Technical White Paper

Dell EMC PowerFlex: Introduction to Replication

Overview and basic configuration of PowerFlex replication

Abstract

Dell EMC PowerFlex™ software-defined provides native asynchronous replication. This paper provides an overview of the PowerFlex replication technology along with deployment and configuration details as well as design considerations for replicating PowerFlex clusters.

June 2021

H18391.2

Revisions

Date	Description
June 2020	Initial release
May 2021	Updates to journal capacity and network recommendations
June 2021	Updates for PowerFlex version 3.6

Acknowledgments

Content Owner: Brian Dean, PowerFlex Technical Marketing

Support: Roy Laverty, Matt Hobbs, Neil Gerren

The information in this publication is provided "as is." Dell Inc. makes no representations or warranties of any kind with respect to the information in this publication, and specifically disclaims implied warranties of merchantability or fitness for a particular purpose.

Use, copying, and distribution of any software described in this publication requires an applicable software license.

This document may contain certain words that are not consistent with Dell's current language guidelines. Dell plans to update the document over subsequent future releases to revise these words accordingly.

This document may contain language from third party content that is not under Dell's control and is not consistent with Dell's current guidelines for Dell's own content. When such third party content is updated by the relevant third parties, this document will be revised accordingly.

Table of contents

Executive summary

PowerFlex™ software-defined infrastructure platform delivers unmatched flexibility, elasticity, and simplicity with predictable performance and resiliency at scale. As PowerFlex continues to evolve, the addition of native asynchronous replication expanded the set of included enterprise storage services. Customers require disaster recovery and replication features to meet business and compliance requirements. Replication can also be leveraged for other use cases, such as offloading demanding analytics workloads, isolating them from mission-critical workloads other business-critical systems. This paper covers:

- The core design principles of PowerFlex replication
- Configuration requirements for pairing storage clusters
- Configuration requirements of Replication Consistency Groups
- Networking Considerations
- Replication use cases

1 Introduction

PowerFlex is a software-defined storage platform designed to reduce operational and infrastructure complexity empowering organizations to move faster by delivering flexibility, elasticity, and simplicity with predictable performance and resiliency at scale. The PowerFlex family of software-defined infrastructure provides a foundation that combines compute and high-performance storage resources in a managed unified fabric. Flexibility is offered as it comes in multiple platform deployment options such as rack, appliance or ready nodes, all of which provide Server SAN, HCI and storage-only architectures.

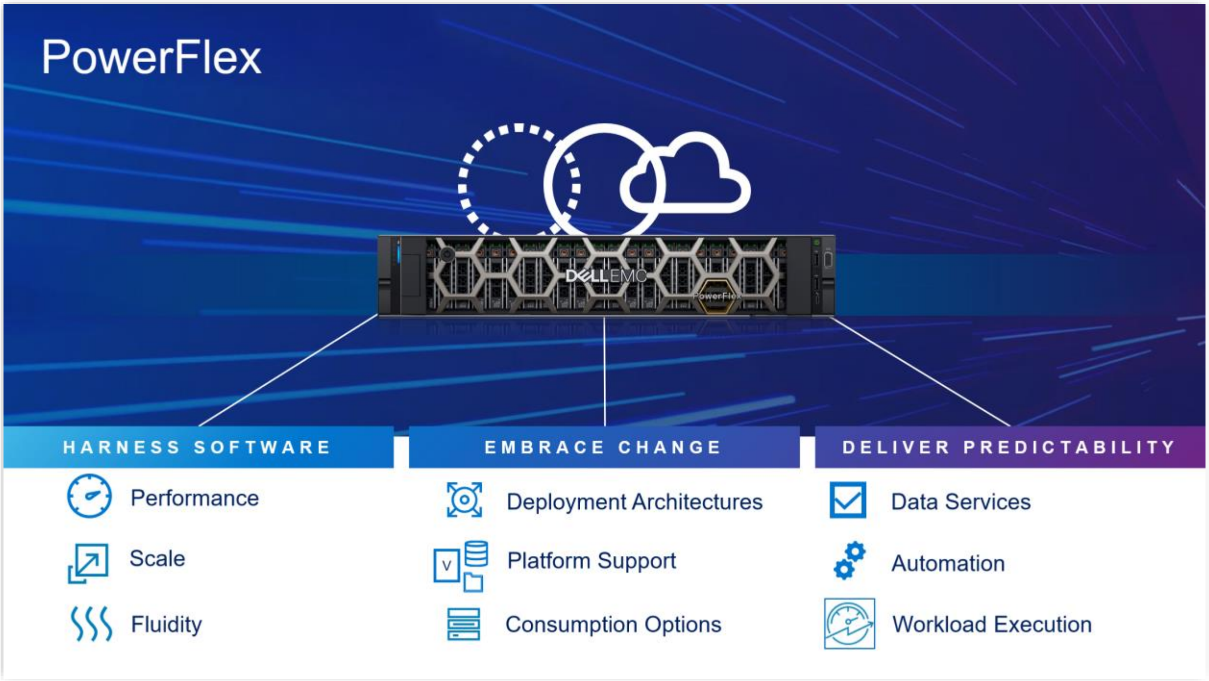


Figure 1 Powerflex Overiew

PowerFlex provides the flexibility and scale demanded by a range of application deployments, whether they’re on bare metal, virtualized, or containerized.

It provides the performance and resiliency required by the most demanding enterprises, demonstrating six 9s or greater of mission-critical availability with stable and predictable latency.

Easily providing millions of IOPs at sub-millisecond latency, PowerFlex is ideal for both high performance applications and for private clouds that desire a flexible foundation with synergies into public and hybrid cloud. It’s also great for organizations consolidating heterogeneous assets into a single system with a flexible, scalable architecture that provides the automation to manage both storage and compute infrastructure.

2 PowerFlex Asynchronous Replication Architecture

To understand how replication works, we must first consider the basic architecture of PowerFlex itself.

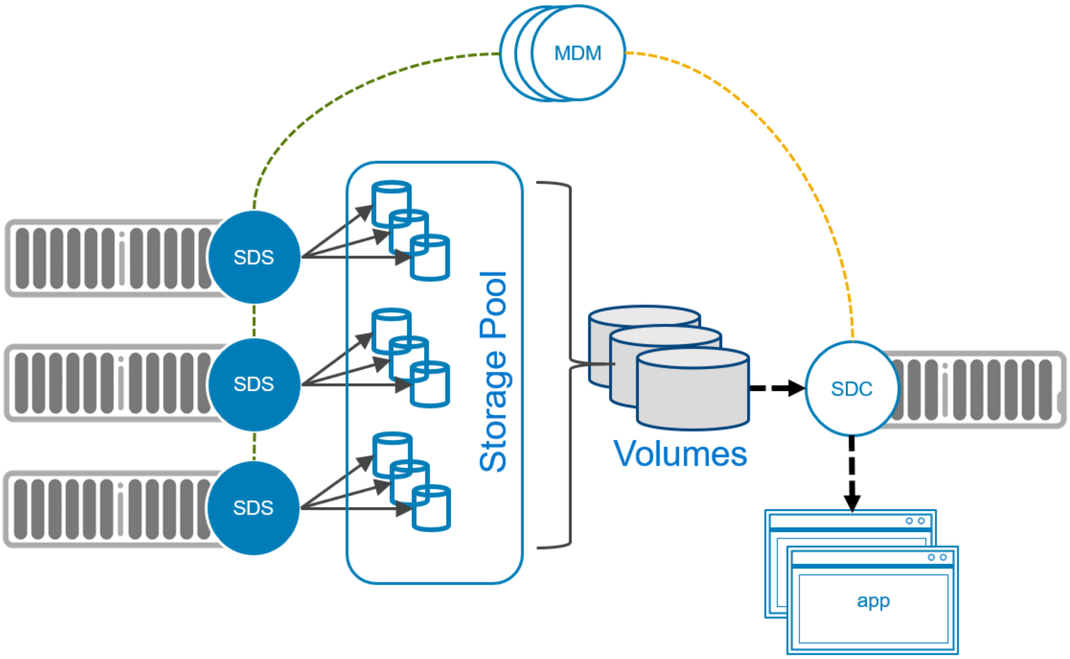


Figure 2 PowerFlex basic components and architecture

Servers contributing media to a storage cluster run the Storage Data Server (SDS) software element which allows PowerFlex to aggregate the media while sharing these resources as one or more unified pools on which logical volumes are created.

Servers consuming storage run the Storage Data Client (SDC) which provides access to the logical volumes via the host SCSI layer. Note that iSCSI is not used, but instead, a resilient load-managing, load-balancing network service which runs on TCP/IP storage networks.

The Metadata Manager (MDM) controls the flow of data through the system but is not in the data path. Instead, it creates and maintains information about volume distribution across the SDS cluster and distributes the mapping to the SDC informing it where to place and retrieve data for each part of the address space.

These three base elements comprise the fundamental parts of best software-defined storage solution today, one that scales linearly to hundreds of SDS nodes.

When considering architectural options for replication, maintaining the scalability and resiliency of PowerFlex was critical. The replication architecture in PowerFlex is a natural extension of the fundamentals just described.

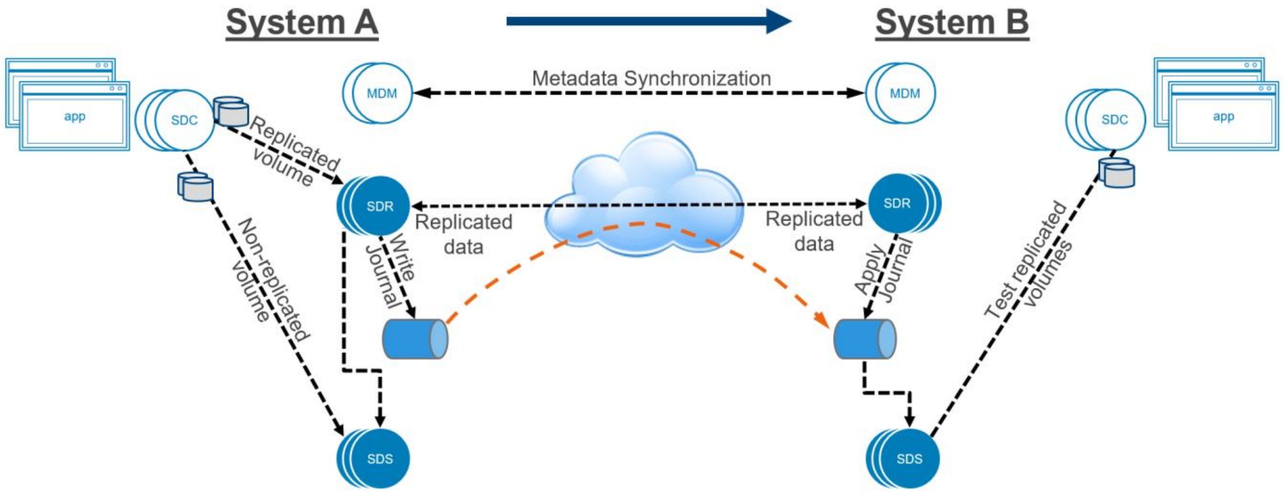


Figure 3 PowerFlex simplified replication architecture

PowerFlex version 3.5 introduced a new storage software component called the Storage Data Replicator (SDR). Figure 3 depicts where the SDR fits into the overall PowerFlex replication architecture. Its role is to proxy the I/O of replicated volumes between the SDC and the SDSs where data is ultimately stored. Write I/Os are split, sending one copy on to the destination SDSs and another to a replication journal volume. Sitting between the SDS and SDC, from the point-of-view of the SDS, the SDR appears as if it were an SDC sending writes. (From a networking perspective, however, the SDR → SDS traffic is still backend/storage traffic.) Conversely, to the SDC, the SDR appears as if it were an SDS to which writes can be sent.

The SDR only mediates the flow of traffic for replicated volumes. (In fact, only actively replicating volumes; the nuance will be covered below). Non-replicated volume I/Os flow, as usual, between SDCs and SDSs directly. As always, the MDM instructs each of the SDCs where to read and write their data. The volume address space mapping, presented to the SDC by the MDM, determines where the volume's data is sent. But the SDC is ignorant of the write-destination as an SDS or an SDR. The SDC is not aware of replication.

2.1 Journaling vs Snapshotting

There are two schools of thought concerning how replication is implemented. Many storage solutions leverage a snapshot-based approach. With snapshots, it's easy to identify the block change delta between two points in time. However, as Recovery Point Objectives get smaller, the number of snapshots required increases dramatically, which places hard limits on how small RPOs can be. Instead, PowerFlex uses a journaling-based approach.

Journal-based replication provides the possibility of very small RPOs, and, importantly, it is not constrained by the maximum number of available snapshots in the system, or on a given volume.

Checkpoints (or intervals) are maintained in journals, and those journals live as PowerFlex volumes in a Storage Pool in the same Protection Domain. However, the journal volume need not reside in the same storage pool as the volume being replicated. The journal volumes resize dynamically as writes are committed, shipped, acknowledged, and deleted. So, the actual capacity used by the journal buffer will vary over time.

2.2 Journal capacity reservations

While the actual capacity varies with usage, the journal reservations (specifying the maximum capacity we will allow the replication processes to consume) must be set manually. Appropriately sizing journal volume reservations is critical to the health of the PowerFlex cluster, especially during WAN outages and other failure scenarios. For example, the journal volume must have enough available capacity to continue ingesting replication data even when the SDR cannot ship the journal intervals to the target site. (Of course, for installations of PowerFlex that do not use replication, no journal space reservations are needed.) If the journal intervals are unable to transmit, the journal buffer capacity will increase, potentially filling it altogether. So, you must consider the maximum cumulative writes that might occur in an outage. If the journal buffer space fills completely, the replica-pair volumes will require re-initialization. More on this topic below.

The administrator sets and adjusts the maximum reservation size of the journal volumes. The minimum requirement for journal capacity is 28GB multiplied by the number of SDR sessions, where SDR sessions equals the number of SDRs installed plus one. However, some additional calculation is required, because the reservation size is stated in the system as a *percentage* of the storage pool in which each journal volume is contained. As a general rule, reserve at least 5% of the storage pools for replication journals.

The reserved capacity for journals may be split into several volumes across multiple storage pools, or the replication journals may all reside in one storage pool of a protection domain. **The performance character of any storage pool in which a journal volume resides must match or exceed the performance requirements of any storage pool in which the replicated volumes reside.**

While the journal capacity must be sufficient to accommodate factors like volume overhead, free space reservations (to sustain node failures, or accommodate Protected Maintenance Mode), etc., the single most important consideration is a possible WAN outage. If we account for this scenario, we end up accounting for all the other considerations.

To begin, assess the journal capacity needed per application. We need to know the maximum application write bandwidth during the busiest hour, because application I/O varies over time, and we cannot predict when an outage might occur. The minimum outage allowance is 1 hour, but we strongly recommend using three hours in the calculations.

Example calculation

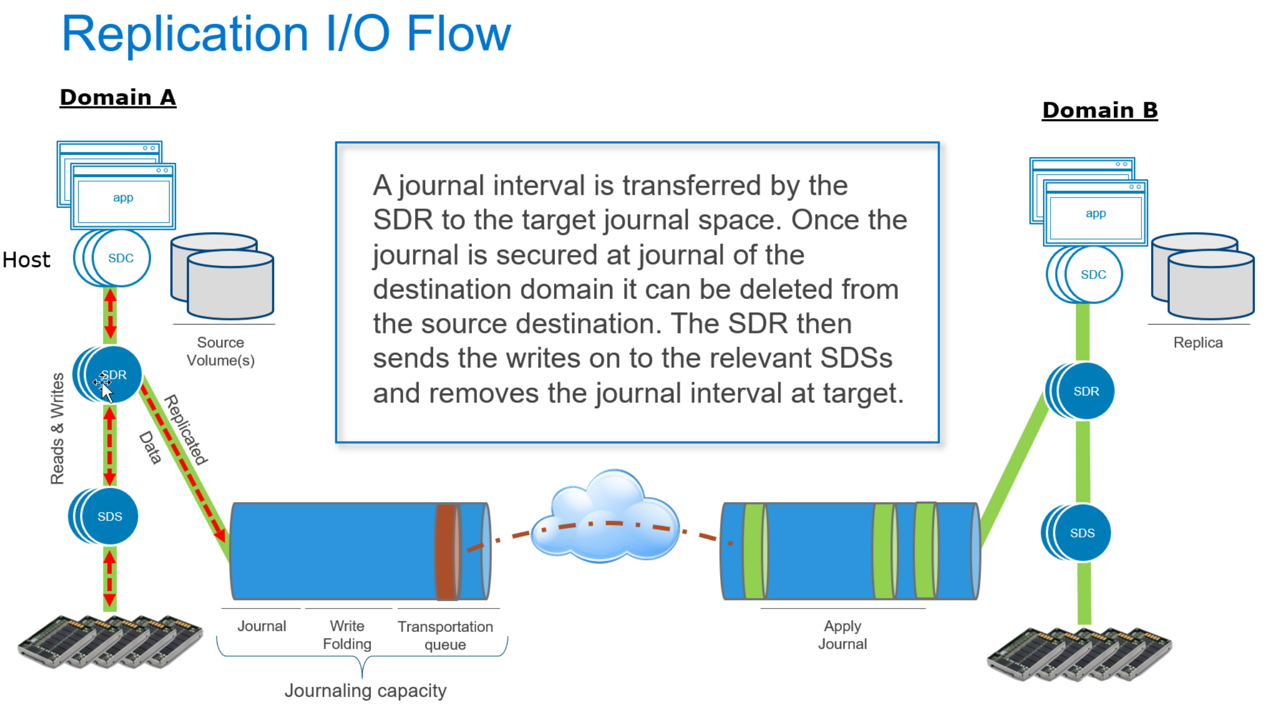
- Our application generates 1GB/s of writes during peak hours
- Using 3 hours as the supported outage, we calculate from 10800 seconds
- The journal capacity reservation needed is 1GB/s * 10800s = ~10.547 TB
- Since journal capacity is calculated as a percentage of Storage Pool capacity, we divide the needed space by the Storage Pool usable capacity. Let’s assume that is 200TB.
- 100 * 10.547TB / 200TB = 5.27%.
    - As a safety margin, we’ll round this up to 6%.

Repeat for each application being replicated.

Note that as the size and capacity of a storage pool changes, the percentage will change. Readjustments to the journal reservation will be necessary as pool capacities vary. Administrators can adjust the journal capacity reservation percentage at any time from the UI, CLI or API.

2.3 Journal intervals and data flow

Each cluster can be both a replication source and a target. This allows customers to split applications between regionally separate clusters, while protecting application availability for either location.



A journal interval is transferred by the SDR to the target journal space. Once the journal is secured at journal of the destination domain it can be deleted from the source destination. The SDR then sends the writes on to the relevant SDSs and removes the journal interval at target.

Figure 4 PowerFlex simplified replication I/O flow

The volume mapping on the source SDC sends writes for replicated data to the SDR, which duplicates the write and forwards it. The local SDSs process those writes normally, while the SDR assembles the journal files which contain checkpoints to preserve write order.

Journals are batched in the journal buffer on the source system. As they near the head of the queue, they are scanned, and duplicate block writes are consolidated (folded) to minimize the volume of data being sent over the wire.

The journal intervals are sent to the remote target journal buffer by the SDR over dedicated subnets on local networks or external WAN networks assigned to replication and, once acknowledged at the target journal, are removed from the source.

On the target system, the journals are processed by the SDR and applied to the target volume, thereby passing the writes on to the relevant SDSs. The SDSs manage the primary and secondary copies as usual.

This raises a frequently asked question: *How is compression affected in replicated volumes?* The short answer is that compressed data is not sent over the WAN. The SDR is a mediator between the SDC and the

SDS, it plays no role in compression. Compression is done by an SDS receiving writes and storing them to disks local to the host on which it runs. Compression in PowerFlex is local to a given SDS. Therefore, data sent over the wire is not compressed.

Once the target-side SDR receives acknowledgement from the target-side SDS, it proceeds to the next write contained in the journal interval being processed. When the last write in a journal interval is processed and acknowledged, the interval is deleted, and the journal capacity is made available for reuse.

There are several other SDR sub-processes working together to protect the integrity of your data, but this description covers all the fundamentals.

There is one limitation worthy of mention related to volume migration. It is not possible to migrate replicated volumes from one Protection Domain to another. This is because the replication journals to not span Protection Domains.

3 Deploying and Configuring PowerFlex clusters for replication

Proper system and storage sizing must be performed before deploying any new PowerFlex clusters. Replication adds additional sizing concerns. Your Dell Technologies technical sales resources have access to a system sizing utility which takes inputs including your workload characterization, your replication footprint, WAN bandwidth and quality, as well as network design and infrastructure.

There are additional cluster setup requirements to consider when adding asynchronous replication. We need

- A way for the clusters participating in replication to communicate securely.
- To group volume pairs together into consistency groups.
- Methods of testing failure, or even distributing workload without impacting the primary application.
- Configuring the physical WAN network for replicating externally when the target cluster is in another datacenter
- Additional IP addresses for replication activity

We cover all these topics in this chapter.

3.1 Deployment and Configuration

When deploying cluster pairs to be used with replication, there are a few required configuration steps.

3.1.1 The exchange of storage cluster Certificate Authority root certificates

PowerFlex system root CA certificates must be exchanged between replicating clusters to protect from possible security attacks. Since this is a security-sensitive issue, this step is performed using the PowerFlex command line interface. On each system, a certificate is created and sent to the other host in the replicated pair. The example command:

```
scli --extract_root_ca --certificate_file /tmp/sys0.cert
```

extracts the certificate from the cluster. The certificate is then manually copied to the partner cluster.

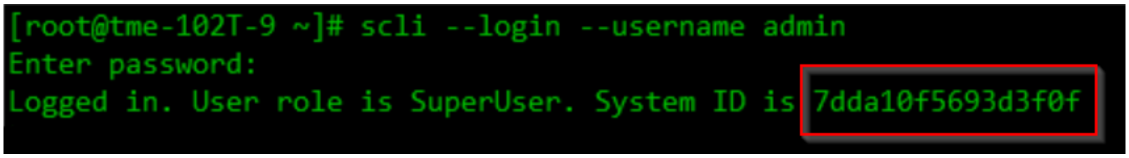
To import the certificate, on the partner system, we use a command of the following form:

```
scli --add_trusted_ca --certificate_file /tmp/sys0.cert --comment Site-A
```

Once the certificates are generated, exchanged, and imported on both systems, the certificate exchange step is complete.

3.1.2 Peering storage clusters

The next step required before configuring replicated volume pairs is Peering. Peering establishes the data paths and communication between the systems. This can be done using the PowerFlex WebUI, but there is one piece of critical information we'll need first. We'll use the PowerFlex CLI to capture the System IDs for both storage clusters. This can be found simply by logging into the PowerFlex CLI. The act of authenticating to the cluster via CLI reveals the cluster ID. You'll need the IDs from both the source and the target systems.



```
[root@tme-102T-9 ~]# scli --login --username admin
Enter password:
Logged in. User role is SuperUser. System ID is 7dda10f5693d3f0f
```

Figure 5 Capturing a PowerFlex System ID

To begin peering, navigate to the PROTECTION side menu in the PowerFlex WebUI and expand it by clicking on it. (This example uses the version 3.6 UI, which is slightly different from the 3.5 UI, but the example can still be easily followed in 3.5.x.)

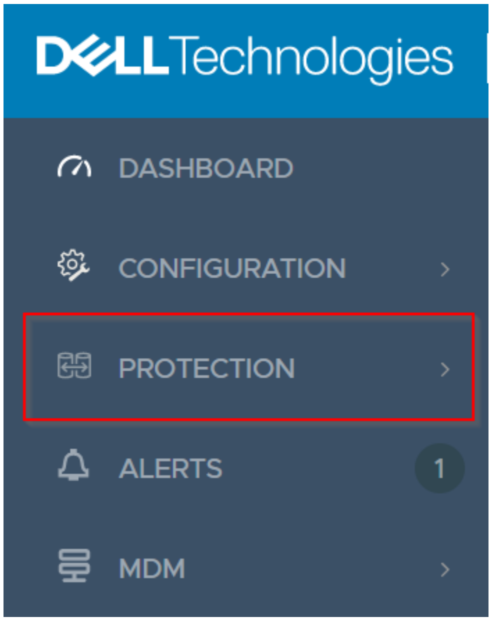


Figure 6

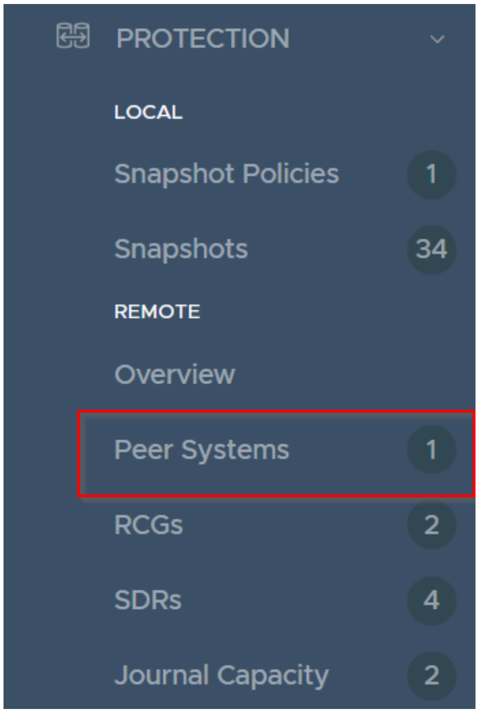
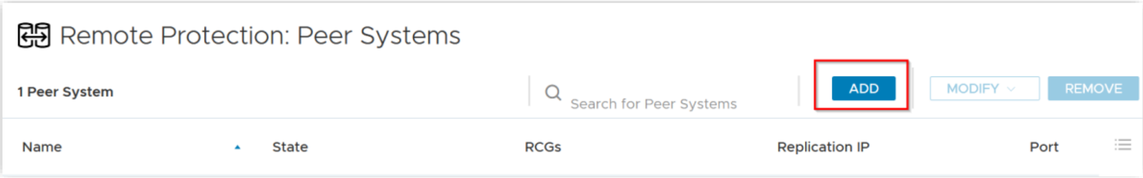


Figure 7

Under the Replication menu choices, select **Peer Systems**.

To add the system peer, click the **ADD** button.



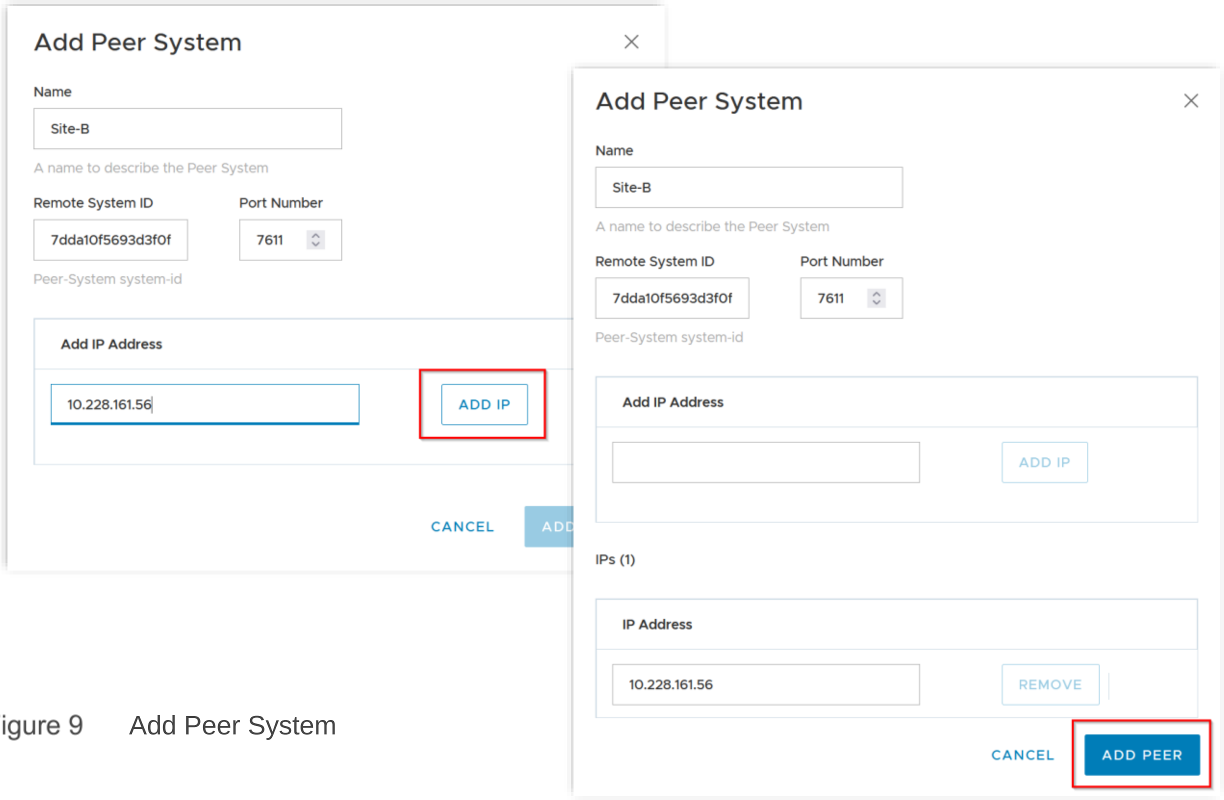
Remote Protection: Peer Systems

1 Peer System

Name	State	RCGs	Replication IP	Port

Figure 8 Add Peer

Complete the Name, Remote System ID, and the IP of the target cluster Primary MDM. Click **Add IP**. Add additional IPs if appropriate. Complete the wizard by clicking **Add Peer**.



Add Peer System

Name: Site-B

A name to describe the Peer System

Remote System ID: 7dda10f5693d3f0f

Port Number: 7611

Peer-System system-id

Add IP Address: 10.228.161.56

Add Peer System

Name: Site-B

A name to describe the Peer System

Remote System ID: 7dda10f5693d3f0f

Port Number: 7611

Peer-System system-id

Add IP Address

IPs (1)

IP Address: 10.228.161.56

Figure 9 Add Peer System

Once you add the peer, repeat the process on the target storage cluster using the same steps, entering the remote system ID of the primary cluster. Once that is complete, the systems are peered in both directions, and you are ready to start pairing your replicated volumes.

3.2 Replication Consistency Groups

Replication Consistency Groups, or RCGs, establish the attributes and behavior of the replication of one or more volume pairs. One such attribute is the target replication storage cluster. While a given RCG can replicate to only one target cluster, in principle other RCGs may replicate to other clusters provided they have exchanged certificates and have been peered. In the 3.5 and 3.6 releases of PowerFlex native asynchronous replication, however, a source site may only be peered with one other site. Future releases will permit additional replication topologies.

Before creating RCGs, our replication volume pairs must exist on both the source and target systems, and they must be the same size. In PowerFlex versions 3.5.x and 3.6, the target volume must be created manually. While the volumes must be identical in size, they are not required to reside in a storage pool of the same type (MG vs. FG), nor must they have the same properties (thick vs. thin, compressed vs. non-compressed). If a volume must be resized, the target volume should be expanded first. Expanding the volumes in this manner prevents any disruptions in replication. This means it's mandatory to know what volumes are being replicated, so that this practice can be followed if data outgrows the source volume.

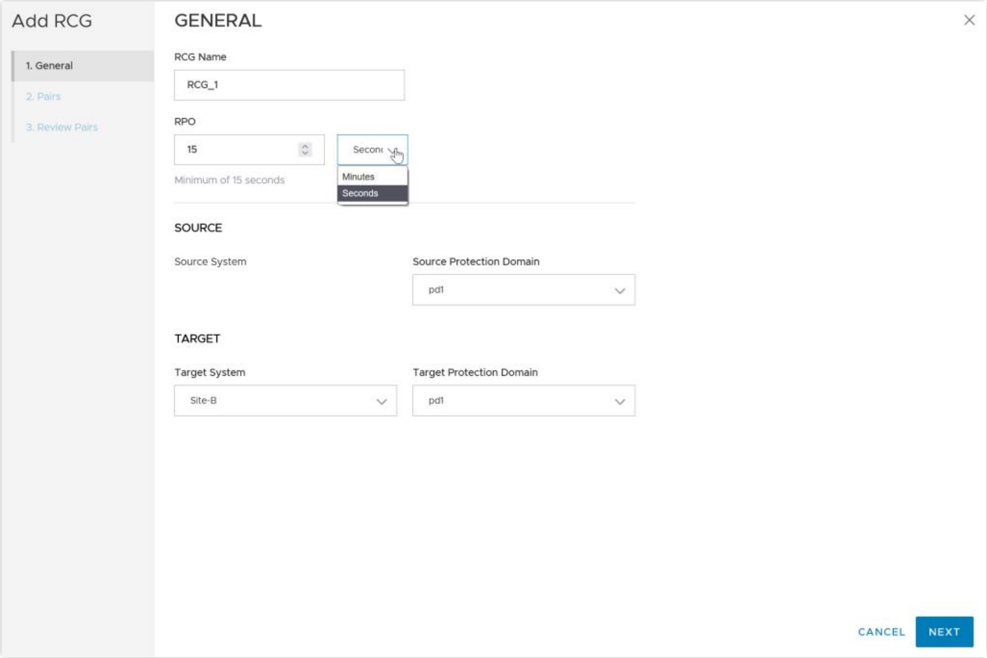
RCGs are very flexible. For some use cases, you might assign all volumes associated with an application to a single RCG. For larger applications, you might create multiple RCGs based on data retention, data type, or related application quiescing procedures to enable read-consistent snapshots when needed. In general, RCGs are crash-consistent. Snapshots can be made read-consistent if application quiescing rules were followed when they were created. This places no special requirements on the storage platform, but generally requires scripting with the application.

Recovery Point Objectives are specified in the RCG configuration. As seen in in Figure 10, below, PowerFlex version 3.6 RPOs can be set between 15 seconds and 60 minutes. (Note: in PowerFlex version 3.5.x, the smallest RPO available was 30 seconds.)

To create an RCG, log into the WebUI and navigate **to PROTECTION > REMOTE > RCGs** and click the **ADD** button.

This first step in creating an RCG involves providing:

- A name for the RCG
- The desired RPO
- The source Protection Domain
- The target system
- The target Protection Domain



GENERAL

RCG Name: RCG_1

RPO: 15 Seconds

Minimum of 15 seconds

SOURCE

Source System

Source Protection Domain: pd1

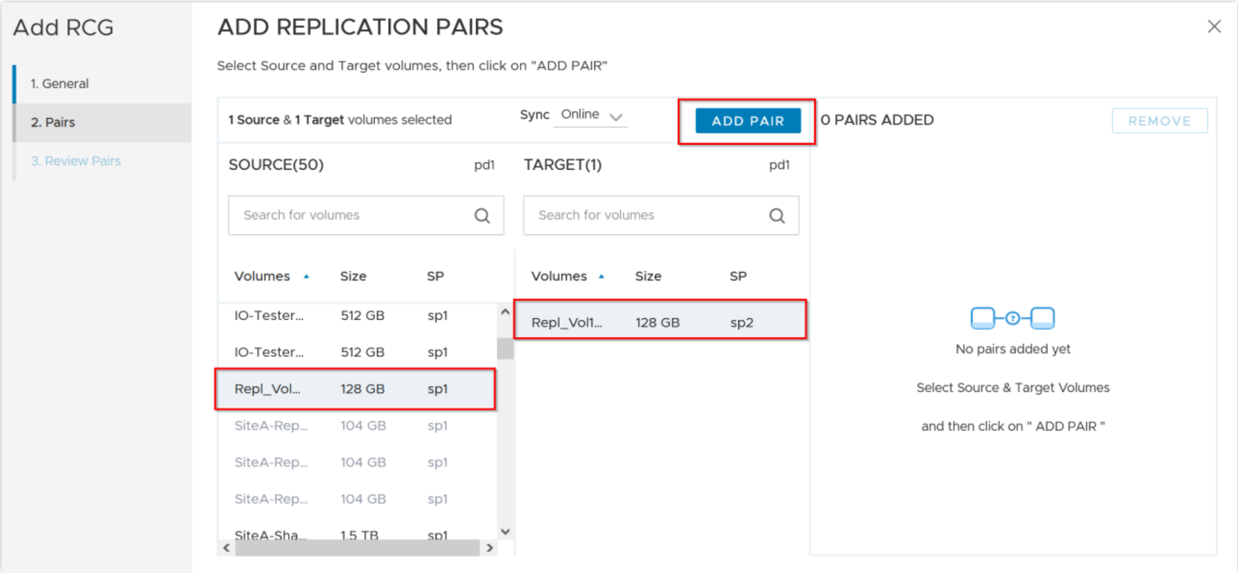
TARGET

Target System: Site-B

Target Protection Domain: pd1

Figure 10 Add RCG – set RPO

To complete the operation, we first match up the desired source and target volumes.



Add RCG

1. General
2. Pairs
3. Review Pairs

ADD REPLICATION PAIRS

Select Source and Target volumes, then click on "ADD PAIR"

1 Source & **1 Target** volumes selected   Sync Online   ADD PAIR   0 PAIRS ADDED   REMOVE

SOURCE(50) pd1

Search for volumes

Volumes	Size	SP
IO-Tester...	512 GB	sp1
IO-Tester...	512 GB	sp1
Repl_Vol...	128 GB	sp1
SiteA-Rep...	104 GB	sp1
SiteA-Rep...	104 GB	sp1
SiteA-Rep...	104 GB	sp1
SiteA-Sha...	1.5 TB	sp1

TARGET(1) pd1

Search for volumes

Volumes	Size	SP
Repl_Vol1...	128 GB	sp2

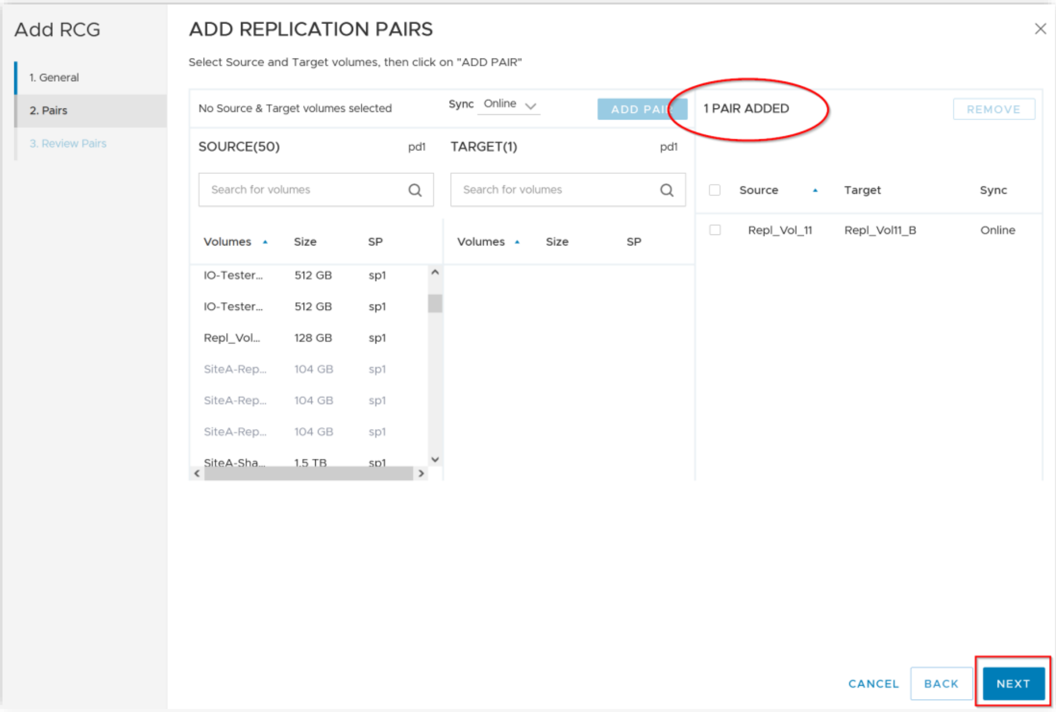
No pairs added yet

Select Source & Target Volumes

and then click on " ADD PAIR "

Figure 11 Add RCG – match source and target volumes

- Unpaired volumes appear in the source and target lists in dark text. When a source volume is selected, un-paired target volumes are shown, provided they have the same capacity as the selected volume. Once the two volumes have been selected, click the **ADD PAIR** button, moving the volume pair into the list appearing on the right. Once all volume pairs have been added, proceed by clicking the **NEXT** button.



Add RCG

1. General
2. Pairs
3. Review Pairs

ADD REPLICATION PAIRS

Select Source and Target volumes, then click on "ADD PAIR"

No Source & Target volumes selected   Sync Online   ADD PAIR   1 PAIR ADDED   REMOVE

SOURCE(50) pd1

Search for volumes

Volumes	Size	SP
IO-Tester...	512 GB	sp1
IO-Tester...	512 GB	sp1
Repl_Vol...	128 GB	sp1
SiteA-Rep...	104 GB	sp1
SiteA-Rep...	104 GB	sp1
SiteA-Rep...	104 GB	sp1
SiteA-Sha...	1.5 TB	sp1

TARGET(1) pd1

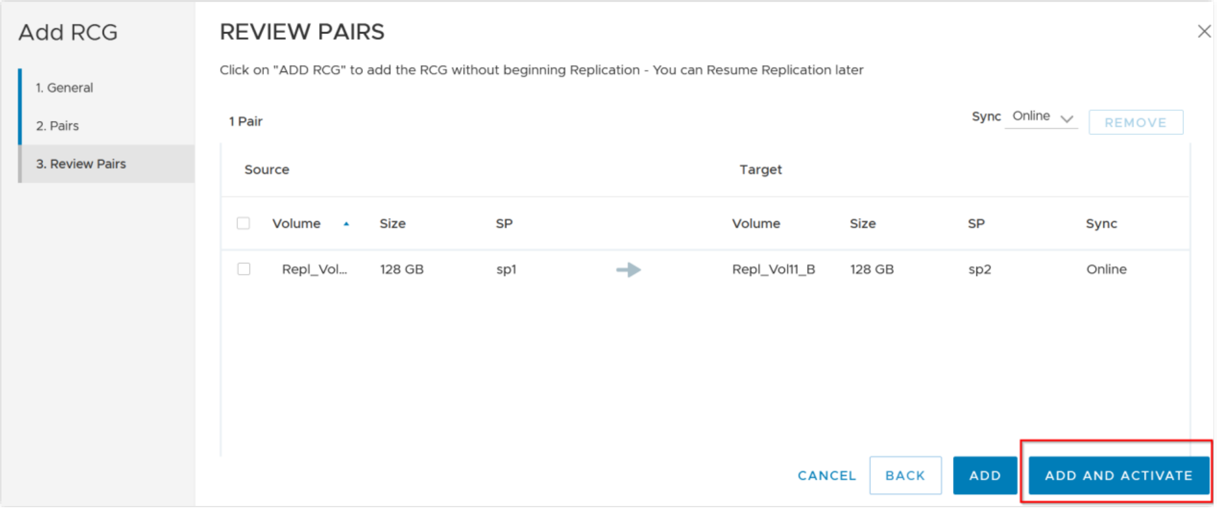
Search for volumes

Volumes	Size	SP

☐	Source	Target	Sync
☐	Repl_Vol_11	Repl_Vol11_B	Online

CANCEL   BACK   NEXT

- A summary is then displayed where you can select pairs and remove them if needed.



REVIEW PAIRS

Click on "ADD RCG" to add the RCG without beginning Replication - You can Resume Replication later

1 Pair

Source				Target			
□ Volume ▲	Size	SP		Volume	Size	SP	Sync
□ Repl_Vol...	128 GB	sp1	→	Repl_Vol11_B	128 GB	sp2	Online

Figure 12 Add RCG – review pairs

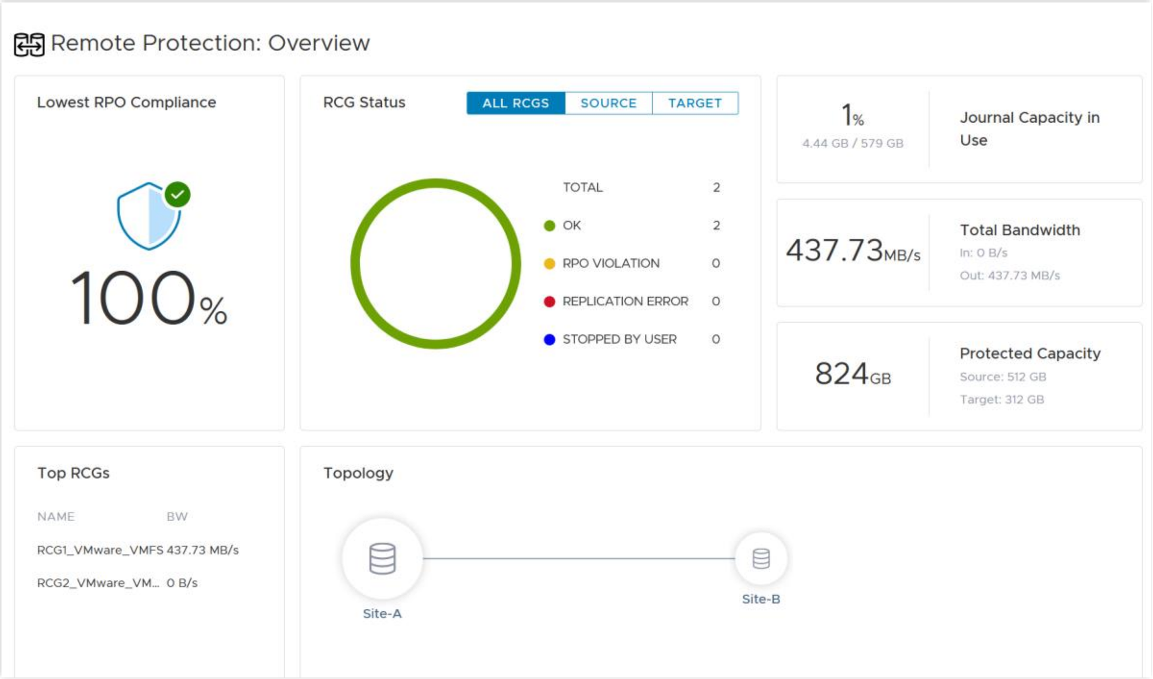
- The final mouse click gives us the option of creating the RCG and activating the volume synchronization immediately or simply adding the configuration without activation. We will discuss Active and Inactive RCG states in more detail below.

- You may add or remove volumes from an RCG at any time. Out of concern for excessive I/O during initial sync, and depending on the size of the volumes, you may elect to add only single volume pairs at a time to the RCG when it’s first created, but this is usually not necessary.

4 Replication Monitoring and Configuration

4.1 The Replication Dashboard

The **PROTECTION → Remote → Overview** area gives us a dashboard to determine the overall health and status of replication in the system.



Remote Protection: Overview

Lowest RPO Compliance

100%

RCG Status

ALL RCGS | SOURCE | TARGET

TOTAL	2
● OK	2
● RPO VIOLATION	0
● REPLICATION ERROR	0
● STOPPED BY USER	0

1%
4.44 GB / 579 GB
Journal Capacity in Use

437.73MB/s
Total Bandwidth
In: 0 B/s
Out: 437.73 MB/s

824GB
Protected Capacity
Source: 512 GB
Target: 312 GB

Top RCGs

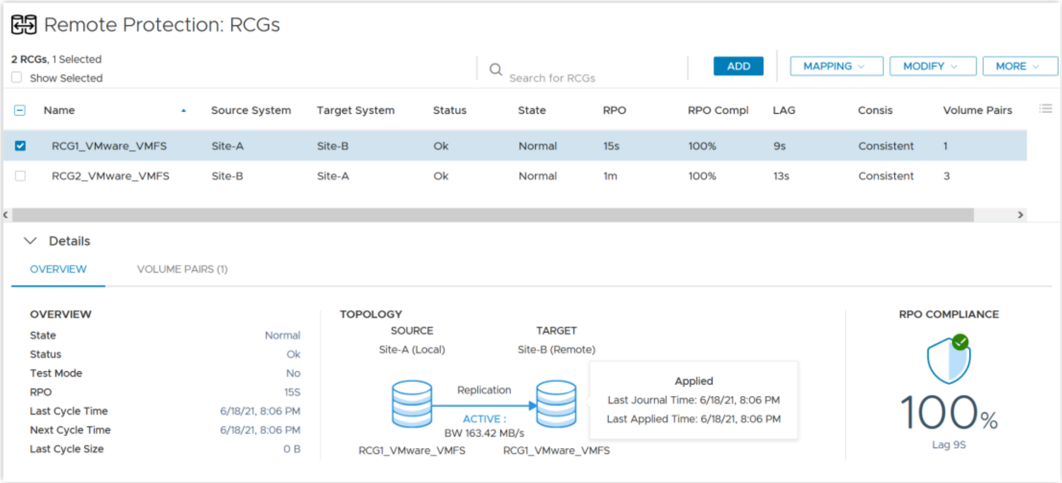
NAME	BW
RCG1_VMware_VMFS	437.73 MB/s
RCG2_VMware_VM...	0 B/s

Topology

Figure 13 Replication Overview

4.2 The Replication Consistency Group view

- The **PROTECTION → Remote → RCGs** view in the WebUI allows us to monitor the health and status of the individual Replication Consistency groups or add new ones.



Remote Protection: RCGs

2 RCGs, 1 Selected
□ Show Selected

	Name	Source System	Target System	Status	State	RPO	RPO Compl	LAG	Consis	Volume Pairs
☑	RCG1_VMware_VMFS	Site-A	Site-B	Ok	Normal	15s	100%	9s	Consistent	1
□	RCG2_VMware_VMFS	Site-B	Site-A	Ok	Normal	1m	100%	13s	Consistent	3

Details

OVERVIEW | VOLUME PAIRS (1)

OVERVIEW

State	Normal
Status	Ok
Test Mode	No
RPO	15S
Last Cycle Time	6/18/21, 8:06 PM
Next Cycle Time	6/18/21, 8:06 PM
Last Cycle Size	0 B

TOPOLOGY

Applied
Last Journal Time: 6/18/21, 8:06 PM
Last Applied Time: 6/18/21, 8:06 PM

RPO COMPLIANCE

100%
Lag 9S

Figure 14 RCG Overview

Selecting the row of any RCG opens the details pane, offering insight into the status of the RCG and it's component volumes.

Selecting an RCG by clicking on its checkbox enables the action menus. Under the MORE menu, we find the following options.

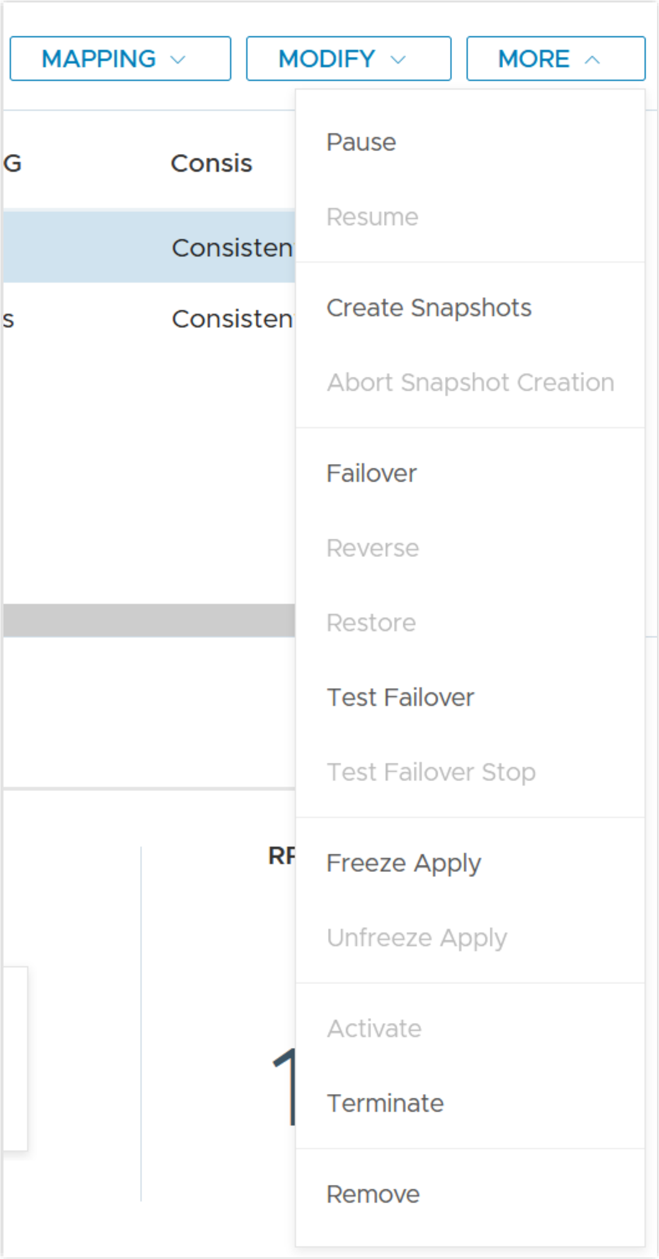


Figure 15 RCG Management Options

Pause: This action pauses replication between source and target. This prevents journals from being shipped to the target cluster until replication is resumed. Writes to the replicated volumes are still collected in the source journal volumes.

Create Snapshots: Generates snapshots of each volume in the RCG on the target system. This can be useful for remotely testing an application or DR activity. There is no RCG menu option to manage or delete the snapshots, so they must be mapped/unmapped and manually deleted on the target system's snapshot list.

Failover: Forces a failover event, passing primary ownership of the volumes within the RCG to the target system. This also switches the Host Access profile on the source-side volumes read-only and on the target to read/write. Once this is done, for planned failovers, you can also select the **Reverse** command to resume protection of the RCG volumes, only now in the opposite direction. If you wish to abort the failover operation, select the **Restore** option to return to the original replication state and direction.

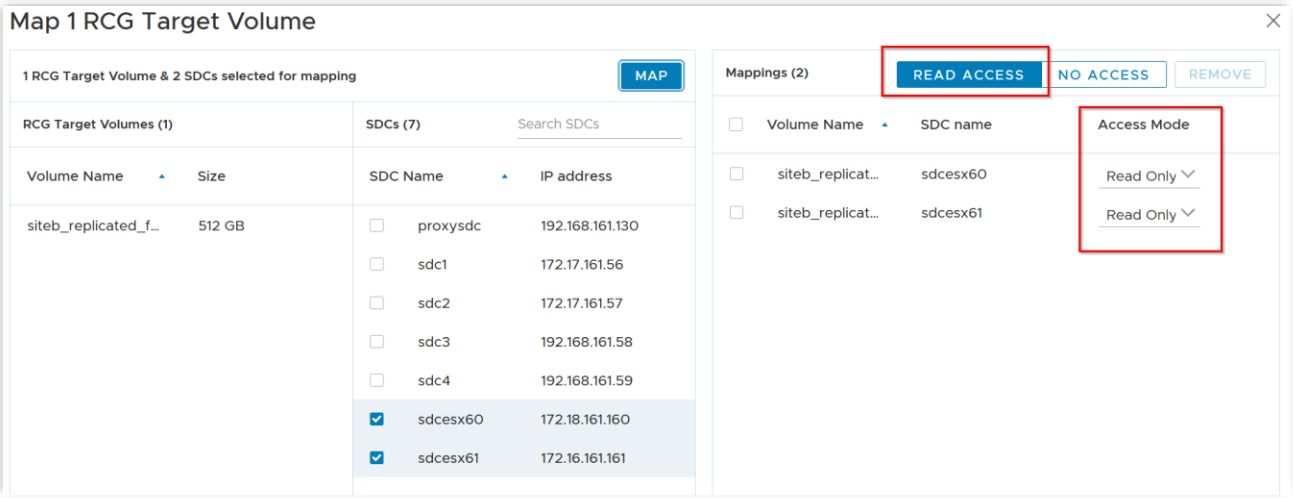
Test Failover: This automatically creates a snapshot on the target system and replaces the original target volume mapping with a mapping to the snapshot. Using this command, you can perform write testing to the volume while preventing the source volume from being corrupted by the test activity.

Freeze Apply: This option freezes the application of writes in the target journal to the target volumes. This does not pause replication between the sites, and the journal intervals will accumulate in the target system's apply journal volumes. When finished, select **Unfreeze Apply** to resume application to the target volumes.

Activate / Terminate: These options are new in PowerFlex version 3.6. If an RCG was created but not activated, or has been placed into an inactive state, then it can be activated here. Activation initiates all of the replication-related processes and begins the flow of I/O through the SDR on the source system. If a user **Terminate**s an RCG, this not only stops the flow of replication data between sites, it also releases the SDR from proxying the I/O and writing to the journal. A terminated or inactive RCG consumes no additional system resources and is merely a configuration placeholder.

4.3 Volume access

Target volumes cannot be set to Read and Write. The default access mode for target volumes with a Replication Consistency Group is "no access". But when mapping target volumes to an SDC, users can choose to map them as Read Only.



Map 1 RCG Target Volume

1 RCG Target Volume & 2 SDCs selected for mapping

RCG Target Volumes (1)

Volume Name	Size
siteb_replicated_f...	512 GB

SDCs (7)

	SDC Name	IP address
□	proxysdc	192.168.161.130
□	sdc1	172.17.161.56
□	sdc2	172.17.161.57
□	sdc3	192.168.161.58
□	sdc4	192.168.161.59
☑	sdcesx60	172.18.161.160
☑	sdcesx61	172.16.161.161

Mappings (2)

□	Volume Name	SDC name	Access Mode
□	siteb_replicat...	sdcesx60	Read Only
□	siteb_replicat...	sdcesx61	Read Only

Figure 16

This can be useful for examining the data on the target volume without taking a snapshot. However, unless the Freeze Apply option is also selected, the volume will continue to receive updated writes from the source.

When configuring replicated systems to also participate in VMware Site Recovery Manager protection groups, the target volumes must be mapped Read Only to the Recovery Site ESXi hosts. For additional information on using VMware SRM to protect VMs in Datastores backed by replicated PowerFlex volumes, see the whitepaper Disaster recovery for virtualized workloads on Dell EMC PowerFlex using VMware Site Recovery Manager.

4.3.1 Test Failover behavior

PowerFlex includes a very useful tool for testing disaster recovery without actually stopping the source-side application and failing over to a secondary site. The act of issuing the **Test Failover** command will:

- Create a snapshot on the target system for all volumes attached to the RCG
- Replace the pointer used by the volume mapping for each volume with a pointer to its snapshot
- Set the access mode of the snapshot/volume mapping of each volume in the target system RCG to **read_write**

These steps all happen in milliseconds, making the volumes immediately write-accessible to the Storage Data Client, if they were previously mapped Read Only. Otherwise, users can map the target volumes to any SDC for testing. Because you are actually mapping the snapshot, you can do whatever you wish with the volumes, whether it’s using them for opening a database, an application, or mounting a filesystem. Since they are snapshots, you can freely test your application, and if the storage pool is of the same type and composition as the source system, your application will perform equally well.

During the Test Failover, replication is paused between the source and target. However, the RCG is still active, so writes are still flowing through the SDRs and accumulating in the source-side journal volumes. While users could use the Test Failover feature to run analytics or perform other test operations on the target volume data, such actions are better done with the Create Snapshot feature, discussed below.

Test Failovers allow administrators to safely run DR scenarios without downtime and a maintenance window. When the **Test Failover Stop** command is given, the target-side pointers are returned to their original state, pointing once again to the replicated volume itself. The snapshot(s) are deleted and any wites made to them are discarded. Finally, replication of data between the source and target is un-paused and the journal intervals will resume shipping.

4.3.2 Failover behavior

When the RCG Failover command is issued, the access mode of the original source volumes switches to **read_only**. This means that in the case of planned fail-over, you are required to shut your applications down. The access mode of the target volumes switches to **read_write**. There is nothing else to do, and the behavior is the same if the failover is issued from the command line interface or REST API. If the failover is planned but the original storage cluster continues functioning, you have the option of initiating the RCG command to **reverse** replication. This keeps the volume pairs in sync, only now in the reverse direction. If you will be shutting down the primary system for a prolonged period, but wish to retain the RCG configuration, you should Terminate the RCG to put it into an inactive state. The RCG volumes will have to undergo an initial sync when later reactivated.

One thing to bear in mind is that each PowerFlex storage system creates unique volume and SCSI IDs, so they will be different for the source and target systems.

4.3.3 Create Snapshots behavior

This RCG command creates snapshots for all volumes attached to the RCG on the target side, but it does not manage the snapshots any further. Consuming the snapshots is done separately and manually. From there, to test your applications or make use of the data contained in them, you would need to:

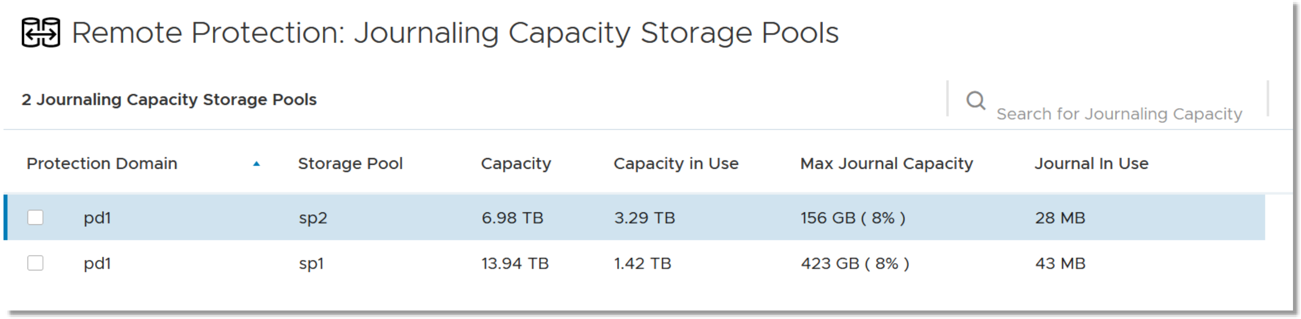
1. Map the volumes to a target SDC compute system(s)
2. Use the volumes as needed
3. Un-map the volumes when they are no longer needed
4. Delete the snapshots

We noted above that the Test Failover feature was not best suited to long running or intensive testing of data in the target side volumes. By making use of writable snapshots, however, users may:

- Perform resource-intensive operations on secondary storage without impacting production
- Test application upgrades on the target system without production impact
- Attach different, and higher-performing compute systems or media in the target environment
- Attach systems with different hardware attributes such as GPUs in the target domain
- Run analytics on the data without impeding your operational systems
- Perform “what-if” actions on the data because that data will not be written back to prod

4.3.4 Monitoring Journal Capacity and Health

By logging onto the WebUI and navigating to PROTECTION → Remote → Journal Capacity you can track the utilization of your Journal space reservation(s).

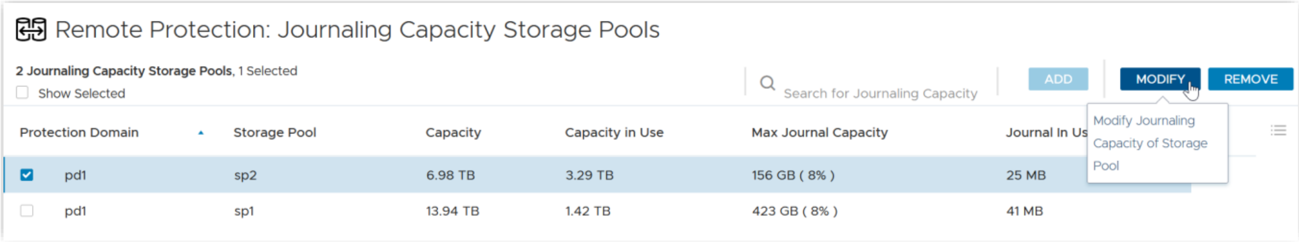


Remote Protection: Journaling Capacity Storage Pools

2 Journaling Capacity Storage Pools

	Protection Domain	Storage Pool	Capacity	Capacity in Use	Max Journal Capacity	Journal In Use
□	pd1	sp2	6.98 TB	3.29 TB	156 GB (8%)	28 MB
□	pd1	sp1	13.94 TB	1.42 TB	423 GB (8%)	43 MB

Here, we see that we’ve reserved 8% or 156GB of Storage Pool sp2 for journaling, and we currently have only 27MB of journal capacity in use. If there is concern that the space reservation is too small or large, you can change it at any time. Select the storage pool checkbox and click on the MODIFY command. Make any needed edits to the reservations.



Remote Protection: Journaling Capacity Storage Pools

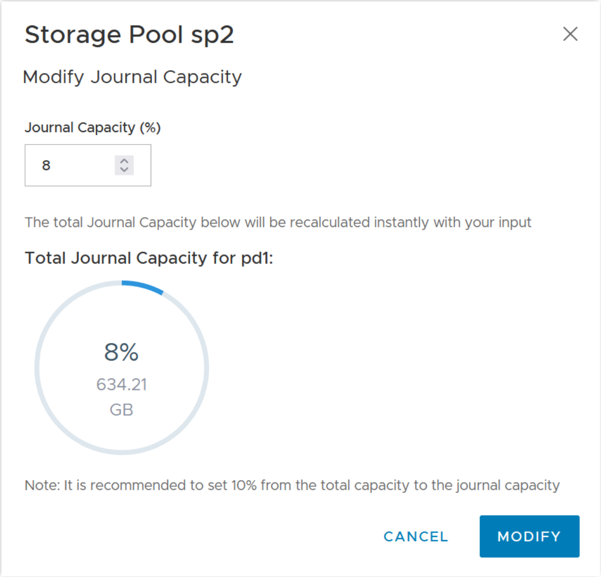
2 Journaling Capacity Storage Pools, 1 Selected

□ Show Selected

Search for Journaling Capacity
ADD
MODIFY
REMOVE
Modify Journaling Capacity of Storage Pool

	Protection Domain	Storage Pool	Capacity	Capacity in Use	Max Journal Capacity	Journal In Us
☑	pd1	sp2	6.98 TB	3.29 TB	156 GB (8%)	25 MB
□	pd1	sp1	13.94 TB	1.42 TB	423 GB (8%)	41 MB

We noted above that changes to the overall storage pool capacity may be one reason to increase or decrease the journal reservation percentage. Another reason might be an increase in volumes or applications that will use replication.



Storage Pool sp2

Modify Journal Capacity

Journal Capacity (%): 8

The total Journal Capacity below will be recalculated instantly with your input

Total Journal Capacity for pd1:

Note: It is recommended to set 10% from the total capacity to the journal capacity

5 PowerFlex replication networking considerations

All networking topologies, availability, and load balancing options previously recommended remain fully supported. However, PowerFlex replication adds a new consideration in the network fabric that must be accounted for. There is additional network overhead associated with replication and related journaling activity. For details, see the PowerFlex Networking Best Practices and Design Considerations white paper.

5.1 TCP/IP port considerations

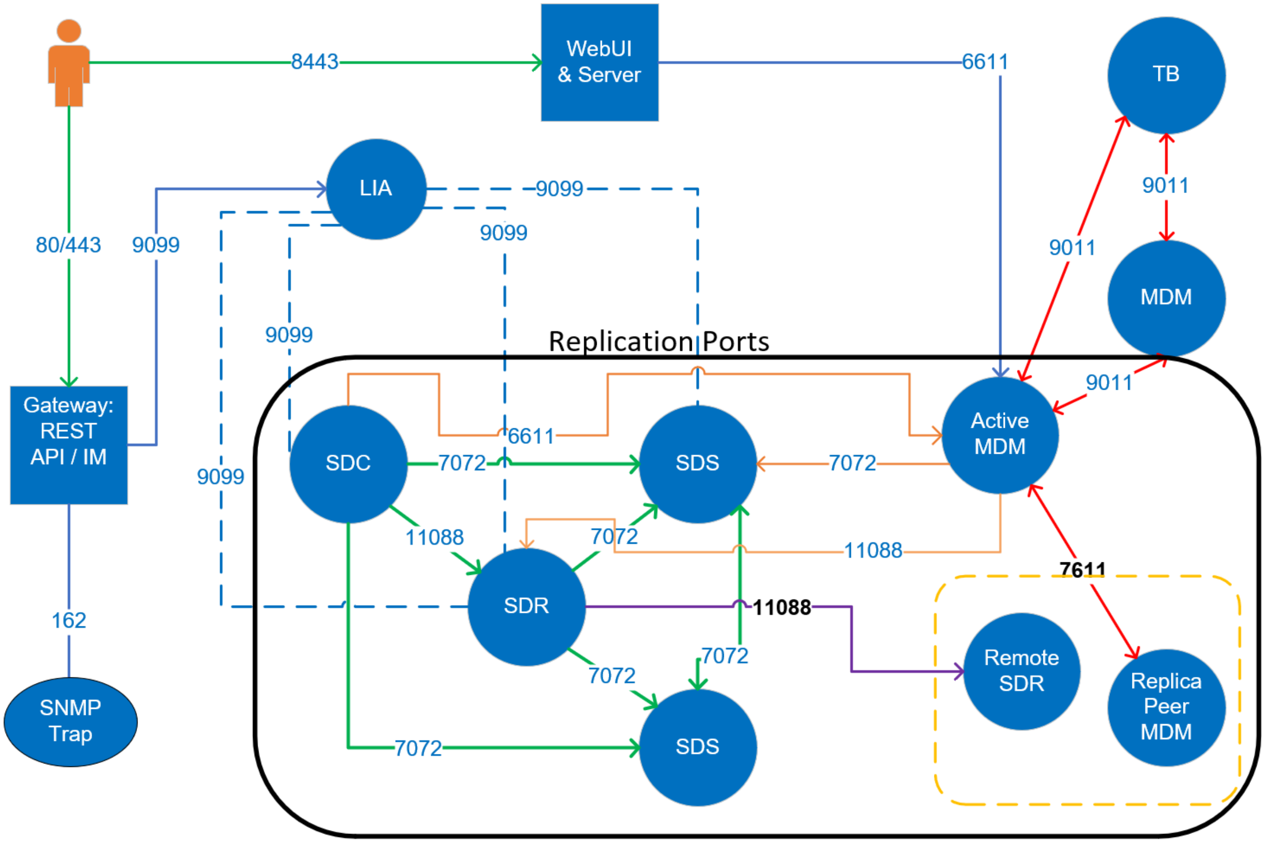


Figure 17 PowerFlex port and traffic overview

The diagram above provides us with a representation of all the logical software components of PowerFlex as well as the TCP/IP ports used by those components. We see the ports which must be associated with firewall rules on our PowerFlex server hosts. We can also see, the ports related specifically to remote replication which include:

1. Port 11088 which links the SDC and MDM to the SDR also links the SDR to the remote SDR.
2. Port 7611 which allows MDM communications between two replicating clusters.

5.2 Additional IP addresses

Within a protection domain, SDRs are installed on the same hosts as SDSs, but the traffic that an SDR writes to a journal volume is sent to all SDSs that host the journal, not only the one is it co-located with on a host. In the backend storage network, each SDR listens on the same node IPs as the SDSs and therefore should be able to reach all SDSs in the protection domain.

The SDRs, however, require additional, distinct IP addresses which will 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.

5.3 Network bandwidth considerations

First, there are general considerations for communications between replicating clusters. **The volume of writes to replicated volumes cannot exceed the networking bandwidth between the clusters**. Plan that at least one path in the network between the clusters will fail and make certain the expected write bandwidth can be sustained with latencies falling within the requirements of your applications and service levels.

5.3.1 Bandwidth within a replicating system

We noted above that when a volume is being replicated I/O is sent from the SDC to the SDR, after which there are subsequent I/Os from the SDR to SDSs on the source system. The SDR first passes on the volume I/O to the associated SDS for processing (e.g., compression) and committal to disk. The associated SDS will probably not be on the same node as the SDR, and bandwidth calculations must account for this. In the second step, the SDR applies incoming writes to the journaling volume. Because the journal volume is just like any other volume within a PowerFlex system, the SDR is sending I/O to the various SDSs backing the storage pool in which the journal volume resides. *This step adds two additional I/Os as the SDR first writes to the relevant primary SDS backing the journal volume and the primary SDS sends a copy to the secondary SDS.* Finally, the SDR makes an extra read from the journal volume before sending to the remote site.

Write operations for replicated volumes therefore require three times as much bandwidth within the source cluster as write operations for non-replicated volumes. **Carefully consider the write profile of workloads that will run on replicated volumes; additional network capacity will be needed to accommodate the additional write overhead**. In replicating systems, therefore, we recommend using 4x 25GbE or 2x 100GbE networks to accommodate the back-end storage traffic.

5.4 Remote replication networking

For networking configurations where access to the remote cluster passes through a static routed WAN, or where the networking baseline latency is greater than 50ms, the greatest concern is latency**. For any configuration, there is a latency limit of 200ms**, which is a potential issue for regionally remote clusters. For network paths exceeding 200ms, it is likely that a path approaching from the other side of the Earth will perform better. This configuration will require at least two subnets connected to the target system, and with the higher latency, bandwidth can become an issue, so thoroughly test the latency and throughput limits of your links and keep the replication bandwidth under your known thresholds.

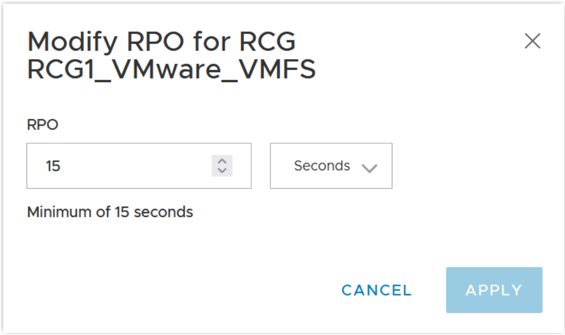
Journal data is shipped between source and target SDRs, first, at the replication pair initialization phase and, second, during the replication steady state phase. Special care should be taken to ensure adequate bandwidth between the source and target SDRs, whether over LAN or WAN. The potential for exceeding available bandwidth is greatest over WAN connections. While write-folding may reduce the amount of data to be shipped to the target journal, this cannot always be easily predicted. *If the available bandwidth is exceeded, the journal intervals will back up, increasing both the journal volume size and the RPO.*

As a best practice, we recommend that the sustained write bandwidth of all volumes being replicated should not exceed 80% of the total available WAN bandwidth. if the peer systems are mutually replicating volumes to one another, the peer SDR←→SDR bandwidth must account for the requirements of both directions simultaneously. Reference and use the latest PowerFlex Sizer for additional help calculating the required WAN bandwidth for specific workloads.

Note: The sizer tool is an internal tool available for Dell employees and partners. External users should consult with their technical sales specialist if WAN bandwidth sizing assistance is needed.

Leaving a 20% margin of safety for replication traffic over a WAN, allows for application I/O bursts and for the initial syncing of new volumes added to or reactivated in RCGs.

In certain cases, when latency is high, you will may need to increase the RPO of your Replication Consistency Groups. This can be done in the RCGs tab. Visit **PROTECTION → Remote → RCGs**, select an RCG, and click the **Modify→ Modify RPO** command to increase the RPO value.



Modify RPO for RCG
RCG1_VMware_VMFS

RPO

15

Seconds

Minimum of 15 seconds

Figure 18

5.4.1 Networking implications for replication health

It is possible to have write peaks that exceed the recommended “0.8 * WAN bandwidth”, but they should be short. The journal size must be large enough to absorb these write peaks.

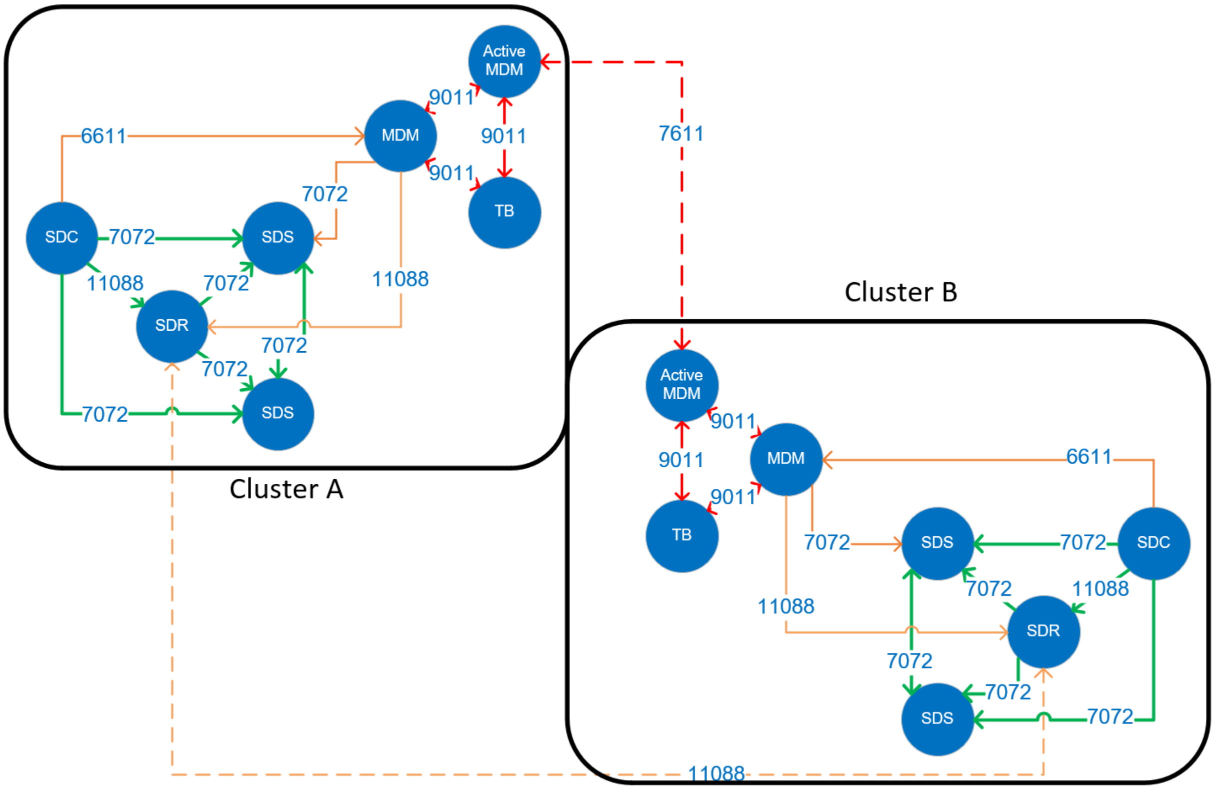
Similarly, the journal volume capacity should be sized to accommodate link outages between peer systems. A one-hour outage might be reasonably expected, but we encourage users to plan for 3 hours. The RPO will obviously increase while the link is down, and one must ensure sufficient journal space to account for the writes during the outage. It is best to use the PowerFlex sizer for such planning, but **in general the journal capacity should be calculated as WAN bandwidth * link down time**. For example, if the WAN link is 2x10Gb (about 2GB/sec) and the planned down time is one hour, the journal size should be 2 x 3600, or approximately 7TB.

When a WAN link is restored, the 20% bandwidth headroom will allow the system to catch-up to its original RPO target.

Note: The volume data shipped in the journal intervals is not compressed. In PowerFlex, compression is for data at rest. In fine-granularity storage pools, data compression takes place in the SDS service after it has been received from an SDC (for non-replicated volumes) or an SDR (for replicated volumes). The SDR is unaware of and agnostic to the data layout on either side of a replica pair. If the destination, or target, volume is configured as compressed, the compression takes place in the target system SDSs as the journal intervals are being applied.

5.4.2 Routing and Firewall considerations for remote replication

Section 5.1 emphasized TCP/IP ports for MDM (7611) communications between replicating clusters as well as SDR (11088) communications used in transporting replication journal logs.



For replication use cases involving distant clusters, we'll need interconnectivity for these IP ports provided over routed networks. The best practice for networking in this situation is to reserve two networks for intra-cluster SDR and MDM communications.

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 should be configured to indicate either the next hop address or the egress interface or both for reaching the remote subnet.

For example: `X.X.X.X/X via X.X.X.X dev interface`

Consider a small system with a few nodes on each side. Each node has four network adapters, two of which are configured with IPs for communication internal to the PowerFlex cluster and two of which are configured with IP addresses for site-to-site, external communication.

In this example, we tell the nodes to access the WAN subnets for the other side through a specified gateway. From source Site A, the network interfaces `enp130s0f0` and `enp130s0f1` are configured with addresses in the `30.30.214.0/24` and the `32.32.214.0/24` ranges, respectively. We can configure a route-interface file for each to direct packets for the remote networks over the specified gateway and interface.

```
route-enp130s0f0 contents →        31.31.0.0/16 via 30.30.214.252 dev enp130s0f0

route-enp130s0f1 contents →        33.33.0.0/16 via 32.32.214.252 dev enp130s0f1
```

Packets intended for the remote network 31.31.214.0/24 are directed through the next hop address at gateway IP 30.30.214.252. And similarly for packets destined for 33.33.214.0/24.

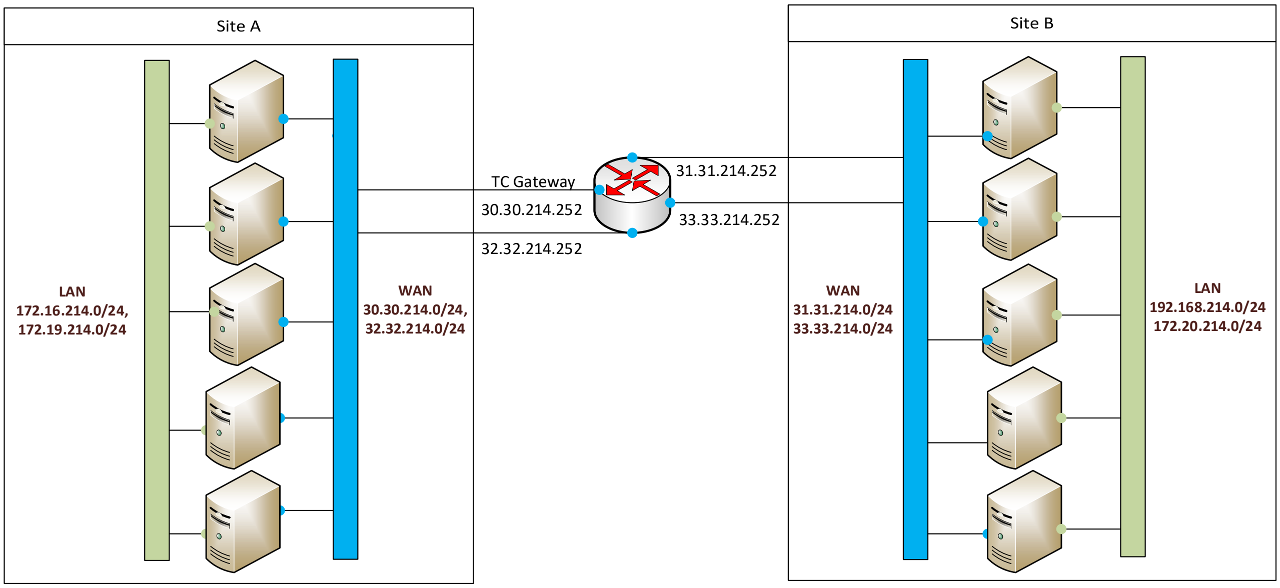


Figure 19 Example WAN topology for PowerFlex replication.

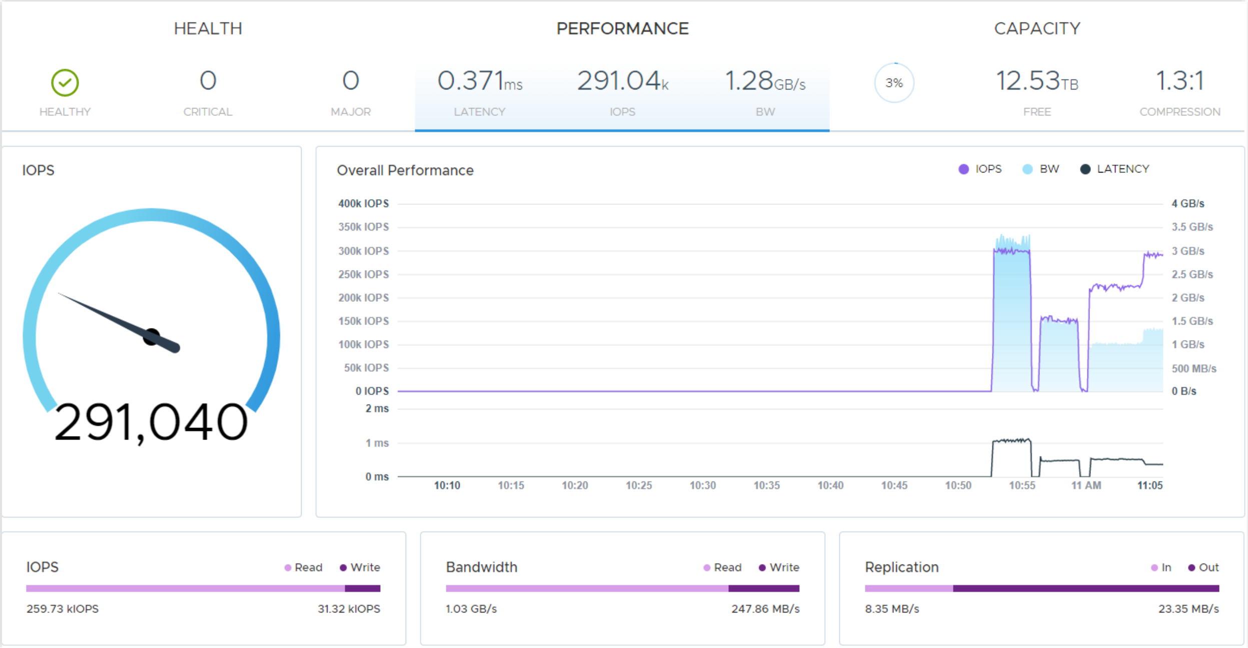
The details of static route configuration will vary with your operating system / hypervisor and overall network architecture, but the general principle is the same.

6 System component, network, and process failure

For our final considerations related to replication, we must face the reality that servers, processes, and network links do periodically fail. The following tests account for a few of these types of failures. The tests were performed on a 6-node R740xd PowerFlex node cluster with three SSDs per storage pool. Replication was active on both storage pools at the time of the failures.

6.1 SDR failure scenarios

Let's start with a baseline workload. We'll go on to fail an SDR, observe the impact, and observe the later impact of restarting it.



HEALTH: ✓ HEALTHY | 0 CRITICAL | 0 MAJOR

PERFORMANCE: 0.371ms LATENCY | 291.04k IOPS | 1.28GB/s BW

CAPACITY: 3% | 12.53TB FREE | 1.3:1 COMPRESSION

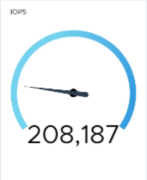
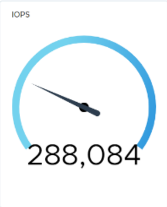
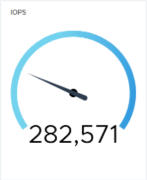
IOPS: 291,040

Overall Performance — ● IOPS ● BW ● LATENCY

IOPS ● Read ● Write — 259.73 kIOPS | 31.32 kIOPS

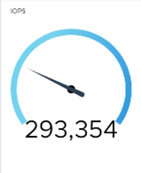
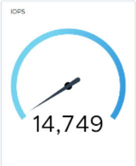
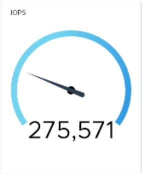
Bandwidth ● Read ● Write — 1.03 GB/s | 247.86 MB/s

Replication ● In ● Out — 8.35 MB/s | 23.35 MB/s

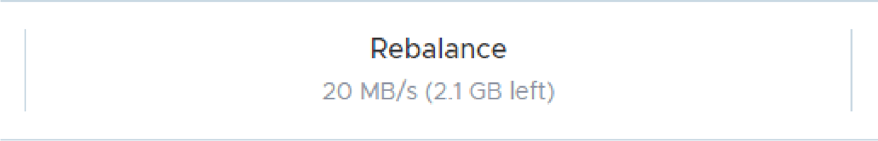
 IOPS 208,187	 IOPS 288,084	 IOPS 282,571
Immediately after failing the SDR, we see a drop in I/O processing	The I/O resumes slightly lower	After restarting, the I/O is slightly impacted, but eventually ramps back up to the baseline.

6.2 SDS failure scenarios

We'll perform the same test for SDS failure.

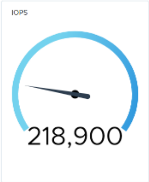
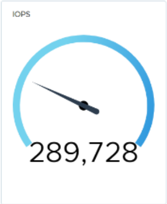
		
Baseline workload	IOPs just after fail the SDS	Five seconds later, it starts ramping up and within 10 seconds resumes the baseline workload indicating the system was more than capable of handling the workload with five active SDS systems

And as expected, we see rebalance activity



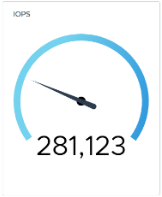
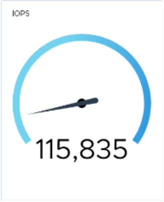
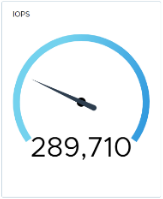
Rebalance

20 MB/s (2.1 GB left)

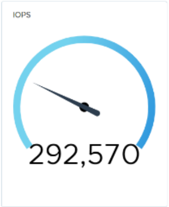
	
Next, we restart the failed SDS, and we see an immediate drop, but not substantial after restarting the SDS	As the rebalance continues, the I/O ramps back up to the baseline

6.3 Network link failure scenarios

Now, we'll fail a network link to demonstrate how the updated native load balancing affects our I/O rate. The system has a network configuration consisting of four data links between systems.

		
Again, we establish a baseline	We fail a link and notice a 3-second drop in I/O	5 seconds later, the baseline returns

After reconnecting the failed port, the baseline I/O level resumes within a few seconds with no noticeable dip.



All of these failure scenarios demonstrate the resilience of PowerFlex. It also shows that the system is well tuned, and that rebuild activity does not severely impact our workload.

7 Replication - Technical Limits

The following table lists the replication-related system limits for PowerFlex 3.6.

Replication Limits	
Number of destination systems for replication	1
Maximum number of SDR per system	128
Maximum number of Replication Consistency Groups (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	8
Maximum number of copies per RCG	1
Recovery Point Objective (RPO)	Min: 15 seconds / Max: 1 hour
Maximum replicated volume size	64 TB

8 Conclusions

You should now have a better understanding of PowerFlex native asynchronous replication including configuration and the journaling method selected.

In summary, it is recommended you start small. Follow the recommendations mentioned. Account for the total replication bandwidth, including all write I/O of all your replicated data. Size your journaling space reservations as recommended. Include margins of error for network and component failure.