Abstract
Dell EMC PowerFlex™ software-defined storage version 3.5 adds a new Protected Maintenance Mode, enhancing the protection of user data during system maintenance and upgrades. This paper provides an overview of the feature, comparing and contrasting it with Instant Maintenance Mode as well as other methods for maintaining SDS systems on production storage.

July 2020
Revisions

<table>
<thead>
<tr>
<th>Date</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>July 2020</td>
<td>Initial release</td>
</tr>
</tbody>
</table>

Acknowledgements

Co-author: Neil Gerren, Senior Principal Engineer, Storage Technical Marketing

Co-author: Brian Dean, Senior Principal Engineer, Storage Technical Marketing

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.

Copyright © 2020 Dell Inc. or its subsidiaries. All Rights Reserved. Dell Technologies, Dell, EMC, Dell EMC and other trademarks are trademarks of Dell Inc. or its subsidiaries. Other trademarks may be trademarks of their respective owners. [7/7/2020] [Technical White Paper] [H18392]
# Table of contents

Revisions .................................................................................................................. 2
Acknowledgements ................................................................................................... 2
Table of contents ....................................................................................................... 3
Executive summary ................................................................................................... 4
  1 Introduction ........................................................................................................ 5
  2 PowerFlex 3.5 New Features ............................................................................. 6
  2.1 Native Asynchronous Replication .................................................................... 6
  2.2 Protected Maintenance Mode .......................................................................... 6
  2.3 SDC Authentication ....................................................................................... 6
  2.4 New WebUI .................................................................................................... 6
  2.5 Secure snapshots ............................................................................................ 7
  2.6 Core Improvements ....................................................................................... 8
  3 Maintenance in PowerFlex systems ................................................................... 9
  3.1 SDS node maintenance options .................................................................... 10
  3.1.1 Remove a node and re-add after maintenance ........................................... 11
  3.1.2 Add additional node before removing node for maintenance .................. 13
  3.1.3 Instant Maintenance Mode (IMM) ............................................................... 13
  3.1.4 Protected Maintenance Mode (PMM) ......................................................... 15
  3.1.5 Maintenance modes and the Storage Data Replicator ............................... 17
  4 Considerations and limitations of PMM ............................................................ 18
  4.1 Mixing maintenance methodologies ............................................................... 18
  4.2 Concurrent PMM operations in a single Protection Domain ......................... 18
  4.3 Initiating Protected Maintenance Mode with degraded capacity .................. 18
  4.4 Considerations for non-disruptive upgrades ................................................ 18
  4.5 Spare capacity considerations ...................................................................... 18
  5 Summary ............................................................................................................ 20
Executive summary

As PowerFlex™ (formerly VxFlex OS) software-defined storage continues to evolve, the 3.5 release adds a variety of core features including Protected Maintenance Mode. Out of concern for having a single available copy of data at any time, this feature to better protect their data when performing maintenance on PowerFlex Storage Data Server nodes. We will start with a review of the previous methods of performing SDS system maintenance and then describe the enhancements made which enable this new maintenance capability.
1 Introduction

PowerFlex is a software-defined storage platform designed to significantly 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 provides a foundation that combines compute as well as high performance storage resources in a managed unified fabric. Flexibility is offered as it comes in multiple hardware deployment options such as integrated rack, appliance or ready nodes, all of which provide Server SAN, HCI and storage only architectures.

![PowerFlex Overview](image)

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 3.5 New Features**  
There is much more to this release than Protected Maintenance Mode, so it is worth mentioning some additional features included in the release.

![Features Diagram]

*Figure 2  New Features*

### 2.1 Native Asynchronous Replication

The flagship feature in version 3.5 is native asynchronous replication. Implemented as a journaling-based replication architecture, PowerFlex replication continues the tradition of vastly scalable, performant, and elastic systems. For additional information see the white papers on *Dell EMC PowerFlex: Introduction to Replication* as well as *Dell EMC PowerFlex: Networking Best Practices and Design Considerations*.

### 2.2 Protected Maintenance Mode

This is the key subject of this paper and will be covered in detail.

### 2.3 SDC Authentication

Authentication of SDCs is better secured with CHAP, or Challenge-Handshake Authentication Protocol. It allows the MDM to validate the authenticity of each SDC when it’s first attached and to establish secrets between the SDCs and SDSs to regulate access to volumes. The MDM regularly refreshes the secrets, forcing the SDCs and SDSs to re-authenticate on a regular basis.

### 2.4 New WebUI

The PowerFlex 3.5 release offers a new, streamlined HTML5-based user interface which is consistent with other Dell Technologies product solutions.
This primary dashboard view displays the majority of system activity at a single glance while also preserving the ability to drill into all PowerFlex elements to view or manage them.

2.5 Secure Snapshots

Secure snapshots were added to meet customer business and statutory requirements for data retention.

Figure 4  Secure snapshot with 1 year expiration time.

Once a snapshot is created with the secure option, it cannot be deleted until the assigned expiration time is reached. For cases where secure snapshots are created by mistake, or must be removed for other reasons, there is a formal process integrated with Dell support that must be followed to delete them. Note also that in 3.5, snapshots now can be created with read-only access, whether they are secure or not.
2.6 Core Improvements

There are several core improvements in 3.5, but a few merit special mentions. Release 3.5 adds a Fine Granularity Metadata cache which eliminates the two-step metadata lookup required for FG volume read I/Os. Up to 32GB of FG pool metadata can be cached per SDS. The cache is not persistent, it resides in DRAM, and it is updated either on new reads after an SDS reboot or upon a cache-miss. This dramatically improves FG read performance for recent and frequently read I/Os.

Data resiliency has been improved with two new features. Persistent checksum is now available for data residing on Medium Granularity storage pools, and this is enabled by default on volumes created after the upgrade to 3.5. Additionally, new Partial Disk Error handling prevents immediate ejections and rebuild of entire drives when only a few sectors fail. This provides a longer useful life of your media.

For more information on the PowerFlex 3.5 release, refer to the *Getting to Know* document in the product documentation bundle found on the Dell EMC support site.
3 Maintenance in PowerFlex systems

To understand how Storage Data Server system maintenance is performed, we must first consider the basic architecture of PowerFlex itself.

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.

These three logical elements comprise the best Software-Defined storage solution today, scaling to hundreds of SDS systems linearly.
3.1 SDS node maintenance options

Sometimes nodes need to be taken offline for planned maintenance. If a PowerFlex node goes offline in an unplanned event, the system will alert the error and initiate a rebuild of the data, reprotecting the user data on the remaining nodes. When undergoing planned maintenance or performing a non-disruptive system upgrade, however, we want to avoid a rebuild and control the process without being in an error state.

![Node requiring maintenance](image)

There are four options for doing maintenance on SDS nodes that are participating in and contributing storage to a PowerFlex storage cluster. They include:

- Remove an SDS node from the cluster and re-add it later after finishing maintenance
- Add a new SDS node and remove the node requiring maintenance
- Instant Maintenance Mode
- Protected Maintenance Mode

We will discuss these options in turn, examining the pros and cons of each.
3.1.1 Remove a node and re-add after maintenance

This method dates back to the earliest PowerFlex versions, and was the default method in PowerFlex Manager prior to version 3.5. In this scenario, when a node is gracefully removed using the UI or CLI, a many-to-many rebalance operation begins among the nodes, ensuring that there are two copies of all data on all of the other nodes before dropping the node-to-be-maintained from the cluster.

![Configuration: SDS](image)

**Figure 7** Graceful removal using the Web UI

![Many-to-many rebuild when the middle node is removed](image)

**Figure 8** Many-to-many rebuild when the middle node is removed

**Note:** Because the rebalancing of the data uses up free spare capacity on the other nodes, users may need to adjust the spare capacity assigned to the cluster overall. For example, if you began with 10 nodes and 10% spare capacity, running with 9 nodes requires 12% spare in order to avoid triggering an insufficient spare capacity alert. Spare capacity in the system must always be equal to or greater than the capacity of the smallest fault unit (node).
During maintenance the cluster functions normally but with one less node, and therefore less capacity and lower performance. Writes are sent to, and mirrored on, the other nodes in the system.

Figure 9  During maintenance, after node removal

The data is fully protected because we always have two available copies of the data. It doesn’t matter how long the maintained node is offline, because it is no longer a part of the cluster. Should the maintenance reveal a problem that prohibits the node from being re-added, there is no exposure or risk of data unavailability.

When maintenance is complete, we can re-add the node to the cluster and a many-to-one rebalance will take place, evenly redistributing the data. Because we’re rehydrating a single node, this many-to-one rebalance is slower than the other steps. This is the primary disadvantage of this maintenance option.
This is the same mechanism that is employed when elastically scaling a system by adding nodes, one-by-one. By default, the I/O priority for rebalance activity is lower than the I/O priority for rebuild/reprotection activity. This is user-configurable, but in general, this should be avoided since we don’t want rebalancing activity to interfere with application I/O.

After adding the node back into the cluster, the user may wish to readjust the spare capacity percentage once again.

### 3.1.2 Add additional node before removing node for maintenance

If you expect the node requiring maintenance to be offline for an extended duration, and you have the luxury of owning a spare system, you can swap the two. If the spare node’s configuration is identical, or at least very similar to the node requiring maintenance, it is possible to add the spare to the cluster before removing the node requiring maintenance. This will take longer, since the system tries to both hydrate the new node and rebalance the data from the node being removed. But PowerFlex can do these steps simultaneously.

This is only a small variation on the preceding option. But it does allow users to maintain the pre-maintenance capacity and performance profile during the maintenance period. In this case, the overall system spare capacity does not need to be modified as long as the spare node’s capacity is roughly equivalent to the maintained node’s capacity.

### 3.1.3 Instant Maintenance Mode (IMM)

Instant Maintenance Mode, or IMM, is designed for quick entry into and exit from a state of maintenance. It is well-suited for cases like non-disruptive, rolling upgrades, where the maintenance window is only a few minutes (for example, a reboot) and there are no known hardware issues. If the maintenance window is expected to be longer than 30 minutes, then an alternative maintenance option should be selected.
In Instant Maintenance Mode, the data on the node undergoing maintenance is not evacuated from the cluster. For the duration of the maintenance activity however, the data on the maintained node is not available for use, so we make use of the extra copies of data residing on the other nodes for serving application reads.

When entering IMM, the existing data on the node is, in effect, frozen on the node. This is a planned operation, and a rebuild is not triggered. Rather, the MDM instructs the SDCs where to read and write IOs that would otherwise have been directed at the node in maintenance.

Specifically, new writes and their mirror copies are directed to other nodes, and any changes that would have affected the node in IMM are tracked. By ensuring that all new writes or updates are written in duplicate to other nodes in the cluster, we ensure that there is no data unavailability (DU) should the node in maintenance fail.

![Diagram of new writes or changes during IMM](image)

**Figure 12** New writes or changes during IMM

When exiting Instant Maintenance Mode, we do not need to rehydrate the node completely. Rather, we need only sync back the relevant changes that have occurred and reuse all the unchanged data still residing on the node. This makes for a very fast exit from maintenance and a quick return to full capacity and performance.

![Diagram of exiting IMM](image)
Maintenance in PowerFlex systems

Figure 13  Only deltas synced back to maintained node

If the node being maintained does fail, then a rebuild is initiated to re-protect the user data on that node. Because of this, there must be enough spare capacity in the system to recreate all of the data mirror copies from the node-in-maintenance elsewhere in the cluster.

In normal IMM operation, we always have two copies of our data, but any copies residing on the node in maintenance are unavailable for the duration of IMM. This is the primary disadvantage to this approach since IMM introduces risk that only a single copy of data is available during the IMM activity.

If a drive or a node elsewhere in the cluster failed while a node is in IMM, then we may have data unavailability (DU). One copy is on the failed component, and the other is temporarily unavailable on the node in IMM. Once we take the node out of IMM and return it to normal use, the MDM will use the data on the node leaving maintenance and create the mirror copies again. If, in more severe case however, another drive or node fails and the node in IMM also fails, we are exposed to the possibility of data loss (DL). In that case both mirrored copies are lost.

For these reasons, IMM should only ONLY be used for short maintenance windows where there are no known or suspected issues with the node undergoing maintenance. In cases where the health of the node is in question, or the expected duration of maintenance exceeds 30 minutes, another option should be used.

3.1.4   Protected Maintenance Mode (PMM)

![Figure 14  Entering PMM from the Web UI](image)

Protected Maintenance Mode, or PMM, is designed to provide the data availability advantages of the first two options, some of the agility of IMM, without the risk of exposure of a accessible single-copy.

Entering into Protected Maintenance Mode initiates the many-to-many rebalancing process as when removing a node from the system. However, there is a significant difference. In the earlier case, we sought to create copies elsewhere in the system in order to remove the node and the data on it. In Protected Maintenance Mode, we preserve the data on the node entering maintenance.
Many-to-many operation to create temporary third copy

Similar to IMM, we freeze the data on the node in maintenance, and this copy of the data is inaccessible during maintenance. The difference from IMM is that we create a temporary third copy on the sustaining nodes. This allows us to maintain two copies of data at all times, avoiding the risks described above resulting from a single available copy.

During maintenance (and again like Instant Maintenance Mode), any new writes or updates that would have affected the node in maintenance are tracked. These writes and their mirrored copies are made on the other cluster nodes.

Writes during PMM
When maintenance is complete and we exit PMM, we need only sync the changes back to the maintained node. Because a third, temporary copy of the data was made on the other nodes in the system, exiting PMM removes these extra copies after the deltas have been synced back to the original.

In effect, PMM combines the entrance phase of the remove/re-add option and the exit phase of IMM while always maintaining two available copies of the data. This means it is slower to initiate PMM than IMM, but just as fast to exit. For maintenance windows longer than 30 minutes, or if there is a suspected problem with the hardware, PMM is the safe choice.

3.1.5 Maintenance modes and the Storage Data Replicator

It should be noted that PMM is an option for maintaining SDSs. In PowerFlex version 3.5, a new software component was introduced to manage and coordinate asynchronous replication of volumes and volume consistency groups to remote target clusters. It is referred to as the Storage Data Replicator, or SDR.

The SDR does not make use of PMM, even though it resides on the same node as the SDS service, and in 3.5 there is a 1:1 ratio of SDSs and SDRs in a PowerFlex cluster. Putting an SDR into maintenance allows the cluster to migrate the SDR operations to a peer SDR within the source system.

Users should be aware that when putting an entire node into maintenance, the order of operations is:

1. the SDS should be put into maintenance first (either PMM or IMM)
2. the SDR is put into maintenance next, transferring replication responsibilities to other SDRs
3. When exiting maintenance, both the SDS and the SDR can resume their functions in parallel.

Depending on the original number of SDRs and the WAN bandwidth to the remote target site, ensure there is enough remaining bandwidth and overhead capacity to accommodate expected replication operations while in maintenance. For more information, see PowerFlex Networking Best Practices and Design Considerations.
4 Considerations and limitations of PMM

In this section, we consider a few things to keep in mind when using Protected Maintenance Mode. We have noted the obvious differences in time to enter and exit maintenance compared with other methods, as well as the one vs. two available copies of data during maintenance. In the following sections, we look at some less-obvious considerations and limitations.

4.1 Mixing maintenance methodologies

One cannot mix Protected Maintenance mode and Instant Maintenance Mode within the same Protection Domain concurrently.

There are, however, no cross-domain concerns when maintaining nodes, so IMM can be used in one Protection Domain while PMM is used in another.

If you are adding or removing a node to a cluster, you can simultaneously initiate PMM on another node in the same Protection Domain.

4.2 Concurrent PMM operations in a single Protection Domain

Within a given Protection Domain, all SDSs concurrently in, or concurrently entering, PMM must belong to the same Fault Set. If you do not use fault sets, this means that you can only have one node (which constitutes a fault unit) in maintenance mode at a time.

Also, provided there is sufficient spare capacity, it is possible to put several nodes in a fault set into PMM at the same time. Nevertheless, for simplicity sake, our suggested guideline is to do nodes one at a time.

4.3 Initiating Protected Maintenance Mode with degraded capacity

The system should be generally healthy with no failed capacity or other critical issues. However, the SDS entering PMM can have degraded capacity (similar to IMM), and other SDSs in the same fault set may also have degraded capacity when a node is entering PMM.

4.4 Considerations for non-disruptive upgrades

Non-disruptive upgrade (NDU) processing is configurable and can use either PMM or IMM.

For back-end software component upgrades (MDM, SDS, SDR, LIA, etc.), it is generally preferable to use IMM. These are very quick, and the entire system's rolling upgrade can be finished in mere minutes. If there are no known or suspected issues with the nodes being upgraded, IMM is a safe and speedy choice.

For upgrade flows that include other node maintenance activities – like firmware or driver upgrades done by PowerFlex Manager on racks or appliances – we recommend using PMM. These have a higher probability of lasting longer and encountering issues. Therefore, PMM is the safer choice.

4.5 Spare capacity considerations

Due to the creation of the temporary third data copy, PMM requires more spare capacity than IMM, due to the creation of the temporary third copies. In addition, because PMM cycles may be long and other elements
could fail, there must be enough spare capacity in the system during PMM to handle at least one other node failure. If you plan to use PMM, account for this spare capacity at the time of deployment.

Note that PMM will make use of both the allocated spare capacity in the system as well as any generally free capacity. This allows PMM to make the best use of all unused, available capacity during maintenance.

The following equation summarizes the minimum requirements:

\[
\text{Free + Spare - 5\% of the Storage Pool} \geq \text{capacity of PMM node(s)}
\]

Unlike IMM, there is no way to override the capacity requirements. With IMM it was possible to ignore the spare capacity needed if the maintained node failed. This is not possible in PMM.
Summary

The addition of Protected Maintenance demonstrates the passion of Dell EMC PowerFlex to prioritize customer experience and the protection of customer data. It adds more flexibility with which to manage your PowerFlex storage clusters and addresses the need for always maintaining two available copies of data.

Which maintenance option is best for you and your current circumstances? You should now have a deeper understanding of the relevant factors to make an informed decision, based on business requirements, data recovery time, rebuild/rebalance times, and your cluster size and relative performance.