Dell PowerFlex 4.5.x

Technical Overview

February 2024
Rev. 1.2

Notes, cautions, and warnings

NOTE: A NOTE indicates important information that helps you make better use of your product.

CAUTION: A CAUTION indicates either potential damage to hardware or loss of data and tells you how to avoid the problem.

WARNING: A WARNING indicates a potential for property damage, personal injury, or death.

Contents

1

Introduction

This document provides a high-level technical overview of PowerFlex. It describes the key PowerFlex features, the management capabilities of the PowerFlex management platform and information regarding the PowerFlex offerings.

The target audience for this document includes presales, Dell Technologies Support, and customers.

For additional PowerFlex software documentation, go to PowerFlex software technical documentation.

2

Revision history

Table 1. Revisions

Date	Document revision	Description of changes
February 2024	1.2	Update to the product limits
November 2023	1.1	Updates for PowerFlex 4.5.1
September 2023	1.0	Initial release

PowerFlex overview

PowerFlex is an enterprise-class, software-defined block and file storage solution that is deployed, managed, and supported as a single system. Highly flexible and scalable, PowerFlex allows customers to have a mixture of two-layer (server SAN), single-layer hyperconverged configurations, storage-only, PowerFlex file services, and other mixed architectures within a single deployment.

With PowerFlex, resources such as storage and compute can be scaled together or separately, non-disruptively, and in small increments. The system can scale from a few nodes to hundreds in a cluster, linearly scaling I/O performance and throughput. In addition, PowerFlex supports a broad set of operating environments that include multiple hypervisors, bare-metal operating systems, container management tools, and PowerFlex file services capabilities (NFS, CIFS, and additional file services offerings). This allows customers to flexibly transition application architectures, while supporting multiple generations of applications with disparate architectures in a single system.

PowerFlex offers integrated product options. PowerFlex appliance is a scalable system with flexible form factors that comes pre-configured and validated for fast, easy deployment. PowerFlex rack is a rack-scale system that is manufactured, managed, supported, and sustained as one system for single-end-to-end lifecycle support. PowerFlex custom node is a validated engineering solution that provides nearly all the flexibility of the PowerFlex software offering.

Terminology for management offerings

The following table describes important PowerFlex Manager terms and concepts for the PowerFlex management offerings.

Table 2. Terms and concepts

Term	Description
Bring your own hypervisor	Allows customer to supply a hypervisor to host the Kubernetes VM or environment. Enables setup of the PowerFlex management platform system for management purposes in ESXi or Kubernetes VM.
Co-residency	The PowerFlex storage-only nodes in a protection domain runs the management of the system.
Management data store (MDS)	The PowerFlex management data store is on the dedicated multi-node controller. It is not the same as the data store used in a co-resident option. Kubernetes VMs host the management containers that reside on this data store.
Management virtual machine (MVM)	The PowerFlex Manager UI is hosted in a Kubernetes environment running across a series of Linux machines. The VM version is called the management VM. The operating system is a customized SLES. version that is maintained by Dell Technologies.
PowerFlex management controller	The infrastructure underlying the PowerFlex management platform in a dedicated multi-node configuration. It is mandatory in the PowerFlex rack, and optional for PowerFlex appliance The physical controllers with ESXi hosting PowerFlex management platform can be a three-node or a five-node ESXi cluster with: ● VSAN (PowerFlex management controller 1.0) ● PowerFlex as shared storage (PowerFlex management controller 2.0) PowerFlex management controller also hosts VMs for vCenter, a jump server, and other optional NSXT or CloudLink managers.
PowerFlex management platform	The management software stack for PowerFlex runs on physical or virtual Linux instances, which are containers that provide services. This management solution can be implemented on various infrastructure options.

Key features

PowerFlex supports the new features described in the following sections.

Audit logging

PowerFlex 4.5.1 enables the redirection of the PowerFlex events and Ingress audit messages to syslog. These events and audit messages are forwarded to an external Security Information and Event Manager (SIEM). You must perform the following configuration steps to enable this feature in PowerFlex Manager:

- Change Ingress settings to emit audit messages.
- Define a notification policy to forward Ingress audit messages in the PowerFlex system to an SIEM.
- Define a notification policy to forward events in the PowerFlex system to an SIEM.

For more information about the configuration steps, see the *Dell PowerFlex 4.5.x Administration Guide*.

PowerFlex cloud support

Dell APEX Block Storage for the public cloud is a deployment of Dell PowerFlex, software-defined block storage innovation, in the public cloud. APEX Block Storage is available for both AWS and Microsoft Azure allowing you to experience the same benefits of enterprise-class storage services in the cloud as with on-premises. It provides higher performance, larger volume sizes, and improved resilience than what is currently available on the public cloud.

Table 3. Capabilities and benefits

Capability	Benefit or Value
Unique multiple availability zone durability	● Enhanced, space efficient data protection through the aggregation of all instances across availability zones for volume provisioning. ● Enables high resilience across availability zones by protecting the data during availability zone failure, without replication of data (space and IOPs).
Extreme performance for mission-critical workloads running in the cloud	Meets and exceeds SLAs with extreme performance (high throughput and low latency) for workloads such as databases and analytics.
Scalable, flexible, and resilient	Meet stringent SLAs and run workloads with confidence and assurance with near linearly scalable performance as workload demand increases.
Data mobility	Seamless data mobility with the ability to easily move data from on-premises to the cloud or across regions within the cloud as workloads demand increases.
Self-healing architecture with rapid rebuild	Ensure high availability and service availability under failure conditions with fast reprotection and rebuild.
Robust ecosystem of automation tools and frameworks (CSM/CSI, Ansible, REST APIs, DDVE, CloudLink, CloudIQ)	Optimal usage and performance of cloud-based block storage with easy to deploy, expand, monitor, and manage capabilities.

NOTE: The following features that are supported by PowerFlex 4.5 are NOT supported with the APEX Block Storage in the public cloud: compression, fine granularity storage pools in the public cloud, SDNAS or PowerFlex file services, and NVMe TCP.

APEX Block Storage for AWS

Dell APEX Block Storage for AWS empowers enterprises to run diverse workloads in the public cloud while ensuring extreme performance, scalability, and a simplified cloud experience. APEX Block Storage can be deployed in two configurations by using Elastic Block Store (EBS) volumes, or native NVMe SSD drives attached to EC2 instances (EC2 Instance Store). Dell APEX Block Storage also provides the proven enterprise data services, such as thin provisioning, snapshots, and asynchronous replication, required to run demanding block-based workloads in the public cloud. Dell APEX Block Storage native asynchronous replication enables data mobility between on-premises and the cloud or across regions in the cloud, for example, AWS East to AWS West.

APEX Block Storage for Microsoft Azure

Dell APEX Block Storage for Public Cloud in Microsoft Azure. APEX Block Storage can be deployed in two configurations using Azure Managed Disks or virtual machines with attached NVMe SSDs based on the use case. Dell APEX Block Storage also provides the proven enterprise data services, such as thin provisioning, snapshots, and asynchronous replication, required to run demanding block-based workloads in the public cloud. Dell APEX Block Storage native asynchronous replication enables data mobility between on-premises and the cloud or across regions in the cloud.

PowerFlex support for Dell PowerEdge R660, R760, R6625, and R7625 servers

PowerFlex 4.5 certifies support of the following platforms for software only and custom node offerings:

- Dell PowerEdge R660 and Dell PowerEdge R760 servers (Intel based)
  - Available with a new persistent memory solution, Software Defined Persistent Memory (SDPM)
- Dell PowerEdge R6625 and Dell PowerEdge R7625 servers (AMD based)

NVMe over TCP connectivity

PowerFlex 4.5 supports the NVMe over Fabrics and NVMe over TCP storage protocol for front-end connectivity, enabling customers to use both NVMe over TCP and SDC hosts.

NVMe over Fabrics and NVMe over TCP connectivity includes:

Host network awareness

PowerFlex 4.5 considers the host networks for resiliency. When a host is losing a network, it can still use other paths through available networks to access the storage.

Network sets

PowerFlex 4.5 supports specifying networks that might fail together. NVMe over TCP connections are allocated to the NVMe over TCP hosts through multiple network sets, offering resiliency to the failure of the network set.

Fault sets support

NVMe over TCP supports fault sets. Connections allocated to the NVMe over TCP are through targets on multiple fault sets that provide path resiliency to a fault set failure.

NVMe reservations support (SCSI-3 equivalent)

PowerFlex 4.5 supports NVMe reservations (SCSI-3 equivalent) with NVMe over TCP.

Persistent discovery controller

NVMe over TCP hosts connect to PowerFlex 4.5 through persistent discovery, which enables automatic updates of the hosts regarding storage side changes.

Migration to NVMe over TCP

PowerFlex 4.5 provides the option to migrate workloads from SDC to NVMe over TCP on ESXi:

1. Online migration using Storage vMotion (VMFS only)
   - The standard way of moving storage is with Storage vMotion. Storage vMotion also supports switching protocols by migrating to a new DataStore. For more information, see the *Dell PowerFlex 4.5.x Administration Guide*.
2. Offline migration (VMFS only)
   - There is a new option for converting an existing VMFS datastore from SCSI (SDC) to NVMe over TCP without having all the data over the network. The offline migration steps are covered in this KB article.

ⓘ **NOTE:** There are not standard ways to migrate Linux environments, ESXi clusters, and RDMs.

PowerFlex file services improvements

PowerFlex Manager introduces the following PowerFlex file services enhancements:

Single namespace

- Ability to create a global namespace (GNS) supported by the NAS cluster with a single export
- Allows all hosts with correct access permission to access existing and newly added file systems to the namespace without needing to explicitly mount them on each client
- Consists of several file systems that may be SMB or NFS

Common Event Publishing Agent

- Support for Common Event Publishing Agent (CEPA), which is part of the Common Event Enabler (CEE) package
- Ability to receive file event notifications
- Ability to use CEPA to see events on some or all of my NFS and SMB file systems.

Operating system highlights

PowerFlex 4.5 supports the following operating systems:

- Red Hat Enterprise Linux 9.2 or 8.8
- SLES 15.5
- ESXi 8.0 for NVMe over TCP client and NAS client

PowerFlex management platform improvements

PowerFlex 4.5 now supports several management improvements.

Stronger security

Use SSH keys and non-root users through sudo in management aspects. For more information about configuration guidelines, see the *Dell PowerFlex 4.5.x Install and Upgrade Guide* and *Dell PowerFlex 4.5.x Administration Guide*.

Webhooks

PowerFlex 4.5 enables Webhooks support by sending alerts to Webhooks servers such as BigPanda.

PowerFlex management platform deployment time

Optimization of the PowerFlex management platform installer run time to 60 minutes. In PowerFlex 4.0.x, the installer run time was 3 hours.

Multi-subnet and multi-VLAN for PowerFlex appliance and PowerFlex rack

PowerFlex 4.5 allows defining multiple subnets and VLANS in a single network. The most common implementation example is with a management or vMotion network that spans multiple racks. Each rack is its own subnet and VLAN.

NOTE: This capability is supported on all networks except for the following:

- PowerFlex data
- vSAN and NSX overlay network
- NSX external networks

Higher flexibility of fault sets

PowerFlex 4.5 defines a subset of nodes in a rack as part of a fault set during deployment and expansion operations. For example, a rack with 20 nodes can have four fault sets, one for every five nodes, for easier maintenance support. Another option is using fault sets in a full rack configuration, such as one fault set per rack.

PowerFlex management controller self-awareness on PowerFlex appliance and PowerFlex rack

PowerFlex 4.5 allows self-awareness of its underlying controller system. Upgrade from end to end and receive alerts from issues in the PowerFlex management controller infrastructure.

Modular (single) component upgrade

PowerFlex 4.5 changes the package in the PowerFlex rack or PowerFlex appliance catalogs to newer or older versions. This upgrade depends on customer cases that want to deviate from the Release Certification Matrix or Intelligent Catalog.

NOTE: By default this feature is turned off and can only be enabled through an RPQ process. If applicable, contact your account team for details.

4

System requirements and product limits

PowerFlex system requirements specify the minimum support required for PowerFlex. The maximum (and, in some cases, minimum) values for key system parameters are specified by the product limits.

Product limits

The following table summarizes PowerFlex supported capabilities and system limits. PowerFlex Manager is not backward compatible.

ⓘ **NOTE:** The supported limits described might differ from system enforced limits in the list output of the `scli --query_system_limits` command.

Table 4. Supported capabilities and system limits

PowerFlex item	Limit
System raw capacity	Up to 16 PB
Maximum data networks in system	8 ⓘ **NOTE:** SDCs can connect through a router and exceed this limit.
Maximum data networks for MDM virtual IPs	4
Device size	Minimum: 240 GB, Maximum: 8 TB (Maximum 15.36 TB for SSDs on medium granularity storage pools)
Minimum number of SDSs in a storage pool	3
Minimum storage pool size	720 GB
Minimum devices (drives) per storage pool	3, one per fault unit
Volume size	Minimum: 8 GB, Maximum: 1 PB
Maximum file system partitions per volume	15
Maximum total number of volumes and snapshots in system	131,072 If more are needed, contact Dell Technologies Support. (up to 10,000 volumes and snapshots when deploying NVMe over TCP)
Maximum total number of volumes and snapshots in protection domain	32,768
Maximum total number of volumes and snapshots per storage pool	32,768
Maximum number of snapshots in a single vTree	126
Maximum raw capacity per protection domain	8 PB
Maximum raw capacity per SDS	160 TB (medium granularity) 128 TB (fine granularity)
Maximum SDCs per system	2048
Maximum SDSs per system	512 If more are needed, contact Dell Technologies Support.

Table 4. Supported capabilities and system limits (continued)

PowerFlex item	Limit
Maximum SDSs per protection domain	128 If more are needed, contact Dell Technologies Support.
Maximum devices (drives) per SDS server ⓘ **NOTE:** Includes NVDIMM devices.	64 On a VMware server, the maximum devices per SDS is 59.
Maximum devices per protection domain	8192
Maximum devices per storage pool	300 If more are needed, contact Dell Technologies Support.
Total size of all volumes per storage pool	4 PB
Maximum acceleration devices of type NVDIMM (maximum supported in node)	6 An acceleration pool can serve multiple storage pools, but a single storage pool cannot use multiple acceleration pools. An acceleration device can serve multiple SDS devices, but SDS devices cannot span across multiple acceleration devices.
Maximum volumes that can be mapped to a single SDC	ESXi 7.0, 1024 volumes Other operating systems: 1024
System over provisioning factor	5x net capacity per MG layout
Fine-granularity maximum compression	10x raw capacity
Maximum protection domains per system	256
Maximum storage pools per system	1024
Maximum storage pools per protection domain	64
Maximum fault sets per protection domain	64 If more are needed, contact Dell Technologies Support.
RMcache	Minimum: 128 MB, Maximum: 300 GB
Maximum number of snapshots a snapshot policy can be defined to retain per vTree (not including locked snapshots)	60
Maximum snapshot policies per system	1000
Maximum user accounts	256 If more are needed, contact Dell Technologies Support.
Maximum number of concurrent logged-in management clients (GUI/REST/CLI)	128
Maximum volumes that can be mapped via API (scli \REST\WebUI) concurrently	1024
Maximum number of configured syslog servers	16 (must be an even number) Network configuration for NAS controller must include link aggregation (bonding/teaming) using LACP.
Volumes per local Consistency Group (snapshot)	1024
Maximum number of volumes to SDC mappings per system	262,143
SCSI-3 reservation type	Write exclusive registrants only

Table 5. PowerFlex file services limits

PowerFlex item	Limit
Maximum NAS cluster size (number of nodes)	16 (must be an even number) Network configuration for NAS controller must include link aggregation (bonding/teaming) using LACP.
Minimum NAS cluster size (number of nodes)	2
Maximum number of NAS clusters per system	1
Maximum file system size	256 TB
Maximum number of file systems	16,384 (1,024*16N)
Maximum number of NAS servers per system	2,048 (128*16N)
Maximum snaps (mounted) per system	46,080
Maximum number of file systems per NAS server	125
Maximum number of file systems plus mounted snaps per NAS server	1,500
Maximum NAS server network interfaces per NAS server	10 + 2 (backup)
Maximum NAS server network interfaces per node	300
Maximum NDMP concurrent sessions per node	20
Maximum NDMP concurrent sessions per cluster	320 (20*16N)
Maximum CIFS servers per system	2,048
Maximum NFS servers per system	2,048
Maximum CIFS shares per node	10,000
Maximum CIFS shares per system	160,000
Maximum NFS exports per node	5,000
Maximum NFS exports per system	80,000
Maximum tree quotas per file system	8,191
Maximum file names per directory	10 million
Maximum sub-directories/files per directory	10 million
Maximum number of home directories	20,000
Maximum CIFS TCP connections	128,000
Maximum NFS TCP connections	128,000
Maximum TCP connections per system	153,600
Maximum unique ACLs per file system	4 million
Maximum CIFS shares + NFS exports per system	160,000 + 80,000
Maximum directories per file system	> 10 billion
Maximum open files/directories	1,024,000
Maximum files per file system	32 billion
Supported NFS versions	v3/v4/v4.1/v4.2
Supported SMB (CIFS) versions	v1/v2/v3/v3.02/v3.1.1
Supported FTP versions	FTP and SFTP

Table 6. NVMe over TCP limits

PowerFlex item	Limit
Maximum volumes mapped to a single NVMe host (Linux)	1024 ESXi 8.0 and ESXi 7.0u3f or higher
Maximum volumes mapped to a single NVMe host (ESXi)	32 with ESXi 7.0u3f 256 using ESXi 8.0
Maximum NVMe hosts connected to system	1024 ⓘ **NOTE:** These hosts are included in the total number of SDCs per system.
Maximum SDTs per protection domain	128
Minimum SDTs per protection domain	2 Using minimum SDTs may block the ability to reach maximum NVMe hosts.
Maximum SDTs per system	512
Maximum paths in multipathing driver per volume	8 4 in ESXi 7.0u3f
Maximum connections per host per protection domain	16
Maximum NVMe host connections (I/O controllers) per SDT	512
Maximum NVMe host connections (I/O controllers) per system	65,519
Maximum I/O controller queue depth	128
Maximum I/O controller queues	32 ⓘ **NOTE:** Number of queues + queue depth is automatically negotiated on connection.
Maximum volume-to-host mappings (SDC/NVMe) per system	262,143

Table 7. Replication limits

PowerFlex item	Limit
Number of destination systems for replication	4
Maximum SDRs per system	128
Maximum number of Replication Consistency Group (RCG)	1024
Maximum replication pairs in RCG with initial copy	1024
Maximum number of volume pairs per RCG	1024
Maximum volume pairs per system	32,000
Maximum number of remote protection domains	6
Maximum number of copies per RCG	1
Maximum ratio of peer SDRs in a source protection domain to a target protection domain	1:6
Maximum ratio of maintenance SDRs in a source protection domain to a target protection domain	1:18
Recovery Point Objective (RPO)	Minimum: 15 seconds, Maximum: 1 hour
Maximum replicated volume size	64 TB

Table 8. PowerFlex Manager limits

PowerFlex management platform capability	Limit
Maximum number of configured users	20
Maximum number of LDAP users	20
Maximum number of LDAP servers	1
Maximum number of DNS servers	1
Maximum number of time sources	1
Maximum number of Syslog servers	1
Maximum number of SMTP servers	1
Maximum number of vCenters	1

Table 9. Deployment and NDU limits

PowerFlex item	Limit
Maximum latency for MDM network	200 msec
SDS IP addresses	Must be unique. For example, 127.0.0.1 must not be used, as it refers to several machines

System requirements

This section lists the requirements for system components and the minimum PowerFlex Manager resource requirements for each consumption model.

This section is specific to PowerFlex software deployments.

Table 10. Software packages and resource requirements

Package name	Resource requirements
RCM for PowerFlex rack	**vCPU**: 42 (14 per node/MVM) **RAM**: 96 GiB (32 GiB per node/MVM) **Disk space**: 600 GB per node
Intelligent Catalog for PowerFlex appliance	
PowerFlex custom node software zip	
PowerFlex software zip	

For more information on software package names and versions, see the *Dell PowerFlex 4.5.x Release Notes* for the applicable version, which is available on the Dell Technologies Support site.

PowerFlex cluster component requirements

Each PowerFlex cluster includes the following components. Before deploying PowerFlex, ensure that the PowerFlex environment meets the server requirements for the cluster type.

Three-node cluster

- One primary MDM
- One secondary MDM
- One tiebreaker
- Minimum of three SDSs (on the same servers as the above components, or on three different servers)
- SDCs, up to the maximum allowed (on the same servers as the above components, or on different servers)

Five-node cluster

- One primary MDM
- Two secondary MDMs
- Two tiebreakers
- Minimum of three SDSs (on the same servers as the above components, or on three different servers) Dell Technologies recommends five SDSs so that the cluster members reside with the SDSs.
- SDCs, up to the maximum allowed (on the same servers as the above components, or on different servers)

PowerFlex management platform requirements

- PowerFlex rack dedicated management cluster—three-nodes or more
- PowerFlex appliance
  - Option 1: Single node using RAID 5 and ESXi 7.0 operating system
  - Option 2: ESXi cluster with three nodes or five nodes, with vSAN, if there is upgrade (PowerFlex management controller 1.0) or PowerFlex as shared storage (PowerFlex management controller 2.0).
- PowerFlex custom node—one node, dedicated PowerFlex management controller

  NOTE: In the one-node configuration, a hypervisor is required, such as ESXi, to host PowerFlex management platform VMs.
- PowerFlex custom node or PowerFlex software: Co-resident PowerFlex management controller — available for PowerFlex software and custom node offerings.
- PowerFlex software or PowerFlex custom node offerings: Bring your own ESXi environment to host PowerFlex management controller.

Physical server requirements

When installing new nodes, the recommended best practice is to keep nodes with similar I/O characteristics in their own protection domain. Different CPUs, memory speeds, drive capacity, drive types, and network speeds can affect the overall node performance. PowerFlex Manager does not prevent the mixing of node types within a protection domain. The lowest performing system dictates performance, which might not always be the preferred option. When replacing newer nodes, add the new nodes to the existing protection domain, move your workloads and MDM roles (if they exist), and then remove the old nodes from the protection domain. Ensure that you wait for the rebuild or rebalance processes to complete before adding or removing a node. For more information about adding and removing nodes or roles, see the *PowerFlex 4.5.x Administration Guide*.

The following table summarizes the requirements of the physical server.

PowerFlex offers a sizing tool to make the complex work of sizing the physical requirement. Go to the PowerFlex sizer client and review the physical sizing considerations on the page. Dell Technologies recommends using the sizer tool for all and not perform manual sizing.

Table 11. Physical server requirements

Component	Requirement
Processor	One of the following: ● Intel or AMD x86 64-bit (recommended) ● Intel or AMD x86 32-bit (for Citrix Hypervisor (Xen) only) **NOTE:** AMD processors are not supported on VxFlex Ready Node and do not support NVDIMMs. For this reason, there is no support for fine granularity storage pools on AMD based servers. CPU threads: ● MDM: 5 ● SDS: default 8/maximum 12 ● CloudLink: 2 ● LIA: 1 ● SDR: 8 ● NVMe target: 8 ● SDC: default 2/maximum 8 ● NVMe initiator: default 2/max 8

Table 11. Physical server requirements (continued)

Component	Requirement
	NOTE: The recommended best practice is to have a different non-uniform memory access (NUMA) domain than SDR if it is on the same operating system.

Physical memory	PowerFlex component requirements:
	RAM (GiB)
	● SDS FG: 10 GiB + ((100*Number_Of_Drives) + (550 * Total_Drive_Capacity_in_TiB))/1024
	● SDS MG: 5 GiB + (210*Total_Drive_Capacity_in_TiB)/1024
	● SDS NVDIMM for FG: NVDIMM_capacity_in_GiB=((100*Number_Of_Drives) + (700 * Total_Drive_Capacity_in_TiB))/1024
	● MDM: 6.5 GiB
	● CloudLink: 4 GiB
	● Operating system: 1 GiB
	● LIA: 0.35 GiB
	● PowerFlex Installer: 8 GiB
	● SDC: 6 GiB
	● NVMe target: 9 GiB
	● NVMe initiator: 2 GiB
	● SDR: 8586 MiB + 550 MiB * (max remote PDs count)
	NOTE: For certain system configurations, the SDS must be configured to be able to use more than one NUMA domain. The SDS is configured by default to have affinity to socket 0 in a server or VM, and by default, the SDS is connected only to NUMA 0, and only has access to the memory in NUMA 0. If NUMA 0 (usually half the total memory) is less than the memory required by the SDS, you must allow the SDS access to the memory in the other NUMA.
	Non-volatile RAM (NVDIMM):
	● Node storage capacity < 49 TiB : NVDIMM size 32 GiB
	● Node storage capacity 49 TiB to 96 TiB: NVDIMM size 64 GiB
	● Node storage capacity > 96 TiB: NVDIMM size 96 GiB
	NOTE: PowerFlex does not support using SWAP on storage-only and SVM nodes. This should be the default setting:
	● Using SWAP with PowerFlex has the following impact:
	○ Slows the performance of the system
	○ Impacts I/O on the system disks which is used for SWAP; which might cause other functions using the system disk such as to have slow I/Os:
	■ MDM repository related operations
	■ PowerFlex logs written to disk. Refer to the following link on how to disable SWAP. In addition, the settings for vm.overcommit must be set to the following settings on nodes that have MDM and SDS process on them: `in/etc/sysctl.conf` file
	`vm.overcommit_memory=2`
	`vm.overcommit_ratio=100`
	For more information on the implication of this setting, see https://community.pivotal.io/s/article/Linux-Overcommit-Strategies-and-Pivotal-GreenplumGPDBPivotal-HDBHDB?language=en_US

Operating system/boot disk	Disk type	RAID	Endurance	Minimum capacity	Sustained performance IOPS (random 4 KB)	Sustained performance bandwidth (sequential)

Table 11. Physical server requirements (continued)

<table>
<tr><th>Component</th><th colspan="6">Requirement</th></tr>
<tr><td rowspan="2"></td><td>BOSS card</td><td>2*M.2 devices in RAID 1</td><td>0.5 DWPD over 5 years or 200 PB life span writes</td><td>120 GB</td><td>50/20K read/ write</td><td>400/80 [MB/sec] read/write</td></tr>
<tr><td colspan="6">ⓘ <b>NOTE:</b> Due to a limitation, do not install any of the following on the same operating system disk as the MDM: PowerFlex Manager, PowerFlex Installer, and PowerFlex Gateway processes.</td></tr>
<tr><td>Disk space</td><td colspan="6">● Disk space formula per node: 1 GB * number of nodes in the system + 1 GB (for log collection)
○ Required path locations: /opt
○ 1 GB for log collection under /tmp (reused every time a log bundle is created)
● For SDC running on Linux or Citrix Hypervisor (XenServer) — 1 GiB
ⓘ <b>NOTE:</b> The minimum operating system size for a three-node cluster is 10 GiB.
● For VMware ESXi — 64 GiB is the minimum for a boot disk for VMware topologies; the PowerFlex Storage VM resides on the boot datastore.
Limit: PowerFlex Gateway cannot be installed on the same operating system disk as the MDM repository (Dell recommends having them on separate hosts).</td></tr>
<tr><td>Connectivity</td><td colspan="6">One of the following:
● 10/25/40/50/100 GbE network
Dual-port network interface cards (recommended)
Ensure the following:
● There is network connectivity between all components.
● Network bandwidth and latency between all nodes is acceptable, according to application demands.
● Ethernet switch supports the bandwidth between network nodes.
● MTU settings are consistent across all servers and switches.
● The TCP ports are not used by any other application, and are open in the local firewall of the server:
ⓘ <b>NOTE:</b> You can change the default ports.</td></tr>
</table>

ⓘ **NOTE:**

- If buying a PowerEdge base solution with a PowerFlex software license, be sure to select hardware that is identical to the engineered solution (PowerFlex appliance, PowerFlex rack, PowerFlex custom node).
- PowerEdge configurations that differ from the engineered solution are at risk of encountering issues. Aligning to the engineered solution is critically important when selecting drives on an SDS device.
- Agnostic drives should be avoided. Vendor specific drives must be selected in order to ensure the nodes ship with PowerFlex validated drives.
- To ensure PowerFlex software compatibility, it is recommended to configure a solution using a PowerFlex base (PowerFlex appliance, PowerFlex rack, PowerFlex custom node.

Operating system requirements

The following is a list of operating systems supported by PowerFlex.

PowerFlex supports Red Hat Enterprise Linux 9.2 or 8.8, SLES 15.5 and ESXi 8.0 for NVMe over TCP client and NAS client. For more information, see the PowerFlex support matrix.

Table 12. Operating system requirements

Operating system	Requirement
Linux	Packages required for all components, all Linux flavors: ● *numactl* ● *libaio1*

Table 12. Operating system requirements (continued)

Operating system	Requirement
	● *wget* ● *apr* ● *libapr1* ● *python2-rpm or python3-rpm* ● *yum-utils* Additional package required for SDC: ● which Additional packages required for MDM components: ● bash-completion (for SCLI completion) ● Latest version of Python 3.X (for newer OSs such as CentOS\RHEL 8.X) ● binutils ● java-11-openjdk-headless Additional packages needed by SDS: ● smartmontools ● smartmontoolslibs ● binutils ● sg3_utils ● hdparm ● pciutils Additional packages needed by Diag-tools: ● sar ● sysstat For supported storage controllers, the following packages are also needed: ● Dell: perccli ● HP raid controllers: hpssacli Additional packages required for SDS components that contain NVDIMMs: ● ndctl ● jq ● daxio ● libpmem To use the secure authentication mode, ensure that OpenSSL 64-bit v1.1 is installed on all servers in the system. ● `libopenssl1_0_0-1.0.1g-0.40.1.x86_64.rpm` ● `openssl-1.0.1g-0.40.1.x86_64.rpm` To use LDAP, ensure that OpenLDAP 2.4 is installed on all servers. Additional packages needed for LIA: pciutils ● NVMe initiator: Linux nvme_tcp and nvme_fabrics modules must be loaded in the OS. Use `Ismod | grep nvme` to validate that they are available.
Windows	● Only SDC, and LIA are supported components on Windows ● To install SDC, ensure that Microsoft Security Update KB3033929 is installed. To use the secure authentication mode, ensure that these are installed on all servers in the system: ● OpenSSL 64-bit v1.0.1 or later (v1.1, however, is not supported) ● Visual C++ redistributable 2010 package, 64-bit

Table 12. Operating system requirements (continued)

Operating system	Requirement
	The Microsoft Windows "Storage Spaces" feature (available from Windows 8.1 and later), is not supported in Microsoft Failover cluster running on PowerFlex volumes.
VMware	Compute-only (SDC) support: VMware ESXi operating systems: 8.0 or 7.0 (SDC only). The latest patches of ESXi and vCenter are tested periodically. For the most updated list, see the PowerFlex support matrix .

Storage virtual memory requirements

Table 13. Storage virtual memory (SVM) requirements

Component	Memory allocation rules
Memory	PowerFlex RAM (GiB) ● SDS FG: 10 GiB + 550 MiB per 1 TiB device size + 100 MiB per device ● SDS MG: 5 GiB + 335 MiB per 1TiB device size ● MD Read Cache: Total_Drive_Capacity_in_TiB * 8% ● MDM: 6.5 GiB ● CloudLink: 4 GiB ● Operating System: 1 GiB ● LIA: 0.35 GiB ● PowerFlex Installer/PowerFlex Gateway: 8 GiB ● SDC: 6 GiB ● SDR: 8586 MiB + 550 MiB * (max remote PDs count) **NOTE:** SDR SVM support is only available as part of a PowerFlex appliance or the PowerFlex rack managed by PowerFlex Manager. ● SDR memory requirement for hyperconverged systems (SVMs): 12GB **NOTE:** For certain system configurations, the SDS needs to be configured to be able to use more than one non-uniform memory access (NUMA) domain. The SDS is configured by default to have affinity to socket 0 in a server or VM, and therefore, by default, the SDS is connected only to NUMA 0, and only has access to the memory in NUMA 0. If NUMA 0 (usually half the total memory) is less than the memory required by the SDS, you need to allow the SDS access to the memory in the other NUMA. Non-Volatile RAM (NVDIMM): ● Node storage capacity < 49 TiB: NVDIMM size 32 GiB ● Node storage capacity 49 to 96 TiB: NVDIMM size 64 GiB ● Node storage capacity > 96 TiB: NVDIMM size 96 GiB **NOTE:** PowerFlex does not support using swap on Storage only and SVM nodes. This should be the default setting. Using swap with PowerFlex has the following impact: 1. Slows the performance of the system. 2. Impacts I/O on the system disks which is used for swap. This may cause other functions that use the system disk to have slow I/Os such as: a. MDM repository related operations b. PowerFlex logs written to disk.

Table 13. Storage virtual memory (SVM) requirements (continued)

Component	Memory allocation rules

	Disable swap and set `vm.overcommit` to the following on nodes that have MDM and SDS process on them in the `in/etc/sysctl.conf` file:
	`vm.overcommit_memory=2`
	`vm.overcommit_ratio=100`

Base SVM	350 MiB
Tiebreaker MDM	50 MiB

Browser support

PowerFlex supports Google Chrome, Mozilla Firefox, Safari, and Microsoft Edge.

Other requirements

PowerFlex requires that you use a minimum of three SDS servers, with a combined free capacity of at least 720 GB. These minimum values are true per system and per storage pool.

PowerFlex supports CloudLink version 7.1.8, a device encryption software.

5

Architecture

This section presents the main components of the PowerFlex architecture.

System

PowerFlex is based on a configuration of hardware and a software components working together.

Hardware

Hardware can be the existing application servers used by the data center, or a new set of nodes (if, for example, you want to dedicate all nodes solely for running the PowerFlex SAN storage system).

The PowerFlex rack system includes a full data center solution including: rack, switches, server nodes, power, and cabling. Built for capacity, performance, or compute per the specific configuration.

- Servers are the basic computer units used to install and run PowerFlex. They can be the same servers used for the applications (server convergence), or a dedicated cluster. In any case, PowerFlex is hardware-agnostic so aside from performance considerations, the type of server is inconsequential.
- The storage media can be any storage media, in terms of the type (HDD, SSD, NVMe SSD or PCIe flash cards) and anywhere (external).

Support for Dell PowerEdge R660, R760, R6625, and R7625 servers

PowerFlex 4.5 certifies support of the following platforms for software only and custom node offerings:

- Dell PowerEdge R660 and Dell PowerEdge R760 servers (Intel based)
  - Available with a new persistent memory solution, Software Defined Persistent Memory (SDPM)
- Dell PowerEdge R6625 and Dell PowerEdge R7625 servers (AMD based)

Software

The PowerFlex virtual SAN consists of the following software components:

- Meta data manager (MDM): Configures and monitors PowerFlex. The MDM can be configured in redundant cluster mode, with three members on three servers or five members on five servers, or in single mode on a single server.

  > **NOTE:** It is not recommended to use single mode in production systems, except in temporary situations. The MDMs contains all the metadata required for system operation. Single mode has no protection, and exposes the system to a single point of failure.
- Storage data server (SDS): Manages the capacity of a single server and acts as a back-end for data access. The SDS is installed on all servers contributing storage devices to PowerFlex. These devices are accessed through the SDS.
- Storage data client (SDC): A lightweight device driver that exposes PowerFlex volumes as block devices to the application that resides on the same server on which the SDC is installed.
- Storage data replication (SDR): A component that acts as a proxy between the SDS and SDC that must be installed for replication.
- NVMe target, also known as storage data target: Manages the translation of I/O from the internal PowerFlex protocol to the NVMe over TCP standard.
- NVMe initiator: NVMe driver that connects to the NVMe over TCP storage system to consume storage, and uses an NVMe multipathing driver for high availability and load balancing.

Depending on the preferred configuration, the software components are installed on the server and create a virtual SAN layer exposed to the applications that reside on the servers.

MDM cluster

The MDM is the monitoring and configuration agent of PowerFlex and is used mainly for management. A multi-MDM environment consists of one primary MDM, while the others function as secondary or tiebreaker.

The MDM is used for management which consists of migration, rebuilds, and all system-related functions. No I/Os run through the MDM.

To support high availability, three or five instances of MDM run on different servers. In a multi-MDM environment, one MDM is given the primary role, and the others act as secondary or tiebreaker MDMs.

The MDM cluster consists of a combination of primary MDM, secondary MDMs, and tiebreaker MDMs. There is also the standby MDM which is not a part of the cluster.

- MDM: An MDM is any server with the MDM package installed and it can be given a manager or a tiebreaker (default) role, during installation. This role cannot be changed later without reinstalling the MDM. MDMs have a unique MDM ID, and can be given unique names. Before the MDM can be part of the cluster, it must be promoted to a standby MDM.
- Standby MDM and tiebreaker: An MDM and a tiebreaker can be added to a system as a standby. Once added, the standby MDM or tiebreaker is attached, or locked, to that specific system. A standby MDM can be called on to assume the position of a manager MDM or tiebreaker MDM according to how it is installed, when it is promoted to be a cluster member.
- Manager MDM: An MDM that can act as a primary or a secondary in the cluster. Manager MDMs have a unique system ID, and can be given unique names. A manager can be a standby or a member of the cluster.
- Tiebreaker MDM: An MDM whose sole role is to help determine which MDM is the primary. A tiebreaker can be a standby or a member of the cluster. A tiebreaker MDM is not a manager. In a three-node cluster, there is one tiebreaker; in a five-node cluster, there are two tiebreakers. This ensures that there are always an odd number of MDMs in a cluster, which guarantees that there is always a majority in electing the primary MDM.

The following terms are relevant to the MDM cluster, specifically:

- Primary MDM: The MDM in the cluster that controls the SDSs and SDCs. The primary MDM contains and updates the MDM repository, the database that stores the SDS configuration, and how data is distributed between the SDSs in the system. This repository is constantly replicated to the secondary MDMs, so they can take over with no delay. Every MDM cluster has one primary MDM.
- Secondary MDM: An MDM in the cluster that is ready to take over the primary MDM role if necessary. In a three-node cluster, there is one secondary MDM, thus allowing for a single point of failure. In a five-node cluster, there are two secondary MDMs, thus allowing for two points of failure. This increased resiliency is a major benefit to enabling the five-node cluster.
- Replica: An MDM that contains a replica of the MDM repository. This includes the primary MDM and any secondary MDMs in the MDM cluster.

The following table describes the available cluster modes:

Table 14. MDM cluster modes

Cluster mode	Members	Description
Three-node (default)	● Primary MDM ● Secondary MDM ● Tiebreaker	Three-node cluster has two copies of the repository, thus can withstand one MDM cluster failure.
Five-node	● Primary MDM ● Two secondary MDM ● Two tiebreakers	Five-node cluster has three copies of the repository, thus can withstand two MDM cluster failure.
Single-node	Primary MDM	Single-node cluster has only one copy of the repository, thus it cannot withstand failure. It is not recommended to use single-node in production systems.

In addition to the cluster members, you can prepare standby managers and tiebreaker nodes, for a total of thirteen cluster and standby MDMs.

The MDM cluster IP address limit is 16 IP addresses, which includes all cluster members (primary, secondary, standby primary, and standby secondary).

The following figure illustrates a five-node MDM cluster:

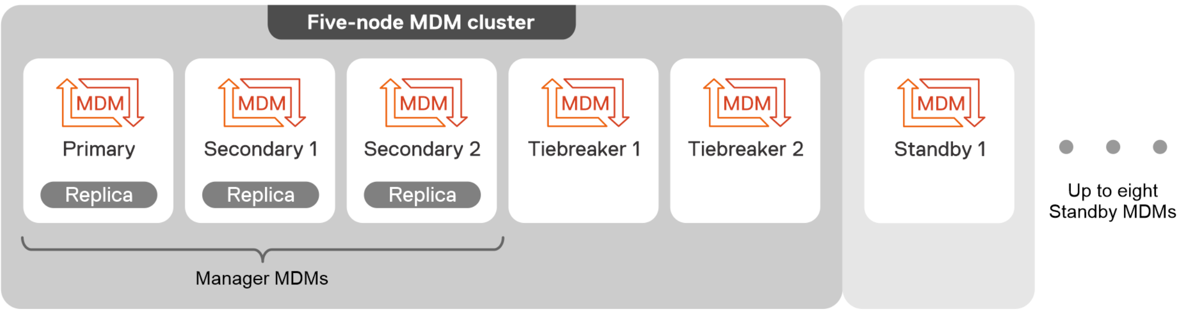


Figure 1. Five-node MDM cluster

All members of the MDM cluster have the same MDM package installed on them.

Before a server makes its way into the MDM cluster, a server must:

1. Install the MDM package on the server.

   During the installation, you determine if the server will be a manager or a tiebreaker (default).

2. Promote the server to standby status, either as a manager or as a tiebreaker according to how it was installed.
3. Add the standby server to the MDM cluster. A manager, once entered into the cluster, can take on the primary or secondary state.

MDM cluster creation is done automatically when deploying a system with any of the automated deployment tools.

The following list includes best practices for adding MDMs to a multiple-rack system:

- For a large system with multiple racks, the MDM cluster members should be distributed in separate racks.
- The MDM network requirements described in the "Networking" section, must also apply to the network connections between the racks.
- You can ensure higher resiliency when the rack is shut down for planned or unplanned reasons, especially if fault sets are used. For more informaation, see section on "Fault sets".
- For simultaneous upgrade of multiple nodes in different racks, it is important to take note of the nodes which have MDMs. Do not simultaneously reboot or shut down the nodes which have MDMs. To ensure users do not shut down the MDMs by mistake, mark them physically with stickers saying "Do not shut down: MDM installed here", or add a meaningful prefix\suffix to the hostname of the server where MDM is installed. For example: *Rack22U32-MDM* You can also make sure a different node in the rack is used per server as shown in the table:

Table 15. Nodes in rack

Node level	Rack number	MDM role
Top node	1	MDM
Bottom node	2	MDM
Middle node	3	TB

System stability

The MDM proxy ability augments the network stability capabilities.

In the PowerFlex architecture, each SDC is required to be fully connected to the MDM for control and management of actions. If the SDC is unable to reach the MDM, the SDC will move to the disconnected state, which at some point could lead to complete I/O failure.

PowerFlex allows clusters that don't use the MDM virtual IP capability to use the secondary MDMs as a proxy for all control traffic, allowing more redundancy in case of network failures.

When such an occurrence occurs and the MDM proxy is in use, the management APIs raise an appropriate alert.

Virtual IP addresses

Virtual IP addresses can be defined for the MDM cluster.

Existing systems may be extended to include additional MDMs to a cluster. The new MDMs should be mapped to the existing virtual IP addresses.

All SDCs will require reconfiguration, to reflect the changes made to the MDM cluster. Otherwise, the SDCs will not be able to communicate with the MDM cluster, and volumes will not be accessible.

A maximum of four MDM virtual IP networks are supported.

Storage definitions

When configuring PowerFlex, you should take the following concepts into account: protection domains, storage pools, and fault sets. Together, these elements link the physical layer to the virtual storage layer.

Protection domains

A protection domain is a logical entity that consists of a group of SDSs that back up each other. Protection domains can be added during installation and modified post-installation.

Each SDS belongs to one protection domain, so each protection domain is a unique set of SDSs.

The maximum recommended number of nodes in a protection domain is 128. This enables the following:

- Optimal performance
- Reduction of theoretical *mean time between failure* issues
- Ability to sustain multiple failures in different protection domains

You can add protection domains during installation. In addition, you can modify protection domains post-installation with all the management clients.

Fault sets

A fault set defines a group of SDSs that are likely to go offline simultaneously. Defining a fault set configures PowerFlex to mirror the devices in the set.

A fault set is a logical entity that contains a group of SDSs within a protection domain that have a higher chance of going offline together such as a group of SDSs that are all powered in the same rack. By grouping them into a fault set, you are configuring PowerFlex to mirror the data for all devices in this fault set. The mirroring should take place on SDSs that are outside of this fault set.

When defining fault sets, the term *fault units* refers to either a fault set, or an SDS not associated with a fault set.

There must be enough capacity within at least three fault units to enable mirroring.

If fault sets are defined, you can use any combination of fault units, for example:

- SDS1, SDS2, SDS3
- FS1, SDS1, SDS2
- FS1, FS2, SDS1
- FS1, FS2, FS3

In the following figure there are three protection domains. The middle one (fully depicted) consists of seven SDSs, each with two storage devices.

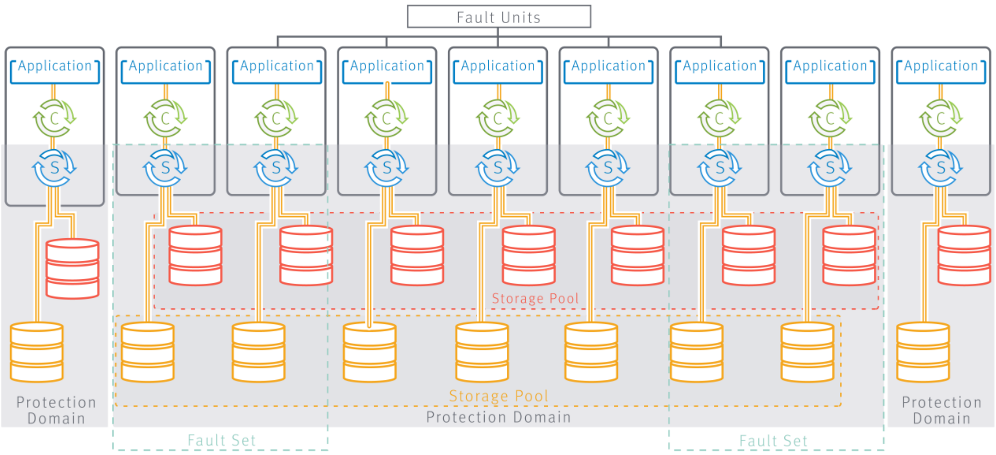


Figure 2. Protection domains, storage pools, and fault sets

To use fault sets, you must work in the following order:

1. Ensure that a protection domain exists, or add a new one.
2. Ensure that a storage pool and fault sets (minimum of three fault units) exist, or add new ones.
3. Add the SDS, designating the protection domain and fault set, and at the same time, adding the SDS devices into a storage pool.

   The automated deployment and installation tools follow this order automatically.

   You can only create and configure fault sets before adding SDSs to the system, and configuring them incorrectly may prevent the creation of volumes. An SDS can be added to a fault set only during the creation of the SDS.

You can also add fault sets when adding SDS nodes after initial installation.

Larger PowerFlex storage-only nodes

SDS maximum is 160 TB (node) on medium granularity storage pools.

- SSD maximum is 15.36 TB (SSD drives)
- Allows higher density systems with larger capacity storage nodes (SDS)
- Initially limited to medium granularity data layout (MG) storage pools only (no expansion)

Storage pools

A storage pool is a set of physical storage devices in a protection domain. A volume is distributed over all devices residing in the same storage pool. You can define a magnetic storage pool (for HDDs) or a high performance storage pool (for SSDs). Storage pools support medium or fine granularity data layouts and allow enabling or disabling zero padding.

Storage pools allow the managing of different storage tiers in PowerFlex. Each storage device belongs to one (and only one) storage pool. The figure shows two storage pools.

When a volume is configured over the virtualization layer, it is distributed over all devices residing in the same storage pool. Each volume block has two copies on two different fault units. This allows the system to maintain data availability following a single-point failure. Two network failures render all the domain in a state where I/O errors are possible on any storage pool.

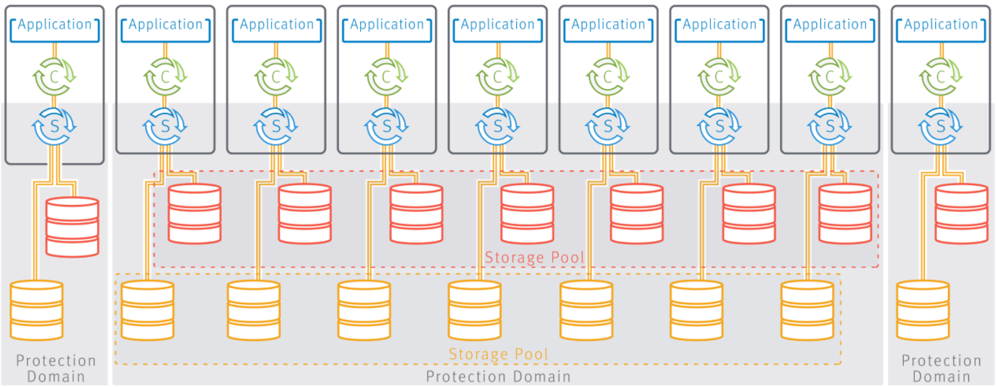


Figure 3. Protection domains and storage pools

You must assign a media type setting to each storage pool. Supported types are: HDD, SSD, and Transitional (allows for migration flows).

If all SDSs in a protection domain have two physical drives associated with them, such as one hard drive, and the other SSD, you should define two storage pools:

- Magnetic storage pool

  Consists of all HDDs in the protection domain

- High performance storage pool

  Consists of all SSDs used for storage purposes in the protection domain

  NOTE: Mixing different types of SSDs is not recommended. Creating a separate storage pool for each type is the recommended best practice, for example: SAS SSD, SATA SSD, NVMe SAS SSD.

  NOTE: PowerFlex might not perform optimally if there are large differences between the sizes of the fault units in the same storage pool. For example, if one device has a much larger capacity than the rest of the devices, performance may be affected. After adding devices, you can define how much of the device capacity is available to PowerFlex by using the SCLI `modify_sds_device_capacity` command.

Storage pools support the following data layouts for HDD or SSD media:

- Medium granularity (MG): Space allocation occurs at 1 MB units.
  - Supports HDD and SSD media
  - Includes persistent checksum for data integrity
- Fine granularity (FG):
  - Requires SSD media and NVDIMM for acceleration
  - Space allocation occurs at 4 KB units
  - Includes persistent checksum for data integrity
  - Supports data compression which reduces the size of data that is stored on the disk
  - Supports thin-provisioned, zero-padded volumes

NOTE: FG and MG storage pools can both exist in a single SDS. You can also migrate volumes across the two layouts.

Each storage pool can work in one of the following modes:

- Zero-padding enabled

  Ensures that every read from an area previously not written to returns zeros. Some applications might depend on this behavior. Furthermore, zero padding ensures that reading from a volume will not return information that was previously deleted from the volume.

  This behavior incurs some performance overhead on the first write to every area of the volume since the area must be filled with zeros first.

  FG is always zero padded.

- Zero-padding disabled (default only for MG)

  A read from an area previously not written to will return unknown content. This content might change on subsequent reads.

Some applications assume that when reading from areas not written to before, the storage will return zeros or consistent data. If you plan to use such applications, then zero padding must be enabled.

You can add storage pools during installation. In addition, you can modify storage pools post-installation with most of the management clients.

NOTE: The zero padding policy cannot be changed after the addition of the first device to a specific storage pool.

Acceleration pools

Acceleration pools enable NVRAM required for fine granularity (FG) storage pools.

NVRAM acceleration pools features and requirements:

- Must reside on NVDIMM devices
- The NVDIMM must be mapped to a DAX Device
- NVRAM acceleration pools cannot be disabled
- Migration to a different NVRAM pool is not supported

Limitation: NVDIMM failure on SDS node causes failure of all SSDs in the FG storage pool for that node.

Data layout

PowerFlex offers a space efficient storage layout with fine granularity (FG).

With FG, since space allocation occurs at 4 KB, the data stored on the disk is significantly reduced. Space allocation for medium granularity (MG) occurs at 1 MB.

For FG, you must have NVDIMM configured as a DAX device. After NVDIMM is all set up in your environment, you need to configure an NVDIMM acceleration pool and then you can select it in the storage pool.

Data-compression is not supported for MG. It supports thin provisioned, zero padded volumes. Zero padded is enabled by default for FG and cannot be disabled. Zero padding is disabled for MG.

FG supports persistent checksum for data integrity.

Volumes

You must add and configure volumes before enabling applications to access them.

The adding and mapping volume process is necessary, as part of the getting started process, before applications can access the volumes. In addition, you may create additional volumes and map them as part of the maintenance of the virtualization layer.

You can configure the caching option when creating the volumes, or you can change the Read RAM Cache feature later. If you want to enable the caching feature, ensure that the feature is also enabled in the backend of the system, for the corresponding storage pool and SDSs.

Define volume names according to the following rules:

- Contains less than 32 characters
- Contains only alphanumeric and punctuation characters
- Is unique within the object type

PowerFlex objects are assigned a unique ID that can be used to identify the object in CLI commands. You can retrieve the ID via a query, or via the property sheet for the object in PowerFlex Manager. It is highly recommended to give each volume a meaningful name associated with its operational role.

Naming PowerFlex objects

Provide meaningful names to all PowerFlex objects to make it easier to perform actions such as defining volumes and associating them with applications.

Storage pools can be named "Capacity_Storage" and "Performance_Storage," which allows you to identify the different tiers.

For protection domains, one example would be separating the SDSs used by the finance department from SDSs used by the engineering department. This separation of different departments is beneficial in many aspects (security being one of them). Thus, you might name the domains "Financial-PD" and "Engineering-PD."

The fault sets could be called "FS_Rack01" and "FS_Rack02."

NVMe over TCP overview

PowerFlex supports the NVMe over TCP storage protocol for front-end connectivity. With NVMe over TCP, PowerFlex provides a simple, optimized, and cost-effective way to consume storage.

The following are the benefits of using NVMe over Fabrics (NVMe-oF):

- NVMe-oF is the new standard protocol for storage, with a growing adoption in the industry.
- NVMe-oF is supported by multiple operating systems and does not require any host agent on the storage.

The following are the benefits of using TCP transport:

- TCP provides a standard, common, and interoperable solution.
- TCP allows the use of existing networks and hardware.
- TCP delivers excellent performance.

Hosts can use standard NVMe over TCP operating system drivers to connect to PowerFlex. These are available in mainstream Linux distributions and in ESXi.

The following figures show how PowerFlex SDC and NVMe over TCP are used on the host. NVMe over TCP uses a newly created NVMe target, the SDT. The SDT runs on the storage side while there is no PowerFlex component running on the host side.

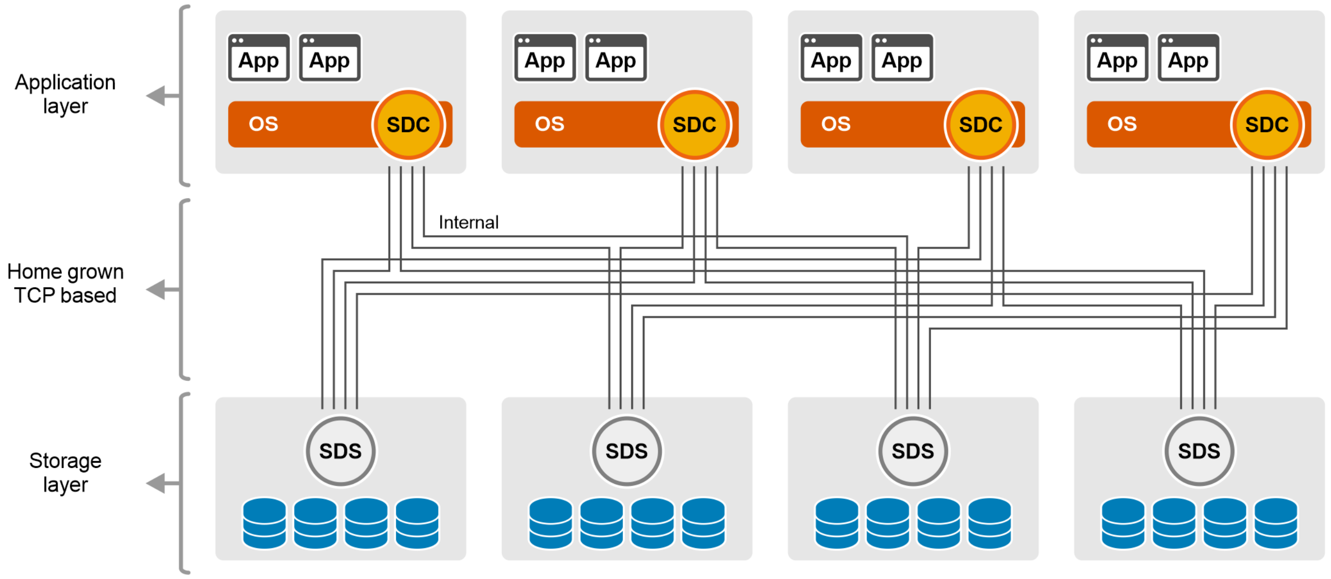


Figure 4. SDC host connectivity

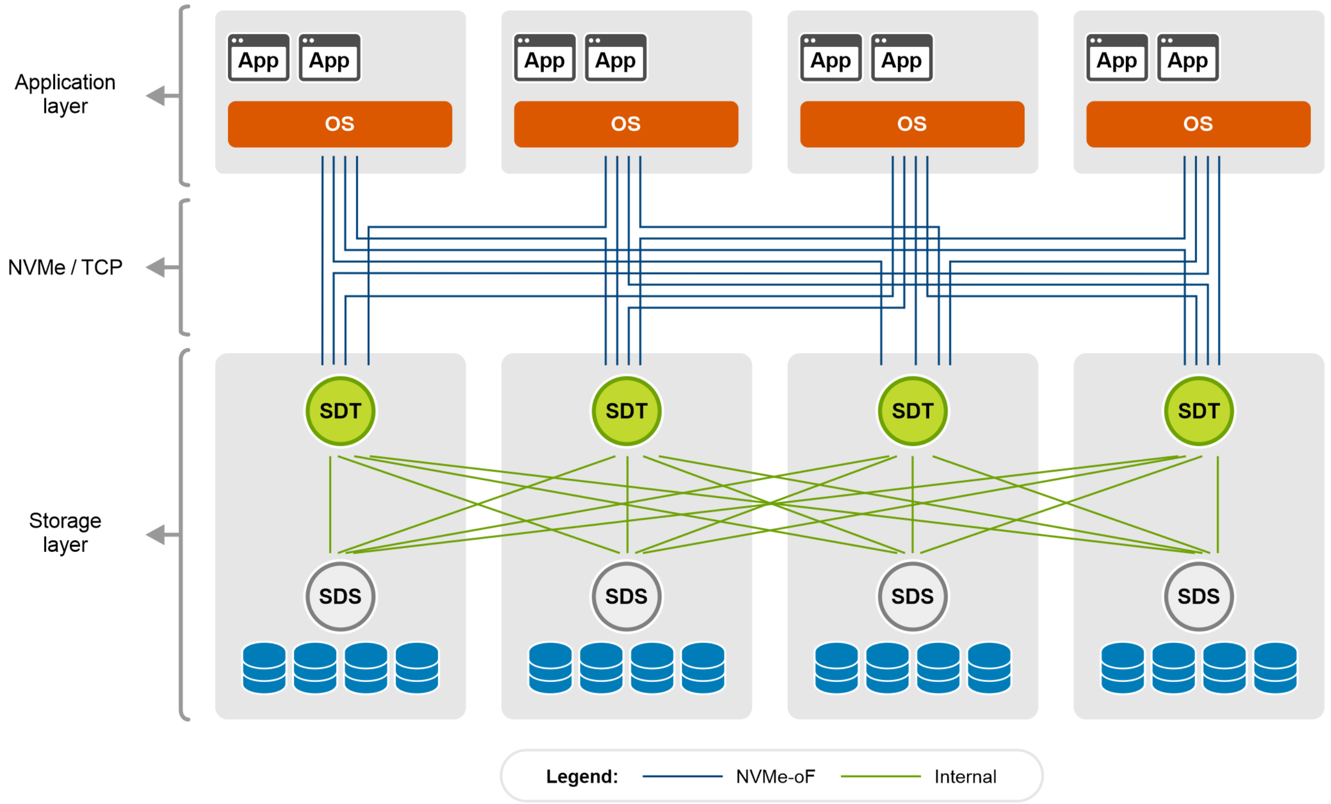


Figure 5. NVMe over TCP host connectivity

NVMe over TCP deployment with PowerFlex enables you to:

- Deploy using PowerFlex Manager (storage nodes)
- Use two-layer deployment
- Work with native operating system drivers and does not require proprietary host agents
- Use the following supported operating systems:
    - ESXi 7.0U 3f and ESXi 8.0
    - Mainstream Linux support for NVMe over TCP is in the tech preview stage (prior to release).

  For the most updated list, see thePowerFlex support matrix.
- Co-exist with PowerFlex SDC. SDC and NVMe over TCP can be used on the same cluster.

  The only exception is that both SDC and NVMe hosts cannot share the same volume at the same time.

PowerFlex supports NVMe reservations (SCSI-3 equivalent) with NVMe over TCP.

NVMe hosts

PowerFlex supports connecting and consuming storage from NVMe hosts. An NVMe host is a host running an operating system that supports NVMe over TCP.

PowerFlex supports connection to storage using SDC or NVMe hosts. Both are supported provided they do not use the same volume.

Once an NVMe host is created in PowerFlex, volumes can be mapped to that host. From that point, the NVMe host can connect to and consume storage.

The following host parameters are configurable for:

- Number of connections

  For hosts that are expected to generate high I/O load, you may increase the default value.
- Number of paths per volume

  Configurable host attribute is up to eight paths. This is mostly an operating system limit per volume and per total paths.

PowerFlex supports specifying networks that might fail together. NVMe over TCP connections are allocated to the NVMe over TCP hosts through multiple network sets, which offer resiliency to the failure of a network set.

NVMe over TCP supports fault sets. Connections allocated to the NVMe over TCP host are through targets on multiple fault sets that provide path resiliency to a fault set failure.

NVMe targets

PowerFlex uses a storage data target to expose NVMe over TCP targets.

The storage data target is deployed with the SDS on each storage server and provides access to the volumes inside that protection domain.

In PowerFlex, every storage data target provides a discovery service. The information returned to the host includes the ports assigned to that host, either by the automatic load balancing or by manual connectivity configuration.

Persistent discovery

PowerFlex supports persistent discovery with the NVMe over TCP target. When hosts are initially discovered through the persistent discovery controller, they remain connected to the discovery service.

When there is a change in the discovery information that is provided to the host, the discovery controller returns an Asynchronous Event Notification and the host requests the updated discovery log page.

Examples of changes in discovery information are the following:

- A new volume is mapped to the host from a new protection domain.
- A new storage data target is added to the system.
- Load balancing has moved the host connection from one storage data target to another storage data target.

When using this functionality, confirm that the host operating system supports the persistent discovery controller and that it is enabled for the host. This feature is supported with ESXi 8.0 and Red Hat Enterprise Linux 9, SLES 15.4, and others. For more information, see the respective operating system for the NVMe over TCP host configuration.

NVMe over TCP host connects to PowerFlex through persistent discovery, which enables automatic updates of the hosts regarding storage side changes.

Load balancing

PowerFlex supports NVMe over TCP connectivity and adds NVMe-oF targets to enable hosts to connect to storage. When hosts are allowed to connect and consume storage, the system ensures that it remains resilient, balanced, and available.

PowerFlex NVMe over TCP load balancer

Utilize the optimal PowerFlex system performance by distributing the NVMe host workloads over the system NVMe targets, for example storage data targets. PowerFlex offers two ways to achieve load distribution: Manual and Automatic. In both cases, system ports are selected for the host and are returned to the host when it performs NVMe discovery.

The automatic load balancing, which is the default policy uses an algorithm that selects the system ports based on the following:

- Access to the mapped volumes of the host
- Path resiliency
- Load balancing

The host connectivity plan is recalculated under the following conditions:

- Host connects.
- Storage data target is added or removed.
- Network set or system network is modified.
- Storage data target is added or removed to a fault set.
- Storage data target IP is reassigned to a different network.

> ⓘ **NOTE:** The load balancer ignores temporary issues such as network failures or if a storage node goes down for a few minutes.

The storage system provides multiple ports through which the host can connect, though every port and node as a limited throughput. Achieve the best performance by ensuring that the combined I/O load of the hosts is distributed well over the available system ports and nodes. The NVMe over TCP load balancer aims to automate the planning and performing of the host-to-storage connectivity task to achieve path resiliency and workload balance.

PowerFlex load balancer considers the host networks for resiliency. When a host is losing a network, it can still use other paths through available networks to access the storage.

NVMe-oF discovery

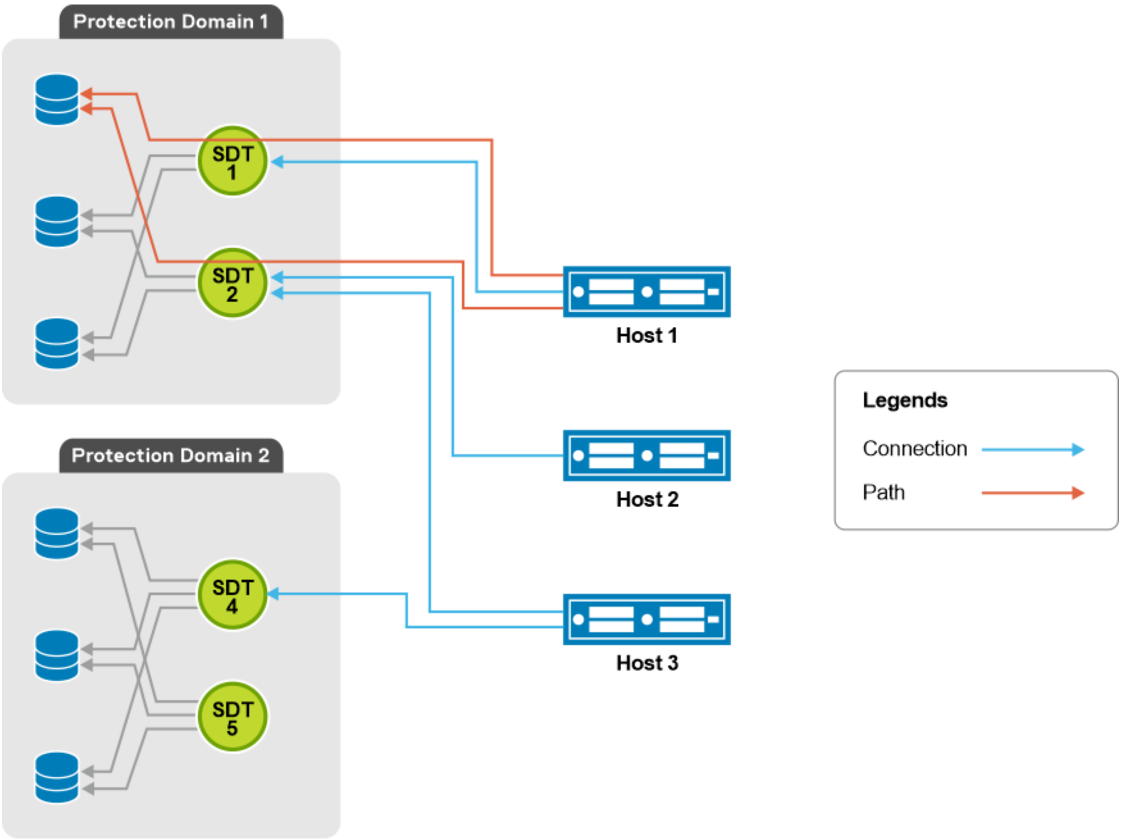
NVMe over Fabric includes a discovery protocol which is a standard method for the host to discover and connect to available subsystems.

The host connects to a discovery service and receives discovery information including all the connection ports to access its volumes. PowerFlex continues to update and host regarding changes using the persistent discovery controller. In PowerFlex, every storage data target provides a discovery service. The information returned to the host includes the ports that are assigned to that host by the load balancer. The host can also be configured manually If the user opted for the manual connectivity configuration.

Balancing host connections

When a host is mapped to the first volume from a protection domain, it assigns ports which might connect to volumes in the protection domain. Each port represents a connection. The number of ports that will be assigned is a configurable host attribute, which is up to 128 per protection domain. The system selects the least occupied ports to ensure fault isolation. The number of paths per volume is a configurable host attribute, which is up to eight. This configurable host attribute is mostly an operating system limit per volume and total paths.

NVMe over TCP connections and paths



A connection is the establishment of the NVMe controller and includes the following:

- Link between a host part and a specific storage data target IP
- Connection couple (host-port and system-port)

- The load balancer provides 10 ports for each host by default and up to 128 ports

A path is a connection that is used to provide access to a volume. This connection makes the path a tuple (host-port, system-port, and volume), which must be a subset of the connections.

By default, the load balancer provides four paths per volume to each host, which is selected from the assigned ports.

Manual load balancing

Users might prefer to manually configure host connectivity. Manual configuration is appropriate for special cases, such as when hosts generate a higher-than-expected traffic load.

To configure manually, the user designates in the host configuration the system ports to which the host must connect.

Migration to NVMe over TCP

PowerFlex provides the option to migrate workloads from SDC to NVMe over TCP on ESXi:

1. Online migration using Storage vMotion (VMFS only)
   - The standard way of moving storage is with Storage vMotion. Storage vMotion also supports switching protocols by migrating to a new DataStore. For more information, see the *Dell PowerFlex 4.5.x Administration Guide*.
2. Offline migration (VMFS only)
   - There is a new option for converting an existing VMFS datastore from SCSI (SDC) to NVMe over TCP without having all the data over the network. The offline migration steps are covered in the following KB article: https://www.dell.com/support/kbdoc/en-us/000213232.

NOTE: There are not standard ways to migrate Linux environments, ESXi clusters, and RDMs.

PowerFlex file services

File system storage

A file system represents a set of storage resources that provide network file storage.

The storage system establishes a file system that Windows users or Linux/UNIX hosts can connect to and use for file-based storage. Users access a file system through its shares, which draw from the total storage that is allocated to the file system.

Table 16. File system storage components

Component	Description
NAS server	A file server configured with its network interfaces and other settings exclusively exporting the set of specified file systems through mount points called shares. Client systems connect to a NAS server on the storage system to get access to the file system shares. A NAS server can have more than one file system, but each file system can only be associated with one NAS server.
File system	A manageable container for file-based storage that is associated with the following properties: ● A specific quantity of storage. ● A particular file access protocol (SMB, NFS, or multiprotocol). ● One or more shares (through which network hosts or users can access shared files or folders).
Share or export	A mountable access point to file system storage that network users and hosts can use for file-based storage.
Windows users or Linux/UNIX hosts	A user, host, netgroup, or subnet that has access to the share and can mount or map the resource. For Windows file systems, access to the share is based on share permissions and ACLs that assign privileges to objects defined in Active Directory. For Linux/UNIX file systems, access is permitted based on NFS access settings.

Shares and exports

Shares represent mount points through which users or hosts can access file system resources. Each share is associated with a single file system and inherits the file system protocol (SMB or NFS) established for that file system. Shares of a multiprotocol file system can be either SMB or NFS.

Access to shares is determined depending on the type of file system:

- Windows (SMB) shares: Access is controlled by SMB share permissions and the ACLs on the shared directories and files. For example, you can configure share permissions using the Microsoft Computer Management utility.
  - Active Directory SMB servers: Configure access for users and groups using Windows directory access controls. User/group authentication is performed through Active Directory.
  - Stand-alone SMB servers: Manage a stand-alone SMB server within a workgroup from a Microsoft Windows host.
- Linux/UNIX (NFS) exports: Hosts access is defined by the NFS access control settings of the NFS export. Use PowerFlex to configure access for individual Linux/UNIX hosts or IP address subnets.

All shares within a single file system draw from the total quantity of storage allocated for the file system. Consequently, storage space for shares is managed at the file system level.

NAS servers

NAS servers provide access to file systems. Each NAS server supports Windows (SMB) file systems, Linux/UNIX (NFS) exports, or both. To provide isolated access to a file system, you can configure a NAS server to function as independent file server with server-specific DNS, NIS, and other settings. The IP address of the NAS server provides part of the mount point that users and hosts use to map to the file system storage resource, with the share name providing the rest. Each NAS server exposes its own set of file systems through the file system share, either SMB or NFS.

Once a NAS server is running, you can create and manage file systems and shares on that NAS server.

NOTE: You can create a file system only if there is a NAS server running on the storage system. The types of file systems that you can create are determined by the file sharing protocols (SMB, NFS, or multiprotocol) enabled for the NAS server.

NAS server physical requirements

PowerFlex file nodes can be configured as small, medium, or large, according to the following specifications.

Table 17. NAS server physical requirements

Configuration	Cores	RAM	NICs	Storage (PERC 755)
Small	2 * 12 (24)	128 GiB	4 * 25G	BOSS 2 * 480 GiB M.2
Medium	2 * 16 (32)	256 GiB	4 * 25G	BOSS 2 * 480 GiB M.2
Large	2 * 28 (56)	256 GiB	4 * 25G/ 4 * 100G	BOSS 2 * 480 GiB M.2

CPU

- 2 * Intel® Xeon® Gold 5317 2.8G, 12C, 150W
- 2 * Intel® Xeon® Gold 6346 3.1G, 16C, 205W
- 2 * Intel® Xeon® Gold 6348 2.6G, 28C, 235W

Memory

- 128 GiB: 16 * 8 GiB RDIMM, 3200 MT/s, single rank
- 256 GiB: 16 * 16 GiB RDIMM, 3200 MT/s, dual rank

Network

The requirements are similar when customers use bring-your-own hardware.

PowerFlex file services support

- Support for PowerFlex file services deployments for PowerFlex rack and PowerFlex appliance

    - Ability to deploy new PowerFlex file services resource groups and import existing PowerFlex file services resource groups for PowerFlex rack and PowerFlex appliance. This enables the capabilities of the **File** menu.
    - Deployment of compute-only resource groups for PowerFlex file services. Before deploying a PowerFlex file services compute-only resource group, deploy a PowerFlex cluster with a hyperconverged or storage-only resource group that uses the block legacy PowerFlex gateway.
    - PowerFlex file services template settings.
    - PowerFlex File sample template, which can be cloned to generate a template that has the correct settings.

  PowerFlex Manager configures the required network on the PowerFlex node and top-of-rack (access) server-facing ports, and the required network on the control plane. When you deploy a PowerFlex file services template, PowerFlex Manager automatically creates control volumes for the deployment.
- Support for new PowerFlex file services deployments for software-only management
    - Ability to deploy new PowerFlex file services resource groups for software management environments. PowerFlex Manager does not support importing existing deployments for software management PowerFlex file services compute-only environments.
    - PowerFlex file services template settings, and support for resource group operations such as node expansion, node removal, and maintenance mode.
    - PowerFlex File - SW Only sample template, which can be cloned to generate a template that has the correct settings.

      NOTE: Installation, network configuration, and setup of the host operating system are prerequisites that the system admin must perform before running the PowerFlex file services deployment from PowerFlex Manager.
- Support for PowerFlex file services operations for PowerFlex rack and PowerFlex appliance
    - Ability to expand the number of nodes, remove nodes from the resource group, and put nodes in maintenance mode. If necessary, you can also remove the resource group as a whole.
    - Ability to expand the cluster up to a maximum of 16 nodes. PowerFlex Manager requires a minimum of two nodes in a PowerFlex file services cluster.
    - Ability to remove PowerFlex file services nodes from the resource group, but not actually delete them. When you remove a resource, it PowerFlex Manager removes the deployment information, but does not make configuration changes to the resource itself. You can remove a PowerFlex file services resource group, which only removes the resource group from the appliance. The cluster still exists and can be added back to the appliance.
- Support for PowerFlex file services serviceability
    - Embedded Service Enabled (ESE) integration for PowerFlex file services alerts
    - SNMP support for PowerFlex file services
    - Backup
    - Troubleshooting bundle
    - CloudIQ integration
- Single namespace
    - Ability to create a global namespace (GNS) supported by the NAS cluster with a single export
    - Allows all hosts with correct access permission to access existing and newly added file systems to the namespace without needing to explicitly mount them on each client
    - Consists of several file systems that may be SMB or NFS
- Common Event Publishing Agent (CEPA)
    - Support for CEPA, which is part of the Common Event Enabler (CEE) package
    - Ability to receive file event notifications
    - Ability to use CEPA to see events on some or all my NFS and SMB file systems

Replication

PowerFlex enables native asynchronous replication for PowerFlex storage-only and PowerFlex hyperconverged configurations.

Replication can be used to quickly and easily recover from a physical or logical disaster, to migrate data, to test data at a remote site, or to offload backup. The PowerFlex implementation is designed to allow a sub-minute recovery point objective (RPO) reducing the data-loss to minimal in case of disaster recovery. As with all other PowerFlex data services, the replication is elastic, can scale online by adding or removing nodes, flexible, and easy to manage. It enables instant test and failover operations.

The following terms and concepts define the replication architecture and its components:

Replication peer systems

Replication occurs between two PowerFlex systems, connected by WAN, and designated as peer systems.

PowerFlex supports up to four peer systems from a single source.

ⓘ **NOTE:** A source volume can be replicated to just a single target.

Storage data replicator

Storage data replicators (SDRs) process all I/Os of replication volumes.

The source SDRs processes all application I/Os of replicated volumes. At the source, the SDC sends application I/Os to the SDR. The I/Os are sent to the target SDRs and stored in their journals. The journals of the target SDRs apply the I/Os to the target volumes. A minimum of two SDRs are deployed at both the source and target systems to maintain high availability. If one SDR fails, the MDM directs the SDC to send the I/Os to an available SDR.

The source SDR processes all application I/Os of replicated volumes. The source SDR saves changes to the journal, and later sends them to the target SDRs. At the target, the SDRs receive the changes, and when a consistent image is available, apply the changes to the target volumes.

A minimum of two SDRs are deployed at the source and target protection domain to ensure high availability.

The ratio of SDRs in a source protection domain to the SDRs in its target protection domain is limited. The maximum ratio of peer SDRs is 1:6, and the maximum ratio of SDRs in maintenance is 1:18. The Add_SDR and Remove_SDR commands fail if the operation results in an unsupported ratio. If the SDRs ratio exceeds the allowed value as a result of a failure or planned maintenance, the Add_RCG command fails. Existing RCGs and replication operations are not impacted.

Journal

SDRs use the journal to hold the replication data.

The source SDRs accumulate the data changes in the journal until sending them to the target. The target SDRs accumulate received data in the target journal until a complete consistent image is received and can be applied to the target volumes.

SDR journal capacity

It is important to assign enough storage capacity for the replication journal. You should consider the following factors when allocating journal capacity.

- The storage pools from which to allocate the journal capacity. The journal is shared between all of the replicated RCGs in the protection domain. Journal capacity should be allocated from storage pools as fast as (or faster than) the storage pool of the fastest replicated application in the protection domain. It should use the same drive technology and about the same drive count and distribution in nodes.
- The minimal requirements needed, for which the default minimum journal size is 400 GB. The minimum journal size can also be calculated as 28 GB per SDR session (where the number of SDR sessions is the number of SDRs + 1) or the capacity needed to sustain an outage, whichever is greater.
- Expected outage time.
  - The minimal outage allowance is one hour, but at least three hours are recommended when computing the minimum size of the replication journal.
  - Journal capacity needed per application is: **`maximal application throughput`** x **`maximum outage interval`**.
- Journal capacity as calculated as a percentage of storage pool capacity, based on the previously calculated needs. Journal capacity should be at least 5% of replicated usable capacity in the protection domain, including volumes used as source and targets.

For example:

- An application generates 1 GB/s of writes.
- The maximal supported outage is 3 hours (3 hours x 3600 seconds = 10800 seconds).
- The journal capacity needed for this application is 1 GB/s x 10800 s = ~10.547 TB.
- Since the journal capacity is expressed as a percentage of the storage pool capacity, divide the 10.547 TB by the size of the storage pool, which is 200 TB: 100 x 10.547TB/200TB = 5.27% Round this up to 6%.
- Repeat this for each application being replicated.

When a protection domain has several storage pools and several replicated applications, the journal capacity should be calculated as in the example above, and the capacity can be divided among all the storage pools (provided they are fast enough). For higher availability, the journal capacity should be allocated from multiple storage pools.

ⓘ **NOTE:** When storage pool capacity is critical, capacity cannot be allocated for new volumes or for expanding existing volumes. This behavior must be taken into account when planning the capacity available for journal usage. The volume usage must leave enough capacity available in the storage pool to allow provisioning of journal volumes. The plan should account for the storage pool staying below critical capacity even when the journal capacity is almost fully used.

It is important to note that since journal capacity is defined as a percentage of the total storage capacity in the storage pool, increasing the total storage capacity by adding devices will increase the journal capacity. Similarly, if you decrease the total storage capacity by removing devices from the storage pool, the journal capacity will automatically decrease.

Replication consistency group

The replication consistency group (RCG) is a logical container for volumes whose application data needs to be replicated consistently to each other.

For example, if you replicate a database and its transaction log, you need the remote copy of each of the database volumes to reflect an image from the same point in time. You can achieve this by placing all of the database volumes in one RCG.

Replication is always defined in the scope of a Protection Domain. All the objects that participate in the replication are contained in the Protection Domain, including the volumes in an RCG. The journal capacity from the Storage Pool in a Protection Domain is shared among all the RCGs in the Protection Domain. The SDRs in the Protection Domain manage I/Os directed to replicated volumes within the Protection Domain.

The following replication attributes are among those that are defined in the RCG:

- RPO (recovery point objective) — The maximum data loss (in units of time) that you are willing to lose. When you set the RPO, this is the goal of replication.
- Direction — Local to remote or remote to local. When there is no replication, for example, following failover or switchover, the direction attribute indicates the direction of the last replication.
- Abstract state — State of replication, for which the options are as follows:
  - OK — Replication is healthy, and the RPO target is met.
  - RPO violation — Replication is healthy, but the RPO target is not met.
  - Error — Replication is not healthy. The Error state attribute provides additional information about the cause for the error.
  - Stopped by user — Replication is not carried out due to user action, such as pause or terminate.
- Error state — Describes the error that is preventing data replication.
- Pause mode — When the RCG is paused, all application I/Os are stored in the source journal. When replication of the RCG is resumed, the source SDR sends the journal contents to the target SDR to be applied to the target volumes. You may want to pause an RCG to handle a network issue between the peer systems or when fixing a hardware issue.
- Activity state — When the RCG is active, replication is enabled. When the RCG is inactive, replication is stopped, but RCG configuration is maintained. The RCG can be automatically inactivated by the system (due to insufficient journal capacity), or can be terminated manually by the user. When activating an inactive RCG, the RCG starts replication from the beginning with a full initialization.
- Failover type — The failover operation requested by the user. Refer to Accessing target volumes for details on the failover type options.
- Failover state — The states are transient, except for None (no failover) or Done (the failover operation has completed).

Replication pairs

A replication pair is a pair of volumes with one volume at the source system and one at the target system. The data from the source volume is replicated to the target volume.

The RCG can contain several replication pairs. A volume can be a member of at most one replication pair. When the replication pair is created, both the source volume and the target volume must be the same size. It is recommended, but not mandatory, that the volumes in the replication pair have the same attributes (including zero padding and granularity). Not doing so can impact performance and capacity.

Recovery point objective

The recovery point objective (RPO) is the maximum data loss, which is measured in units of time, that would be lost in a disaster or outage.

PowerFlex uses journal-based replication to achieve very low RPO and ensure minimal data loss. To ensure RPO compliance, PowerFlex replicates at least twice for every RPO period. For example, setting RPO to one minute means that PowerFlex can immediately return to operation at the target system with loss of only one minute of data. In order to achieve an RPO of one minute, replication takes place at least every 30 seconds. RPO compliance is calculated according to when the application data arrives at the target journal. Achieving RPO is the most important consideration in replication management. The user-defined RPO is the goal of replication. Even so, RPO compliance is not guaranteed. The system reports RPO compliance for each RCG.

Data protection enables replication to multiple systems from one source system to up to four target systems.

Replication I/O flow

For replicated volumes, application data is processed by SDRs at the source and passed to the SDRs at the target.

Application I/Os (both reads and writes) intended for replication volumes are sent from an SDC to an SDR. The source SDR packages the data and distills it so that only the most recent writes are included. The source SDR sends the data over the WAN to the target SDR. At the target system, the SDR processes the replicated data and applies it to the volumes.

Replication auto-provision

PowerFlex supports automatic creation and mapping of a volume in the peer system.

For a given volume on the source system, the user can:

- Create a volume on the destination system.
- Create a replication pair for the given source volume and the new destination volume.

In addition, the user may request to map the destination volume to an SDC.

Replication initial copy

When replication is first activated for an RCG, the target volumes need to be synchronized with the source volumes.

For each replication pair, the entire contents of each source volume are copied to the corresponding target volume when replication is active. This occurs when the RCG containing the replication pair is activated, or when the replication pair is added to an already active RCG. When there is more than one replication pair in the RCG, the order in which the volumes are synchronized is determined by the order in which the replication pairs were activated. You can manually override this order, and also pause and resume the initial copy process, using CLI commands.

The initial synchronization is carried out while the applications are running and performing I/O. Any writes to an area of the volume that has already been synchronized will be sent to the journal. Writes to an area of the volume that has not already been synchronized will be ignored, as the updated content will be copied over eventually as part of the synchronization.

The system limits the number of volumes that undergo Initial Copy at any point in time. New volume pairs added to existing RCGs are given priority over new RCGs. This means that replication pairs in a newly activated RCG may not start Initial Copy immediately.

Activating or terminating a replication consistency group

When a replication consistency group is active, replication is enabled. When a replication consistency group is inactive, replication is stopped and changes are not tracked, but the replication consistency group maintains its configuration.

By default, a new replication consistency group is created as active. You may, however, specify that it be created in the inactive state.

When the replication consistency group is in an inactive state, replication is stopped and the journal capacity is released. The I/Os are handled by the SDS and not the SDR and no journal capacity is consumed. All replication consistency group configurations are maintained, including source and target roles, pair configuration, and volume access modes.

The replication consistency group can be terminated automatically by the system when the journal capacity is fully consumed. You can also manually inactivate the replication consistency group using the terminate command. When the replication consistency group is terminated, volume I/Os bypass the SDR and go directly to the SDS. You can terminate the replication consistency group at both sites separately. Alternatively, you can terminate the replication consistency group at one site; once connectivity is restored, the peer will automatically terminate the replication consistency group at the other site.

Failover can be performed when the replication consistency group is inactive, however consider that the data at the target may be old. While in failover mode, mappings at both the source and target can be activated and volumes can be written to, just like when the replication consistency group is in an active state. There is no data synchronization when the replication consistency group is inactive. You can activate the replication consistency group only once it is no longer in failover mode.

ⓘ **NOTE:** Activating a terminated replication consistency group requires full initialization. The initial copy process may take a long time.

The active or inactive replication consistency group state and the commands to terminate or activate a replication consistency group are only available if both peer systems are running PowerFlex v3.6.x or later.

Pause replication

Replication can be paused for a given RCG.

While paused, any application I/O is stored in the source journal. Once replication is resumed, the journal contents are sent to the target to be applied to the target volumes.

Using replicated volumes

While replication is active, the target volumes remain inaccessible to the hosts with only limited read-only access permitted. This is done to maintain data integrity and consistency between the two peer systems. The data at the target can be accessed to test replication, create snapshots, and to fail over.

Accessing target volumes

There are several situations in which the target volumes are made accessible to the hosts.

- Failover — Fail over the RCG, used especially for disaster recovery. During failover, the application I/Os are stopped at the source and the access mode of the source volumes is changed to inaccessible. The target volumes are brought to the newest consistent image available on the destination and their access mode is changed to read-write. Recovery from a failed over state to a normal operation state can be achieved by either restoring the replication, or reversing it.
- Test failover — Test a failover of the RCG. This allows you to test failover mode, and provides the application with access to the target volumes, without interrupting data synchronization.
- Switchover — Switch over the RCG. Switchover is a planned failover. The application I/Os are stopped at the source and the data is synchronized so that the target volume is consistent with the source volume. The access mode of the original source volumes is changed to unavailable or read-only. The access mode of the original target volumes is changed so that the hosts can read and write to the volumes.

Target volumes can be accessed in R/W only after failover or switchover has been performed. Once the RCG is in failover or switchover mode, you can decide how to continue with replication:

- Restore replication — This maintains the replication direction from the original source and overwrites all data at the target.
- Reverse replication — This changes the direction so that the original target becomes the source. All data at the original source is overwritten by the data at the target.

ⓘ **NOTE:** Prior to executing the restore or reverse operation, you should consider creating a snapshot on the new target volumes in order to preserve the consistent image during the resynchronization process.

Replication direction and mapping

Use the table as a reference for replication direction and default access to volumes according to a subsequent RCG operation and action of the PowerFlex system.

The replication involves two peers, system A and system B. When the replication is set up, system A is set as source and system B is set as target. The following replication direction refers to the initial direction A->B and to changes to that direction.

Table 18. Replication direction and mapping

Subsequent RCG operations	Possible actions	Replication direction/access	Access to volumes
Normal	Switchover/test failover/ failover Remove	A->B	Access to the volumes is allowed only through the source (system A)
After failover	Reverse/restore Remove	N/A - data is not replicated	By default access to the volume is allowed through the original target (system B). It is possible to enable access through the original source (system A).
After failover + reverse ⓘ **NOTE:** Switchover and test failover are only possible after the peers are synchronized.	Switchover/test failover/ failover Remove	B->A	Access to the volumes is allowed only through the original target (system B)
After failover + restore ⓘ **NOTE:** Switchover and test failover are only possible after the peers are synchronized.	Switchover/test failover/ failover Remove	A->B	Access to the volumes is allowed only through the source (system A)
After switchover	Switchover/test failover stop/failover Remove	B->A	Access to the volumes is allowed only through the original target (system B)
After test failover	Switchover/test failover/ failover Remove	A->B	Access to the volumes is allowed through both systems (system A and system B)

Setting up replication

You are required to initially prepare the infrastructure to use replication in PowerFlex.

The following is an overview of the necessary steps.

1. Add peer systems, at both the source and target systems. Add the corresponding certificates to each system.
2. Add journal capacity, ensuring sufficient capacity.
3. Add SDRs, at least two per protection domain that has replicated volumes.

Perform the following steps to configure replication:

1. Create Replication Consistency Group. Define RPO.
2. Add volume pairs. Volumes in a volume pair must have the same size. It is not necessary but is recommended that they have the same attributes (medium/fine granularity, zero padded/non-zero padded).

   If both systems are 4.x or higher then the target volumes may be automatically provisioned when the pair is created.
3. If the RCG was created as active (the default) then initial copy is started automatically. Otherwise, the RCG must be activated before initial copy is started.

For additional information about replication, see the *PowerFlex 4.5.x Administration Guide*.

Dell APEX Block Storage

Dell APEX Block Storage for public cloud is a deployment of Dell PowerFlex, software-defined block storage innovation, in the public cloud. APEX Block Storage is available for both AWS and Microsoft Azure allowing you to experience the same benefits of enterprise-class storage services in the cloud as with on-premises. It provides higher performance, larger volume sizes, and improved resilience than what is currently available on the public cloud.

Table 19. Capabilities and benefits

Capability	Benefit or Value
Unique multi-availability zone durability	● Enhanced, space efficient data protection through the aggregation of all instances across availability zones for volume provisioning. ● Enables high resilience across availability zones by protecting the data during availability zone failure, without replication of data (space and IOPs).
Extreme performance for mission-critical workloads running in the cloud	Meets and exceeds SLAs with extreme performance (high throughput and low latency) for workloads such as databases and analytics.
Scalable, flexible, and resilient	Meets stringent SLAs and run workloads with confidence and assurance with near linearly scalable performance as workload demand increases.
Data mobility	Seamless data mobility with the ability to easily move data from on-premises to the cloud or across regions within the cloud as workloads demand increases.
Self-healing architecture with rapid rebuild	Ensure high availability and service availability under failure conditions with fast reprotection and rebuild.
Robust ecosystem of automation tools and frameworks (CSM/CSI, Ansible, REST APIs, DDVE, CloudLink, CloudIQ)	Optimal usage and performance of cloud-based block storage with easy to deploy, expand, monitor, and manage capabilities.

NOTE: The following features that are supported by PowerFlex are NOT supported with the APEX Block Storage in the public cloud: compression, fine granularity storage pools in the public cloud, SDNAS or PowerFlex file services, and NVMe TCP.

APEX Block Storage for AWS

Dell APEX Block Storage for AWS empowers enterprises to run diverse workloads in the public cloud while ensuring extreme performance, scalability, and a simplified cloud experience. APEX Block Storage can be deployed in two configurations by using Elastic Block Store (EBS) volumes, or native NVMe SSD drives attached to EC2 instances (EC2 Instance Store). Dell APEX Block Storage also provides the proven enterprise data services, such as thin provisioning, snapshots, and asynchronous replication, required to run demanding block-based workloads in the public cloud. Dell APEX Block Storage native asynchronous replication enables data mobility between on-premises and the cloud or across regions in the cloud, for example, AWS East to AWS West.

APEX Block Storage for Microsoft Azure

Now available Dell APEX Block Storage for Public Cloud in Microsoft Azure. APEX Block Storage can be deployed in two configurations using Azure Managed Disks or virtual machines with attached NVMe SSDs based on the use case. Dell APEX Block Storage also provides the proven enterprise data services, such as thin provisioning, snapshots, and asynchronous replication, required to run demanding block-based workloads in the public cloud. Dell APEX Block Storage native asynchronous replication enables data mobility between on-premises and the cloud or across regions in the cloud.

Protection and load balancing

PowerFlex maintains user data in a distributed mesh mirrored layout. Each piece of data is stored on two different fault units. The copies are distributed over the storage devices according to an algorithm that ensures uniform load of each fault unit by terms of capacity and expected network load. Rebuild and rebalance processes are fully automated, but are configurable.

Rebuild

PowerFlex initiates a rebuild process in response to failure. Forward rebuild refers to creating a new copy of the data on another server. Backward rebuild refers to re-synchronizing one of the copies.

When a failure occurs, such as on a server, device or network failure, PowerFlex immediately initiates a process of protecting the data. This process is called *rebuild*, and comes in two flavors:

- *Forward rebuild* is the process of creating another copy of the data on a new server. In this process, all the devices in the storage pool work together, in a many-to-many fashion, to create new copies of all the failed storage blocks. This method ensures an extremely fast rebuild.
- *Backward rebuild* is the process of re-synchronization of one of the copies. This is done by passing to the copy only changes made to the data while this copy was inaccessible. This process minimizes the amount of data transferred over the network during recovery.

PowerFlex automatically selects the type of rebuild to perform. This implies that in some cases, more data will be transferred to minimize the time that the user data is not fully protected.

Rebuild throttling

The rebuild throttling policy determines the priority of rebuild I/Os versus application I/Os when accessing SDS devices. The possible rebuild throttling policies are no limit on rebuild I/Os, limit concurrent I/Os per SDS device, favor application I/Os and dynamic bandwidth throttling.

Rebuild throttling sets the rebuild priority policy for a storage pool. The policy determines the priority between the rebuild I/O and the application I/O when accessing SDS devices. Please note that application I/Os are continuously served.

Applying rebuild throttling will on one hand increase the time the system is exposed with a single copy of some of data, but on the other hand, will reduce the impact on the application. You must attempt to choose the right balance between the two.

The following priority policies may be applied:

- No Limit: No limit on rebuild I/Os. Any rebuild I/O is submitted to the device immediately, without further queuing.

  NOTE: Rebuild I/Os are relatively large and hence setting this policy will speed up the rebuild, but will have the maximal effect on the application I/O.
- Limit Concurrent I/O: Limit the number of concurrent rebuild I/Os per SDS device (default). The rebuild I/Os are limited to a predefined number of concurrent I/Os. When the limit is reached, the next incoming rebuild I/O waits until the completion of a currently executed rebuild I/O. This will complete the Rebuild quickly for best reliability, however, there is a risk of host application impact.
- Favor Application I/O: Limit rebuild in both bandwidth and concurrent I/Os. The rebuild I/Os are limited both in bandwidth and in the amount of concurrent I/Os. As long as the number of concurrent rebuild I/Os, and the bandwidth they consume, do not exceed the predefined limits, rebuild I/Os will be served. Once either threshold is reached, the rebuild I/Os wait until both I/O and bandwidth are below their thresholds. For example, setting the value to "1" will guarantee the device will only have one concurrent rebuild I/O at any given moment, which will ensure the application I/Os only wait for 1 rebuild I/O at worst case. This imposes bandwidth on top of the Limit Concurrent I/Os option, which is a prerequisite to using this policy.
- Dynamic Bandwidth Throttling: This policy is similar to Favor Application I/O, but extends the interval in which application I/Os are considered to be flowing by defining a minimal quiet period. This quiet period is defined as a certain interval in which no application I/Os occurred. Note that the limits on the rebuild bandwidth and concurrent I/Os are still imposed.
- Default Values:
  - The default policy for rebuild is: Limit Concurrent I/O
  - Rebuild concurrent I/O Limit: 1 concurrent I/O

    NOTE: Rebuild throttling affects system performance and should only be used by advanced users.

Rebalance

When PowerFlex detects that user data is not balanced across devices in the storage pool, it initiates a process to restore the balance in which data copies are moved from the most used devices to the least used.

Rebalance is the process of moving one of the data copies to a different server. It occurs when PowerFlex detects that the user data is not evenly balanced across the fault units in a storage pool. This can occur as a result of several conditions such as: SDS addition or removal, device addition/removal, or following a recovery from a failure. PowerFlex moves copies of the data from the most used devices to the least used ones.

Both rebuild and rebalance compete with the application I/O for the system resources, which include network, CPU and disks. PowerFlex provides a rich set of parameters that can control this resource consumption. While the system is factory-tuned for balancing between speedy rebuild or rebalance and minimization of the effect on the application I/O, the user has fine-grain control over the rebuild and rebalance behavior.

Rebalance throttling

The rebalance throttling policy determines the priority of rebalance I/Os versus application I/Os when accessing SDS devices.

Rebalance throttling sets the rebalance priority policy for a storage pool. The policy determines the priority between the rebalance I/O and the application I/O when accessing SDS devices. Please note that application I/Os are continuously served. Rebalance, unlike rebuild, does not impact the reliability of the system, and therefore reducing its impact is not risky.

ⓘ **NOTE:** Rebalance throttling affects the performance of the system, and should be used only by advanced users.

The following possible priority policies may be applied:

- No Limit: No limit on rebalance I/Os. Any rebalance I/O is submitted to the device immediately, without further queuing. Please note that rebalance I/Os are relatively large and hence setting this policy will speed up the rebalance, but will have the maximal effect on the application I/O.
- Limit Concurrent I/O: Limit the number of concurrent rebalance I/Os per SDS device. The rebalance I/Os are limited to a predefined number of concurrent I/Os. Once the limit is reached, the next incoming rebalance I/O waits until the completion of a currently executed rebalance I/O. For example, setting the value to "1" will guarantee that the device will only have one rebalance I/O at any given moment, which will ensure that the application I/Os only wait for 1 rebalance I/O in the worst case.
- Favor Application I/O: Limit rebalance in both bandwidth and concurrent I/Os. The rebalance I/Os are limited both in bandwidth and in the amount of concurrent I/Os. As long as the number of concurrent rebalance I/Os, and the bandwidth they consume, do not exceed the predefined limits, rebalance I/Os will be served. Once either limiter is reached, the rebalance I/Os wait until such time that the limits are not met again. This imposes a bandwidth limit on top of the Limit Concurrent I/Os option.
- Dynamic Bandwidth Throttling: This policy is similar to Favor Application I/O, but extends the interval in which application I/Os are considered to be flowing by defining a minimal quiet period. This quiet period is defined as a certain interval in which no application I/Os occurred. Note that the limits on the rebalance bandwidth and concurrent I/Os are still imposed.
- Default Values:
  - The default policy for rebalance: Favor Application I/O
  - Rebalance concurrent I/O Limit: 1 concurrent I/O per SDS device
  - Rebalance bandwidth limit: 10240 KB/s

For information on instant maintenance mode that uses rebuild and rebalance, see Instant maintenance mode (IMM).

Checksum protection

PowerFlex calculates and validates the checksum value for the payload during transit to protect data in-flight. Checksum protection is applied to all I/Os.

This feature addresses errors that change the payload during the transit through PowerFlex. PowerFlex protects data in-flight by calculating and validating the checksum value for the payload at both ends.

ⓘ **NOTE:**

- The feature is off by default for medium granularity.
- The checksum feature can have a significant impact on large block I/O latency. Contact Customer Support for more information.

- During write operations, the checksum is calculated when the SDC receives the write request from the application. This checksum is validated just before each SDS writes the data on the storage device.
- During read operations, the checksum is calculated when the data is read from the SDS device, and is validated by the SDC before the data returns to the application. If the validating end detects a discrepancy, it will initiate a retry. The checksum will be done in the granularity of a sector (1/2 KB).
- Pools with fine granularity with or without compression, have a persistent checksum by default. This cannot be changed.

Each I/O goes through compression, and the checksum is calculated before it is written to the disk. There are two types of checksum:

- Fine granularity layout saves checksum data before and after processing to guarantee data integrity (compressed or not).
- There are also system checksums for metadata.

This feature applies to all I/Os: application, rebuild, rebalance, and migrate. The checksum is also kept in RMcache (read memory cache), protecting every block that is maintained in SDS memory against memory corruption. The checksum feature can be enabled at the protection domain level, and defined at the storage pool level. The feature is T10/DIF-ready.

The fine granularity data layout has a default checksum whether it is compressed or not.

Fine granularity layout checksum implementation

The fine granularity data layout has a default checksum whether it is compressed or not.

Cache

PowerFlex supports RAM Read Cache (using DRAM server memory), to enhance system performance.

The following table summarizes information about the caching modes provided by the system.

Table 20. Caching modes

Mode	Description	Considerations	Default Setting
RAM Read Cache (RMcache)	Read-only caching performed by server RAM.	RAM Read cache, the fastest type of caching, uses RAM that is allocated for caching. Its size is limited to the amount of allocated RAM. **NOTE:** The amount that may be allocated is limited, and can never be the maximum available RAM.	Disabled, except when storage-only nodes are deployed.

The following table illustrates the caching support matrix:

Table 21. Caching support matrix

System	RMcache
PowerFlex	Yes
VxFlex Ready Node PowerEdge 13G servers	Yes
VxFlex Ready Node PowerEdge 14G servers	Yes

Networking

This section describes the various considerations for managing inter-node communication: through a separate network with access to the PowerFlex components, or on the same network.

In PowerFlex, inter-node communication manages data locations, rebuild and rebalance, and access applications to stored data on one IP network, or on separate IP networks. Use any of the management interfaces to perform management tasks in the following ways:

- Through a separate network with access to the other PowerFlex components
- On the same network

Configure these options in the following ways:

- During deployment in the full PowerFlex Installer (using the CSV topology file) and using the VMware plug-in
- After deployment with the CLI.

The following table includes the list of PowerFlex networking requirements and supported architectures or configurations.

Table 22. Networking requirements and limitations

PowerFlex minimum network requirements and supported architectures	PowerFlex network limitations
● Ethernet-based ● Minimum supported speed: 2*10 GbE ● Recommended speed: 2*25 GbE ● Certified configurations: 4*25 GbE, 2*100 GbE, 4*100 GbE ● Supported MTU: 1500/9000 (jumbo frames recommended for optimal performance) ● High availability and load balancing use at least two physical ports with: ○ PowerFlex native multipath across two or more subsets: ■ Supports one to eight subnets ○ Link aggregation: ■ A single IP subnet can be used if link aggregation is used, otherwise a minimum of two subnets is required. ○ It is recommended to use MDM virtual IP for simplified network stability if there are failures. ● IP addresses: See guidelines about IP address calculation in IP addresses for R640. R740xd, and R840 servers ● Data network latency requirements: ○ MDM: Maximum of 300 milliseconds ● Switch redundancy: ○ Switches must be redundant. This provides continued access to components inside the rack in the network in case a top of rack (ToR) switch fails. ○ Flat network is supported. ○ Spine-leaf architecture is recommended for cross rack traffic. ○ VXLAN is supported. ● SDC connection through a router is supported and required for a low latency connection for optimal performance. ● PowerFlex has a single management subnet which can be a dedicated port or a VLAN on a trunk connection. For best practices, see: ○ VxFlex Ready Node R640-R740xd-R840 Operating System Installation and Configuration Guide for Linux ○ VxFlex Ready NodeR640-R740xd-R840 Operating System Installation and Configuration Guide for ESXi	● RMDA is not supported. ● InfiniBand is not supported. ● IPv4 and IPv6 are supported on PowerFlex software only and offering. ● IPv4 is supported on PowerFlex appliance and PowerFlex rack offering. ● DHCP must not be deployed on a network where PowerFlex MDMs, SDS, SDR, or SDCs reside. ● Firewall rules must allow PowerFlex ports. For port information, see the *Dell PowerFlex 4.5.x Security Configuration Guide*.

The following table includes the list of PowerFlex networking requirements and supported architectures for systems with replication.

Table 23. Networking requirements and limitations for systems with replication

PowerFlex network requirements and supported architectures for systems with replication	PowerFlex network limits for systems with replication
● Minimum best practice with replication: 4*25 GbE or 2*100 GbE ● Additional IP addresses for site to site routing include:	● MDM to MDM peer metadata synchronization should take place over a WAN with less than 200-millisecond latency.

Table 23. Networking requirements and limitations for systems with replication

PowerFlex network requirements and supported architectures for systems with replication	PowerFlex network limits for systems with replication
○ Within a protection domain, SDRs are installed on the same hosts as SDSs. The traffic that an SDR writes to the journal volume, is sent to all SDSs that host the journal, not only the one that is co-located on the host. In the backend storage network, each SDR listens on the same node IPs as the SDSs and should be able to reach all SDSs within the protection domain. ○ The SDRs require additional, distinct IP addresses which allow them to communicate with remote SDRs. In most cases, these should be routable addresses with a properly configured gateway. For redundancy, each SDR should have two. ● Static routes: ○ PowerFlex asynchronous replication usually happens over a WAN between physically remote clusters that do not share the same address segments. If the default route itself is not suitable to properly direct packets to the remote SDR IPs, static routes must be configured to indicate either the next top address or the egress interface or both for reaching the remote subnet. ○ PowerFlex Manager can add static routes for replication use cases.	● Local SDR to remote SDR: Latency is not as sensitive in SDR > SDR traffic, but round-trip time should not be greater than 200ms. ● Firewall rules must allow PowerFlex ports. For port information, see the *Dell PowerFlex 4.5.x Security Configuration Guide*.

See the following configuration guidelines used by VxFlex Ready Node for:

- Linux examples: R640 and R740xd nodes see Configuring network ports on Linux servers
- ESXi examples: R640, R740xd, and R840 nodes see Configuring network ports on ESXi hosts

This section describes how to choose from these options, depending on the requirements of your organization, security considerations, performance needs, and IT environment.

PowerFlex networking considerations:

- Single IP network: All communications and IOs used for management and for data storage are performed on the same IP network. This setup offers the following benefits:
  - Ease of use.
  - Fewer IP addresses required
- Multiple separate IP networks: Separate networks are used for management and for data storage, or separate networks are used within the data storage part of the system. This setup offers the following benefits:
  - Security
  - Redundancy
  - Performance
  - Separate IP roles in order to separate between customer data and internal management

  NOTE: Network high availability can be implemented by using NIC-bonding (see https://www.dellemc.com/collateral/white-papers/h17332-dellemc-vxflex-os-networking-best-practices.pdf) or by using several data networks in PowerFlex.

  For more information about MTU performance considerations and best practices, see the *PowerFlex Configure and Customize Guide*.

  NOTE: The MDM cluster IP address limit is 16 IP addresses, which include all cluster members (primary, secondary, and standby).

Table 24. Range of potential IP address configurations

Column in CSV file	MDM Mgmt IP	MDM IPs	SDS All IPs	SDS-SDS Only IPs	SDS-SDC Only IPs
Comments	Management Access	Control Network	Rebuild and Data Path Network	Rebuild Network	Data Path Network
	Optional, but recommended; not applicable for tiebreaker IP addresses that can be used to provide access to PowerFlex management applications, such as PowerFlex Manager, CLI, REST API, OpenStack. This IP address must be externally accessible.	Mandatory IP addresses used for MDM control communications with SDSs and SDCs, used to convey data migration decisions, but no user data passes through the MDM. Must be on the same network as the data network. Must be externally accessible if no MDM Management IP addresses are used. MDM Virtual IP is supported on the data network.	IP addresses used for both SDS-SDS and SDS-SDC communications. These IP addresses will also be used to communicate with the MDM.	IP addresses used for SDS-SDS communication only. These addresses are used for rebuild and rebalance operations. These IP addresses will also be used to communicate with the MDM.	IP addresses used for SDS-SDC communication. These addresses are only used for read/write user data operations.

The following combinations can be used for SDS or SDC:

- Only *SDS All IPs*
- Only *SDS-SDS Only IPs* + *SDS-SDC Only IPs*
- *SDS All IPs* + either *SDS-SDS Only IPs* or *SDS-SDC Only IPs* (can be used in cases of multiple networks; ensure that you do not use the same IP address more than once in the networks).
- *SDS All IPs* + both *SDS-SDS Only IPs* and *SDS-SDC Only IPs* (can be used in cases of multiple networks; ensure that you do not use the same IP address more than once in the networks).

Networking limitation

When deployed on VMware, PowerFlex, the same IP Subnet can be configured on different VM kernel interfaces\virtual switches.

For more information, see the VMware ESXi limitation.

PowerFlex only supports the following network configurations when deployed on VMware:

- A single data storage network
- Two or more data networks, each on separate IP subnets
- A single IP data network using several NIC-bonding configurations, or vSwitch load balancing

Network stability

Application level multipathing configurations:

Resilient network connectivity is standard for PowerFlex. Nevertheless the actual implementation may vary. In some cases, the resiliency is attained through LACP or bonding, while other cases through application level multipathing.

In PowerFlex v3.6, additional capabilities were added to support the following error scenario use cases:

Use case one:

- Partial failures on existing connections: While many network issues are manifested as a complete disconnection, some are not, PowerFlex is now aware of degraded connections. Types of degraded connections are:
  - Stable connection and yet very slow vs others.
  - Unstable connection (occasional disconnections).
  - Combination of the above two.

Use case two:

- SDC to SDS Proxy connection (in the PowerFlex architecture, each SDC is required to be fully connected to all the SDSs on each of the PDs it consumes volumes from. In case the SDC is unable to reach one or more of the SDSs, the MDM will reflect that and the management APIs will raise an alert. In PowerFlex v3.6 a capability was introduced:
  - In case the SDC cannot access the SDS it will send the I/O to another SDS within the same PD (or subset of this PD). We will refer to this SDS as Proxy-SDS since its only role is to forward the I/O to the SDS which owns the data, receive the I/O result and return it to the initiating SDC.
- PowerFlex supports using the MDM secondary IPs as a proxy connection to the Primary MDM.

NOTE: This capability is applicable only when MDM VIP is not in use, since it requires the SDCs to have all the IPs configured to attempt a proxy connection.

Networking requirements for PowerFlex file deployment

The following lists the networking requirements for a PowerFlex file node.

- Ensure that a NIC with an IP address is configured in the same range as the following networks:
  - PowerFlex management
  - PowerFlex data (1 to 4)
- Create a bonded interface on the PowerFlex node. No sub interfaces or IP addresses should be assigned to this bonded interface dedicated to NAS traffic.

Sample configurations

A sample network configuration:

```
eth0 - PowerFlex Management (Node management)
eth1 - PowerFlex Data
eth2 - PowerFlex Data
bond0 (there should not be any tagged interface under bond0)
```

A sample IP address configuration on the PowerFlex file node:

```
1: lo: <LOOPBACK,UP,LOWER_UP> mtu 65536 qdisc noqueue state UNKNOWN group default qlen
1000
    link/loopback 00:00:00:00:00:00 brd 00:00:00:00:00:00
    inet 127.0.0.1/8 scope host lo
       valid_lft forever preferred_lft forever
    inet6 ::1/128 scope host
       valid_lft forever preferred_lft forever
2: eth0: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1500 qdisc mq state UP group default qlen
1000
    link/ether 00:50:56:b0:95:da brd ff:ff:ff:ff:ff:ff
    altname enp11s0
    altname ens192
    inet 10.1.1.226/20 brd 10.234.223.255 scope global eth0
       valid_lft forever preferred_lft forever
3: eth1: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1500 qdisc mq state UP group default qlen
1000
    link/ether 00:50:56:b0:bb:8c brd ff:ff:ff:ff:ff:ff
    altname enp19s0
    altname ens224
    inet 172.16.221.226/16 brd 172.16.255.255 scope global eth1
       valid_lft forever preferred_lft forever
    inet6 fe80::250:56ff:feb0:bb8c/64 scope link
       valid_lft forever preferred_lft forever
4: eth2: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1500 qdisc mq state UP group default qlen
1000
    link/ether 00:50:56:b0:3f:09 brd ff:ff:ff:ff:ff:ff
    altname enp27s0
    altname ens256
    inet 172.17.221.226/16 brd 172.17.255.255 scope global eth2
       valid_lft forever preferred_lft forever
    inet6 fe80::250:56ff:feb0:3f09/64 scope link
       valid_lft forever preferred_lft forever
5: docker0: <NO-CARRIER,BROADCAST,MULTICAST,UP> mtu 1500 qdisc noqueue state DOWN group
default
    link/ether 02:42:14:01:33:16 brd ff:ff:ff:ff:ff:ff
    inet 172.29.200.1/24 brd 172.29.200.255 scope global docker0
       valid_lft forever preferred_lft forever
6: eth3: <BROADCAST,MULTICAST,SECONDARY,UP,LOWER_UP> mtu 1500 qdisc mq primary bond0
state UP group default qlen 1000
    link/ether 00:50:56:b0:65:04 brd ff:ff:ff:ff:ff:ff
    altname enp12s0
    altname ens193
7: eth4: <BROADCAST,MULTICAST,SECONDARY,UP,LOWER_UP> mtu 1500 qdisc mq primary bond0
state UP group default qlen 1000
    link/ether 00:50:56:b0:65:04 brd ff:ff:ff:ff:ff:ff
    altname enp4s0
    altname ens161
8: bond0: <BROADCAST,MULTICAST,PRIMARY,UP,LOWER_UP> mtu 1500 qdisc noqueue state UP
group default qlen 1000
    link/ether 00:50:56:b0:65:04 brd ff:ff:ff:ff:ff:ffa
```

NOTE: The switch port connected to interfaces that are assigned to NAS bond interface should be configured to accept NAS management VLAN as untagged and NAS data VLANs as tagged.

A sample Cisco Nexus switch configuration:

```
interface ethernet1/1/4
  no shutdown
  switchport mode trunk
  switchport access vlan 100
  switchport trunk allowed vlan 200,201
  mtu 9216
  flowcontrol receive off
  channel-group 8 mode active
interface port-channel10
  switchport mode trunk
  switchport trunk native vlan 100
  switchport trunk allowed vlan 200.201
  mtu 9216
  no lacp suspend-individual
  lacp vpc-convergence vpc 8
```

SDC connection monitoring

The system monitors all connections between SDCs and SDSs and sends out an alert when an active connection between an SDC and an SDS goes down.

To monitor SDC and SDS connections effectively, the MDM collects connectivity updates from all of the SDCs. The MDM posts events whenever an SDC connects to or disconnects from a specific SDS IP address. The MDM frequently analyzes the connectivity status to determine the current system state. The system does not send out alerts for temporary connectivity issues that are resolved in less than 10 seconds.

The following are the possible connectivity states between SDCs and SDSs in the system:

- All connected
- One SDC is disconnected from one SDS
- One SDC is disconnected from one SDS IP address
- One SDC is disconnected from all SDSs
- All SDCs are disconnected from one SDS
- All SDCs are disconnected from one SDS IP address
- All SDCs are disconnected from all SDSs
- Multiple disconnections

When the connectivity state of the system changes to any state other than `All Connected`, an alert is displayed in PowerFlex Manager and is written to the MDM event log. Once an alert is generated, you can use the SCLI to query details on the disconnection using the command `scli --query_sdc_to_sds_disconnections`.

The MDM does not monitor the connectivity state of SDCs or SDSs in the following scenarios:

- SDS is in maintenance mode
- SDS is disconnected from the MDM
- SDS is in the process of being removed
- SDC is disconnected from the MDM for more than two minutes
- SDC is not approved

SMART hardware monitoring

The PowerFlex bare-metal solution provides monitoring capabilities for RAID controllers and storage devices compatible with SMART (Self-Monitoring, Analysis and Reporting Technology) protocols.

In Linux-based environments, SMART-compatible HDDs, SSDs and RAID storage controllers can be monitored for SMART attributes such as temperature, SSD wear level, and error counters. LEDs can also be lit on these hardware devices, to simplify physical identification for maintenance purposes.

Each hardware vendor defines specific thresholds for the SMART attributes. This feature currently supports storage devices controlled by LSI, HP and Dell RAID controllers, and stand-alone devices. During system deployment, an external monitoring tool is installed as part of the LIA on each node. Additional RAID controller tools must be installed manually after system deployment: storcli for LSI RAID controllers, hpssacli for HP RAID controllers, or perccli for Dell RAID controllers. These tools are used by the system to collect the counters that are returned to the MDM.

ⓘ **NOTE:** In some cases, LSI RAID controllers may report vendor information as "AVAGO" instead of LSI.

The MDM queries the SDSs at set intervals, and stores the returned information. This information can be viewed using CLI queries. In addition, when thresholds are crossed for SMART attributes, alerts are generated by the system.

When the CLI is used to query device information, physical device information, such as serial number, disk slot information, model name, vendor etc., temperature, and wear level information (for SSDs only) is included in the returned response.

You can use PowerFlex Manager to monitor SMART-related alerts in the **Alerts** view.

In addition, SNMP traps and Secure Connect Gateway alert codes can be used to monitor alerts triggered by devices compatible with SMART

.

List of approved RAID controllers

High-level specifications of RAID controllers, which are tested and certified by PowerFlex.

Table 25. PowerFlex-certified RAID controllers

Manufacturer	Specifications
Dell	● Model Name: PERC H730 Mini ● Firmware Version: 25.5.5.0005 ● Driver Version: 07.700.52.00 and above ● Driver Name: megaraid_sas
	● Model Name: H740p ● Firmware Version: See VxFlex Ready Node Firmware and Driver support matrix. ● Driver Version: See VxFlex Ready Node Firmware and Driver support matrix. ● Driver Name: megaraid_sas
	● Model Name: HBA330 ● Firmware Version: See VxFlex Ready Node Firmware and Driver support matrix. ● Driver Version: See VxFlex Ready Node Firmware and Driver support matrix. ● Driver Name: mpt3sas

Table 25. PowerFlex-certified RAID controllers (continued)

Manufacturer	Specifications
HP	● Model Name: Smart Array P440ar ● Firmware Version: 3.56 ● Driver Version: 3.4.10 ● Driver Name: hpsa

NVDIMM hardware awareness feature

From PowerFlex, the customer can enable the NVDIMM hardware awareness feature on the SDS.

When this feature is enabled, the SDS detects bad blocks on NVDIMM devices used for fine granularity acceleration and takes appropriate actions based on how and when the corruptions in the NVDIMM devices are detected.

Snapshots

PowerFlex enables you to create snapshots, that is, instantaneous copies of volumes. A snapshot can be manipulated as any volume exposed by the storage system. Snapshots that are taken simultaneously form a consistency group and can be manipulated as a group.

The PowerFlex storage system enables you to take snapshots of existing volumes, up to 128 per volume. The snapshots are thin provisioned and are extremely quick. For more information about thin provisioning, see "SAN virtualization layer."

Once a snapshot is generated, it becomes a new, unmapped volume in the system. You can manipulate it in the same manner as any other volume exposed by the PowerFlex storage system.

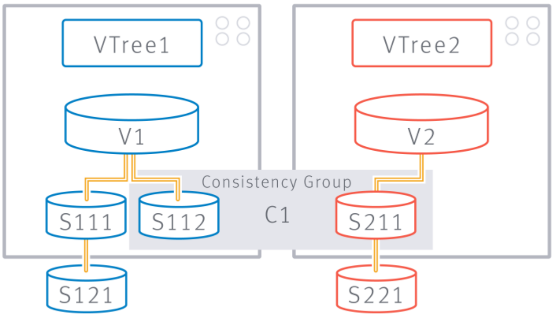


Figure 6. Snapshot operations

The structure related to all the snapshots resulting from one volume is referred to as a vTree (short for volume tree). When taking a snapshot in the system, you can specify more than one volume. All snapshots that are taken together form a consistency group. They are consistent in the sense that they were all taken simultaneously. So if there is a contextual relationship between the data that is contained by all the snapshot members, then that set is meaningful. The consistency group allows manipulation of the entire set.

If you remove an entire consistency group, all snapshots that were taken together are removed. In RED, S211 is a snapshot of V2. Since S112 and S211 were taken together, they compose a consistency group that is designated as C1.

NOTE:

- The consistency group is only for convenience purposes. There are no protection measures done by PowerFlex to conserve the consistency group. For example, you can remove a snapshot that is a member of a consistency group.
- Support of removal of a parent snapshot in the system. You can merge the parent to a child snapshot.
- It is recommended not to trim volumes that contain snapshots as the logical capacity may not shrink.

Volume trees (vTrees)

A vTree consists of the root volume and all of its descendant volumes and snapshots.

A vTree is the structure comprised of a volume and the snapshots resulting from that volume. It is a tree spanning from the source volume at its root, whose descendants are either snapshots of the volume itself or snapshots of a snapshot. In the vTree diagram, S_{111} and S_{112} are snapshots of V_1. S_{121} is a snapshot of snapshot S_{111}. Together, V_1 and S_{1xy} are the vTree of V_1.

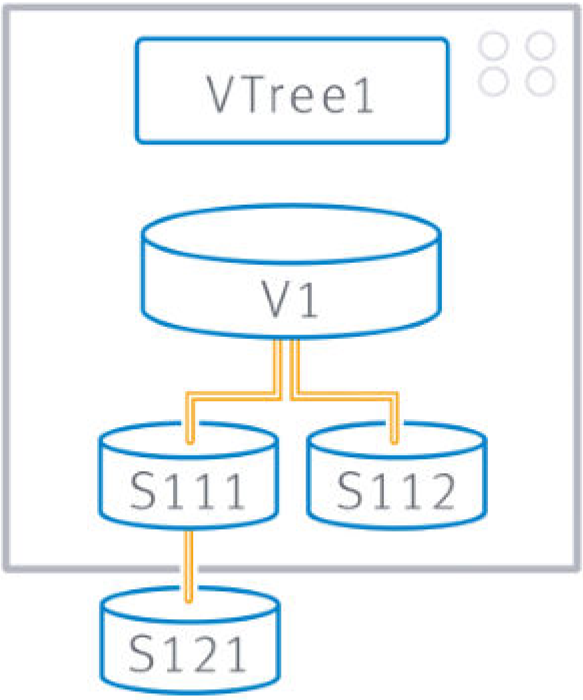


Figure 7. vTree diagram

Snapshot policies

Snapshot policies enable you to define policies where you can configure the number of snapshots to take at a given time for one or more volumes.

Snapshots are taken according to the rule defined. You can define the time interval in-between two rounds of snapshots as well as the number of snapshots to retain, in a multi-level structure. For example, take snapshots every x minutes/hours/days/weeks. There are one to six levels, with the first level having the most frequent snapshots.

Example:

Rule: Take snapshots every 60 minutes

Retention Levels:

- 24 snapshots
- 7 snapshots
- 4 snapshots

After defining the parameters, select the source volume to add to the snapshot policy. You can add multiple source volumes to a snapshot policy, but only a single policy per source volume is allowed. Only one volume per vTree may be used as a source volume of a policy (any policy). When you remove the source volume from the policy, you must choose how to handle snapshots. Snapshots created by the policy are referred to as auto snapshots. In PowerFlex Manager, this is indicated if a snapshot policy is displayed for the snapshot.

- If the source volume has no auto snapshots, you cannot unassign the source volume from the snapshot policy. You can remove auto snapshots from **Snapshots** in PowerFlex Manager.
- If the source volume has auto snapshots but none of them are locked, you are prompted to confirm that you would like to delete all auto snapshots. If any of the auto snapshots are locked, the locked auto snapshots are just detached from the

snapshot policy, but not deleted. You need to manually remove auto snapshots. It does not matter if the auto snapshot is locked or not, you can still delete.

Performance

PowerFlex is a scale-out, high-performance storage system. The performance of the system is almost perfectly scalable, as shown in the graph. This means that adding nodes adds capacity and also increases the performance of the system.

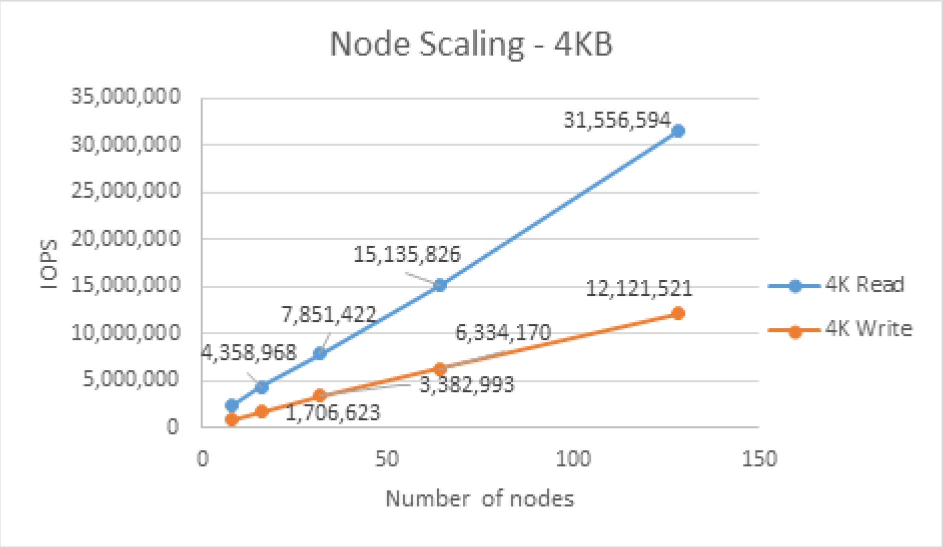


Figure 8. PowerFlex performance

Migrating volumes

You can migrate volumes to a different storage pool or protection domain within your PowerFlex system.

You can migrate volumes from the following storage layers:

- From one storage pool to another within the same protection domain
- From one storage pool to another across different protection domains
- From the volume tree and all of its snapshots when they are migrated together
- From thick to thin
- From fine or medium to medium or fine granularity

Requirements to migrate volumes:

- Medium granularity pools should have zero padding enabled.

Implementing PowerFlex

Implementing PowerFlex is, in general, a two-step process: first build the physical storage layer, then configure the virtual SAN layer on top of it.

Communication is done over the existing LAN data network using standard TCP/IP. The MDM and SDS nodes can be assigned to up to eight IP addresses, enabling wider bandwidth and better I/O performance and redundancy.

To build the physical storage layer:

1. Physically assemble the server nodes in the rack and perform all cabling, according to the instructions for your PowerFlex product.
2. Configure the operating system, including all required IP addresses, based on network best practice guidelines.
3. Using PowerFlex Installer, deploy the PowerFlex software.

The physical layer is now ready, and you can expose a virtual storage layer.

PowerFlex appliance and PowerFlex rack offer a more automated installation experience using PowerFlex Installer. For more information regarding PowerFlex Manager management and operation capabilities, see PowerFlex appliance and PowerFlex rack product information.

PowerFlex implementation in an ESXi-based system

In the VMware environment, the MDM, and SDS components are installed on a dedicated SVM, and the SDC is installed directly on the ESXi host.

Minimum requirements

These are the minimum system requirements for a PowerFlex implementation in an ESXi-based system:

- Three ESXi servers with 240 GB of free capacity per server
- 10 Gbps network

Table 26. Minimum system requirements

Configuration type	Supported VMware ESXi versions	Supported PowerFlex offering
VMware hyperconverged nodes with SDC ⓘ **NOTE:** Compute nodes may also consume storage from VMware Hyperconverged nodes.	ESXi 7.0	PowerFlex appliance PowerFlex rack
VMware compute nodes with SDC connecting to a PowerFlex storage-only system	ESXi 7.0	PowerFlex appliance PowerFlex rack PowerFlex software PowerFlex custom node VxFlex Ready Node
VMware compute nodes with NVMe over TCP connecting to a PowerFlex storage-only system	ESXi 7.0	PowerFlex appliance PowerFlex rack PowerFlex software PowerFlex custom node VxFlex Ready Node

ⓘ **NOTE:** For a full list of supported operating systems, see the PowerFlex support matrix .

Components

The PowerFlex hyperconverged configuration consists of a dedicated storage virtual machine (SVM) and these software components:

- Storage virtual machine (SVM): A Linux-based virtual machine dedicated to PowerFlex and is used to host the different PowerFlex software components described here.
- Meta data manager (MDM): Configures and monitors the PowerFlex system. The MDM is configured in a redundant cluster mode. The MDM is installed on the SVM. The MDM can be configured as a three-node cluster (primary MDM, secondary MDM, and tiebreaker MDM) or as a five-node cluster (primary MDM, two secondary MDMs, and two tiebreaker MDMs) to provide greater resiliency.
- Storage data server (SDS): Manages the capacity of a single server and acts as a back-end for data access. The SDS is installed on all servers contributing storage devices to the PowerFlex system. These devices are accessed using SDS. The SDS is installed on the SVM.
- Storage data replicator (SDR): Handles volumes replication activities between PowerFlex systems through journal management. The SDR is located alongside the SDS.

- Storage data client (SDC): A lightweight device driver that presents PowerFlex volumes as block devices to the application on the server which the SDC is installed. The SDC creates a logical adapter, which is an ESXi kernel construct. The adapter informs ESXi about the arrival and disappearance of SCSI devices. These LUNs can be formatted with VMFS and then exposed using the ESXi host to the virtual machine or can be used as RDM devices.

NOTE: NVMe target (SDT) is not supported in VMware hyperconverged configuration.

This implementation is shown in the figure. The SDC is installed within the ESXi kernel similar to any other VIB.

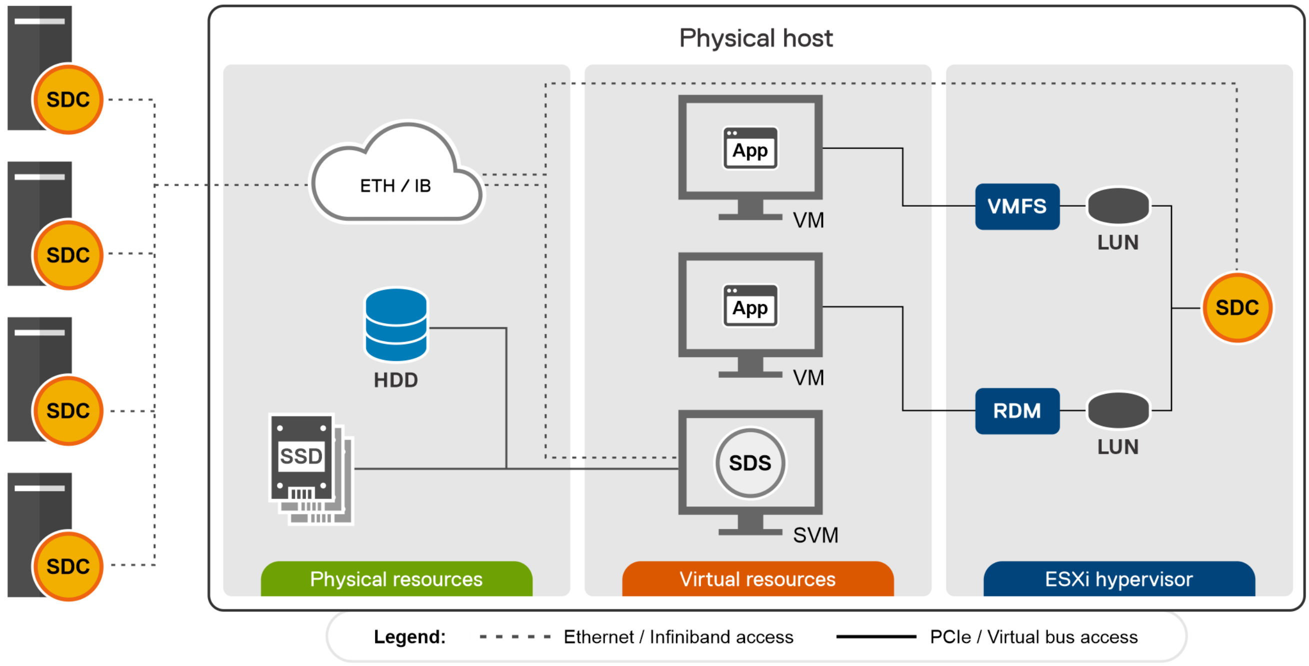


Figure 9. PowerFlex implementation in ESXi with SDC in VMkernel

The LUNs in this example can be formatted with VMFS, and then exposed using the ESXi host to the virtual machine, or the LUNs (volumes) that are NVMe target based can be used as RDM devices. When the LUNs are used as RDM devices, the VMFS layer is omitted.

Supported features

The following VMware features are supported by PowerFlex:

- vMotion
- Storage vMotion
- Fault tolerance
- DRS
- Storage DRS
- VAAI (except full copy primitive)

Device management

Devices can be managed with VMDirectPath I/O.

Pre-deployment considerations

You should take these considerations into account before deploying the system:

ESXi Integrated feature support

ESXi vStorage APIs for Array Integration

ESX vStorage APIs for Array Integration provides hardware acceleration functionality. It allows the host to offload specific virtual machine and storage management operations to compliant storage hardware. With the assistance of the storage hardware, the host performs these operations faster, and consumes less CPU, memory, and storage fabric bandwidth..

VAAI uses these fundamental operations:

- Atomic Test & Set, which is used during creation and locking of files on the VMFS volume
- Clone Blocks/Full Copy/XCOPY, which is used to copy or migrate data within the same physical array
- Zero Blocks/Write Same, which is used to zero-out disk regions
- Thin Provisioning in ESXi 5.x and later hosts, which allows the ESXi host to tell the array when the space previously occupied by a virtual machine (whether it is deleted or migrated to another datastore) can be reclaimed on thin provisioned LUNs.
- Block Delete in ESXi 5.x and later hosts, which allows for space to be reclaimed using the SCSI UNMAP feature.

The PowerFlex supported VAAI features are:

- Atomic Test & Set
- Zero Blocks/Write Same
- Thin Provisioning in ESXi 6.x and later hosts
- Block Delete in ESXi 6.x and later hosts

The following output is an example of typical output:

```
esxcli storage core device vaai status get -d
eui.7dbf14034834bbe01bf7e55800000002
eui.7dbf14034834bbe01bf7e55800000002
VAAI Plugin Name:
ATS Status: supported
```

`Clone Status: unsupported` This means that Clone Block/Full Copy/Xcopy is not supported.

`Zero Status: supported` This means that write same is supported.

`Delete Status: supported` This means that UNMAP is supported.

NOTE: Thin provisioning is not shown in VAAI output.

VMWare SIOC

While Storage I/O Control (SIOC), Storage I/O Statistics Collection (SIOSC) and Network I/O Control (NIOC) are particularly useful for vSan environments, their implementation may actually cause significant issues in a PowerFlex environment, so use of these options is not supported.

PowerFlex provides built-in capabilities to limit network bandwidth and IOPS limits for each volume for each SDC.

Consumption models

PowerFlex is offered in the following consumption models: PowerFlex appliance, PowerFlex rack, PowerFlex software, PowerFlex custom node, and VxFlex Ready Node.

PowerFlex appliance

PowerFlex appliance is a scalable system with flexible form factors that comes pre-configured and validated for fast, easy deployment.

Table 27. PowerFlex appliance components

Component	PowerFlex appliance Intelligent Catalog
PowerFlex software	Included

Table 27. PowerFlex appliance components (continued)

Component	PowerFlex appliance Intelligent Catalog
Operating system and drivers	Included
System firmware ● BIOS ● Cards ● Drives	Included
Infrastructure firmware ● Cabinet / PDU ● IPI Appliance	N/A
CloudLink	Included
VMware NSX	Customer-managed
Network switch firmware	Full network automation is included Partial network automation: Support Matrix

PowerFlex rack

PowerFlex rack is a rack-scale system that is manufactured, managed, supported, and sustained as one system for single-end-to-end lifecycle support.

Table 28. PowerFlex rack components

Component	PowerFlex rack Release Certificate Matrix
PowerFlex software	Included
Operating system and drivers	Included
System firmware ● BIOS ● Cards ● Drives	Included
Infrastructure firmware ● Cabinet/PDU ● IPI Appliance	Included
CloudLink	Included
VMware NSX	Customer-managed
Network switch firmware	Included

PowerFlex software

Table 29. PowerFlex software components

Component	Software management catalog
PowerFlex software	Included
Operating system	Support Matrix
System firmware and drivers ● BIOS ● Cards ● Drives	Customer-managed

Table 29. PowerFlex software components (continued)

Component	Software management catalog
Infrastructure firmware ● Cabinet / PDU ● IPI Appliance	N/A
CloudLink	Support Matrix
VMware NSX	Customer-managed
Network switch firmware	Customer-managed

PowerFlex custom node and VxFlex Ready Node

Table 30. PowerFlex custom node and VxFlex Ready Node components

Component	Software management catalog
PowerFlex custom node and VxFlex Ready Node	Included
Operating system	Support Matrix
System firmware and drivers ● BIOS ● Cards ● Drives	Support Matrix
Infrastructure firmware ● Cabinet/PDU ● IPI Appliance	N/A
CloudLink	Support Matrix
VMware NSX	Customer-managed
Network switch firmware	Customer-managed

Additional functions

This section lists additional functions of PowerFlex.

Physical layer

This section describes the PowerFlex physical layer and outlines the steps for implementing it.

The physical layer consists of the hardware (servers with storage devices and the network between them) and the software that is installed on them.

Typically, each SDS is physically on a separate server.

To implement the physical layer, perform the following steps:

1. Install the MDM component on the MDM nodes in one of the following configurations:
   - Three-node redundant cluster (one primary MDM, one secondary MDM, and one tiebreaker).
   - Five-node redundant cluster (one primary MDM, two secondary MDMs, and two tiebreakers).
   - Single node (one primary MDM).

     > **NOTE:** It is not recommended to use single mode in production systems, except in temporary situations. The MDM contains all the metadata required for system operation. Single Mode has no protection, and exposes the system to a single point of failure.

MDMs do not require dedicated nodes. They can be installed on nodes hosting other PowerFlex components.

2. Install the SDS component on all nodes that contributes some, or all, of their physical storage.
   - Divide the SDS nodes into protection domains. Each SDS can be a member of only one protection domain.
   - Per protection domain, divide the physical storage units into storage pools, and optionally, into fault sets.
3. Install the SDC component on all nodes on which the application accesses the data exposed by the PowerFlex volumes.
   - In the following figure, the gray boxes represent a storage node which hosts the SDS, SDC, and MDM where applicable. In the cases where the gray boxes host the SDS, SDC, and MDM it is a three-node MDM cluster.

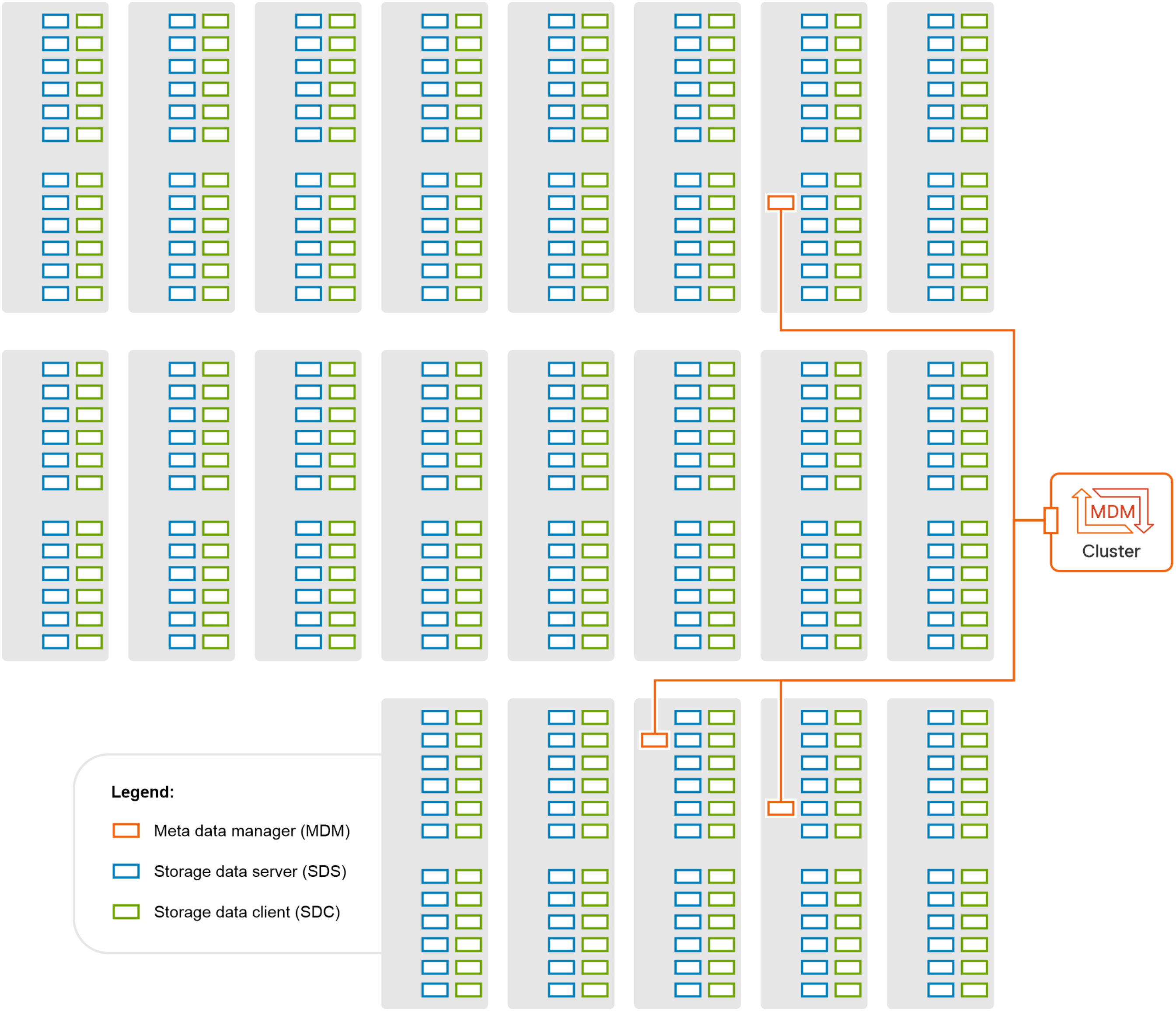


Figure 10. Physical layout example—three-node MDM cluster

Communication is done over the existing LAN using standard TCP/IP. The MDM and SDS nodes can be assigned to up to eight IP addresses, enabling wider bandwidth and better I/O performance and redundancy.

SAN virtualization layer

This section details the steps for exposing the virtual SAN devices to servers with installed and configured SDCs.

The MDM cluster manages the entire system. It aggregates the entire storage that is exposed to it by all the SDSs to generate a virtual layer - virtual SAN storage. Volumes can now be defined over the storage pools and can be exposed to the applications as a local storage device using the SDCs.

To expose the virtual SAN devices to your servers (the ones on which you installed and configured SDCs), perform the following:

1. Define volumes. Each volume that is defined over a storage pool is evenly distributed over all members using a RAID protection scheme. By having all SDS members of the storage pool participate, PowerFlex ensures:
   - Highest and most stable and consistent performance possible
   - Rapid recovery and redistribution of data
   - Massive IOPS and throughput

   You can define volumes as follows:

   - Thick

     Capacity is allocated immediately, even if not actually used. This can cause capacity to be allocated, but never used, leading to wasted capacity.

     Thick capacity provisioning is limited to available capacity.

   - Thin

     Capacity is “on reserve,” but not allocated until actually used. This policy enables more flexibility in provisioning.

     Whereas thick capacity is limited to available capacity, thin capacity provisioning can be oversubscribed, as follows:

     Maximum thin capacity provisioning = 5 * (total raw capacity - used capacity)

     When capacity usage reaches the level where it may cause I/O errors, alerts are generated. At certain higher capacity levels, volumes (even thin volumes) can no longer be created.

     Example:

     In a system with 3 SDSs, each with 10 TB, there are 30 TB of storage.

     In the system, there is already a thick-provisioned volume that takes up 15 TB of the total raw capacity (created by adding a 7.5 TB volume).

     MDM will allow a total of 300 TB total raw capacity to be provisioned, and since 15 TB are already allocated, you can add a thin-provisioned volume of 285 TB total raw capacity (by adding a 142.5 TB volume) or a thick-provisioned volume of 15 TB total raw capacity.

     ⓘ **NOTE:** This example uses 10x fine granularity over provisioning plus compression support.
2. Map volumes. Designate which SDCs can access the given volumes. This gives rise to the following:
   - Access control per volume exposed
   - Shared nothing or shared everything volumes

     Once an SDC is mapped to a volume, it immediately gets access to the volume and exposes it locally to the applications as a standard block device. These block devices appear as `/dev/scini`*X* where *X* is a letter, starting from “a.”

     For example:

     - `/dev/scinia`
     - `/dev/scinib`
   - The maximum amount of partitions for the scini disk is 15.
3. The maximum amount of volumes that can be mapped to an SDC is listed in the “Supported capabilities and system limits” table.

   ⓘ **NOTE:** SDC mapping is similar to LUN mapping, in the sense that it only allows volume access to clients that were explicitly mapped to the volume.

Audit logging

PowerFlex 4.5.1 enables the redirection of the PowerFlex events and Ingress audit messages to syslog. These events and audit messages are forwarded to an external Security Information and Event Manager (SIEM). You must perform the following configuration steps to enable this feature in PowerFlex Manager:

- Change Ingress settings to emit audit messages.
- Define a notification policy to forward Ingress audit messages in the PowerFlex system to an SIEM.
- Define a notification policy to forward events in the PowerFlex system to an SIEM.

For more information about the configuration steps, see the *Dell PowerFlex 4.5.x Administration Guide*.

Maintenance

Maintenance of PowerFlex is primarily limited to configuration changes of the physical and virtual layers. It requires minimal user attention. When maintenance or planned restart of an SDS is required, the maintenance mode feature can be used to streamline system operation.

Maintain the physical layer

Adding/removing hardware units and configuring them into PowerFlex can result from scale out, hardware failure, OS patching/upgrade and hardware expansion.

In the physical layer, maintenance is limited to adding and removing hardware units and configuring them into PowerFlex. These operations are usually a result of:

- Scaling out when there is a need for additional capacity. This usually results in adding more storage media to the existing servers, or adding additional servers.
- Hardware failure. In cases where there is a hardware (storage media or server) failure and it needs to be replaced.

  In all of the above cases, the operation will require adding or removing storage capacity from the system. In some cases, it may include adding or removing an entire server, and its associated storage media, from the configuration. As far as PowerFlex is concerned, all of these activities translate to SDS reconfigurations.

  If the removed server is an SDC node, or the server to be added requires exposing storage locally, SDC reconfiguration will happen as well.
- Adding or removing storage media. Add or remove the media from the SDS with which it is associated. PowerFlex will redistribute the data accordingly and seamlessly.
- OS patching/Upgrade or Hardware expansions. Graceful shutdown\reboot operations with return to fully operation state flow.

Instant maintenance mode (IMM)

Using instant maintenance mode, you can restart a server hosting an SDS without requiring shutdown or interrupting application I/Os. Storing all writes created during maintenance to a dedicated fault set prevents data loss in case of a single failure.

Instant maintenance mode enables you to restart a server that hosts an SDS, with minimal impact on PowerFlex, thus bypassing the disruption and effort caused by disorderly shutdown, protection domain shutdown, and orderly shutdown.

Whereas PowerFlex always uses two copies of user data, invoking maintenance mode introduces an additional copy that stores all writes created during maintenance to an SDS or fault set (created during maintenance) in both a primary location and a new location. This copy prevents data loss if a single failure occurs.

When the SDS or fault set is returned from maintenance mode, only the new writes are required to be resynchronized, thus minimizing data transfer during and after the update.

Instant maintenance mode does not interrupt application I/Os; it can be run on any amount of members of a fault set; and it can run in parallel on different protection domains. While an SDS is in maintenance mode, most PowerFlex operations (like adding a volume) cannot be performed in the fault set, protection domain, or storage pool in which the SDS and its devices reside.

To invoke maintenance mode, the following conditions are required:

- Only one Fault Unit (or standalone SDS) can be in maintenance mode at any given time.
- No other SDSs can be in degraded or failed state (force override can be used).
- There must be adequate space on other SDSs for the additional backup (force override can be used).

  ⓘ **NOTE:** Use of force override options when entering maintenance mode can lead to data unavailability while maintenance mode is activated.

  While an SDS is in maintenance mode, it can be shut down with no danger to data.

Protected maintenance mode (PMM)

PMM is anther type of maintenance mode that ensures no data loss during down time.

PMM is another type of maintenance state where a third copy is created. If there should be a node failure during IMM, then there is no SDS up to serve I/Os for the data. With PMM, a third copy is created before maintenance mode, which ensures that if there is a node failure where the second copy of SDS is required, there is still a full backup of the SDS.

Auto abort PMM

PMM consumes a significant amount of capacity. In some cases, this capacity may be required for more important tasks such as user I/O or rebuild.

The purpose of this capability is to abort PMM when approaching the end of capacity, to avoid impacting I/O or rebuild. For example, if a single node is entering PMM and a third copy is being created, then a second node fails. Auto abort will make sure that the third copy space is made available to perform a rebuild.

Maintain the virtualization layer

This section lists the operations that can be performed on volumes exposed by the PowerFlex virtual SAN.

The following operations may be performed on volumes that are exposed by the PowerFlex virtual SAN:

- Add or remove a volume:

  Create or delete a volume in the system.

- Increase volume size:

  Add capacity to a given volume, as needed. The change in volume size occurs seamlessly without interrupting I/O.

- Map and unmap volumes to an SDC:

  This enables or disables access to a volume by an SDC, and thus by an application residing on the same node.

- MDM operations:
  - Move the MDM role to another role
  - Add MDM VIP
  - Upgrade of MDM, SDS, SDC and LIA

Decommisioned features

Read Flash Cache (RFCache) has been removed from the Powerflex 4.0 product offering.

When upgrading a system to 4.0 that has RFcache, you must remove the RFCache configuration and files before proceeding with the upgrade.

PowerFlex management

PowerFlex management covers all product functionalities, including block storage, file storage, hardware management, network management, reporting, and alerts.

- The management can be accessed through a Web UI or REST API.
- PowerFlex Manager handles user management for all systems, including:
  - PowerFlex local users
  - LDAP/AD users
- PowerFlex Manager supports system security, as follows:
  - PowerFlex management platform exposes an HTTPS interface for Web UIs and REST APIs.
  - Users must have an account defined in PowerFlex.
  - Upon login, users receive an access token.
  - PowerFlex Manager uses OpenID Connect (OIDC).
  - PowerFlex management platform communication with PowerFlex file and block storage systems is secured using mTLS.

- PowerFlex management platform is a cluster of three or more self-managed Kubernetes nodes, providing resiliency and high availability.
  - PowerFlex management platform run time optimized to 60 minutes.
- The following CLI tools are available:
  - A CLI that provides full control over the block storage system.
  - A CLI with expanded support for management, reporting, alerting, and file storage.
- Enhanced security strength by using SSH keys and non-root users through sudo in management aspects.

  For deployment and post-deployment procedures, see the *Dell PowerFlex 4.5.x Install and Upgrade Guide* and *Dell PowerFlex 4.5.x Administration Guide* for configuration guidelines.

- Webhooks support enables sending alerts to the Webhooks servers such as BigPanda.
- Multi-Subnet and Multi-VLAN for PowerFlex appliance and PowerFlex rack allows defining multiple subnets and VLANs in a single network.

  The most common implementation example is with a management or vMotion network that spans multiple racks. Each rack is its own subnet and VLAN.

  ⓘ **NOTE:** Supported on all networks except for the following: PowerFlex Data, vSAN, and NSX external networks.
- Higher flexibility of fault sets defines a subset of nodes in a rack as part of a fault set during deployment and expansion operations.

  For example, a rack with 20 nodes can have four fault sets, one for every five nodes, for easier maintenance support. Another option is using fault sets in a full rack configuration, such as one fault set per rack.

- PowerFlex management controller self-awareness on PowerFlex appliance and PowerFlex rack allows self-awareness of its underlying controller system.
  - Upgrade from end to end and receive alerts from issues in the PowerFlex management controller infrastructure.
- Modular (single) component upgrade changes the package in the PowerFlex rack and PowerFlex appliance catalogs to newer or older versions.

  This upgrade depends on customer cases that want to deviate from the Release Certification Matrix or Intelligent Catalog.

  ⓘ **NOTE:** By default this feature is turned off and can only be enabled through an RPQ process. If applicable, contact your account team for details.

PowerFlex Manager

PowerFlex Manager is a unified tool that is used for the management and operations of PowerFlex. PowerFlex is a software-defined storage platform that is designed to reduce operational and infrastructure complexity empowering organizations with predictable performance and resiliency at scale.

PowerFlex Manager provides IT operations management for PowerFlex systems. It increases efficiency by reducing time-consuming manual tasks that are required to manage system operations. PowerFlex Manager can also be used to deploy and manage new and existing PowerFlex rack and PowerFlex appliance systems.

The PowerFlex Manager interface provides PowerFlex block and file capabilities, along with lifecycle mode and resource management capabilities.

The PowerFlex Manager unified management interface includes the following set of tabs:

- A dashboard that enables you to see a summary of performance and inventory data.
- A block tab that provides a user interface for managing PowerFlex block storage.
- A file tab that provides a user interface which incorporates PowerFlex file services capabilities for managing file storage.
- A protection tab that provides a user interface for performing PowerFlex storage protection operations.
- A lifecycle tab that allows you to create and clone templates and deploy them as resource groups (previously known as services).
- A resources tab that allows you to manage all resource types within the system.
- A monitoring tab that allows you to track events, alerts, and jobs running.
- A settings tab that enables you to configure system settings for PowerFlex Manager.

PowerFlex Manager is part of all consumption models whether that be PowerFlex rack, PowerFlex appliance, VxFlex Ready Node, PowerFlex custom node, or PowerFlex software deployments.

PowerFlex Manager is designed as microservices, specialized modules working together and based on the Rancher Kubernetes Engine.

Management capabilities

PowerFlex Manager provides the unified interface and full lifecycle mode (LCM) services. Role-based and scope-based access control enables common PowerAPIs to improve ecosystem and customer integration of the PowerFlex storage solution.

PowerFlex provides one tool for all consumption models.

- PowerFlex Manager (UX, API, CLI)
  - Unified dashboard
  - REST API for PowerFlex rack, PowerFlex appliance, and PowerFlex software
  - PowerFlex Block and PowerFlex file services
  - SDC and NVMe over TCP hosts
- Unified services:
  - Security
  - Single sign-on (SSO) and role-based access control (RBAC)
  - Deployment and LCM
  - Events, alerts, and reporting
  - Logging and metrics collection
  - SNMP, Secure Remote Services, SMTP, CloudIQ
- Unified RBAC and user management
  - No separate users or LDAP configuration
  - Single set of roles
  - Meta data manager (data-path) does not maintain RBAC from PowerFlex 4.0.
- Single WebUI
- Single REST endpoint for all APIs: Legacy and modern, ISG-standard PowerAPI

  NOTE: For login information, see the PowerFlex 4.5 API documentation, at https://developer.dell.com.
- Events (log) and alerts (state) including meta data manager (MDM) events.
- Resilient management stack (HA)

Serviceability

PowerFlex Manager provides support through integration with the secure connect gateway ensuring better alignment with Dell Technologies Services initiatives to enhance the user experience.

Registering your device with SupportAssist through the secure connect gateway, enables you to do the following:

- Collect system state information and telemetry.
- Automate issue detection and support case creation.
- Resolve common issues remotely.
- Provide critical information to Dell Technologies remote support to resolve complex issues.

Deployment options for each management offering

PowerFlex management platform is an integrated part of the PowerFlex system. PowerFlex management platform can be set up in multiple deployment options, regardless of the consumption model.

Depending on which consumption model you choose, you have different management options within the solution stack. PowerFlex Manager manages some aspects of the solution while other elements are customer-managed.

The supported deployment options available are as follows:

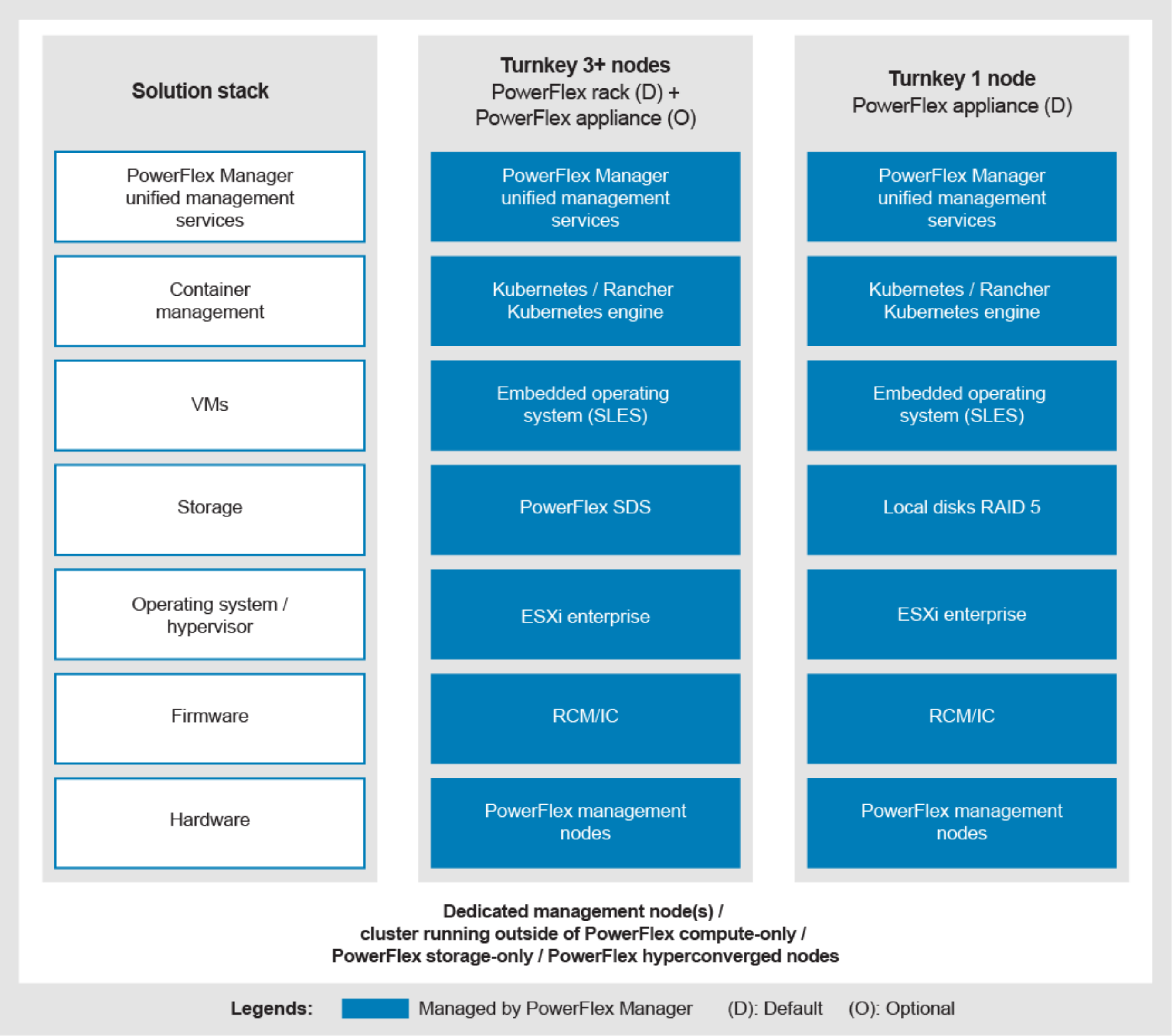
- Co-resident: Production nodes and storage hosts management instances of Kubernetes and containers.
- Bring your own hypervisor: Your hypervisor and shared storage hosts management VMs with Kubernetes.
- Single management node: A few disks in RAID 5 for shared storage hosts management VMs with Kubernetes.
- This option uses three or mode nodes that are configured to create a Management ESXi cluster running a management PowerFlex cluster which provides the shared storage datastore for management operations.

Table 31. Deployment option per offering

Supported offering	Co-resident	Bring your own hypervisor	Single management node	Three or more management nodes
PowerFlex software	Yes	Yes	No	No
PowerFlex custom node or VxFlex Ready Node	Yes	Yes	Yes	No
PowerFlex appliance	No	Yes Not available by default. Contact Dell Technologies Support for more information.	Yes	Yes
PowerFlex rack	No	Yes Not available by default. Contact Dell Technologies Support for more information.	No	Yes

Dedicated management cluster model

PowerFlex Manager can be deployed on dedicated management nodes or clusters running outside of PowerFlex storage-only nodes, PowerFlex compute-only nodes, or PowerFlex hyperconverged nodes. When you deploy a dedicated management cluster, PowerFlex Manager manages the complete solution stack.



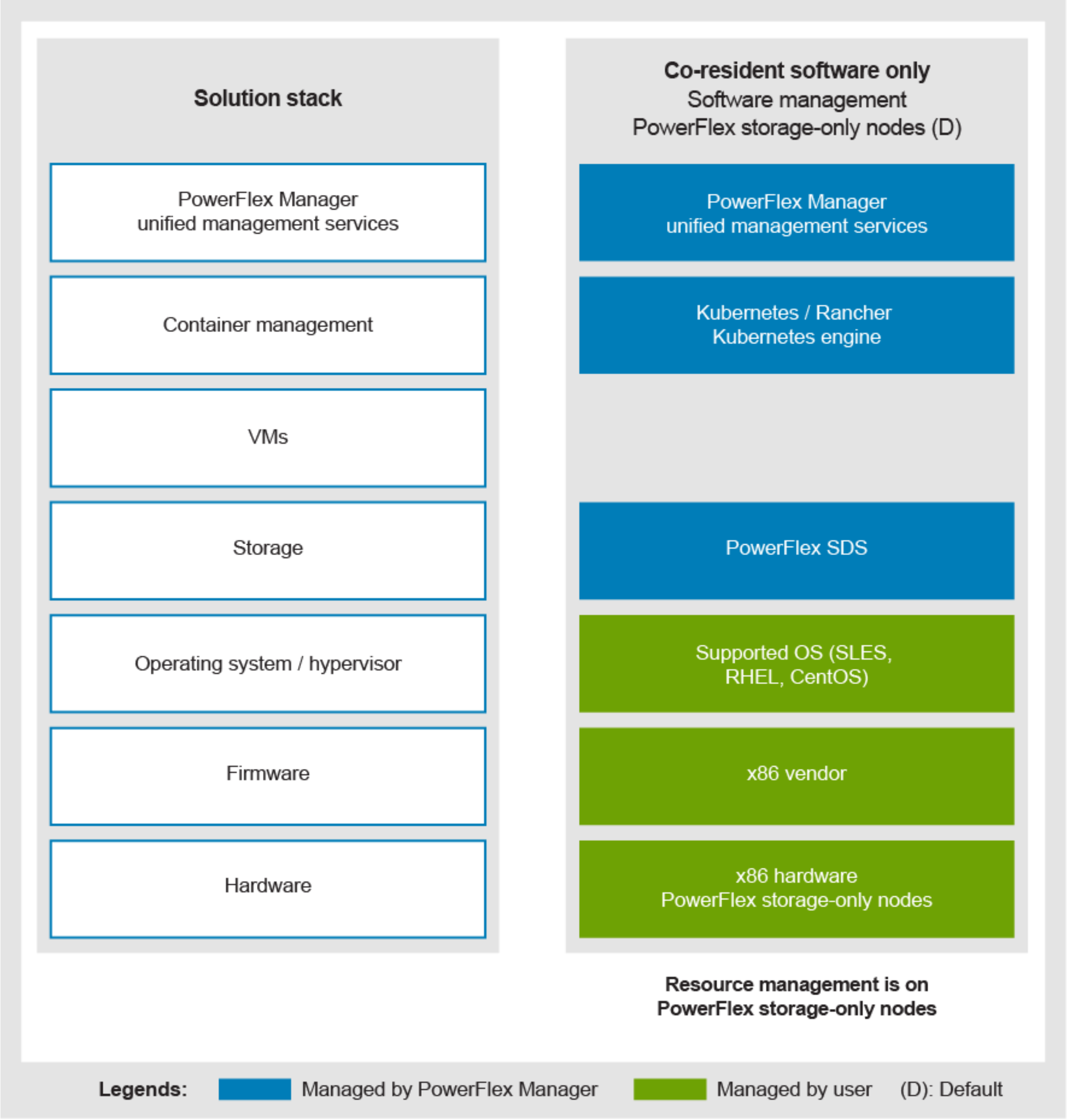
Solution stack	Turnkey 3+ nodes PowerFlex rack (D) + PowerFlex appliance (O)	Turnkey 1 node PowerFlex appliance (D)
PowerFlex Manager unified management services	PowerFlex Manager unified management services	PowerFlex Manager unified management services
Container management	Kubernetes / Rancher Kubernetes engine	Kubernetes / Rancher Kubernetes engine
VMs	Embedded operating system (SLES)	Embedded operating system (SLES)
Storage	PowerFlex SDS	Local disks RAID 5
Operating system / hypervisor	ESXi enterprise	ESXi enterprise
Firmware	RCM/IC	RCM/IC
Hardware	PowerFlex management nodes	PowerFlex management nodes

**Dedicated management node(s) /
cluster running outside of PowerFlex compute-only /
PowerFlex storage-only / PowerFlex hyperconverged nodes**

Legends: 🟦 Managed by PowerFlex Manager (D): Default (O): Optional

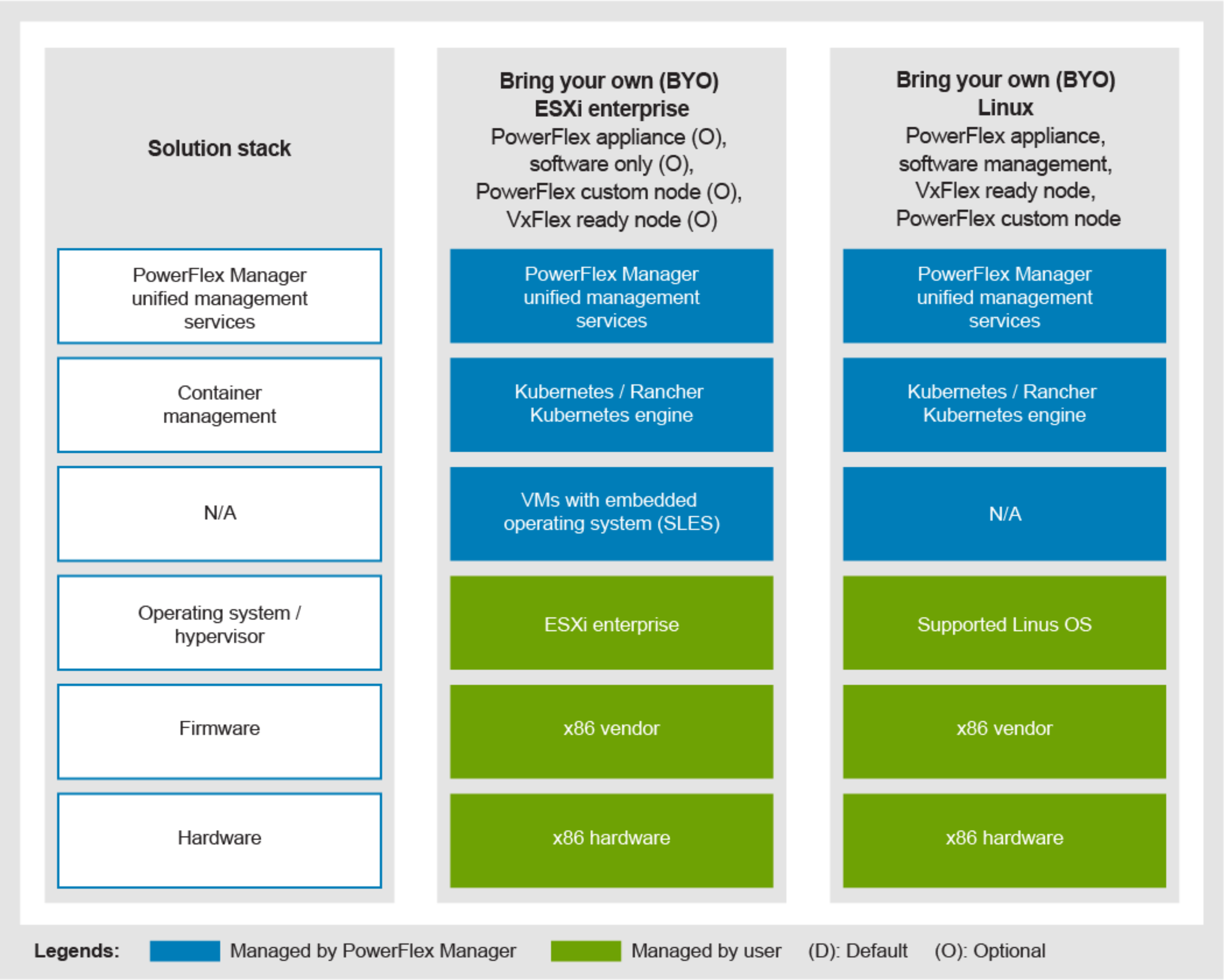
Co-resident software management model

PowerFlex Manager supports co-residency where it can be deployed on PowerFlex storage-only nodes. In this management option, PowerFlex Manager manages only the containers and PowerFlex as illustrated in the diagram below:


Solution stack
PowerFlex Manager unified management services
Container management
VMs
Storage
Operating system / hypervisor
Firmware
Hardware
Co-resident software only
Software management
PowerFlex storage-only nodes (D)
PowerFlex Manager unified management services
Kubernetes / Rancher Kubernetes engine
PowerFlex SDS
Supported OS (SLES, RHEL, CentOS)
x86 vendor
x86 hardware PowerFlex storage-only nodes
Resource management is on PowerFlex storage-only nodes
Legends:
Managed by PowerFlex Manager
Managed by user
(D): Default

Non co-resident bring your own management model

PowerFlex Manager supports non co-residency where it can be deployed on a supported operating system of your choice. In this management option, PowerFlex Manager manages only the containers and PowerFlex.



Solution stack	**Bring your own (BYO) ESXi enterprise** PowerFlex appliance (O), software only (O), PowerFlex custom node (O), VxFlex ready node (O)	**Bring your own (BYO) Linux** PowerFlex appliance, software management, VxFlex ready node, PowerFlex custom node
PowerFlex Manager unified management services	PowerFlex Manager unified management services	PowerFlex Manager unified management services
Container management	Kubernetes / Rancher Kubernetes engine	Kubernetes / Rancher Kubernetes engine
N/A	VMs with embedded operating system (SLES)	N/A
Operating system / hypervisor	ESXi enterprise	Supported Linus OS
Firmware	x86 vendor	x86 vendor
Hardware	x86 hardware	x86 hardware

Legends: 🟦 Managed by PowerFlex Manager 🟩 Managed by user (D): Default (O): Optional

Configure direct attached storage

When adding direct attached storage to SDS, it is recommended to configure the raw devices as stand-alone devices. Upon adding a device to an SDS PowerFlex verifies that the device is clear and returns an error if it is otherwise.

PowerFlex works with any free capacity—internal or direct-attached devices, either magnetic hard disk drives or flash-based devices such as solid state drive (SSD) and PCIe cards. Although PowerFlex can work with any device topology, it is recommended to configure the raw devices as stand-alone devices.

Device data is erased when devices are added to SDS. When adding a device to an SDS, PowerFlex checks that the device is clear before adding it. An error is returned, per device, if it is found not to be clear. You can override this check by using the force device takeover option.

The following devices are considered to be not "clear," and thus cannot be added to SDS:

- Linux - A complete device with either a file system or partition, or a partitioned device with a file system.
    - Multipath devices cannot be added as SDS devices.
    - Devices in an LVM group cannot be added to an SDS.
- ESXi - Same as above, depending on the operating system of the SVM where the SDS is installed.

Limitations:

- SAN devices will not be prevented from being added.
- Devices in an LVM group cannot be added to an SDS.
- Within the database devices, only Oracle ASM devices can be detected and blocked.

ⓘ **NOTE:**

If the server has a RAID controller, PowerFlex prefers to use the caching abilities of the controller for better performance, but is better used when all devices are configured as stand-alone (that is setting each of the devices to RAID 0 separately). Dell Technologies recommends enabling RAID controller caching for hard disk drive devices. Caching is not needed for SSD devices.

ⓘ **NOTE:**

For hard disk drives: It is recommended to use RAID controller caching when available as follows:

- READ/WRITE: If the cache is battery-backed
- READ ONLY: If the cache is NOT battery-backed

For SSDs (such as Flash): setting should be write-through