to top-of-rack switches

After an appliance is configured, 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. Since 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 *Dell EMC PowerStore: Replication Technologies* on Dell.com/StorageResources.

After completing cabling, add IP addresses in PowerStore Manager. Figure 30 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. Figure 31 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: Click **Hardware** > **Appliance Details** > **Ports**.
- PowerStore X models: Click **Hardware** > **Appliance Details** > **Virtual Ports.**

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 being 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 *Dell EMC PowerStore: Virtualization Integration* on Dell.com/StorageResources.

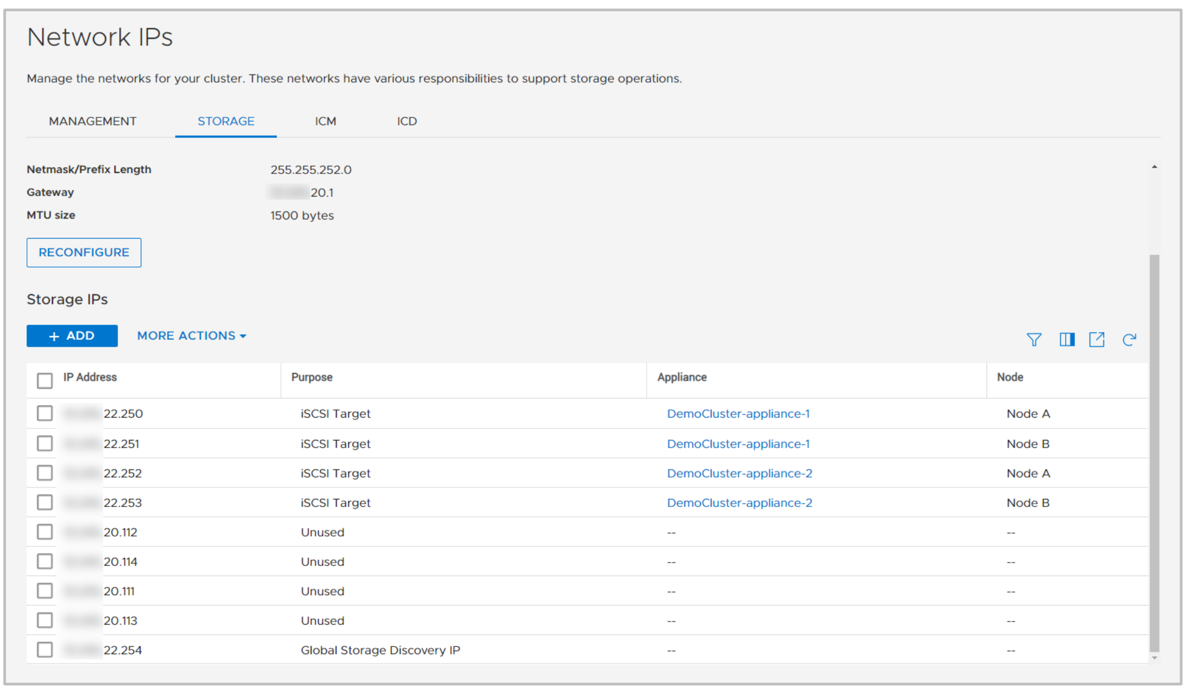

Figure 30 Verify unused storage network IP addresses

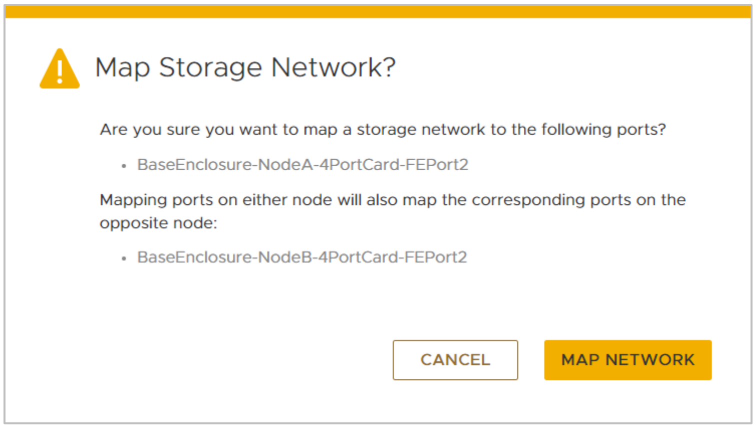

Figure 31 Map Storage Network

3.3.2 Fibre Channel configuration

To achieve high availability with Fibre Channel (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, zoning must be configured 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, more FC ports can be zoned to provide additional paths to the PowerStore appliance.

There are two locations that are used to locate the port World Wide Names (WWNs) in the PowerStore Manager. The first location is in the **Hardware** > **Ports** page, as shown in Figure 32. The second location is found in the **Hardware** > **Rear** view page, as shown in Figure 33. For additional details on configuring hosts, see the document *Dell EMC PowerStore Host Configuration Guide* on the PowerStore Info Hub.

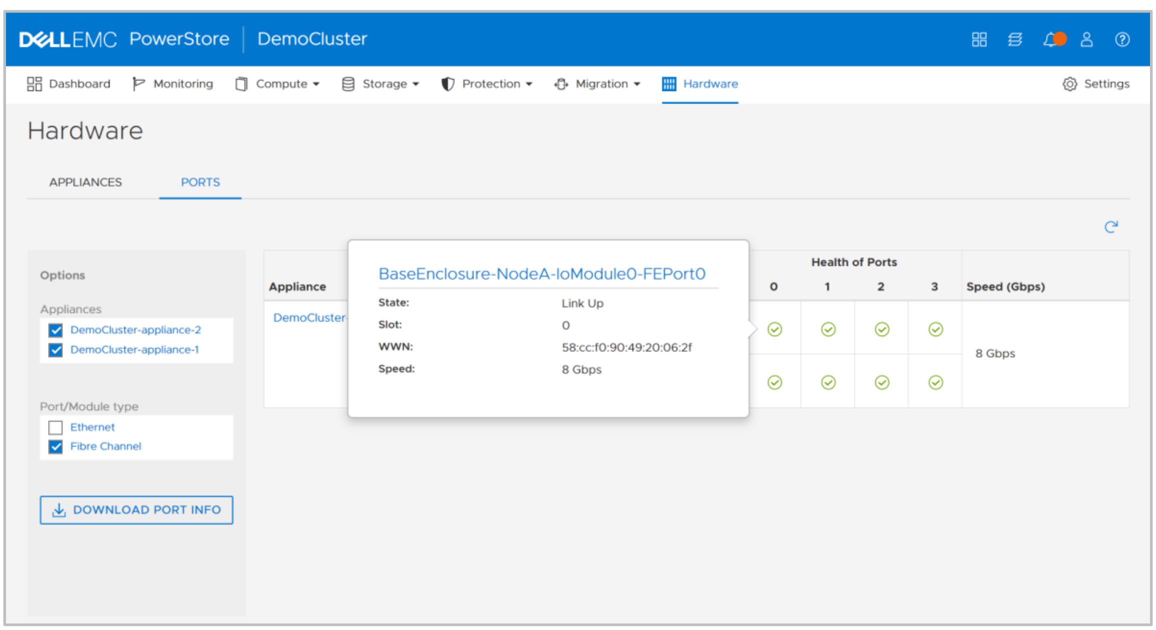

Figure 32 PowerStore Manager hardware ports page

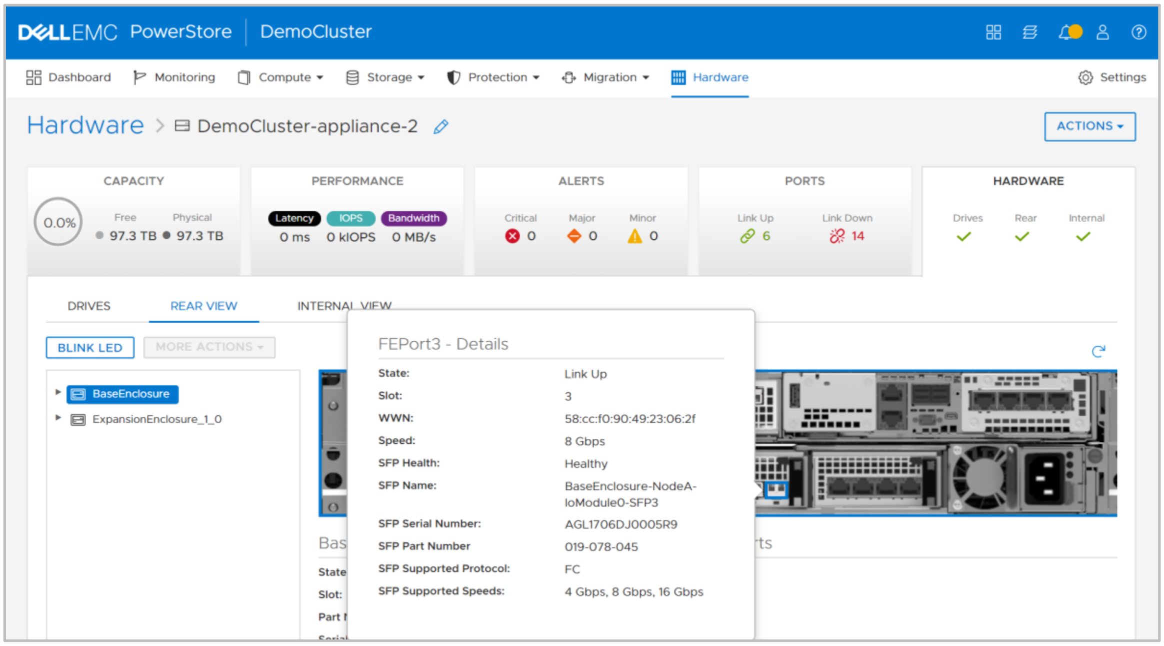

Figure 33 PowerStore Manager hardware back view

3.3.3 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. Figure 34 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 PowerStore Info Hub for detailed information about cabling and configuration of the network infrastructure.

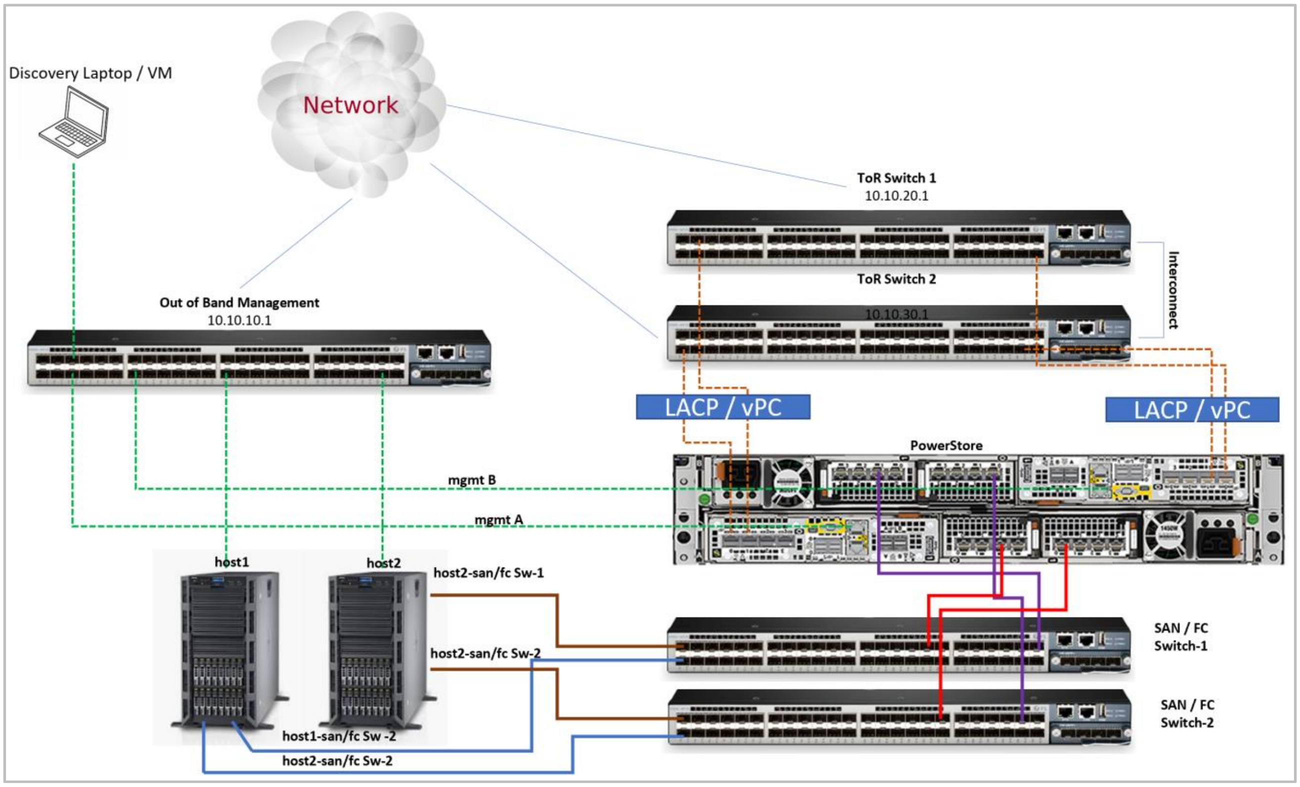

Figure 34 Highly available block configuration

3.4 File storage

Before file-level resources are shared from a PowerStore T model system, the system creates all NAS server interfaces automatically on the first two bonded ports of the 4-port card. The ability to assign NAS interfaces to certain ports is not supported. An administrator would then create a NAS server which 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 that the NAS server should be running on, based on this algorithm. Once provisioned, the preferred node for a NAS server never changes. The current node indicates the node that the NAS server is running on. Changing the current node moves the NAS server to a different node, which can be used for loading-balancing purposes. When a NAS server is moved to a new node, all file systems on the NAS server are moved along with it.

Figure 35 shows the current and preferred node columns in PowerStore Manager.

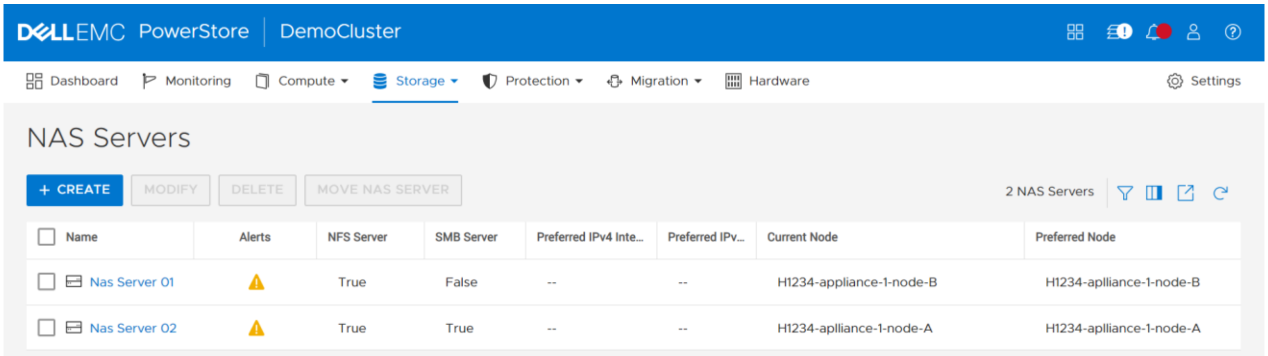

Figure 35 Current Node and Preferred Node

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. Once 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 completes, the NAS servers return to the node that they were assigned to at the beginning of the upgrade.

3.5 VMware integration

The following subsections within VMware Integration describe some of the HA capabilities with the PowerStore X model unless stated otherwise. For more details on PowerStore integration with VMware, reference *Dell EMC PowerStore: Virtualization Integration* on Dell.com/StorageResources.

3.5.1 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. Since these are not shared datastores, in the event of a node failure, 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. It is important to note that the controller VM should not be moved and should not have any settings changed.

3.5.2 vSphere HA

When administrators are leveraging virtual machines on either the PowerStore T or PowerStore X model appliances, it is recommended that vSphere HA be enabled on the cluster. In the event of power loss for a node or host in the cluster, virtual machines are restarted on a surviving host or node in the cluster.

3.5.3 vCenter connection

Both PowerStore T model and PowerStore X model appliances offer the capability of 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 for through vCenter may be temporarily unavailable. However, it is important to note that I/O continues to be processed for all storage objects in the PowerStore appliance.

3.6 Replication

To protect against outages at a system or data-center level, it is recommended to use replication to a remote site. This use includes planned maintenance events, unplanned power outages, or natural disasters. PowerStore supports replication to simplify disaster recovery. For details on how to set up and use the replication features on PowerStore, see the document *Dell EMC PowerStore: Replication Technologies* on Dell.com/StorageResources.

3.7 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.

3.8 Cluster high availability

PowerStore appliances implement a Quorum management feature that is used for cluster resource management and split brain handling to ensure data consistency in response to changes in the cluster. This Quorum management feature 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

3.8.1 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. Figure 36 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. Since 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

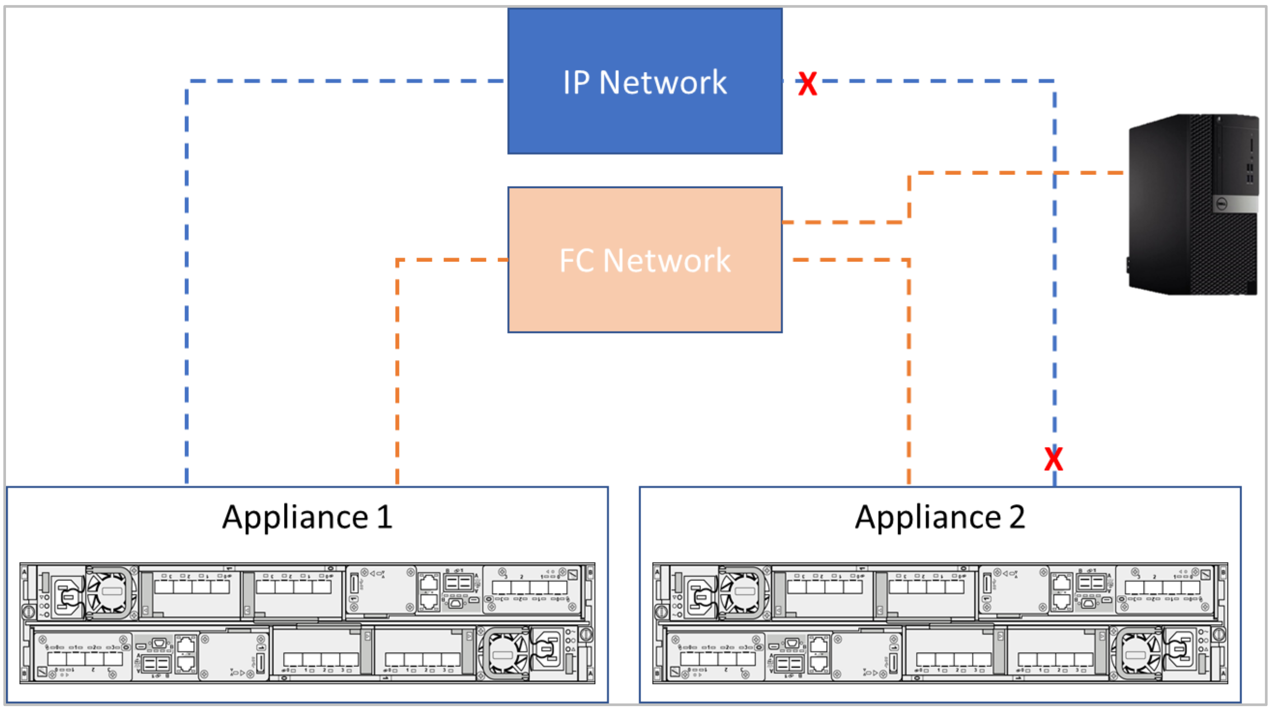

Figure 36 Two-appliance cluster with no quorum

In Figure 37, this example 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.

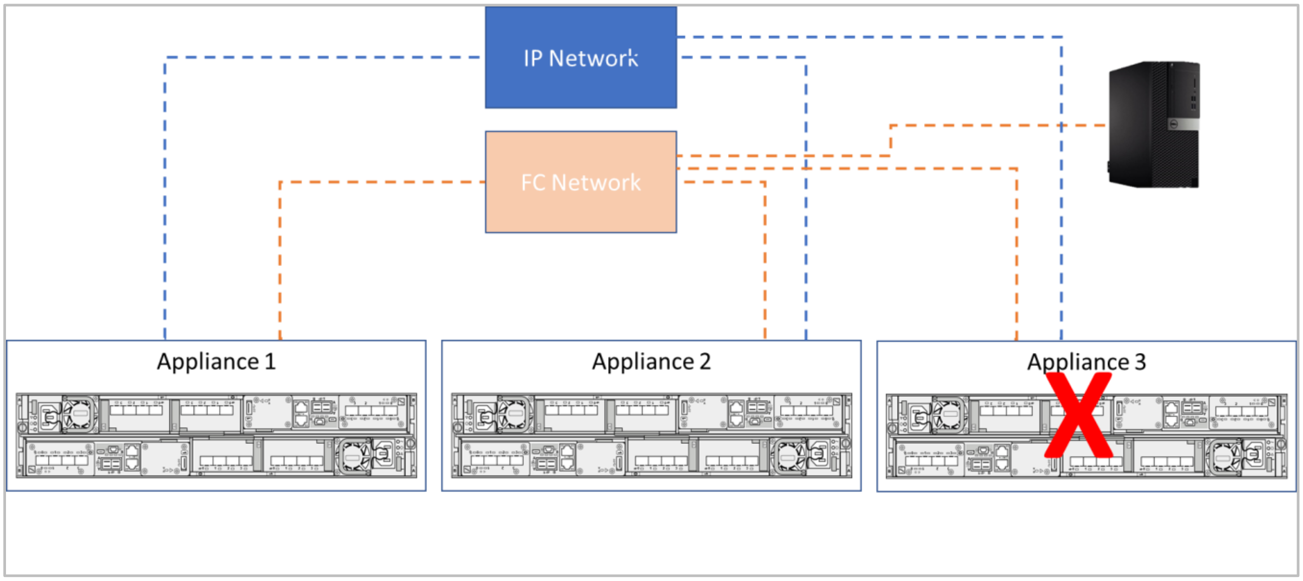

Figure 37 Three-appliance cluster with quorum

3.8.2 Cluster IP

The cluster IP address is a required field 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 found 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 may take several minutes to regain access to the PowerStore Manager.

3.8.3 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 hosts only require one GSIP and can discover all the storage paths for all the appliances in the cluster. Otherwise, iSCSI hosts require a list of storage IPs so that if one IP is down, the host can try the next IP.

4 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.[1] By using the clustering and high availability features in PowerStore, organizations can minimize the risk of data unavailability.

A Technical support and resources

Dell.com/support is focused on meeting customer needs with proven services and support.

Storage technical documents and videos provide expertise that helps to ensure customer success on Dell EMC storage platforms.

The PowerStore Info Hub provides detailed documentation on how to install, configure, and manage Dell EMC PowerStore systems.

[1] Based on the Dell Technologies specification for Dell EMC PowerStore, April 2020. Actual system availability may vary.