IBM Elastic Storage System 6.1.4

Problem Determination Guide

SC27-9875-00

Note

Before using this information and the product it supports, read the information in "Notices" on page 125.

This edition applies to Version 6 release 1 modification 4 of the following product and to all subsequent releases and modifications until otherwise indicated in new editions:

- IBM Spectrum® Scale Data Management Edition for IBM® ESS (product number 5765-DME)
- IBM Spectrum Scale Data Access Edition for IBM ESS (product number 5765-DAE)

IBM welcomes your comments; see the topic "How to submit your comments" on page xiv. When you send information to IBM, you grant IBM a nonexclusive right to use or distribute the information in any way it believes appropriate without incurring any obligation to you.

Contents

Tables

About this information

Who should read this information

This information is intended for administrators of IBM Elastic Storage® System (ESS) that includes IBM Spectrum Scale RAID.

IBM Elastic Storage System information units

IBM Elastic Storage System 5147-102 documentation consists of the following information units.

Information unit	Type of information	Intended users
Hardware Planning and Installation Guide	This unit provides ESS 5147-102 information including technical overview, planning, installing, troubleshooting, and cabling.	System administrators and IBM support team
Quick Deployment Guide	This unit provides ESS information including the software stack, deploying, upgrading, setting up call home, and best practices.	System administrators, analysts, installers, planners, and programmers of IBM Spectrum Scale clusters who are very experienced with the operating systems on which each IBM Spectrum Scale cluster is based
Service Guide	This unit provides ESS 5147-102 information including servicing and parts listings.	System administrators and IBM support team
Problem Determination Guide	This unit provides ESS 5147-102 information including events, replacing servers, issues, maintenance procedures, and troubleshooting.	System administrators and IBM support team
Command Reference	This unit provides information about ESS commands and scripts.	System administrators and IBM support team
IBM Spectrum Scale RAID: Administration	This unit provides IBM Spectrum Scale RAID information including administering, monitoring, commands, and scripts.	• System administrators of IBM Spectrum Scale systems • Application programmers who are experienced with IBM Spectrum Scale systems and familiar with the terminology and concepts in the XDSM standard

IBM Elastic Storage System (ESS) 3500 documentation consists of the following information units.

Information unit	Type of information	Intended users
Hardware Planning and Installation Guide	This unit provides ESS 3500 information including technical overview, planning, installing, troubleshooting, and cabling.	System administrators and IBM support team

Information unit	Type of information	Intended users
Quick Deployment Guide	This unit provides ESS information including the software stack, deploying, upgrading, setting up call home, and best practices.	System administrators, analysts, installers, planners, and programmers of IBM Spectrum Scale clusters who are very experienced with the operating systems on which each IBM Spectrum Scale cluster is based
Service Guide	This unit provides ESS 3500 information including servicing and parts listings.	System administrators and IBM support team
Problem Determination Guide	This unit provides ESS 3500 information including events, replacing servers, issues, maintenance procedures, and troubleshooting.	System administrators and IBM support team
Command Reference	This unit provides information about ESS commands and scripts.	System administrators and IBM support team
IBM Spectrum Scale RAID: Administration	This unit provides IBM Spectrum Scale RAID information including administering, monitoring, commands, and scripts.	• System administrators of IBM Spectrum Scale systems • Application programmers who are experienced with IBM Spectrum Scale systems and familiar with the terminology and concepts in the XDSM standard

IBM Elastic Storage System (ESS) 3200 documentation consists of the following information units.

Information unit	Type of information	Intended users
Hardware Planning and Installation Guide	This unit provides ESS 3200 information including technical overview, planning, installing, troubleshooting, and cabling.	System administrators and IBM support team
Quick Deployment Guide	This unit provides ESS information including the software stack, deploying, upgrading, setting up call home, and best practices.	System administrators, analysts, installers, planners, and programmers of IBM Spectrum Scale clusters who are very experienced with the operating systems on which each IBM Spectrum Scale cluster is based
Service Guide	This unit provides ESS 3200 information including servicing and parts listings.	System administrators and IBM support team
Problem Determination Guide	This unit provides ESS 3200 information including events, replacing servers, issues, maintenance procedures, and troubleshooting.	System administrators and IBM support team
Command Reference	This unit provides information about ESS commands and scripts.	System administrators and IBM support team

Information unit	Type of information	Intended users
IBM Spectrum Scale RAID: Administration	This unit provides IBM Spectrum Scale RAID information including administering, monitoring, commands, and scripts.	• System administrators of IBM Spectrum Scale systems • Application programmers who are experienced with IBM Spectrum Scale systems and familiar with the terminology and concepts in the XDSM standard

IBM Elastic Storage System (ESS) 3000 documentation consists of the following information units.

Information unit	Type of information	Intended users
Hardware Planning and Installation Guide	This unit provides ESS 3000 information including technical overview, planning, installing, troubleshooting, and cabling.	System administrators and IBM support team
Quick Deployment Guide	This unit provides ESS information including the software stack, deploying, upgrading, and best practices.	System administrators, analysts, installers, planners, and programmers of IBM Spectrum Scale clusters who are very experienced with the operating systems on which each IBM Spectrum Scale cluster is based
Service Guide	This unit provides ESS 3000 information including events, servicing, and parts listings.	System administrators and IBM support team
Problem Determination Guide	This unit provides ESS 3000 information including setting up call home, replacing servers, issues, maintenance procedures, and troubleshooting.	System administrators and IBM support team
Command Reference	This unit provides information about ESS commands and scripts.	System administrators and IBM support team
IBM Spectrum Scale RAID: Administration	This unit provides IBM Spectrum Scale RAID information including administering, monitoring, commands, and scripts.	• System administrators of IBM Spectrum Scale systems • Application programmers who are experienced with IBM Spectrum Scale systems and familiar with the terminology and concepts in the XDSM standard

IBM Elastic Storage System (ESS) 5000 documentation consists of the following information units.

Information unit	Type of information	Intended users
Hardware Guide	This unit provides ESS 5000 information including system overview, installing, and troubleshooting.	System administrators and IBM support team

Information unit	Type of information	Intended users
Quick Deployment Guide	This unit provides ESS information including the software stack, deploying, upgrading, and best practices.	System administrators, analysts, installers, planners, and programmers of IBM Spectrum Scale clusters who are very experienced with the operating systems on which each IBM Spectrum Scale cluster is based
Model 092 storage enclosures	This unit provides information including initial hardware installation and setup, and removal and installation of field-replaceable units (FRUs), customer-replaceable units (CRUs) for ESS 5000 Expansion – Model 092, 5147-092.	System administrators and IBM support team
Model 106 storage enclosures	This unit provides information including hardware installation and maintenance for ESS 5000 Expansion – Model 106.	System administrators and IBM support team
Problem Determination Guide	This unit provides ESS 5000 information including setting up call home, replacing servers, issues, maintenance procedures, and troubleshooting.	System administrators and IBM support team
Command Reference	This unit provides information about ESS commands and scripts.	System administrators and IBM support team
IBM Spectrum Scale RAID: Administration	This unit provides IBM Spectrum Scale RAID information including administering, monitoring, commands, and scripts.	• System administrators of IBM Spectrum Scale systems • Application programmers who are experienced with IBM Spectrum Scale systems and familiar with the terminology and concepts in the XDSM standard

ESS Legacy documentation consists of the following information units.

Information unit	Type of information	Intended users
Quick Deployment Guide	This unit provides ESS information including the software stack, deploying, upgrading, and best practices.	System administrators, analysts, installers, planners, and programmers of IBM Spectrum Scale clusters who are very experienced with the operating systems on which each IBM Spectrum Scale cluster is based
Problem Determination Guide	This unit provides information including setting up call home, replacing servers, issues, maintenance procedures, and troubleshooting.	System administrators and IBM support team
Command Reference	This unit provides information about ESS commands and scripts.	System administrators and IBM support team

Information unit	Type of information	Intended users
IBM Spectrum Scale RAID: Administration	This unit provides IBM Spectrum Scale RAID information including administering, monitoring, commands, and scripts.	• System administrators of IBM Spectrum Scale systems • Application programmers who are experienced with IBM Spectrum Scale systems and familiar with the terminology and concepts in the XDSM standard

Related information

Related information

For information about:

- IBM Spectrum Scale, see IBM Documentation.
- mmvdisk command, see mmvdisk documentation.
- Mellanox OFED (MLNX_OFED_LINUX-4.9-5.1.0.2) Release Notes, go to https://docs.nvidia.com/networking/display/MLNXOFEDv494170/MLNX_OFED+Documentation+Rev+4.9-4.1.7.0+LTS.
- Mellanox OFED (MLNX_OFED_LINUX-5.4-3.0.3.0) Release Notes, go to https://docs.nvidia.com/networking/display/MLNXOFEDv562090/Release+Notes. (The Mellanox OFED 5.5.x is shipped with ESS 6.1.4.)
- IBM Elastic Storage System, see IBM Documentation.
- IBM Spectrum Scale call home, see Understanding call home.
- Installing IBM Spectrum Scale and CES protocols with the installation toolkit, see Installing IBM Spectrum Scale on Linux® nodes with the installation toolkit.
- Detailed information about the IBM Spectrum Scale installation toolkit, see Using the installation toolkit to perform installation tasks: Explanations and examples.
- CES HDFS, see Adding CES HDFS nodes into the centralized file system.
- Installation toolkit ESS support, see ESS awareness with the installation toolkit.
- IBM POWER8® servers, see https://www.ibm.com/docs/en/power-sys-solutions/0008-ESS?topic=P8ESS/p8hdx/5148_22l_landing.htm
- IBM POWER9™ servers, see https://www.ibm.com/docs/en/ess/6.1.0_ent?topic=guide-5105-22e-reference-information.

For the latest support information about IBM Spectrum Scale RAID, see the IBM Spectrum Scale RAID FAQ in IBM Documentation.

Conventions used in this information

Table 1 on page xiii describes the typographic conventions used in this information. UNIX file name conventions are used throughout this information.

Table 1. Conventions

Convention	Usage
bold	Bold words or characters represent system elements that you must use literally, such as commands, flags, values, and selected menu options. Depending on the context, **bold** typeface sometimes represents path names, directories, or file names.

Table 1. Conventions (continued)

Convention	Usage
bold underlined	bold underlined keywords are defaults. These take effect if you do not specify a different keyword.
constant width	Examples and information that the system displays appear in constant-width typeface. Depending on the context, constant-width typeface sometimes represents path names, directories, or file names.
italic	*Italic* words or characters represent variable values that you must supply. *Italics* are also used for information unit titles, for the first use of a glossary term, and for general emphasis in text.
<key>	Angle brackets (less-than and greater-than) enclose the name of a key on the keyboard. For example, <Enter> refers to the key on your terminal or workstation that is labeled with the word *Enter*.

****	In command examples, a backslash indicates that the command or coding example continues on the next line. For example:
	`mkcondition -r IBM.FileSystem -e "PercentTotUsed > 90" \`
	`-E "PercentTotUsed < 85" -m p "FileSystem space used"`

{item}	Braces enclose a list from which you must choose an item in format and syntax descriptions.
[item]	Brackets enclose optional items in format and syntax descriptions.
<Ctrl-x>	The notation <Ctrl-*x*> indicates a control character sequence. For example, <Ctrl-c> means that you hold down the control key while pressing <c>.
item...	Ellipses indicate that you can repeat the preceding item one or more times.
|	In *synopsis* statements, vertical lines separate a list of choices. In other words, a vertical line means *Or*. In the left margin of the document, vertical lines indicate technical changes to the information.

How to submit your comments

To contact the IBM Spectrum Scale development organization, send your comments to the following email address:

scale@us.ibm.com

Chapter 1. Best practices for troubleshooting

Following certain best practices makes the troubleshooting process easier.

For information on IBM Spectrum Scale issues and their resolution, see the *Troubleshooting* section in the IBM Spectrum Scale documentation.

How to get started with troubleshooting

Troubleshooting the issues that are reported in the system is easier when you follow the process step-by-step.

When you experience some issues with the system, go through the following steps to get started with the troubleshooting:

1. Check the events that are reported in various nodes of the cluster by using the **mmhealth cluster show** and **mmhealth node show** commands.
2. Check the user action corresponding to the active events and take the appropriate action. For more information on the events and corresponding user action, see "Events" on page 71.
3. Check for events that happened before the event you are trying to investigate. They might give you an idea about the root cause of problems. For example, if you see an event nfs_in_grace and node_resumed a minute before you get an idea about the root cause why NFS entered the grace period, it means that the node resumed after a suspend.
4. Collect the details of the issues through logs, dumps, and traces. You can use various CLI commands like gpfs.snap, esssnap and the **Settings** > **Diagnostic Data** GUI page to collect the details of the issues reported in the system.
5. Based on the type of issue, browse through the various topics that are listed in the troubleshooting section and try to resolve the issue.
6. If you cannot resolve the issue by yourself, contact IBM Support.

Example command output

```
# essinstallcheck -N localhost
Start of install check
nodelist:  localhost
Getting package information.
[WARN] Package check cannot be performed other than on EMS node.
Checking nodes.
================= Summary of node: localhost ============================
[INFO] Getting system profile setting.
Installed version:              ess3000_6.1.0.0_0315-22_dme
[OK] Linux kernel installed:            4.18.0-193.40.1.el8_2.x86_64
[OK] Systemd installed:             239-29.el8.x86_64
[OK] Networkmgr installed:          1.22.8-4.el8.x86_64
[OK] Mellanox OFED level:               MLNX_OFED_LINUX-4.9-2.2.4.0
[OK] System Firmware:               2.02.000_0B0G_1.73_FB300052_0C32.official_FW1255_FW1255
[OK] System profile setting:            scale
[OK] System profile verification PASSED.
[OK] Memory inspection passed for ESS 3000 nodes: localhost
[OK] CPU inspection passed for ESS 3000 nodes: localhost
[OK] Boot drive health inspection passed for ESS 3000 nodes: localhost
[OK] Drive partition health inspection passed for ESS 3000 nodes: localhost
[OK] Drive format correct for ESS 3000 nodes: localhost
[OK] GNR Level:                 5.1.0.2 efix4
Performing Spectrum Scale RAID configuration check.
[OK] mmvdisk settings match best practices.
[OK] GNR callback events:
postRGTakeover,postRGRelinquish,rgOpenFailed,rgPanic,pdFailed,pdRecovered,pdReplacePdisk,pdPathD
own,daRebuildFailed
[OK] Network adapter MT4121 firmware: 16.28.2006, net adapter count: 4
[OK] Network adapter firmware
Obtaining storage firmware versions from IO nodes. May take a long time...
[OK] Enclosure 5141-AF8 firmware: 1111, enclosure count: 1
[OK] Drive a8241014065c firmware: SN5ASN5A, drive count: 1
[OK] Drive a8221014063b firmware: SN1OSN1O, drive count: 11
[OK] Storage system firmware
```

```
[OK] Node is not reserving KVM memory.
[WARN] IBM Electronic Service Agent (ESA) test cannot be checked other than EMS node.
[OK] Software Callhome group defined.
[OK] Software callhome configured correctly on ESS system.
[OK] Software Callhome Connectivity Verification Passed.
End of install check
[PASS] essinstallcheck passed successfully

# mmhealth cluster show
Component            Total         Failed       Degraded        Healthy          Other
-------------------------------------------------------------------------------------
NODE                     4              0              0              2              2
GPFS                     4              0              0              2              2
NETWORK                  4              0              0              4              0
FILESYSTEM               1              0              0              1              0
DISK                     4              0              0              4              0
CALLHOME                 1              0              0              1              0
FILESYSMGR               1              0              0              1              0
GUI                      1              0              1              0              0
NATIVE_RAID              2              0              0              2              0
PERFMON                  3              0              0              3              0
THRESHOLD                3              0              0              3              0

# mmhealth node show
Node name:      fab3d-hs.example.net
Node status:    HEALTHY
Status Change:  8 hours ago
Component       Status        Status Change     Reasons
-------------------------------------------------------------
GPFS            HEALTHY       8 hours ago       -
NETWORK         HEALTHY       8 hours ago       -
FILESYSTEM      HEALTHY       8 hours ago       -
DISK            HEALTHY       8 hours ago       -
NATIVE_RAID     HEALTHY       8 hours ago       -
PERFMON         HEALTHY       8 hours ago       -
THRESHOLD       HEALTHY       8 hours ago       -

# gnrhealthcheck
################################################################################
# gnrhealthcheck invoked: Wed Mar 17 10:09:11 MST 2021
################################################################################
################################################################################
# Beginning topology checks.
################################################################################
Topology checks successful.
################################################################################
# Beginning enclosure checks.
################################################################################
Enclosure checks successful.
################################################################################
# Beginning recovery group checks.
################################################################################
Recovery group checks successful.
################################################################################
# Beginning pdisk checks.
################################################################################
Pdisk checks successful.
################################################################################
# Beginning IBM Power RAID checks.
################################################################################
IBM Power RAID checks successful.
################################################################################
# Beginning the NVMe Controller checks.
################################################################################
The NVMe Controller checks are successful.
################################################################################
# Beginning SSD endurance checks
################################################################################
The SSD endurance checks are successful.
```

```
# essinstallcheck -N ems1-ib,essio11-ib,essio12-ib --get-version
Start of install check
nodelist:  ems1-ib  essio11-ib  essio12-ib
Node: ems1-ib        Installed version:          ess5000_6.1.0.0_0311-20_dme
Node: essio11-ib         Installed version:          ess5000_6.1.0.0_0311-20_dme
Node: essio12-ib         Installed version:          ess5000_6.1.0.0_0311-20_dme
[PASS] essinstallcheck passed successfully

# mmhealth cluster show
Component            Total         Failed       Degraded        Healthy          Other
-------------------------------------------------------------------------------------
```

```
NODE                    4               0               1               2               1
GPFS                    4               0               0               3               1
NETWORK                 4               0               3               1               0
FILESYSTEM              2               0               0               2               0
DISK                    8               0               0               8               0
CALLHOME                1               1               0               0               0
CES                     1               0               0               1               0
CESIP                   1               0               0               1               0
FILESYSMGR              1               0               0               1               0
NATIVE_RAID             2               0               2               0               0
PERFMON                 3               1               0               2               0
THRESHOLD               3               0               0               3               0

# mmhealth node show
Node name:      ems1-ib.example.net
Node status:    DEGRADED
Status Change:  5 days ago
Component      Status        Status Change     Reasons
------------------------------------------------------------------------------------------------
------
CALLHOME       FAILED        5 days ago        callhome_heartbeat_failed
GPFS           HEALTHY       5 days ago        -
NETWORK        DEGRADED      5 days ago        ib_rdma_port_width_low(mlx5_0/1, mlx5_1/1)
FILESYSTEM     HEALTHY       5 days ago        -
PERFMON        FAILED        5 days ago        pmcollector_down
THRESHOLD      HEALTHY       5 days ago        -

# gnrhealthcheck
################################################################################
# gnrhealthcheck invoked: Wed Mar 17 12:20:26 CDT 2021
################################################################################
################################################################################
# Beginning topology checks.
################################################################################
Topology checks successful.
################################################################################
# Beginning enclosure checks.
################################################################################
Enclosure checks successful.
################################################################################
# Beginning recovery group checks.
################################################################################
Recovery group checks successful.
################################################################################
# Beginning pdisk checks.
################################################################################
Found recovery group ess5k_7894E4A pdisk n001v001 has 0 paths.
################################################################################
# Beginning IBM Power RAID checks.
################################################################################
IBM Power RAID checks successful.
################################################################################
# Beginning the NVMe Controller checks.
################################################################################
The NVMe Controller checks are successful.
################################################################################
# Beginning SSD endurance checks
################################################################################
The SSD endurance checks are successful.
```

Back up your data

You need to back up data regularly to avoid data loss. It is also recommended to take backups before you start troubleshooting. IBM Spectrum Scale provides various options to create data backups.

Follow the guidelines in the following sections to avoid any issues while creating backup:

- *GPFS(tm) backup data* in *IBM Spectrum Scale: Concepts, Planning, and Installation Guide*
- *Backup considerations for using IBM Spectrum Protect* in *IBM Spectrum Scale: Concepts, Planning, and Installation Guide*
- *Configuration reference for using IBM Spectrum Protect with IBM Spectrum Scale(tm)* in *IBM Spectrum Scale: Administration Guide*
- *Protecting data in a file system using backup* in *IBM Spectrum Scale: Administration Guide*
- *Backup procedure with SOBAR* in *IBM Spectrum Scale: Administration Guide*

The following best practices help you to troubleshoot the issues that might arise in the data backup process:

1. Enable the most useful messages in the **mmbackup** command by setting the **MMBACKUP_PROGRESS_CONTENT** and **MMBACKUP_PROGRESS_INTERVAL** environment variables in the command environment prior to issuing the **mmbackup** command. Setting **MMBACKUP_PROGRESS_CONTENT=7** provides the most useful messages. For more information on these variables, see the *mmbackup command* in the *IBM Spectrum Scale: Command and Programming Reference*.
2. If the **mmbackup** process is failing regularly, enable debug options in the backup process:

   Use the **DEBUGmmbackup** environment variable or the **-d** option that is available in the **mmbackup** command to enable debugging features. This variable controls what debugging features are enabled. It is interpreted as a bitmask with the following bit meanings:

   0x001
   Specifies that basic debug messages are printed to STDOUT. There are multiple components that comprise **mmbackup**, so the debug message prefixes can vary. Some examples include:

   ```
   mmbackup:mbackup.sh
   DEBUGtsbackup33:
   ```

   0x002
   Specifies that temporary files are to be preserved for later analysis.

   0x004
   Specifies that all **dsmc** command output is to be mirrored to STDOUT.

   The **-d** option in the **mmbackup** command line is equivalent to **DEBUGmmbackup=1** .
3. To troubleshoot problems with backup subtask execution, enable debugging in the **tsbuhelper** program.

   Use the **DEBUGtsbuhelper** environment variable to enable debugging features in the **mmbackup** helper program **tsbuhelper**.

Resolve events in a timely manner

Resolving the issues in a timely manner helps to get attention on the new and most critical events. If there are a number of unfixed alerts, fixing any one event might become more difficult because of the effects of the other events. You can use either the CLI or the GUI to view the list of issues that are reported in the system.

You can use the **mmhealth node eventlog** to list the events that are reported in the system.

The **Monitoring** > **Events** GUI page lists all events reported in the system. You can also mark certain events as read to change the status of the event in the events view. The status icons become gray in case an error or warning is fixed or if it is marked as read. Some issues can be resolved by running a fix procedure. Use the action **Run Fix Procedure** to do so. The **Events** page provides a recommendation for which fix procedure to run next.

Keep your software up to date

Check for new code releases and update your code on a regular basis.

This can be done by checking the IBM support website to see if new code releases are available: IBM Elastic Storage Server support website. The release notes provide information about new functions in a release plus any issues that are resolved with the new release. Update your code regularly if the release notes indicate a potential issue.

Note: If a critical problem is detected on the field, IBM may send a flash, advising the user to contact IBM for an efix. The efix when applied might resolve the issue.

Subscribe to the support notification

Subscribe to support notifications so that you are aware of best practices and issues that might affect your system.

Subscribe to support notifications by visiting the IBM support page on the following IBM website: http://www.ibm.com/support/mynotifications.

By subscribing, you are informed of new and updated support site information, such as publications, hints and tips, technical notes, product flashes (alerts), and downloads.

Know your IBM warranty and maintenance agreement details

If you have a warranty or maintenance agreement with IBM, know the details that must be supplied when you call for support.

For more information on the IBM Warranty and maintenance details, see Warranties, licenses and maintenance.

Know how to report a problem

If you need help, service, technical assistance, or want more information about IBM products, then you can find a wide variety of sources available from IBM to assist you.

IBM maintains pages on the web where you can get information about IBM products and fee services, product implementation and usage assistance, break and fix service support, and the latest technical information. The following table provides the URLs of the IBM websites where you can find the support information.

Table 2. IBM websites for help, services, and information

Website	**Address**
IBM home page	http://www.ibm.com
Directory of worldwide contacts	http://www.ibm.com/planetwide
Support for ESS	IBM Elastic Storage Server support website
Support for IBM System Storage® and IBM Total Storage products	http://www.ibm.com/support/entry/portal/product/system_storage/

Note: Available services, telephone numbers, and web links are subject to change without notice.

Before you call

Make sure that you have taken steps to try to solve the problem yourself before you call. Some suggestions for resolving the problem before calling IBM Support include:

- Check all hardware for issues beforehand.
- Use the troubleshooting information in your system documentation. The troubleshooting section of the IBM Documentation contains procedures to help you diagnose problems.

To check for technical information, hints, tips, and new device drivers or to submit a request for information, go to the IBM Elastic Storage Server support website.

Using the documentation

Information about your IBM storage system is available in the documentation that comes with the product. That documentation includes printed documents, online documents, readme files, and help files in addition to the IBM Documentation.

Chapter 2. Collecting information about an issue

To begin the troubleshooting process, collect information about the issue that the system is reporting.

Collect the output of the **gpfs.snap** and **essinstallcheck** commands from each I/O canister node. Run the **esssnap** on the EMS to collect critical deployment related logs. Both the **gpfs.snap** command and the **essinstallcheck** command can also be run within this command.

From the EMS, issue the following command:

```
esssnap -i -g -N <IO  node1>,<IO node 2>,..,<IO node X>
```

The system returns a **gpfs.snap**, an **essinstallcheck**, and the data from each node.

The following example output displays a snap taken only on the EMS. All I/O server logs are redirected to the EMS thus it is only required to take the snap from the management node.

```
/opt/ibm/ess/tools/bin/esssnap
esssnap [INFO]: Collecting sosreports for node(s): essems1.gpfs.net
esssnap [INFO]: Collecting iprsosreports for node(s): essems1.gpfs.net
esssnap [INFO]: Collecting ESS snap
################################################################################################
tar file: /tmp/esssnap.20211026T025200Z.tgz
SHA256 file: /tmp/esssnap.20211026T025200Z.tgz.sha256
Please provide tar file to IBM service
################################################################################################
```

For more information, see *gsssnap script* in *Deploying the Elastic Storage Server* and *esssnap* command in *Elastic Storage Server: Command Reference*.

Chapter 3. Servicing log tip

This section gives information about servicing log.

Re-creating the NVR partitions for ESS Legacy

The Non-Volatile Random-Access Memory (NVRAM) physically resides within the IPR-Raid adapter that is installed on the EMS and each of the I/O nodes. The NVR partitions are created on the local sda drive that is installed on the ESS I/O nodes to hold data for the log tip pdisks.

Although a total of 6 partitions are created, only 2 are actually used per I/O node, one for each NVR pdisk. In some cases, the NVRAM partitions might need to be recreated. For example, a hardware or OS failure would be one such case.

Before re-creating the NVR partitions, list all the existing partitions for sda. To list all partitions for sda, run the following command:

```
parted /dev/sda unit KiB print
```

This command gives a similar output:

```
Model: IBM IPR-10 749FFB00 (scsi)
Disk /dev/sda: 557727744kiB
Sector size (logical/physical): 512B/4096B
Partition Table: msdos
Disk Flags:

Number  Start         End           Size          Type      File system     Flags
1      1024kiB       9216kiB       8192kiB       primary                   boot, prep
2      9216kiB       521216kiB     512000kiB     primary   xfs
3      521216kiB     176649216kiB  176128000kiB  primary   xfs
4      176649216kiB  557727744kiB  381078528kiB  extended
5      176651264kiB  279051264kiB  102400000kiB  logical   xfs
6      279052288kiB  381452288kiB  102400000kiB  logical   xfs
7      381453312kiB  483853312kiB  102400000kiB  logical   xfs
8      483854336kiB  535054336kiB  51200000kiB   logical   xfs
9      535055360kiB  543247360kiB  8192000kiB    logical   linux-swap(v1)
```

For optimal alignment, each partition must be exactly 2048000 KiB in size, and must be 1024 KiB apart from each other.
In the sample output, the last end size pertains to Partition # 9, and has a value of 543247360 KiB.
To get the NVR partition's new start value, add 1024 KiB to the last end size value, and add 2048000 KiB to the start value to determine the new end as shown:

1. NVR Partition 1 new start value = Last end size value + 1024 KiB = 543247360 KiB + 1024 KiB = 543248384 KiB
2. NVR Partition 1 new end = NVR Partition 1 new start value + 2048000 KiB = 543248384 KiB + 2048000 KiB = 545296384 KiB

To create the first NVR partition, run the following command:

```
parted /dev/sda mkpart logical 543248384KiB 545296384KiB
```

To get the new start for the second partition, you need to add 1024 KiB to the end size value of partition 1. Repeat the steps to calculate the start and end positions for the second partition as shown:

1. NVR Partition 2 new start = NVR Partition 1 end value + 1024 KiB = 545296384 KiB + 1024 KiB = 545297408 KiB
2. NVR Partition 2 new end = NVR Partition 2 new start value + 2048000 KiB = 545297408 KiB + 2048000 KiB = 547345408 KiB

Repeat the above steps four times to create a total of six partitions. When complete, the partitions list for sda will look similar to the following:

```
 [root@ems1 ~]# parted /dev/sda unit KiB print
Model: IBM IPR
-
10 749FFB00 (scsi)

Disk /dev/sda: 557727744kiB
Sector size (logical/physical): 512B/4096B
Partition Table: msdos
Disk Flags:
Number   Start         End           Size          Type      File system     Flags
1        1024kiB       9216kiB       8192kiB       primary                   boot, prep
2        9216kiB       521216kiB     512000kiB     primary   xfs
3        521216kiB     176649216kiB  176128000kiB  primary   xfs
4        176649216kiB  557727744kiB  381078528kiB  extended
5        176651264kiB  279051264kiB  102400000kiB  logical   xfs
6        279052288kiB  381452288kiB  102400000kiB  logical   xfs
7        381453312kiB  483853312kiB  102400000kiB  logical   xfs
8        483854336kiB  535054336kiB  51200000kiB   logical   xfs
9        535055360kiB  543247360kiB  8192000kiB    logical   linux-swap(v1)
10       543248384kiB  545296384kiB  2048001kiB    logical   xfs
11       545297408kiB  547345408kiB  2048001kiB    logical   xfs
12       547346432kiB  549394432kiB  2048001kiB    logical   xfs
13       549395456kiB  551443456kiB  2048001kiB    logical   xfs
14       551444480kiB  553492480kiB  2048001kiB    logical   xfs
15       553493504kiB  555541504kiB  2048001kiB    logical   xfs
```

Re-creating NVRAM disks for ESS Legacy systems

NVRAM pdisks are used to store the log tip data, which is eventually migrated to the log home vdisk. Although ESS can continue to function without NVRAM pdisks, the performance is impacted without their presence. Therefore, it is important to ensure that the NVRAM pdisks are functioning at all times.

The NVRAM pdisks may stop functioning and go into a `missing` state. This could be due to hardware failure of the IPR card, or corrupt or missing NVR OS partition caused by an OS failure. To fix this problem, the NVRAM pdisks must be recreated.

Attention: For recovery groups under the management of the **mmvdisk** command, contact IBM® support for assistance in re-creating the NVRAM disks.

You can find the pdisks that are in a missing state by running the **mmlsrecoverygroup** command as shown:

```
mmlsrecoverygroup rg_gssio1 -L --pdisk | grep NVR
NVR       no            1       2     0,0          1    3632 MiB   14 days  inactive  0%  low
n1s01              0,  0      NVR          1816 MiB    missing
n2s01              0,  0      NVR          1816 MiB    missing

mmlsrecoverygroup rg_gssio2 -L --pdisk | grep NVR
NVR        no            1       2     0,0          1    3632 MiB   14 days  inactive  0%  low
n1s02               0,  0      NVR          1816 MiB    missing
n2s02               0,  0      NVR          1816 MiB    missing
```

Before recreating the pdisks, ensure that all six NVRAM partitions exist on the sda by using the following command:

```
parted /dev/sda unit KiB print

Model: IBM IPR
-
10 749FFB00 (scsi)
Disk /dev/sda: 557727744kiB
Sector size (logical/physical): 512B/4096B
Partition Table: msdos
Disk Flags:
Number  Start         End           Size          Type      File system     Flags
1       1024kiB       9216kiB       8192kiB       primary                   boot, prep
2       9216kiB       521216kiB     512000kiB     primary   xfs
```

```
3       521216kiB     176649216kiB  176128000kiB  primary   xfs
4       176649216kiB  557727744kiB  381078528kiB  extended
5       176651264kiB  279051264kiB  102400000kiB  logical   xfs
6       279052288kiB  381452288kiB  102400000kiB  logical   xfs
7       381453312kiB  483853312kiB  102400000kiB  logical   xfs
8       483854336kiB  535054336kiB  51200000kiB   logical   xfs
9       535055360kiB  543247360kiB  8192000kiB    logical   linux-swap(v1)
10      543248384kiB  545296384kiB  2048001kiB    logical   xfse*/
11      545297408kiB  547345408kiB  2048001kiB    logical   xfs
12      547346432kiB  549394432kiB  2048001kiB    logical   xfs
13      549395456kiB  551443456kiB  2048001kiB    logical   xfs
14      551444480kiB  553492480kiB  2048001kiB    logical   xfs
15      553493504kiB  555541504kiB  2048001kiB    logical   xfs
```

Note: In case the partitions are not present, you must recreate the 6 NVR partitions. For more information, see "Re-creating the NVR partitions" .

After you have verified the 6 NVR partitions, create a stanza file for each of the NVRAM devices that are missing, and save it.

```
gssio1.stanza:
%pdisk: pdiskName=n1s01 device=//gssio1/dev/sda10 da=NVR rotationRate=NVRAM
%pdisk: pdiskName=n2s01 device=//gssio2/dev/sda10 da=NVR rotationRate=NVRAM
```

Run the **mmaddpdisk** command using the stanza file that was created to replace the missing pdisks.

```
mmaddpdisk rg_gssio1 -F gssio1.stanza --replace
```

The following pdisks will be formatted on the node `gssio.ess.com`:

- //gssio1/dev/sda10
- //gssio2/dev/sda10

Run the **mmlsrecoverygroup** command to confirm the current state of the pdisks.

```
mmlsrecoverygroup rg_gssio1 -L --pdisk | grep NVR
n1s01                 1,  1      NVR          1816 MiB    ok
n2s01                 1,  1      NVR          1816 MiB    ok
```

Run the **mmaddpdisk** command to recreate the other missing NVRAM pdisks.

Re-creating NVRAM disks for ESS 5000

The Non-Volatile Random-Access Memory (NVRAM) is supported by NVDIMM block devices that are installed on each of the I/O nodes. The NVRAM pdisks are created from NVDIMM block devices that are installed on the Enterprise Storage Server® I/O nodes to hold data for the log tip pdisks. There are two NVDIMM block drives, `/dev/pmem0s` and `/dev/pmem1s`, available on each of the ESS I/O nodes. NVRAM pdisks are used to store the log tip data, which is eventually moved to the log home vdisk. Although ESS can continue to function without NVRAM pdisks, the performance is impacted without their presence. Therefore, it is important to ensure that the NVRAM pdisks are functioning always.

The NVRAM pdisks might stop functioning, and go into a `missing` state. This might be due to one of the following:

- Loss of high-speed network between the I/O Server nodes of the building block
- Hardware failure of the NVDIMM card
- Corrupted or missing NVDIMM block devices caused by an OS failure

To fix this problem, the NVRAM pdisks must be re-created.

You can find the pdisks that are in a `missing` state by running the **mmvdisk pdisk list** command for NVR declustered array of individual recovery groups by specifying the `--not-ok` option, as shown:

```
# mmvdisk  pdisk list --recovery-group BB01L  --declustered-array  NVR --not-ok
```

```
                          declustered
recovery group   pdisk        array      paths   capacity   free space   FRU (type)   state
--------------   ---------    ------     -----   --------   ----------   ----------   -----
BB01L            n002v001     NVR        0       31 GiB      31 GiB      34GB NVRAM    missing/
undrainable

#
```

The state of the pdisk n002v001 is missing or undrainable. If an NVRAM pdisk state is missing, then its corresponding NVDIMM block device is either missing or has encountered some hardware errors, or the NVDIMM block devices are not formatted in the `sector` mode.

Use the **`mmhealth`** command with following syntax to display the drives that are in a `missing` state:

```
# mmhealth node show NATIVE_RAID PHYSICALDISK
```

The system displays an output similar to the following:

```
Node name:      c145f03zn02.gpfs.net

Component          Status        Status Change     Reasons
..
  BB01L/e2s105     HEALTHY       2 days ago        -
  BB01L/n001v001   HEALTHY       2 days ago        -
  BB01L/n002v001   DEGRADED      2 days ago        gnr_pdisk_missing(BB01L/n002v001)
  BB01R/e1s098     HEALTHY       2 days ago        -
  BB01R/e1s100     HEALTHY       2 days ago        -
..
```

The **`ndctl`** command can be run on each I/O node to verify that the NVDIMM block devices are available and properly formatted in `sector` mode on the I/O nodes. The **`ndctl list`** command displays an output similar to the following:

```
# ndctl list
[
  {
    "dev":"namespace1.0",
    "mode":"sector",
    "size":34325135360,
    "uuid":"df8d9a0f-115d-4ff1-8367-efdbac6a3684",
    "sector_size":4096,
    "blockdev":"pmem1s"
  },
  {
    "dev":"namespace0.0",
    "mode":"sector",
    "size":34325135360,
    "uuid":"927c8c54-15c6-4612-b2b5-9d99faf9adaf",
    "sector_size":4096,
    "blockdev":"pmem0s"
  }
]
#
```

Verify that the NVDIMM block device mode is `sector`, and the sector-size is 4096 bytes. If the block device is not in `sector` mode, then it needs to be converted to the `sector` mode before you add the NVDIMM block device as NVRAM pdisks to the recovery group. For example, the NVDIMM block device, whose namespace is `namespace0.0`, can be in `fsdax` mode as shown:

```
# ndctl list -n namespace0.0
[
  {
    "dev":"namespace0.0",
    "mode":"fsdax",
    "map":"dev",
    "size":34324086784,
    "uuid":"ecf1092c-5576-4002-892a-7c49dde54f43",
    "sector_size":512,
    "align":2097152,
    "blockdev":"pmem0"
  }
```

```
]
#
```

To convert the NVDIMM block device mode to `sector` mode, run the **`ndctl`** command as shown:

```
#ndctl create-namespace -e namespace0.0 -m sector -l 4096 -f
```

The system displays an output similar to the following:

```
# ndctl create-namespace -e namespace0.0 -m sector -l 4096 -f
{
  "dev":"namespace0.0",
  "mode":"sector",
  "size":"31.97 GiB (34.33 GB)",
  "uuid":"6efda885-698d-41a3-9c84-16e7c89fd1e2",
  "sector_size":4096,
  "blockdev":"pmem0s"
}
#
```

After the NVDIMM block device is converted to `sector` mode, verify it again using **`ndctl list`** command as shown:

```
# ndctl list -n namespace0.0
[
  {
    "dev":"namespace0.0",
    "mode":"sector",
    "size":34325135360,
    "uuid":"6efda885-698d-41a3-9c84-16e7c89fd1e2",
    "sector_size":4096,
    "blockdev":"pmem0s"
  }
]
#
```

Verify that the NVDIMM block device is changed to `sector` mode, and that the block device name now appears with the character `s` appended at the end. The NVDIMM drives also can be listed from the `/dev` directory by using the **`ls`** command as shown:

```
# ls -l  /dev/pmem*s
brw-rw----. 1 root disk 259, 0 Jul 29 03:27 /dev/pmem0s
brw-rw----. 1 root disk 259, 1 Jul 29 03:27 /dev/pmem1s
#
```

Sometimes all the NVDIMM block devices are available on an I/O node, and the mode of these devices is set to `sector`, but the NVRAM pdisks are still missing. In such cases, the NVDIMM devices might encounter some hardware errors for which a call home event is generated. Similarly, if one or both of the NVDIMM devices are missing from the **`ndctl list`** command, then the NVDIMM devices encounters hardware issues for which a call home event is generated.

For NVDIMM drive hardware errors that require replacement, refer to the ESS 5000 I/O Server node hardware component replacement procedures. If an NVDIMM drive was reformatted or replaced, then the associated log tip pdisk must be re-created to make the NVDIMM drive usable by the recovery group. After all the issues are resolved, contact IBM support for assistance in re-creating the NVRAM disks.

Replacing the logtip backup solid state drive

The solid state drive (SSD) is used to hold the `logtip backup vdisk`. In a normal operation, when both the copies of the primary logtip vdisk on the NVDIMM are available, the logtip backup vdisk is not used during the write. However, when a write to the logtip is not able to make copies on every replica of the vdisk, GNR also writes the logtip data to the logtip backup vdisk.

Ensure that the following checks are done before you begin to replace the SSD DA:

1. Ensure that the system is healthy by running the **essrun healthcheck** and **mmhealth node show** commands.
2. Ensure that the hardware does not have any issues.
3. Stop or suspend any long running tasks.
4. Find a maintenance window with a light load. For example, weekends or off-hour shifts, etc.

Contact IBM support to perform the logtip backup replacement procedure. When the logtip backup replacement procedure is completed, perform a final health check by running the **essrun healthcheck** and **mmhealth node show** commands.

Steps to restore an I/O node for ESS Legacy

If an I/O node fails due to a hardware or OS problem, and the OS is no longer accessible, you must restore the node by using the existing configuration settings that are stored in xCAT, which typically stored on the EMS node.

This process restores the OS image and the required ESS software, drivers, and firmware.

Note: For the following steps, assume that the `gssio1` node is the node that is being restored.

1. Disable the GPFS auto load by using the `mmchconfig` command.

   Note: This prevents GPFS from restarting automatically upon reboot.

```
[ems]# mmlsconfig autoload
    autoload yes
 [ems]# mmchconfig autoload=no
 [ems]# mmlsconfig autoload
    autoload no
```

2. List the recovery groups by using the `mmlsrecoverygroup` command to verify that the replacement node is not an active recovery group server currently.

```
    [ems1]# mmlsrecoverygroup
     recovery group        vdisks     vdisks  servers
     ------------------  -----------  ------  -------
     rg_gssio1                   3      18  gssio1,gssio2
     rg_gssio2                   3      18  gssio2,gssio1
```

   List the current active recovery group server for each recovery group.

```
[ems1]# mmlsrecoverygroup rg_gssio1 -L | grep "active recovery" -A2
     active recovery group server                    servers
     ------------------------------------------------  -------
     gssio1                                          gssio1,gssio2

[ems1]# mmlsrecoverygroup rg_gssio2 -L | grep "active recovery" -A2
     active recovery group server                    servers
     ------------------------------------------------  -------
     gssio2                                          gssio2,gssio1
```

   Note: When you restore `gssio1`, the primary and active recovery group server for `rg_gssio1` must be `gssio2`. If the server is not set to `gssio2`, you must run the **mmchrecoverygroup** command. If the recovery group is under the mmvdisk control, you must run the **mmvdisk** command to change the server.

```
[ems1]# mmchrecoverygroup <RG> --servers <NEW PRIMARY NODE>,<OLD PRIMARY NODE> -v no
[ems1]# mmchrecoverygroup <RG> --active <NEW PRIMARY NODE>
or
[ems1]# mmvdisk rg change --rg <RG> --primary <NEW PRIMARY NODE> --backup <OLD PRIMARY
NODE> -v no
[ems1]# mmvdisk rg change --rg <RG> --active <NEW PRIMARY NODE>

[root@gssio1 ~]# mmchrecoverygroup rg_gssio1 --servers gssio2,gssio1 -v no
[root@gssio1 ~]# mmchrecoverygroup rg_gssio1 --active gssio2
[ems1]# mmlsrecoverygroup rg_gssio1 -L | grep "active recovery" -A2
     active recovery group server                    servers
```

```
     ---------------------------------------------  -------
     gssio2                                         gssio2,gssio1

[ems1]# mmlsrecoverygroup rg_gssio2 -L | grep "active recovery" -A2
     active recovery group server                   servers
     ---------------------------------------------  -------
     gssio2                                         gssio2,gssio1
```

3. Create a backup of the replacement node's network file.

```
[ems]# rm -rf /tmp/replacement_node_network_backup
[ems]# mkdir /tmp/replacement_node_network_backup
[ems]# scp <REPLACEMENT NODE>:/etc/sysconfig/network-scripts/ifcfg-*
 /tmp/replacement_node_network_backup/
    [ems]# scp gssio2:/etc/sysconfig/network-scripts/ifcfg-*
 /tmp/replacement_node_network_backup/
```

Note: This is an optional step, and can be taken only when the replacement node can be accessed.

4. Check for the RHEL images available for installation on the EMS node.

   The RHEL image is needed to re-image the node that is being restored. The OS image must be located on the EMS node under the following directory:

```
[ems]# ls /tftpboot/xcat/osimage/
    rhels7.3-ppc64-install-gss
```

5. Configure the replacement node's boot state to `Install` for the specified OS image.

```
[ems]# nodeset <REPLACEMENT NODE> osimage=<OS_ISO_image>
    [root@ems1 ~]# nodeset gssio2 osimage=rhels7.3-ppc64-install-gss
    gssio2: install rhels7.3-ppc64-gss
```

6. Ensure that the remote console is properly configured on the EMS node.

```
[ems]# makeconservercf <REPLACEMENT NODE>
 [root@ems1 ~]# makeconservercf gssio2
```

7. Reboot the replaced node to initiate the installation process.

```
[ems]# rnetboot <REPLACEMENT NODE> -V
[root@ems1 ~]# rnetboot gssio2 -V
lpar_netboot Status: List only ent adapters
lpar_netboot Status: -v (verbose debug) flag detected
lpar_netboot Status: -i (force immediate shutdown) flag detected
lpar_netboot Status: -d (debug) flag detected
node:gssio2
Node is gssio2
...
    # Network boot proceeding - matched BOOTP, exiting.
    # Finished.
    sending commands ~. to expect
    gssio2: Success
```

   Monitor the progress of the installation, and wait for the `xcatpost/yum/etc` script to finish.

```
[ems]# watch "nodestat <REPLACEMENT NODE>; echo; tail /var/log/consoles/<REPLACEMENT NODE>"
    [root@ems1 ~]# watch "nodestat gssio2; echo; tail /var/log/consoles/gssio2"
    gssio2: noping
    ...
    gssio2: install rhels7.3-ppc64-gss
    ...
    gssio2: sshd
```

```
[ems]# watch -n .5 "ssh <REPLACEMENT NODE> 'ps -eaf | grep -v grep' |
 egrep 'xcatpost|yum|rpm|vpd'"
    [root@ems1 ~]# watch -n .5 "ssh gssio2 'ps -eaf | grep -v grep' |
egrep 'xcatpost|yum|rpm|vpd'"
```

Note: Depending on what needs to be updated, the node might reboot more than once. You must wait until there is no process output before taking the next step.

8. Verify that the upgrade files are copied to the I/O node sync directory, `/install/gss/sync/ppc64/`.

```
[ems]# ssh <REPLACEMENT NODE> "ls /install/gss/sync/ppc64/"
    [root@ems1]# ssh gssio2 "ls /install/gss/sync/ppc64/"
    gssio2: mofed
```

Wait for the directory to sync. After the `mofed` directory is created, you can take the next step.

9. Copy the host files from the healthy node to the replacement node.

```
[ems]# scp /etc/hosts <REPLACEMENT NODE>:/etc/
    [root@ems1 mofed]# scp /etc/hosts gssio2:/etc/
```

10. Configure the network on the replacement node.

    If you created a backup of the network files previously, you can copy them over to the node, and restart the node. Verify that the names of the devices are consistent with the names in the backup file before you replace the files.

    You can also apply the Red Hat® updates not included in the xCAT image, if necessary.

11. Rebuild the GPFS kernel extensions on the replacement node.

    If the kernel patches were applied, it might be necessary to rebuild the GPFS portability layer by running the **mmbuildgpl** command.

```
[ems]# ssh <REPLACEMENT NODE> "/usr/lpp/mmfs/bin/mmbuildgpl"
    [root@ems1 ~]# ssh gssio2 "/usr/lpp/mmfs/bin/mmbuildgpl"
    -----------------------------------------------------
    mmbuildgpl: Building GPL module begins at Wed Nov  8 17:18:21 EST 2017.
    -----------------------------------------------------
    Verifying Kernel Header...
      kernel version = 31000514 (3.10.0-514.28.1.el7.ppc64, 3.10.0-514.28.1)
      module include dir = /lib/modules/3.10.0-514.28.1.el7.ppc64/build/include
      module build dir   = /lib/modules/3.10.0-514.28.1.el7.ppc64/build
      kernel source dir  = /usr/src/linux-3.10.0-514.28.1.el7.ppc64/include
      Found valid kernel header file under /usr/src/kernels/3.10.0-514.28.1.el7.ppc64/include
    Verifying Compiler...
      make is present at /bin/make
      cpp is present at /bin/cpp
      gcc is present at /bin/gcc
      g++ is present at /bin/g++
      ld is present at /bin/ld
    Verifying Additional System Headers...
      Verifying kernel-headers is installed ...
        Command: /bin/rpm -q kernel-headers
        The required package kernel-headers is installed
    make World ...
    make InstallImages ...
    -----------------------------------------------------
    mmbuildgpl: Building GPL module completed successfully at Wed Nov  8 17:18:39 EST 2017.
```

12. Restore the GPFS configuration from an existing healthy node in the cluster.

```
[ems]# ssh <REPLACEMENT NODE> "/usr/lpp/mmfs/bin/mmsdrrestore -p <GOOD NODE>"
    [root@ems ~]# ssh gssio2 "/usr/lpp/mmfs/bin/mmsdrrestore -p ems1"
    mmsdrrestore: Processing node gssio1
    mmsdrrestore: Node gssio1 successfully restored.
```

Note: This code is run on the replacement node, and the `-p` option is applied to an existing healthy node.

13. Start GPFS on the recovered node, and enable the GPFS auto load.

    a. Before you start GPFS, verify that the replacement node is still in the `DOWN` state.

```
[ems]# mmgetstate -aL
     Node number  Node name  Quorum  Nodes up  Total nodes  GPFS state  Remarks
    ------------------------------------------------------------------------------------
           1      gssio1        2        2          5       active      quorum node
           2      gssio2        0        0          5       down        quorum node
           3      ems1          2        2          5       active      quorum node
           4      gsscomp1      2        2          5       active
           5      gsscomp       2        2          5       active
```

    b. Start GPFS on the replacement node.

```
[ems]# mmstartup -N <REPLACEMENT NODE>
    mmstartup: Starting GPFS ...
```

c. Verify that the replacement node is active.

```
[ems]# mmgetstate -aL
     Node number  Node name  Quorum  Nodes up  Total nodes  GPFS state  Remarks
    ------------------------------------------------------------------------------------
           1      gssio1        2        3          5       active      quorum node
           2      gssio2        2        3          5       active      quorum node
           3      ems1          2        3          5       active      quorum node
           4      gsscomp1      2        3          5       active
           5      gsscomp2      2        3          5       active
```

d. Ensure that all the file systems are mounted on the replacement node.

```
[ems]# mmmount all -N <REPLACEMENT NODE>
[ems]# mmlsmount all -L
```

e. Re-enable the GPFS auto load.

```
[ems]# mmlsconfig autoload
    autoload no

[ems]# mmchconfig autoload=yes
    mmchconfig: Command successfully completed

[ems]# mmlsconfig autoload
    autoload yes
```

14. The primary and active recovery group server for `rg_gssio1` must be `gssio1`. If the server is not set to `gssio1`, you must run the **mmchrecoverygroup** command to set the server. If the recovery group is under the mmvdisk control, you must run the **mmvdisk** command to change the server.

```
[ems1]# mmchrecoverygroup <RG> --servers <NEW PRIMARY NODE>,<OLD PRIMARY NODE> -v no
[ems1]# mmchrecoverygroup <RG> --active <NEW PRIMARY NODE>
or
[ems1]# mmvdisk rg change --rg <RG> --primary <NEW PRIMARY NODE> --backup <OLD PRIMARY
NODE> -v no
[ems1]# mmvdisk rg change --rg <RG> --active <NEW PRIMARY NODE>

[root@gssio1 ~]# mmchrecoverygroup rg_gssio1 --servers gssio1,gssio2 -v no
[root@gssio1 ~]# mmchrecoverygroup rg_gssio1 --active gssio1
[ems1]# mmlsrecoverygroup rg_gssio1 -L | grep "active recovery" -A2
     active recovery group server                    servers
     ----------------------------------------------  -------
     gssio1                                          gssio1,gssio2

[ems1]# mmlsrecoverygroup rg_gssio2 -L | grep "active recovery" -A2
     active recovery group server                    servers
     ----------------------------------------------  -------
     gssio2                                          gssio2,gssio1
```

15. Verify that the NVRAM partition exists, and ensure that the following conditions are met:
    - There must be 11 partitions.
    - Partitions 6 through 11 must be 2 GB.
    - Partitions 6 through 9 are marked as `xfs` for file system.
    - Partitions 10 and 11 must not have a file system that is associated with it.
    - After re-imaging, the node that was re-imaged will have an `xfs` file system as shown:

```
[ems]# ssh gssio1 "lsblk | egrep 'NAME|sda[0-9]'"
    NAME    MAJ:MIN RM   SIZE RO TYPE MOUNTPOINT
    ├─sda1    8:1    0     8M  0 part
    ├─sda2    8:2    0   500M  0 part /boot
    ├─sda3    8:3    0 246.1G  0 part /
    ├─sda4    8:4    0     1K  0 part
    ├─sda5    8:5    0   3.9G  0 part [SWAP]
    ├─sda6    8:6    0     2G  0 part
    ├─sda7    8:7    0     2G  0 part
    ├─sda8    8:8    0     2G  0 part
    ├─sda9    8:9    0     2G  0 part
    ├─sda10   8:10   0     2G  0 part
    └─sda11   8:11   0     2G  0 part
```

```
[ems1]# ssh gssio1 "parted /dev/sda -l | egrep 'boot, prep' -B 1 -A 10"
    Number  Start   End     Size    Type      File system     Flags
     1      1049kB  9437kB  8389kB  primary                   boot, prep
     2      9437kB  534MB   524MB   primary   xfs
     3      534MB   265GB   264GB   primary   xfs
     4      265GB   284GB   18.9GB  extended
     5      265GB   269GB   4194MB  logical   linux-swap(v1)
     6      269GB   271GB   2097MB  logical   xfs
     7      271GB   273GB   2097MB  logical   xfs
     8      273GB   275GB   2097MB  logical   xfs
     9      275GB   277GB   2097MB  logical   xfs
    10      277GB   279GB   2097MB  logical
    11      279GB   282GB   2097MB  logical

[ems1]# ssh gssio2 "lsblk | egrep 'NAME|sda[0-9]'"
    NAME    MAJ:MIN RM   SIZE RO TYPE MOUNTPOINT
    ├─sda1    8:1    0     8M  0 part
    ├─sda2    8:2    0   500M  0 part /boot
    ├─sda3    8:3    0 246.1G  0 part /
    ├─sda4    8:4    0     1K  0 part
    ├─sda5    8:5    0   3.9G  0 part [SWAP]
    ├─sda6    8:6    0     2G  0 part
    ├─sda7    8:7    0     2G  0 part
    ├─sda8    8:8    0     2G  0 part
    ├─sda9    8:9    0     2G  0 part
    ├─sda10   8:10   0     2G  0 part
    └─sda11   8:11   0     2G  0 part

[ems1]# ssh gssio2 "parted /dev/sda -l | egrep 'boot, prep' -B 1 -A 10"
    Number  Start   End     Size    Type      File system     Flags
     1      1049kB  9437kB  8389kB  primary                   boot, prep
     2      9437kB  534MB   524MB   primary   xfs
     3      534MB   265GB   264GB   primary   xfs
     4      265GB   284GB   18.9GB  extended
     5      265GB   269GB   4194MB  logical   linux-swap(v1)
     6      269GB   271GB   2097MB  logical   xfs
     7      271GB   273GB   2097MB  logical   xfs
     8      273GB   275GB   2097MB  logical   xfs
     9      275GB   277GB   2097MB  logical   xfs
    10      277GB   279GB   2097MB  logical   xfs
    11      279GB   282GB   2097MB  logical   xfs
```

If the partitions do not exist, you need to create them. For more information, see “Re-creating the NVR partitions for ESS Legacy” on page 9

16. View the current NVR device status.

```
[ems1]# mmlsrecoverygroup rg_gssio1 -L --pdisk | egrep "n[0-9]s[0-9]"
    n1s01                1,  1      NVR          1816 MiB    ok
    n2s01                0,  0      NVR          1816 MiB    missing

[ems1]# mmlsrecoverygroup rg_gssio2 -L --pdisk | egrep "n[0-9]s[0-9]"
    n1s02                1,  1      NVR          1816 MiB    ok
    n2s02                0,  0      NVR          1816 MiB    missing
```

Note: The missing NVR devices must be re-created or replaced. For more information, see “Re-creating NVRAM disks for ESS Legacy systems” on page 10.

Chapter 4. ESS deployment troubleshooting: Helpful podman, Ansible, and log information

This section details the following podman, Ansible®, and log information.

Creating CES shared root file system required for deploying protocol nodes

Use the following command to create a small CES shared root file system, which is required for protocol nodes.

```
essrun -N prt1 filesystem --suffix=-hs --ces
```

The following is a high-level instruction set for using the installation toolkit to create a cluster with protocol nodes and start CES services.

```
./Spectrum_Scale_Data_Management-5.1.0.2-ppc64LE-Linux-install --silent
cd /usr/lpp/mmfs/5.0.5.1/installer/
./spectrumscale node list
./spectrumscale setup -s EMSNodeHighSpeedIP -i /root/pem_key/id_rsa
./spectrumscale config populate -N EMSNodeHighSpeedName
./spectrumscale setup -s EMSNodeHighSpeedIP -i /root/pem_key/id_rsa
./spectrumscale node add EMSNodeHighSpeedIP -a
./spectrumscale node add ProtocolNodeHighSpeedIP  -p
./spectrumscale node list
./spectrumscale install -pr
./spectrumscale install
./spectrumscale config protocols -e CESIP1,CESIP2,...
./spectrumscale config protocols -f cesSharedRoot -m /gpfs/cesSharedRoot
./spectrumscale enable nfs
./spectrumscale enable smb
./spectrumscale node list
./spectrumscale deploy --precheck
./spectrumscale deploy
```

For more information, see the *Installing IBM Spectrum Scale on Linux nodes with the installation toolkit* section in the IBM Spectrum Scale documentation

Adding an additional ESS 3000 storage to an existing file system

Before doing these steps, follow the steps in *ESS 3000 initial setup instructions* in *ESS 3000: Quick Deployment Guide*. Make sure that you update the `/etc/hosts` file with the new node names and IP addresses. Copy the updated `/etc/hosts` to all nodes before starting. Stop after creating the network bonds.

1. Add ESS 3000 nodes to the current file system.

```
essrun -N NodeAlreadyinCluster cluster --add-nodes NewNode1,NewNode2 --suffix=Suffix
```

2. Configure the **mmvdisk** node class. A unique node class name is required for a new building block.

```
mmvdisk server configure --nc ChosenNodeClassName --recycle one
```

3. Create the recovery group.

```
essrun -N NewNode1,NewNode2 vdisk --name ChosenVdiskSetName --suffix=Suffix --code RAIDCode
--bs BlockSize --size SetSize --extra-vars "--nsd-usage dataOnly --sp data"
```

Note: For this example command, it is assumed that you are adding data-only vdisks to the existing file system. You might have a different use case, so adjust options accordingly. For help and default values, use **essrun vdisk --help**.

4. Define the vdisk set.

```
mmvdisk vs define --vs ChosenVdiskSetName --rg ChosenRGName --code RAIDCode --bs BlockSize
--ss SetSize --nsd-usage dataOnly --sp data
```

Note: For this example command, it is assumed that you are adding data only vdisks to the existing file system. You might have a different use case, so adjust options accordingly.

```
Example values (adjust to meet needs of existing filesystem):
--code 8+2p
--bs 4M
--ss 80%
```

5. Create the vdisk set.

```
mmvdisk vs create --vs ChosenVdiskSetName
```

6. Add the vdisk set to the file system.

```
ssh NodeAlreadyinCluster
mmvdisk fs add --file-system FileSystem --vdisk-set ChosenVdiskSetName
```

FileSystem is the name of the file system that you are adding the storage to.

7. Add the new nodes to performance monitoring.

```
mmchnode --perfmon -N NewNode1,NewNode2
```

8. Fix the compDB.

```
mmaddcompspec default --replace
```

9. Start or restart the GUI on the EMS node.

```
systemctl restart gpfsgui
```

Adding an additional ESS 5000 storage to an existing file system

Before doing these steps, follow the steps in *ESS 5000 Common setup instructions* in *ESS 5000: Quick Deployment Guide*. Make sure that you update the `/etc/hosts` file with the new node names and IP addresses. Copy the updated `/etc/hosts` to all nodes before starting. Stop after creating the network bonds.

1. Fix the SSH keys between new nodes and the current cluster.

```
essrun -N NewNode1,NewNode2,NodesAlreadyinCluster config load -p ibmesscluster
```

2. Add ESS 5000 nodes to the current file system.

```
essrun -N NodeAlreadyinCluster cluster --add-nodes NewNode1,NewNode2 --suffix=Suffix
```

3. Create the recovery group.

```
essrun -N NewNode1,NewNode2 vdisk --name ChosenVdiskSetName --suffix=Suffix --code RAIDCode
--bs BlockSize --size SetSize --extra-vars "--nsd-usage dataOnly --sp data"
```

Note: For this example command, it is assumed that you are adding data-only vdisks to the existing file system. You might have a different use case, so adjust options accordingly. For help and default values, use **essrun vdisk --help**.

4. Add the vdisk set to the file system.

```
ssh NodeAlreadyinCluster
mmvdisk fs add --file-system FileSystem --vdisk-set ChosenVdiskSetName
```

FileSystem is the name of the file system that you are adding the storage to.

5. Add the new nodes to performance monitoring.

```
mmchnode --perfmon -N NewNode1,NewNode2
```

6. Fix the compDB.

```
mmaddcompspec default --replace
```

7. Start or restart the GUI on the EMS node.

```
systemctl restart gpfsgui
```

Adding ESS 3000 to an ESS for Power environment

Before adding ESS 3000 to an existing ESS for Power® environment, the existing ESS system must already be converted to mmvdisk.

Before doing these steps, follow the steps in *ESS 3000 initial setup instructions* in *ESS 3000: Quick Deployment Guide*. Make sure that you update the `/etc/hosts` file with the new node names and IP addresses. Copy the updated `/etc/hosts` to all nodes before starting. Stop after creating the network bonds.

1. Add ESS 3000 nodes to the existing ESS system by running the following command from one of the canister nodes.

```
essaddnode -N NewNode1,NewNode2 --suffix=Suffix --accept-license --no-fw-update --cluster-
node NodeAlreadyinClusterWithSuffix --nodetype ess3k
```

For this example command, it is assumed that the new ESS 3000 system has two canister nodes called NewNode1 and NewNode2.

2. Configure the mmvdisk node class. A unique node class name is required for a new building block.

```
mmvdisk server configure --nc ChosenNodeClassName --recycle one
```

3. Create the recovery group.

```
essrun -N NewNode1,NewNode2 vdisk --name ChosenVdiskSetName --suffix=Suffix --code RAIDCode
--bs BlockSize --size SetSize --extra-vars "--nsd-usage dataOnly --sp data"
```

Note: For this example command, it is assumed that you are adding data-only vdisks to the existing file system. You might have a different use case, so adjust options accordingly. For help and default values, use **essrun vdisk --help**.

4. Define the vdisk set.

```
mmvdisk vs define --vs ChosenVdiskSetName --rg ChosenRGName --code RAIDCode \
--bs BlockSize --ss SetSize --nsd-usage dataOnly --sp data
```

Note: For this example command, it is assumed that you are adding data only vdisks to the existing file system. You might have a different use case, so adjust options accordingly.

```
Example values (adjust to meet needs of existing filesystem):
--code 8+2p
--bs 4M
--ss 80%
```

5. Create the vdisk set.

```
mmvdisk vs create --vs ChosenVdiskSetName
```

6. Add the vdisk set to the file system.

```
ssh NodeAlreadyinCluster
mmvdisk fs add --file-system FileSystem --vdisk-set ChosenVdiskSetName
```

FileSystem is the name of the file system that you are adding the storage to.

7. Add the new nodes to performance monitoring.

```
mmchnode --perfmon -N NewNode1,NewNode2
```

8. Fix the compDB.

```
mmaddcompspec default --replace
```

9. Start or restart the GUI on the EMS node.

```
systemctl restart gpfsgui
```

Adding ESS 5000 to an ESS for Power environment

Before adding ESS 5000 to an existing ESS for Power environment, the existing ESS system must already be converted to mmvdisk.

Before doing these steps, follow the steps in *ESS 5000 Common setup instructions* in *ESS 5000: Quick Deployment Guide*. Make sure that you update the `/etc/hosts` file with the new node names and IP addresses. Copy the updated `/etc/hosts` to all nodes before starting. Stop after creating the network bonds.

1. Add ESS 5000 nodes to the existing ESS system by running the following command from one of the canister nodes.

```
essaddnode -N NewNode1,NewNode2 --suffix=Suffix --accept-license --no-fw-update --cluster-
node NodeAlreadyinClusterWithSuffix --nodetype ess5k
```

For this example command, it is assumed that the new ESS 5000 system has two canister nodes called NewNode1 and NewNode2.

2. Create the recovery group.

```
essrun -N NewNode1,NewNode2 vdisk --name ChosenVdiskSetName --suffix=Suffix --code RAIDCode
--bs BlockSize --size SetSize --extra-vars "--nsd-usage dataOnly --sp data"
```

Note: For this example command, it is assumed that you are adding data-only vdisks to the existing file system. You might have a different use case, so adjust options accordingly. For help and default values, use **essrun vdisk --help**.

3. Add the vdisk set to the file system.

```
ssh NodeAlreadyinCluster
mmvdisk fs add --file-system FileSystem --vdisk-set ChosenVdiskSetName
```

FileSystem is the name of the file system that you are adding the storage to.

4. Add the new nodes to performance monitoring.

```
mmchnode --perfmon -N NewNode1,NewNode2
```

5. Fix the compDB.

```
mmaddcompspec default --replace
```

6. Start or restart the GUI on the EMS node.

```
systemctl restart gpfsgui
```

Cleaning up an existing mmvdisk environment

1. Unmount the file system:

```
mmumount FileSystem -a
```

2. Delete the file system:

```
mmdelfs FileSystem
```

You can also delete the file system by using **mmvdisk** (including vdisk set and recovery group):

```
mmvdisk filesystem delete --file-system FileSystem
```

This command also deletes the vdisk set.

3. List the vdisk sets:

```
mmvdisk vdiskset list
```

4. Delete the vdisk set for the deleted file system:

```
mmvdisk vdiskset delete --vdisk-set VdiskSet
```

This command also deletes the NSDs and data and metadata vdisk.

5. Undefine vdisk sets:

```
mmvdisk vdiskset undefine --vdisk-set VdiskSet
```

6. List the recovery groups:

```
mmvdisk recoverygroup list
```

7. Delete the recovery groups:

```
mmvdisk recoverygroup delete --recovery-group RecoveryGroup
```

8. List the mmvdisk servers:

```
mmvdisk server list
```

9. Unconfigure the servers:

```
mmvdisk server unconfigure --node-class ServerNodeClass
```

10. Delete the node class:

```
mmvdisk nodeclass delete --node-class ServerNodeClass
```

Troubleshooting issues when running the container

If you are facing issues when running the container with **essmgr -r**, you can try these steps.

1. Re-create the bridge.

```
nmcli c del mgmt_bridge
nmcli c del fsp_bridge
nmcli c del bridge-slave-fsp
nmcli c del bridge-slave-mgmt
ifup mgmt
ifup fsp
cd to extracted image directory (.dir)
./essmgr -n
./essmgr -r
```

Debugging deployment issues

When the **essrun** is used, it issues Ansible commands to the target. You can check the following logs to debug the progress of those commands.

- On the canister, run this command: **grep -i ansible-* /var/log/messages**

  Example output:

```
Feb 28 17:21:59 fab3a ansible-command[7300]: Invoked with _raw_params=ofed_info -n warn=True
_uses_shell=False stdin_add_newline=True
strip_empty_ends=True argv=None chdir=None executable=None creates=None removes=None
stdin=None
Feb 28 17:27:01 fab3a ansible-command[4884]: Invoked with _raw_params=/xcatpost/
ess_ofed.ess3k warn=True _uses_shell=False stdin_add_newline=True
strip_empty_ends=True argv=None chdir=None executable=None creates=None removes=None
stdin=None
Feb 28 17:41:43 fab3a ansible-command[44520]: Invoked with _raw_params=/usr/lpp/mmfs/bin/
mmlscluster warn=True _uses_shell=False stdin_add_newline=True
strip_empty_ends=True argv=None chdir=None executable=None creates=None removes=None
stdin=None
Feb 28 17:41:44 fab3a ansible-command[44636]: Invoked with _uses_shell=True
_raw_params=/usr/lpp/mmfs/bin/mmcommon showLocks | grep CCR warn=True stdin_add_newline=True

strip_empty_ends=True argv=None chdir=None executable=None creates=None removes=None
stdin=None
Feb 28 17:46:47 fab3a ansible-command[5133]: Invoked with _raw_params=/usr/lpp/mmfs/bin/
mmbuildgpl warn=True _uses_shell=False
stdin_add_newline=True strip_empty_ends=True argv=None chdir=None executable=None
creates=None removes=None stdin=None
```

- On the container, view the `essansible.json` file.

```
/var/log/ess/6.1.2.x/essansible.json
```

- The default log location for the ESS commands is `/var/log/ess/6.1.2.x/`.

  Use this location to debug details of the various python based commands running under Ansible control.
- To debug OS or package upgrades, you can view the DNF log on respective nodes.

```
/var/log/dnf.log
```

- If you add `-v` to any **essrun** command, you can see the verbose output. This might be helpful, additional debug information.

Customizing file system parameters for ESS

If you want to customize the file system parameters from the defaults, do the following steps from within the container before running the **essrun filesystem** command:

1. Open the `/opt/ibm/ess/tools/ansible/vars.yml` file.

```
vim /opt/ibm/ess/tools/ansible/vars.yml
```

2. Edit these values based on the platform you are using:
   - For ESS Legacy systems:

```
Node_Class_5x: "ess5x_ppc64le_mmvdisk"
Recovery_Group_5x: "ess5x"
Code_5x: "8+2p"
Block_Size_5x: "16M"
Size_5x: "100%"
```

   - For ESS 3000 systems:

```
Node_Class_3k: "ess_x86_64_mmvdisk"
Recovery_Group_3k: "ess3k"
Mount_Point_3k: "/gpfs"
Code_3k: "8+2p"
Block_Size_3k: "4M"
Size_3k: "80%"
```

   - For ESS 5000 systems:

```
Node_Class_5k: "ess5k_ppc64le_mmvdisk"
Recovery_Group_5k: "ess5k"
Code_5k: "8+2p"
Block_Size_5k: "16M"
Size_5k: "100%"
```

   - For ESS 3200 systems:

```
Node_Class_3200: "ess3200_x86_64_mmvdisk"
Recovery_Group_3200: "ess3200"
Code_3200: "8+2p"
Block_Size_3200: "4M"
Size_3200: "80%"
```

   Note: You must use a **`Size`** value of lower than or equal to 80%.

3. Save the file and quit.

```
:wq
```

Alternately, you can also pass the following optional arguments to the Ansible file system through using the **`essrun`** command::

```
optional arguments: -h, --help show this help message and exit
--name FS_NAME Provide filesystem name (Default "fs5k")
--code RaidCode Provide Raid Code (Default "8+2p")
--bs BlockSize Provide Block Size (Default "16M")
--size {n% | n | nK | nM | nG | nT} Provide Vdiskset Size (Default "100%")
```

Turning on syslog redirection

Use these steps to redirect the `/var/log/messages` file on each canister node to the EMS node. Doing this allows you to access logs from a centralized location to debug any issues that might occur.

1. Log in to each canister node.
2. Edit the `/etc/rsyslogd.conf` file to add the IP address of the EMS node at the bottom of the file.

   For example:

```
*.*  @@192.168.20.1:514
```

Where 192.168.20.1 is the IP of the EMS node (bridge IP address).

3. Save the file and restart **rsyslogd**.

```
systemctl restart rsyslog
```

Note: Syslog redirection is automatically setup when essrun config load is executed.

Restoring the backup files and SSH keys

Note:

- For the following command example, it is assumed that the backup location is `/home/backup/6xxx/xcatdb`.

```
/tmp/cems_restore.sh /home/backup/6xxx/xcatdb
cp -a /home/backup/6xxx/hostkeys /etc/xcat/hostkeys
```

Helpful podman commands

- List installed images:

```
podman images
```

- List containers:

```
podman ps -a
```

- Stop container:

```
podman stop  ContainerName
```

- Remove container:

```
podman rm ContainerName
```

- Remove image:

```
podman image rm ContainerName -f
```

- Re-create network bridge:

```
From within the ESS extracted directory run ./essmgr -n
```

- Re-run container:

```
From within the ESS extracted directory run ./essmgr -r
```

- Re-attach to running container:

```
podman attach ContainerName
```

- Start a container:

```
podman start ContainerName
```

- Exit container without stopping it:

```
Ctrl + p then Ctrl + q
```

- Enter container quietly:

```
podman exec -it ContainerName /bin/bash
```

Troubleshooting issues during an essrun update

The system might face issues when the partner node is active during an `essrun` update. This could be caused due to one of the following:

- There is no recovery group being hosted on the node that is being updated.

  Note: This issue is observed during an online update only.

  To resolve this issue, update the node list in offline mode using the following command:

  ```
  essrun -N essio1,essio2 update --offline
  ```

- The admin node name is not the same as the daemon node name in the **`mmlscluster`** command output.

  ```
  [root@ems01 ~]# mmlscluster
  GPFS cluster information
  ========================
    GPFS cluster name:         test01.gpfs.net
    GPFS cluster id:           1284427297386954425
    GPFS UID domain:           test01.gpfs.net
    Remote shell command:      /usr/bin/ssh
    Remote file copy command:  /usr/bin/scp
    Repository type:           CCR
   Node  Daemon node name     IP address  Admin node name      Designation
  ----------------------------------------------------------------------
     1   essio1Highspeed.gpfs.net   10.0.0.101  essio1-hs.gpfs.net   quorum-manager-perfmon
  ```

  To resolve this issue, change either the admin node or the daemon node name using the **`mmchnode`** command.

  ```
  mmchnode --daemon-interface={essio1-hs.gpfs.net} -N essio1-hs
  ```

  Note: This issue is resolved if the admin node name and daemon node name have the same value.

Troubleshooting for Ansible issues

The following table details the cause and solution for Ansible issues.

Table 3. Troubleshooting for Ansible issues and errors

Problem	Cause	Solution
Seeing several timestamps when the **`essrun`** command is run. For example, `Wednesday 08 July 2020 16:39:17 +0000 (0:00:01.808) 0:11:51.631 ********`	The Ansible is skipping some tasks in the target node, which might be an I/O, EMS or Protocol node. The Ansible is skipping this task because these tasks are not applicable to these target nodes. The timestamps allow users to check the start and end time of these tasks.	To remove these timestamps, follow these steps: 1. In the container, go to the `/etc/ansible/ansible.cfg` file. 2. Remove `profile_tasks` from line number 7. 3. Save the file, and quit.
Failure to obtain the enclosure device name with rc=2	The `pemsmod` module is not running correctly in the ESS 3000 canister.	Run the following command to reinstall the platform rpm: `yum reinstall gpfs.ess.platform.ess3k`
Viewing the following error message: `Please be sure you have set the HMC1 port IP and ipmitool is installed in your node.`	The `ipmitool` is not installed, or you are not using HMC1 port as the default FSP/BMC interface.	Connect the FSP/BMC network in the HMC1 port in the back of your P9 nodes.

Table 3. Troubleshooting for Ansible issues and errors (continued)

Problem	Cause	Solution

Viewing the following error message: `Failed to download metadata for repo <Repository name>`.	`http` is not running in the container.	To resolve this error, follow these steps:
		1. Check whether the `http` is running in the container by using the following command:
		`ps aux | grep http`
		2. If there is no output, then run the following command:
		`httpd`

Viewing the following error message: `Failed to connect to the host via ssh.`	There was a timeout during the Ansible execution. There are many reasons for this. The most common are: • Kernel changed while the update was running, and reboot took more than 20 minutes. • The kernel crashed, and the connection was lost. This is more likely in ESS 3000, but might happen in ESS 5000 also. • The `var/crash` contains a recent crash file when you ssh to the node after the Ansible fails.	To resolve this error, follow these steps: 1. From the container, save the `/var/log/ansible.json` file. 2. Run the **`gpfs.snap`** command from the failed node, and save the output for reference. 3. Contact IBM Support for further investigation.
The following message is displayed: `Failure to obtain interface details on node.`	The `/etc/hosts` file does not contain valid entries.	Confirm that the `/etc/hosts` file contains the correct entries on each node.

Execution might hang after the following step when you run the **`essrun -N nodelist cluster --sufix=SUFFIX`** command: `TASK [cluster : install | Initialize gpfs profile]`	Check that the FQDNs (Full Qualified Domain Names) are part of the `know_hosts` entries across all the nodes.	Run the following commands from each node in the cluster:
		`ssh ems1.localdomain date`
		`ssh essio1.localdomain date`
		`ssh essio2.localdomain date`

The following message is displayed while running the `./essmkyml` script: `DNS domain is not configured in the system`.	The EMS node required a domain name to run `esskyml`.	Run the following command to assign a domain in the EMS hostname: `hostnamectl set-hostname <EMS hostnameame>.<domain name>` The same information must be added to the `/etc/resolv.conf` file.

Table 3. Troubleshooting for Ansible issues and errors (continued)

Problem	Cause	Solution
The connection is lost after you run the following script from the extracted build directory: **./essmgr -n**	The same IP is used for the enP1p8s0f0 interface and MGMT_BRIDGE_IP.	Use another IP for MGMT_BRIDGE_IP.
The current FSP password is not working, or the following error message is displayed: `Please use correct BMC password for <node name> node.`	The password is not set properly.	Change the FSP password by using the following command: `ipmitool user set password 1 <New Password>`
The SSR port is not giving any IP.	The DHCP service is not correctly running in the port, and IP is not set properly.	Set the IP address manually as explained in the *Assigning the management IP address* section in the *IBM Elastic Storage System 5000: Hardware Installation Guide*.
Cannot monitor your installation.	During the I/O and Protocol nodes' initial deployment you can ssh to the node IP, which prompts you to the anaconda installer.	To resolve this error, follow these steps: 1. Log in to the node by using ssh. 2. Run the following script to detach from the installation screen: `tmux attach` 3. Press the key combo `Ctrl`+b, release them, and then press d.
Cannot create a file system after the following message is displayed: `Can't create additional CES fileystem because there are no mmvdisk servers configured. Please create an ESS file system before creating a CES file system.`	CES file system requires an ESS file system to be created before you can create a CES file system.	To resolve this error, follow these steps: 1. Remove the `/tmp/vslist` file from the container. 2. Create an ESS file system, and then a CES file system.

Table 3. Troubleshooting for Ansible issues and errors (continued)

Problem	Cause	Solution

If you see the following messages in essrun update:
`- ATTENTION:`
`- VerbsRDMA is enabled in the cluster (For more info check: mmlsconfig verbsRDMA).`
`- Unable to run update in offline mode because there are 6 active nodes in cluster.`
`- There should not be any node active in the cluster for 6.1.2.0 update if verbsRDMA is enabled in the cluster.`
`- Please manually shutdown all nodes in cluster using 'mmshutdown -a'.`
	Cannot update nodes to ESS 6.1.2.0 if the value `verbRDMA` is enabled in any node in the cluster due to new MOFED version.	If you chose to disable verbsRDMA in the cluster for performing an online update, run the following command in the EMS node:
		`mmchconfig verbsRDMA=disable -N all`

An Ansible task (especially during essrun update) takes a lot of time, which is more than 20 minutes.	Some tasks run for more than 10 minutes, especially during the OFED installation or upgradation phase.
	This can be a real issue where the node crashes and the Ansible execution remains idle until it times out.
		To check whether your node is in the idle or failed state, issue the following commands on the node from another SSH session:
		1. Check whether one or two processes are running on the Ansible execution PID by using the following command:
		`ps aux | grep -i ansible`
		2. Check whether the latest Ansible tasks are executed by using the following command:
		`grep ansible- /var/log/messages`
		3. Check for a crash folder by using the following command:
		`ls -lrth /var/crash/`
		If there is a crash folder, then contact the IBM support.
		If there is no crash folder and Ansible PID is running, and Ansible is executing, then a long-run task is performed.

Chapter 5. Debugging yum update issues from the container

During the EMS or I/O node update from the container, specifically in the `yum update` task, you might encounter some issues. Use the following information to resolve these issues.

When running the update playbook with the `essrun` command, if there is a dependency issue, you might encounter a failure. This could be due to the following reasons:

Migrating from Legacy xCAT deployment (5.3.x.x) to the containerized method (6.x.x.x)

The legacy method (xCAT) was used to deploy POWER8 nodes until ESS 3000 was introduced. For ESS 3000, ESS 5000, and ESS 3200, the only method of deployment is by using the new containerized method. Starting from ESS 6.1.0.0, POWER8 nodes can also move to this container approach, provided that the IBM Spectrum Scale version is 5.1.x.x or higher. If the conversion is done, customers can use the POWER8 or POWER9 EMS to run the container.

Note: This migration option is only applicable to POWER8 environments, and must be done for users who want IBM Spectrum Scale 5.1.x.x or higher.

The Quick Deployment Guide contains an appendix with the best practices and high-level steps for migrating to the containerized deployments. For more information, see the *Tips for migrating from xCAT based (5.3.7.x)* section in the *Elastic Storage Server: Quick Deployment Guide*.

Issue caused due to manually installed packages from the full Red Hat Server ISO

Customers might manually install packages from the full Red Hat Server ISO that can cause issues with future updates. The system could display the following message:

```
essrun -N Node update --offline

---> Package java-1.8.0-openjdk-headless.ppc64le 1:1.8.0.222.b03-1.el7 will be updated
---> Processing Dependency: java-1.8.0-openjdk-headless(ppc-64) = 1:1.8.0.222.b03-1.el7 for
package: 1:java-1.8.0-openjdk-1.8.0.222.b03-1.el7.ppc64le
---> Package kernel.ppc64le 0:3.10.0-1062.21.1.el7 will be erased
---> Package kernel-devel.ppc64le 0:3.10.0-1062.21.1.el7 will be erased
--> Finished Dependency Resolution
You could try using --skip-broken to work around the problem
You could try running: rpm -Va --nofiles --nodigest
('[ERROR]', '2021-03-18T00:57:43.612088', 'Update failed in ess3k')
```

Follow these steps to resolve this issue:

1. Leave the container and log in to the node in question.
2. Run the **yum update** command to display any RPM related issues.

   Note: Do not use the `-y` option or start the actual update.
3. Resolve any problems that occur. Usually, this involves removing any problematic RPMs that are found. Considering the preceding command output as an example, use the following command:

   ```
   yum remove java-1.8.0-openjdk
   ```
4. Rerun the **yum update** command to check that there are no RPM related issues.

   Note: Do not use the `-y` option or start the actual update.
5. Return to the container and re-try the update.

Chapter 6. GUI Issues

When troubleshooting GUI issues, it is recommended to view the logs that are located under `/var/log/cnlog/mgtsrv`. By default, the GUI is installed on the EMS node. It is possible that the customer installed it in another node. In such cases, the GUI logs are stored in the node where the GUI is installed.

The following logs can be viewed to troubleshoot the GUI issues:

mgtsrv-system-log

Logs everything that runs in background processes such as refresh tasks. This is the most important log for GUI.

mgtsrv-trace-log

Logs everything that is directly triggered by the GUI user. For example, starting an action, clicking a button, executing a GUI CLI command, etc.

wlp-messages.log

This log covers the underlying Websphere Liberty. The log is mostly relevant during the startup phase.

gpfsgui_trc.log

Logs the issues related to incoming requests from the browser. Users must check this log if the GUI displays the error message:

```
Server was unable to process request.
```

Issue with loading GUI

If there are problems in loading the GUI, you can reconfigure the GUI to see if that resolves the problem.

Follow these steps to reconfigure the GUI:

1. Run the following command to force the GUI to launch the wizard after the next login:

   ```
   /usr/lpp/mmfs/gui/cli/debug enablewizard
   systemctl restart gpfsgui
   ```

2. Run the following command to force the GUI to no longer display the wizard after login:

   ```
   /usr/lpp/mmfs/gui/cli/debug disablewizard
   systemctl restart gpfsgui
   ```

3. If the problem persists, reinstall the GUI RPM that can be found on the EMS node using the following command:

   ```
   yum -Uvh /opt/ibm/gss/install/rhel7/<arch>/gui/gpfs.gui*
   ```

4. If there is a possibility that the GUI database has become corrupt or has inconsistencies that are preventing the GUI from loading properly, take the following steps.

   CAUTION: This should be done as a last resort since the GUI configuration settings will be lost after you execute the following steps:

   a. Stop the GUI service.

   ```
   systemctl stop gpfsgui
   ```

   b. Drop the GUI schema from the postgres database.

```
psql postgres postgres -c "DROP SCHEMA FSCC CASCADE"
```

c. Start the GUI service.

```
systemctl start gpfsgui
```

Chapter 7. Recovery Group Issues

The following sections describe the recovery group issues and their solutions for the different ESS platforms.

Recovery group issues for shared recovery groups

An ESS 3000 or ESS 3200 recovery group is called a `shared recovery group` because the enclosure disks are shared by both the canister servers in the building block. These building block contains two canister servers and an NVMe enclosure, and configures as a single recovery group that is simultaneously active on both canister servers.

The single shared recovery group structure is necessitated because the ESS system can have as few as 12 disks, which is the smallest number of disks a recovery group can contain. Having 12 disks allows for one equivalent spare and 11-wide 8+3P RAID codes.

The following example displays a canister server pair of a representative ESS building block that is using the individual building block node class ESSNC:

```
 # mmvdisk server list --node-class ESSNC
 node
number  server                          active   remarks
------  --------------------------------  -------  -------
     3  canister1.gpfs.net                yes      serving ESSRG: LG002, LG004
     4  canister2.gpfs.net                yes      serving ESSRG: root, LG001, LG003
```

For these ESS systems, each server is simultaneously serving the same single recovery group, ESSRG. The server workload within the building block is balanced by subdividing the single shared recovery group into the following log groups: LG001, LG002, LG003, LG004, and the lightweight root or master log group. The non-root log groups are called `user log groups`. Only the user log groups contain the file system vdisk NSDs.

All recovery groups in a cluster can be listed by using the **`mmvdisk recoverygroup list`** command:

```
# mmvdisk recoverygroup list
                                                          needs    user
recovery group  active   current or master server         service  vdisks  remarks
--------------  -------  -------------------------------  -------  ------  -------
ESSRG           yes      canister2.gpfs.net               no           16
ESSRG1          yes      server1.gpfs.net                 no            8
ESSRG2          yes      server2.gpfs.net                 no            8
```

The `needs service` column in all the IBM Spectrum Scale RAID commands is narrowly defined to mean whether a disk in the recovery group is called out for replacement. The **`mmvdisk recoverygroup list --not-ok`** command can be used to show other recovery group issues, including those involving log groups or servers:

```
# mmvdisk recoverygroup list --not-ok
recovery group  remarks
--------------  -------
ESSRG       server canister2.gpfs.net 'down'
#
```

If one server of an ESS shared recovery group is down, all the log groups must failover to the remaining server:

```
 # mmvdisk recoverygroup list --server --recovery-group ESSRG
 node
number  server                          active   remarks
------  --------------------------------  -------  -------
     3  canister1.gpfs.net                yes      serving ESSRG: root, LG001, LG002, LG003,
LG004
     4  canister2.gpfs.net                no       configured
```

When the down server is brought back up, the Recovery Group Configuration Manager (RGCM) process that is running on the cluster manager node assigns it two of the user log groups. The two user log groups are used to rebalance the recovery group server workload. For more information, see “Server failover for shared recovery groups” on page 43.

Other than cases where a failover occurs or while servers are rejoining a recovery group, RGCM must always keep two user log groups on each server. In the unlikely event that both servers are active but each server does not have two user log groups, you can shut down one of the servers and restart it. Shutting down the servers and restarting them causes the RGCM to redistribute the user log groups to the servers.

For example, consider a situation where the following allocation of log groups lasts for five or more minutes:

```
 # mmvdisk recoverygroup list --server --recovery-group ESSRG
 node
number  server                          active   remarks
------  --------------------------------  -------  -------
     3  canister1.gpfs.net                yes      serving ESSRG: root, LG001, LG002, LG003
     4  canister2.gpfs.net                yes      serving ESSRG: LG004
```

In such cases, shutting down `canister2` and starting it back up restores the log group workload balance in the building block within five or fewer minutes:

```
# mmshutdown -N canister2.gpfs.net
# mmstartup -N canister2.gpfs.net
# sleep 300
# mmvdisk recoverygroup list --server --recovery-group ESSRG
 node
number  server                          active   remarks
------  --------------------------------  -------  -------
     3  canister1.gpfs.net                yes      serving ESSRG: root, LG002, LG003
     4  canister2.gpfs.net                yes      serving ESSRG: LG001, LG003
```

Recovery group issues for paired recovery groups

The recovery groups in an ESS 5000 or ESS Legacy system are called `paired recovery groups` and always come in pairs. These pairs divide ownership of the enclosure disks in half, with one recovery group primary to each of the two servers in the ESS building block. The ESS building blocks always contain a minimum of 24 disks, which can therefore be divided into two paired recovery groups of at least 12 disks

Use the **`mmvdisk recoverygroup list`** command to check which recovery groups are available:

```
mmvdisk recoverygroup list
```

The command gives an output similar to the following:

```
recovery group  active   current or master server          service  vdisks  remarks
--------------  -------  --------------------------------  -------  ------  -------
rg_rchgss1-hs    yes      rchgss1-hs.gpfs.rchland.ibm.com   no            5
rg_rchgss2-hs    yes      rchgss2-hs.gpfs.rchland.ibm.com   no            5
```

Each of the recovery groups must be served by its own server. If the server is unavailable due to maintenance or other issues, the recovery group must be served by an available server. After a failure or maintenance event, when the recovery group’s primary server becomes active again, it must automatically begin serving its recovery group. You will find the following information under the `/var/adm/ras/mmfs.log.latest` file under in the recovery group server:

- Now serving recovery group rg_rchgss1-hs.
- Reason for takeover of rg_rchgss1-hs: 'primary server became ready'.

If the recovery group is not being served by its respective server, examine the gpfs log on that server for errors that might prevent the server from serving the recovery group. If there are no issues, you can

manually activate the recovery group. For example, to allow `rchgss1-hs.gpfs.rchland.ibm.com` to serve the `rg_rchgss1-hs` RG, execute:

```
mmvdisk recoverygroup change --recovery-group rg_rchgss1-hs --active rchgss1-
hs.gpfs.rchland.ibm.com
```

For more information, see "Server failover for paired recovery groups" on page 44

Manually starting GPFS disks in response to recovery group issues

In certain situations, if an ESS server node experiences a pdisk failure, the GPFS disks might be marked down, and does not automatically start. This can prevent the recovery group from becoming active. For more information on troubleshooting disk problems, see the *Disk Issues* section in the IBM Spectrum Scale documentation.

Before troubleshooting further, ensure that GPFS is in the `active` state for the node in question by running the **`mmgetstate`** command:

```
mmgetstate -a
```

The command gives an output similar to the following:

```
 Node number  Node name        GPFS state
------------------------------------------
       1      rchgss1-hs       active
       2      rchgss2-hs       active
       3      rchems1          active
```

Execute the **`mmlsdisk`** command to check the status of the disks. The `-e` option will only display disks with errors.

```
mmlsdisk gpfs0 -e
```

The command gives an output similar to the following:

```
disk                         driver   sector  failure     holds    holds                            storage
name                         type       size  group       metadata data  status        availability pool
---------------------------- -------- ------  ----------  -------- ----- ------------- ------------ ------------
rg_rchgss1_hs_MetaData_1M_3W_1 nsd         512   30          Yes      No    to be emptied  up           system
```

Attention: Due to an earlier configuration change the file system might contain data that is at risk of being lost.

In the previous example, the disk is in the suspended state, hence the `to be emptied` status. Other disks might be in the non-ready state or might be unavailable, so this prevents the disks from being used by GPFS or ESS.

```
disk                                driver   sector  failure   holds     holds                                storage
name                                type     size    group     metadata  data  status    availability  disk id pool   remarks
----------------------------------- -------- ------  --------- --------  ----- --------- ------------ --------------  --------
rg_rchgss1_hs_MetaData_1M_3W_1      nsd      512     30        Yes       No    ready          up           1           system
rg_rchgss1_hs_Data_16M_2p_1         nsd      512     30        No        Yes   ready          up           2           data desc
rg_rchgss2_hs_MetaData_1M_3W_1      nsd      512     30        Yes       No    ready          up           3           system
desc
rg_rchgss2_hs_Data_16M_2p_1         nsd      512     30        No        Yes   ready          up           4           data desc
```

You can try to manually start the disks by running the **`mmchdisk`** command.

```
mmchdisk gpfs0 start -d rg_rchgss1_hs_MetaData_1M_3W_1
mmnsddiscover:  Attempting to rediscover the disks.  This may take a while ...
mmnsddiscover:  Finished.
rchgss1-hs.gpfs.rchland.ibm.com:  Rediscovered nsd server access to
rg_rchgss1_hs_MetaData_1M_3W_1
```

If multiple disks are down, you can run the command:

```
mmchdisk gpfs0 start -a
```

Note: Depending on the number of disks that are down and their size, the **mmnsddiscover** command might take a while to complete.

Chapter 8. Maintenance procedures

Very large disk systems, with thousands or tens of thousands of disks and servers, will likely experience a variety of failures during normal operation.

To maintain system productivity the vast majority of these failures must be handled automatically without loss of data, without temporary loss of access to the data, and with minimal impact on the performance of the system. Some failures require human intervention, such as replacing failed components with spare parts or correcting faults that cannot be corrected by automated processes.

You can also use the ESS GUI to perform various maintenance tasks. The ESS GUI lists various maintenance-related events in its event log in the **Monitoring > Events** page. You can set up email alerts to get notified when such events are reported in the system. You can resolve these events or contact the IBM Support Center for help as needed. The ESS GUI includes various maintenance procedures to guide you through the fix process.

Updating the firmware for host adapters, enclosures, and drives

After you create a GPFS cluster, install the most current firmware for host adapters, enclosures, and drives only if instructed to do so by IBM support.

You can update the firmware either manually or with the help of directed maintenance procedures (DMP) that are available in the GUI. The ESS GUI lists events in its event log in the **Monitoring** > **Events** page if the host adapter, enclosure, or drive firmware is not up-to-date, compared to the firmware packages on the servers that are currently available. Select **Action** > **Run Fix Procedure** for the firmware-related event to start the corresponding DMP in the GUI. For more information on the available DMPs, see *Directed maintenance procedures* in the *Elastic Storage System: Problem Determination Guide*.

The most current firmware is packaged as the `gpfs.ess.firmware` RPM. You can find the most current firmware on Fix Central.

Follow these steps to perform the update:

1. Sign in with your IBM ID and password.
2. On the **Find product** tab:
   a. In the **Product selector** field, type one of the following based on your platform, and click the right arrow:
      - ESS Legacy: IBM Spectrum Scale RAID
      - ESS 3000, ESS 3200, or ESS 5000: `IBM Elastic Storage System(ESS)`
   b. On the **Installed Version** menu, select: `6.0.1`
   c. Based on your platform select one of the following from the **Platform** menu:
      - ESS Legacy: Linux 64-bit, pSeries.
      - ESS 3000 and ESS 3200: Linux 64-bit, x86_64.
      - ESS 5000: Linux Power PC 64, Little Endian.
   d. Click **Continue**.
3. On the **Select fixes** page, select the most current fix pack.
4. Click **Continue**.
5. On the **Download options** page, select your preferred downloading method. Make sure the check box to the left of **Include prerequisites and co-requisite fixes** has a check mark in it. You can select the ones you need later.
6. Click **Continue** to go to the **Continue** page and download the fix pack files.

The following RPMs needs to be installed on all ESS server nodes based on your platform:

- For ESS Legacy and ESS 5000: `gpfs.gss.firmware`
- for ESS 3000 and ESS 3200: `gpfs.ess.firmware`

It contains the most current updates of the following types of supported firmware for an ESS configuration:

- Host adapter firmware
- Enclosure firmware
- Drive firmware
- Firmware loading tools

For command syntax and examples, see the **mmchfirmware command** in *IBM Spectrum Scale RAID: Administration*.

Enclosure firmware troubleshooting for ESS 3000

This section describes the common issues that the enclosure firmware encounters and how to resolve them.

BIOS update requires power cycle of the canister

Following a BIOS update, in order for the new BIOS version to take effect, the IBM Elastic Storage System 3000 canister needs to be power cycled. A simple restart of the operating system is not enough. A canister power cycle can be accomplished by physically reseating the canister module, or by the following these steps:

1. Run the following command to identify the sg device name associated with the enclosure:

   ```
   lsscsi -g |grep 5141-AF8
   ```

   The last column of the output is the sg device as shown:

   ```
   Example:
   [root@fab3a ~]# lsscsi -g |grep 5141-AF8
   [13:0:0:0]   enclosu IBM-ESS  5141-AF8         1111  -          /dev/sg5
   [root@fab3a ~]#
   ```

2. Identify which canister you want to power cycle, top or bottom.

   Note: The bottom canister is often identified as a and the top canister is identified as b.

3. Run the following command by using the sg device to perform a low-level power cycle of the canister to reset the canister:

   ```
   [root@fab3a ~]# /usr/lpp/mmfs/bin/tsplatformctl -d /dev/sg5 -r --canister=bottom --i-know-
   what-i-am-doing
   ```

   CAUTION: The parameter **--i-know-what-i-am-doing** is a safety mechanism to make sure that the user is aware that the activity must be taken seriously as it causes a canister to power cycle. Ensure that you are power cycling the correct canister.

   The system gives an output similar to the following:

   ```
   tsplatformctl:log:0:[I] Resetting bottom canister:
   Hex Dump for comp reset cmd to send
   0000  3B 11 C2 07 10 00 00 00  04 76                    |;........v      |
   [root@fab3a ~]#
   ```

BMC or BIOS update failures due to IPMI BMC USB connection issues

IBM Elastic Storage System 3000 enclosure firmware update for components BMC and BIOS uses Intelligent Platform Management Interface (IPMI) through low-level BMC USB communication path.

Check whether the enclosure firmware update fails for those components and the `/var/log/ess/platform/ess3kfwLoader.log` file has the following messages:

```
../../Common/main.c-4785 LIBIPMI_Create_IPMI20_Session using USB
../../Common/main.c-4795 Enabling USB
-----------------------------------------------
Open IPMI Drivers
-----------------------------------------------
Un-Loading ipmi_devintf
Un-Loading ipmi_si
Un-Loading ipmi_msghandler
-----------------------------------------------
Open IPMI Drivers
-----------------------------------------------
Loading Open IPMI Driver:ipmi_devintf
Parsing.RebootFirmware:1,FlashSelected:0
Loading Open IPMI Driver:ipmi_si
Loading Open IPMI Driver:ipmi_msghandler

Enable USB failed retry cnt:4 BMC has been reset, please waiting for 5 minutes to be restarted
again
Wed Jun 17 07:09:31 2020

20 01 01 73 02 bf 02 30 00 d1 03 00 00 00 00
Wed Jun 17 07:15:21 2020
```

In such cases, follow these steps:

1. Run the following script to load the `usb_storage` module:

   ```
   insmod /root/usb-storage.ko.xz
   ```

2. Run the following script to load the `uas kernel` module:

   ```
   modprobe uas
   ```

3. Retry the enclosure firmware update for BMC and BIOS.
4. When the update is complete, use the following commands to unload the `usb modules` to restore the previous state:

   ```
   rmmod uas
   rmmod usb_storage
   ```

You must also check whether the CD-ROM setting of the BMC USB port is enabled for the IPMI communication through BMC USB connection. Follow these steps to check the CD-ROM settings:

1. Run the following command to read the current CD-ROM disable value:

   ```
    ipmitool raw 0x32 0xca 0
   ```

   Note: On running this command, you could get one of the following outputs:

   00
   Indicates that the USB CD-ROM capability is enabled. No further action is required.

   01
   Indicates that the USB CD-ROM capability is not enabled. Continue to step 2.

2. Run the following command to set the `USB CDROM disable` bit to 0 to prevent disablement:

   ```
   ipmitool raw 0x32 0xcb 0
   ```

   Note: Wait for 30 seconds to allow the command to complete.

3. After the command run is completed, run the following command to verify whether the value to validate it is now `00`:

   ```
   ipmitool raw 0x32 0xca 0
   ```

4. Retry the enclosure firmware update for BMC and BIOS.

Disk diagnosis

For information about disk hospital, see *Disk hospital* in *IBM Spectrum Scale RAID: Administration*.

When an individual disk I/O operation (read or write) encounters an error, IBM Spectrum Scale RAID completes the NSD client request by reconstructing the data (for a read) or by marking the unwritten data as stale and relying on successfully written parity or replica strips (for a write), and starts the disk hospital to diagnose the disk. While the disk hospital is diagnosing, the affected disk will not be used for serving NSD client requests.

Similarly, if an I/O operation does not complete in a reasonable time period, it is timed out, and the client request is treated just like an I/O error. Again, the disk hospital will diagnose what went wrong. If the timed-out operation is a disk write, the disk remains temporarily unusable until a pending timed-out write (PTOW) completes.

The disk hospital then determines the exact nature of the problem. If the cause of the error was an actual media error on the disk, the disk hospital marks the offending area on disk as temporarily unusable, and overwrites it with the reconstructed data. This cures the media error on a typical HDD by relocating the data to spare sectors reserved within that HDD.

If the disk reports that it can no longer write data, the disk is marked as `readonly`. This can happen when no spare sectors are available for relocating in HDDs, or the flash memory write endurance in SSDs was reached. Similarly, if a disk reports that it cannot function at all, for example not spin up, the disk hospital marks the disk as `dead`.

The disk hospital also maintains various forms of *disk badness*, which measure accumulated errors from the disk, and of *relative performance*, which compare the performance of this disk to other disks in the same declustered array. If the badness level is high, the disk can be marked `dead`. For less severe cases, the disk can be marked `failing`.

Finally, the IBM Spectrum Scale RAID server might lose communication with a disk. This can either be caused by an actual failure of an individual disk, or by a fault in the disk interconnect network. In this case, the disk is marked as `missing`. If the relative performance of a disk falls below a particular threshold, the disk is declared as `slow` in the pdisk state, and the disk is prepared for replacement. To check the current value, run the **`mmlsconfig nsdRAIDDiskPerformanceMinLimitPct`** command.

If a disk would have to be marked `dead`, `missing`, or `readonly`, and the problem affects individual disks only (not a large set of disks), the disk hospital tries to recover the disk. If the disk reports that it is not started, the disk hospital attempts to start the disk. If nothing else helps, the disk hospital power-cycles the disk (assuming the JBOD hardware supports that), and then waits for the disk to return online.

Before actually reporting an individual disk as `missing`, the disk hospital starts a search for that disk by polling all disk interfaces to locate the disk. Only after that fast poll fails is the disk actually declared `missing`.

If a large set of disks has faults, the IBM Spectrum Scale RAID server can continue to serve read and write requests, provided that the number of failed disks does not exceed the fault tolerance of either the RAID code for the vdisk or the IBM Spectrum Scale RAID vdisk configuration data. When any disk fails, the server begins rebuilding its data onto spare space. If the failure is not considered *critical*, rebuilding is throttled when user workload is present. This ensures that the performance impact to user workload is minimal. A failure might be considered critical if a vdisk has no remaining redundancy information, for example three disk faults for 4-way replication and 8 + 3p or two disk faults for 3-way replication and 8 + 2p. During a critical failure, critical rebuilding will run as fast as possible because the vdisk is in imminent danger of data loss, even if that impacts the user workload. Because the data is declustered, or spread out over many disks, and all disks in the declustered array participate in rebuilding, a vdisk will remain in critical rebuild only for short periods of time (several minutes for a typical system). A double or triple fault is extremely rare, so the performance impact of critical rebuild is minimized.

In a multiple fault scenario, the server might not have enough disks to fulfill a request. More specifically, such a scenario occurs if the number of unavailable disks exceeds the fault tolerance of the RAID code. If some of the disks are only temporarily unavailable, and are expected back online soon, the server will

stall the client I/O and wait for the disk to return to service. Disks can be temporarily unavailable for any of the following reasons:

- The disk hospital is diagnosing an I/O error.
- A timed-out write operation is pending.
- A user intentionally `suspended` the disk, perhaps it is on a carrier with another failed disk that has been removed for service.

If too many disks become unavailable for the primary server to proceed, it will fail over. In other words, the whole recovery group is moved to the backup server. If the disks are not reachable from the backup server either, then all vdisks in that recovery group become unavailable until connectivity is restored.

A vdisk will suffer data loss when the number of permanently failed disks exceeds the vdisk fault tolerance. This data loss is reported to NSD clients when the data is accessed.

Background tasks

While IBM Spectrum Scale RAID primarily performs NSD client read and write operations in the foreground, it also performs several long-running maintenance tasks in the background, which are referred to as *background tasks*.

The background task that is currently in progress for each declustered array is reported in the long-form output of the **`mmvdisk recoverygroup list --dammvdisk recoverygroup list --da`** command. Table 4 on page 43 describes the long-running background tasks.

Table 4. Background tasks

Task	Description
`repair-RGD/VCD`	Repairing the internal recovery group data and vdisk configuration data from the failed disk onto the other disks in the declustered array.
`rebuild-critical`	Rebuilding virtual tracks that cannot tolerate any more disk failures.
`rebuild-1r`	Rebuilding virtual tracks that can tolerate only one more disk failure.
`rebuild-2r`	Rebuilding virtual tracks that can tolerate two more disk failures.
`rebuild-offline`	Rebuilding virtual tracks where failures exceeded the fault tolerance.
`rebalance`	Rebalancing the spare space in the declustered array for either a `missing` pdisk that was discovered again, or a new pdisk that was added to an existing array.
`scrub`	Scrubbing vdisks to detect any silent disk corruption or latent sector errors by reading the entire virtual track, performing checksum verification, and performing consistency checks of the data and its redundancy information. Any correctable errors found are fixed.

Server failover

This section contains information about how to overcome server failover based on your platform.

Server failover for shared recovery groups

Each of the two servers in a shared recovery group is capable of serving the entire recovery group if the other canister is not available. When only one canister server is available, all of the log groups are served by the remaining server. When an unavailable server becomes active again, it takes back two of the user log groups from the other server.

During a normal operation both the ESS servers are active, and each serves two of the user log groups:

```
# mmvdisk recoverygroup list --recovery-group ESS3000RG --server
 node
number  server                          active   remarks
------  ------------------------------  -------  -------
     3  canister1.gpfs.net              yes      serving ESS3000RG: LG001, LG003
     4  canister2.gpfs.net              yes      serving ESS3000RG: root, LG002, LG004
```

If canister2 fails or is shutdown, its two user log groups transparently switch to being served by canister1. The root log group also fails over if it is located on canister2. Application workload to the affected log groups is paused while the log groups are recovered on canister1, but are not otherwise affected.

When an ESS recovery group is operating with server failover, all the log groups are located on one server, and the recovery group is reported as `not OK`:

```
# mmvdisk recoverygroup list --recovery-group ESS3000RG --server
 node
number  server                          active   remarks
------  ------------------------------  -------  -------
     3  canister1.gpfs.net              yes      serving ESS3000RG: root, LG001, LG002,
LG003, LG004
     4  canister2.gpfs.net              no       configured
# mmvdisk rg list --not-ok
recovery group  remarks
--------------  -------
ESS3000RG       server ccanister2.gpfs.net 'down'
```

Server failover for paired recovery groups

If the primary IBM Spectrum Scale RAID server loses connectivity to a sufficient number of disks, the recovery group attempts to fail over to the backup server.

If the backup server is also unable to connect, the recovery group becomes unavailable until connectivity is restored. If the backup server had taken over, it relinquishes the recovery group to the primary server when it becomes available again.

Data checksums

IBM Spectrum Scale RAID stores checksums of the data and redundancy information on all disks for each vdisk. Whenever data is read from disk or received from an NSD client, checksums are verified. If the checksum verification on a data transfer to or from an NSD client fails, the data is retransmitted. If the checksum verification fails for data read from disk, the error is treated similarly to a media error:

- The data is reconstructed from redundant data on other disks.
- The data on disk is rewritten with reconstructed good data.
- The disk badness is adjusted to reflect the silent read error.

Disk replacement for ESS

You can use the ESS GUI for detecting failed disks and for disk replacement.

When one disk fails, the system rebuilds the data that was on the failed disk onto spare space and continues to operate normally. However, the performance is slightly reduced because the same workload is shared among fewer disks. With the default setting of two spare disks for each large declustered array, failure of a single disk would typically not be a sufficient reason for maintenance.

The system might continue to operate in the presence of several disk failures even if there is no remaining spare space as long as the number of failures is less than the configured vdisk fault tolerance. The next disk failure would make the system unable to maintain the redundancy that the user requested during vdisk creation. A service request is sent to a maintenance management application that requests replacement of the failed disks, and specifies the disk FRU numbers and locations.

Call home for disk maintenance is requested when the number of failed disks in a declustered array reaches the disk replacement threshold. By default, the replace threshold is one if the number of data

spares is zero or one, or two if the number of spares is two or greater. The maximum value is one more than the number of spares.

You can replace the disk either manually or with the help of directed maintenance procedures (DMP) that are available in the GUI. The ESS GUI lists events in its event log in the **Monitoring** > **Events** page if a disk failure is reported in the system. Select the *gnr_pdisk_replaceable* event from the list of events, and then select **Action** > **Run Fix Procedure** from the menu to start the `replace disk DMP` in the GUI. For more information, see "Replace disks" on page 58.

Disk maintenance is done by using the **`mmvdisk pdisk replace`** command with the `--prepare` option for ESS recovery groups, which:

- Suspends any functioning disks on the carrier if the multi-disk carrier is shared with the disk that is being replaced.
- If possible, powers down the disk to be replaced or all of the disks on that carrier.
- Turns on indicators on the disk enclosure and disk or carrier to help locate and identify the disk that needs to be replaced.
- If necessary, unlocks the carrier for disk replacement.

After the disk is replaced and the carrier is reinserted, the **`mmvdisk pdisk replace`** command powers on the replacement disk, and integrates it into the ESS recovery group.

Commandless disk replacement

Commandless disk replacement automates the process of replacing a failed or bad drive with a new drive.

The commandless disk replacement feature helps in automating the process of replacing a failed drive with a new drive. The disk hospital begins moving data off that drive whenever the drive fails in preparation for replacement. When data is drained off the drive, or if the drive is undrainable or dead, the disk hospital marks the drive as `replaceable`. Then, the commandless disk replacement component runs the `prepare for replacement` operation on the drive. As a part of the `prepare` operation, the LED on the drive for replacement is turned on, indicating that it is ready for removal.

If the `prepare` operation fails, then use the **`mmvdisk pdisk replace –-prepare`** command to prepare the disk for replacement, and proceed with manual disk replacement For more information, see Disk Replacement.

Identify the bad drive that needs to be replaced and verify that the `replace` LED is turned on. Next, remove the bad drive, and insert a new drive. The new drive is identified by commandless disk replacement within minutes, as defined by the configuration keyword `nsdRAIDDiskDiscoveryInterval`. Commandless disk replacement then runs the `replace` operation. This operation makes the new drive ready for use. If it runs successfully without any errors, then no further action is required.

If the `replace` operation on a new drive fails for any reason, error messages, indicating the failure, are logged in the `mmfs.log` file. You can then replace the drive manually by following the manual disk replacement procedure. For more information, see Disk Replacement.

Replacing bad drives with new drives by using commandless disk replacement

The following section describes the procedure to replace bad drives with new drives by using commandless disk replacement.

You can replace bad drives with good drives by using the following commandless disk replacement process.

1. To enable commandless disk replacement, use the **`mmchconfig enableAutomaticDiskReplacement=yes -i`** command.

   Note: This step needs to be performed only once.
2. Wait for any drive to fail and become replaceable.

3. List out all failed drives by using the **mmvdisk pdisk list –-recovery-group all –-replace** command.
4. When a drive fails and is ready to be replaced, the `release` operation runs automatically.
5. When the drive becomes replaceable, two things happen:
   - The replace LED is turned on.
   - The following message is logged in the `mmfs.log` file:

     ```
     [I] Automatic Pdisk Release for pdisk:e1s19 in RG:BB01L succeeded
     ```

   Note:

   - If the automatic `release` on a bad drive fails, then an error message is logged in the `mmfs.log` file. Also, the `replace` LED is not turned on.

     The following message is logged in the `mmfs.log` file:

     ```
     [E] Automatic Pdisk Release for pdisk:e1s19 in RG:BB01L failed with err:<Error code>
     ```

     If the automatic `release` fails, then follow the manual procedure to replace the failed drive. For more information, see Disk Replacement.
6. When the failed drive is identified and its `replace` LED is turned on, remove the failed drive and insert the new drive in the same slot.
7. When the new drive is successfully accepted, data is rebalanced onto the new drive automatically. On the successful acceptance of the new drive, the following events occur:
   - The `replace` LED is turned off.
   - The following message is logged to `mmfs.log` file:

     ```
     [I] Automatic Pdisk Replace for pdisk:e1s19 in RG:BB01L succeeded
     ```

     If the automatic `replace` on the newly replaced drive fails, the following error message is logged in `mmfs.log` file, and the `replace` LED is not turned off.

     ```
     [E] Automatic Pdisk Replace for pdisk:e1s19 in RG:BB01L failed with err:<Error code>
     ```

     If the new drive is not accepted successfully, then the drive is automatically released again, allowing for replacement to be retried with a different drive. Review the `mmfs.log` file for the drive replacement error messages listed above, and determine if retrying replacement with a different drive would resolve the issue.

   Note:

   You might need to wait for the maximum time, in minutes, defined by the configuration keyword `nsdRAIDDiskDiscoveryInterval` before the replacement drive is discovered and the `replace` LED is turned off.

   The command **mmvdisk pdisk list --recovery-group all --replace** can be used to verify that the replaced drive is no longer in the list.

   If the drive still appears in the replace list after the discovery interval, check the `mmfs.log` file to determine if the replacement drive was rejected and then automatically released from the system. If the drive was not discovered by the system, then follow the manual procedure to replace the failed drive. For more information, see “Disk replacement for ESS” on page 44.

Use cases for disk replacement

The following section describes some use cases for disk replacement.

Replacing failed disks in an ESS recovery group: a sample scenario for ESS 3000

This scenario shows how to detect and replace failed disks in a recovery group that is built on an ESS 3000 building block.

Detecting failed disks in your ESS 3000 enclosure

The recovery group contains one declustered array DA1 containing log home and user data VDisk.

The data declustered array is defined as follows:

- 24 pdisks per data declustered array
- Default disk replacement threshold value set to two

The replacement threshold of two means that IBM Spectrum Scale RAID requires disk replacement only when two or more disks fail in the declustered array. Otherwise, rebuilding onto spare space or reconstruction from redundancy is used to supply affected data. This configuration can be seen in the output of **mmvdisk recoverygroup list** for the recovery groups, which are shown here for RG1:

```
declustered   needs               vdisks       pdisks           capacity
   array     service  type  trim  user log  total spare rt  total raw free raw  background task
-----------  -------  ----  ----  ---- ---  ----- ----- --  --------- --------  ---------------
DA1          no       NVMe  yes      4   5     24     2  2     76 TiB 7826 GiB  scrub 14d (32%)

mmvdisk: Total capacity is the raw space before any vdisk set definitions.
mmvdisk: Free capacity is what remains for additional vdisk set definitions.

                  declustered array                                        block size and
vdisk               and log group    activity  capacity  RAID code        checksum granularity  remarks
------------------  -------  -------  --------  --------  ---------------  --------- ----------  -------
RG001LG001LOGHOME   DA1      LG001    normal    4096 MiB  4WayReplication      2 MiB       4096  log home
RG001LG002LOGHOME   DA1      LG002    normal    4096 MiB  4WayReplication      2 MiB       4096  log home
RG001LG003LOGHOME   DA1      LG003    normal    4096 MiB  4WayReplication      2 MiB       4096  log home
RG001LG004LOGHOME   DA1      LG004    normal    4096 MiB  4WayReplication      2 MiB       4096  log home
RG001ROOTLOGHOME    DA1      root     normal    4096 MiB  4WayReplication      2 MiB       4096  log home
RG001LG001VS001     DA1      LG001    normal      13 TiB  8+2p                 4 MiB       8192
RG001LG002VS001     DA1      LG002    normal      13 TiB  8+2p                 4 MiB       8192
RG001LG003VS001     DA1      LG003    normal      13 TiB  8+2p                 4 MiB       8192
RG001LG004VS001     DA1      LG004    normal      13 TiB  8+2p                 4 MiB       8192
```

The indication that disk replacement is called for in this recovery group is the value of no in the needs service column for declustered array DA1.

The fact that DA1 is undergoing rebuild of its IBM Spectrum Scale RAID tracks that can tolerate one strip failure is by itself not an indication that disk replacement is required. This just indicates that data from a failed disk is being rebuilt onto the spare space. Only if the replacement threshold is met, the disks are marked for replacement and the declustered array are flagged as needing service.

IBM Spectrum Scale RAID provides the following indications that disk replacement is required:

- Entries in the Linux syslog.
- The pdReplacePdisk callback, which can be configured to run an administrator-supplied script at the moment a pdisk is marked for replacement.
- The output from the following commands, which can be run from the CLI on any IBM Spectrum Scale RAID cluster node. Consider the following example:

  1. **mmvdisk recoverygroup list --rg** with the --declusterd-array flag shows yes in the needs service column.
  2. **mmvdisk recoverygroup list --rg** and the --pdisk flags shows the states of all pdisks, which might be examined for the replace pdisk state.
  3. **mmvdisk pdisk list** with the --replace flag, which lists only those pdisks that are marked for replacement.

Note: Because the output of **mmvdisk recoverygroup list --rg rg1 --pdisk** is long, this example shows only some of the disks, but includes the disks that are marked for replacement:

```
# mmvdisk recoverygroup list --rg rg1 --pdisk
            declustered     paths                                AU
pdisk       array       active total capacity free space log size state
----------- ----------- ------ ----- -------- ---------- -------- -----
e1s01        DA1         2      2    3576 GiB  2264 GiB   256 MiB  ok
e1s02        DA1         0      0    3576 GiB  2334 GiB   256 MiB  simulatedDead/draining/replace
e1s03        DA1         2      2    3576 GiB  2266 GiB   256 MiB  ok
e1s04        DA1         2      2    3576 GiB  2262 GiB   256 MiB  ok
e1s05        DA1         2      2    3576 GiB  2262 GiB   256 MiB  ok
e1s06        DA1         2      2    3576 GiB  2264 GiB   256 MiB  ok
e1s07        DA1         2      2    3576 GiB  2264 GiB   256 MiB  ok
e1s08        DA1         2      2    3576 GiB  2264 GiB   256 MiB  ok
e1s09        DA1         2      2    3576 GiB  2264 GiB   256 MiB  ok
e1s10        DA1         2      2    3576 GiB  2264 GiB   256 MiB  ok
e1s11        DA1         2      2    3576 GiB  2264 GiB   256 MiB  ok
e1s12        DA1         0      0    3576 GiB  2318 GiB   256 MiB  simulatedDead/draining/replace
e1s13        DA1         2      2    3576 GiB  2264 GiB   256 MiB  ok
e1s14        DA1         2      2    3576 GiB  2264 GiB   256 MiB  ok
e1s15        DA1         2      2    3576 GiB  2264 GiB   256 MiB  ok
e1s16        DA1         2      2    3576 GiB  2266 GiB   256 MiB  ok
e1s17        DA1         2      2    3576 GiB  2264 GiB   256 MiB  ok
e1s18        DA1         2      2    3576 GiB  2262 GiB   256 MiB  ok
e1s19        DA1         2      2    3576 GiB  2264 GiB   256 MiB  ok
e1s20        DA1         2      2    3576 GiB  2264 GiB   256 MiB  ok
e1s21        DA1         2      2    3576 GiB  2266 GiB   256 MiB  ok
e1s22        DA1         2      2    3576 GiB  2264 GiB   256 MiB  ok
e1s23        DA1         2      2    3576 GiB  2266 GiB   256 MiB  ok
e1s24        DA1         2      2    3576 GiB  2262 GiB   256 MiB  ok
```

The preceding output shows that the following pdisks are marked for replacement:

- e1s02 in DA1
- e1s12 in DA1

The naming convention that is used during recovery group creation indicates that these disks are in Enclosure 1 Slot 2 and Enclosure 1 Slot 12. To confirm the physical locations of the failed disks, use the **mmvdisk pdisk list** command to list information about the pdisks in declustered array DA1 of recovery group Brg1 that are marked for replacement:

```
# mmvdisk pdisk list --recovery-group all --replace

recovery group  pdisk         priority  FRU (type)       location
--------------  ------------  --------  ---------------  --------
BB01L           e2s11             1.15  00W1240          Enclosure 2 Drive 11
BB01L           e3s01             1.15  00W1240          Enclosure 3 Drive 1

mmvdisk: A lower priority value means a higher need for replacement.
```

The physical locations of the failed disks are confirmed to be consistent with the pdisk naming convention and with the IBM Spectrum Scale RAID component database:

```
------------------------------------------------------------------------------
Disk Location User Location
------------------------------------------------------------------------------
pdisk e1s02 78E00KW-2 Slot 2
------------------------------------------------------------------------------
pdisk e1s12 78E00KW-12 Slot 12
------------------------------------------------------------------------------
```

This example shows how the component database provides an easier-to-use location reference for the affected physical disks. The pdisk name e1s02 means Enclosure 1 Slot 2. Additionally, the location provides the serial number of enclosure 1, 78E00KW, with the slot number. But the user location that is defined in the component database can be used to precisely locate the disk in an equipment rack and a named disk enclosure. There is no external enclosure for an ESS 3000 system. All of the NVMe devices are in the canisters.

The relationship between the enclosure serial number and the user location can be seen with the **mmlscomp** command:

```
mmlscomp --serial-number 78E00KW
Storage Enclosure Components
Comp ID Part Number  Serial Number   Name             Node Number
------- ----------- -------------   ---------------- ----------
3        5141-AF8    78E00KW        5141-AF8-78E00KW     55
```

Replacing failed disks in a recovery group

Note: In this example, it is assumed that two new disks with the appropriate Field Replaceable Unit (FRU) code are obtained as replacements for the failed pdisks e1s02 and e1s12. In this case, the FRU attribute of the FRU is 3.84TB NVMe G3.

Replacing each disk is a three-step process:

1. Use the **mmvdisk pdisk replace** command with the --prepare flag to inform IBM Spectrum Scale to locate the disk, suspend it, and allow it to be removed.
2. Locate and remove the failed disk and replace it with a new one.
3. Use the **mmvdisk pdisk replace** command to use the new disk.

IBM Spectrum Scale RAID assigns a priority to the pdisk replacement. Disks with smaller values for the replacementPriority attribute must be replaced first. In this example, the only failed disks are in DA1 and both have the same replacementPriority value. Disk e1s02 is chosen to be replaced first.

1. Release the pdisk e1s02 in recovery group rg1 by using the following command:

```
# mmvdisk pdisk replace --prepare --recovery-group rg1 --pdisk e1s02
mmvdisk: Suspending pdisk e1s02 of RG rg1 in location 78E00KW-2.
mmvdisk: Location 78E00KW-2 is Enclosure 5141-AF8-78E00KW Drive 2.
mmvdisk:
mmvdisk: Carrier released.
mmvdisk:   - Remove carrier.
mmvdisk:   - Replace disk in location 78E00KW-2 with type '3.84TB NVMe G3 '.
mmvdisk:   - Reinsert carrier.
mmvdisk:   - Issue the following command:
mmvdisk:
mmvdisk: mmvdisk pdisk replace --recovery-group rg1 --pdisk 'e1s02'
```

2. Unlatch and pull the handle for the failed disk in slot 2. Slide out the failed disk and set it aside.

   Note: The amber LED is turned on for the failed disk. In this example, the failed disk is the disk in slot 2. The drive LEDs turn off when the slot is empty.
3. Insert the new disk with FRU 3.84TB NVMe G3 in place, push its handle forward, and latch it.
4. Finish the replacement of pdisk e1s02, by using the following command:

```
# mmvdisk pdisk replace --recovery-group rg1 --pdisk e1s02
mmvdisk:
mmvdisk: Preparing a new pdisk for use may take many minutes.
mmvdisk:
mmvdisk: mmchcarrier : [I] The following pdisks will be formatted on node c202f06fs03a:
mmvdisk: /dev/nvme11n1
mmvdisk:
mmvdisk: mmchcarrier : [I] Pdisk e1s02 of RG rg1 successfully replaced.
mmvdisk: mmchcarrier : [I] Resuming pdisk e1s02#026 of RG rg1.
mmvdisk: mmchcarrier : [I] Carrier resumed
```

When the **mmvdisk pdisk replace** command returns successfully, IBM Spectrum Scale RAID begins rebuilding and re balancing the IBM Spectrum Scale RAID strips onto the new disk, which assumes the pdisk name e1s02. The failed pdisk might remain in a temporary form, until all data from it rebuilds, at which point it is deleted. The temporary form is indicated in this example by the name e1s02#026. Only one block device name is mentioned as being formatted as a pdisk; the second path is discovered in the background.

Disk e1s12 is still marked for replacement, and DA1 of rg1 still needs service. This is because the IBM Spectrum Scale RAID replacement policy expects all failed disks in the declustered array to be replaced after the replacement threshold is reached.

To replace pdisk e1s12 following these steps:

1. Release pdisk e1s12 in recovery group rg1:

```
# mmvdisk pdisk replace --prepare --recovery-group rg1 --pdisk e1s12
mmvdisk: Suspending pdisk e1s12 of RG rg1 in location 78E00KW-12.
mmvdisk: Location 78E00KW-12 is Enclosure 5141-AF8-78E00KW Drive 12.
mmvdisk:
mmvdisk: Carrier released.
mmvdisk:    - Remove carrier.
mmvdisk:    - Replace disk in location 78E00KW-12 with type '3.84TB NVMe G3 '.
mmvdisk:    - Reinsert carrier.
mmvdisk:    - Issue the following command:
mmvdisk:
mmvdisk: mmvdisk pdisk replace --recovery-group rg1 --pdisk 'e1s12'
```

2. Find the enclosure and drawer, unlatch and remove the disk in slot 4, place a new disk in slot 4, push in the drawer, and replace the enclosure bezel.
3. Finish the replacement of pdisk e1s12, run the following command:

```
# mmvdisk pdisk replace --recovery-group rg1 --pdisk e1s12
[I] The following pdisks will be formatted on node c202f06fs03a:
/dev/nvme0n1
[I] Pdisk e1s12 of RG rg1 successfully replaced.[I] Resuming pdisk e1s12#029 of RG rg1.
[I] Carrier resumed.
```

The disk replacements can be confirmed by using the **mmvdisk recoverygroup list --rg rg1 --pdisk** command:

```
# mmvdisk recoverygroup list --rg rg1 --pdisk --declustered-array

declustered    needs        vdisks             pdisks    replace      capacity
   array       service      type    user log   total spare    threshold  total raw  free raw  background
task
-----------    -------      ----    ---- ---    ----- -----    ---------  --------  --------  ------------
DA1            no           NVMe    4     5      24    2      2         76 TiB   786 GiB   scrub 14d
(0%)

mmvdisk: Total capacity is the raw space before any vdisk set definitions.
mmvdisk: Free capacity is what remains for additional vdisk set definitions.

                       declustered        paths                                                  AU
pdisk                  array        active    total    capacity     free space     log size     state
------                 -----        ------    -------  ----------   ------------   ----------   ------
e1s01                  DA1               2        2     3576 GiB       342 GiB     256 MiB    ok
e1s02                  DA1               2        2     3576 GiB       342 GiB     256 MiB    ok
e1s02#026                DA1             0        0       3576 GiB       342 GiB     256 MiB    simulatedDead/deleting/ draining/01008.6c0
e1s03                  DA1               2        2     3576 GiB       344 GiB     256 MiB    ok
e1s04                  DA1               2        2     3576 GiB       340 GiB     256 MiB    ok
e1s05                  DA1               2        2     3576 GiB       342 GiB     256 MiB    ok
e1s06                  DA1               2        2     3576 GiB       344 GiB     256 MiB    ok
e1s07                  DA1               2        2     3576 GiB       342 GiB     256 MiB    ok
e1s08                  DA1               2        2     3576 GiB       340 GiB     256 MiB    ok
e1s09                  DA1               2        2     3576 GiB       338 GiB     256 MiB    ok
e1s10                  DA1               2        2     3576 GiB       344 GiB     256 MiB    ok
e1s11                  DA1               2        2     3576 GiB       340 GiB     256 MiB    ok
e1s12                  DA1               2        2     3576 GiB       340 GiB     256 MiB    ok
e1s12#029                DA1             0        0       3576 GiB       342 GiB     256 MiB    simulatedDead/deleting/draining/01008.6c0
e1s13                  DA1               2        2     3576 GiB       344 GiB     256 MiB    ok
e1s14                  DA1               2        2     3576 GiB       342 GiB     256 MiB    ok
e1s15                  DA1               2        2     3576 GiB       344 GiB     256 MiB    ok
e1s16                  DA1               2        2     3576 GiB       344 GiB     256 MiB    ok
e1s17                  DA1               2        2     3576 GiB       340 GiB     256 MiB    ok
e1s18                  DA1               2        2     3576 GiB       342 GiB     256 MiB    ok
e1s19                  DA1               2        2     3576 GiB       342 GiB     256 MiB    ok
e1s20                  DA1               2        2     3576 GiB       344 GiB     256 MiB    ok
e1s21                  DA1               2        2     3576 GiB       340 GiB     256 MiB    ok
e1s22                  DA1               2        2     3576 GiB       344 GiB     256 MiB    ok
e1s23                  DA1               2        2     3576 GiB       342 GiB     256 MiB    ok
e1s24                  DA1               2        2     3576 GiB       342 GiB     256 MiB    ok
```

During replacement, the new disks take the name of the replaced pdisks. In the event that replaced pdisks have not completely drained, they are given a temporary name consisting of the old pdisk name with a suffix of the form #nnnn, and are counted toward the total number of pdisks in the recovery group rg1 and the declustered array DA1. The temporary pdisk exists until IBM Spectrum Scale RAID rebuild completes the reconstruction of the data that they carried onto other disks, including their replacements. When rebuild completes, the temporary pdisks disappear, and the number of disks in DA1 becomes 24 again.

Replacing failed disks in an ESS recovery group: a sample scenario for ESS 5000

The scenario presented here shows how to detect and replace failed disks in a recovery group built on an ESS building block.

Detecting failed disks in your ESS enclosure

Assume an SL2 building block on which the following two recovery groups are defined:

- BB01L, containing the disks in the left side of each enclosure.
- BB01R, containing the disks in the right side of each enclosure.

Each recovery group contains the following:

- One NVR declustered array (LOGTIP)
- One SSD declustered array (LOGTIPBACKUP)
- A log portion of the data declustered array, DA1(LOGHOME)
- One data declustered array, DA1

The data declustered array is defined according to SL2 best practices as follows:

- 91 pdisks per data declustered array
- Default disk replacement threshold value set to 2

The replacement threshold of 2 means that IBM Spectrum Scale RAID only requires disk replacement when two or more disks fail in the declustered array; otherwise, rebuilding onto spare space or reconstruction from redundancy is used to supply affected data. This configuration can be seen in the output of **`mmvdisk recoverygroup list`** for a recovery group, which is shown here for BB01L:

```
# mmvdisk recoverygroup list --recovery-group BB01L --declustered-array --vdisk

declustered   needs               vdisks       pdisks           capacity
   array     service  type  trim  user log  total spare rt  total raw free raw  background task
-----------  -------  ----  ----  ---- ---  ----- ----- --  --------- --------  ---------------
NVR          no       NVR   -        0   1      2     0  1           -        -  scrub 14d (35%)
SSD          no       SSD   -        0   1      1     0  1           -        -  scrub 14d (35%)
DA1          yes      HDD   no       2   1     91     2  2    798 TiB  119 TiB  scrub 14d (11%)

mmvdisk: Total capacity is the raw space before any vdisk set definitions.
mmvdisk: Free capacity is what remains for additional vdisk set definitions.

                  declustered                                          block size and
vdisk             array        activity  capacity  RAID code        checksum granularity  remarks
----------------  -----------  --------  --------  ---------------  ---------  ---------  -------
RG001LOGHOME      DA1          normal     144 GiB  4WayReplication      2 MiB       4096  log home
RG001LOGTIP       NVR          normal     192 MiB  2WayReplication      2 MiB       4096  log tip
RG001LOGTIPBACKUP SSD          normal     192 MiB  Unreplicated         2 MiB       4096  log tip backup
RG001VS001        DA1          normal      19 TiB  4WayReplication      1 MiB      32 KiB
RG001VS003        DA1          normal     433 TiB  8+3p                16 MiB      32 KiB
```

The indication that disk replacement is called for in this recovery group is the value of `yes` in the needs service column for declustered array DA1.

The fact that DA1 is undergoing rebuild of its IBM Spectrum Scale RAID is by itself not an indication that disk replacement is required; it merely indicates that data from a failed disk is being rebuilt onto spare space. Only if the replacement threshold has been met will disks be marked for replacement and the declustered array be marked as needing service.

IBM Spectrum Scale RAID provides several indications that disk replacement is required:

- Entries in the Linux syslog
- The pd`ReplacePdisk` callback, which can be configured to run an administrator-supplied script at the moment when a pdisk is marked for replacement
- The output from the following commands, which can be run from the CLI on any IBM Spectrum Scale RAID cluster node. Consider the following examples:

  1. **`mmvdisk recoverygroup list --rg`** with the `--declustered-array` flag shows `yes` in the `needs service` column.

2. **mmvdisk recoverygroup list --rg** with the `--pdisk` flag shows the states of all pdisks, which might be examined for the replace pdisk state.
3. **mmvdisk pdisk list --rg** with the `--replace` flag, which lists only those pdisks that are marked for replacement.

Example:

Note: Because the output of **mmvdisk rg list –rg BB01L --pdisk** is long, this example shows only some of the pdisks but includes those marked for replacement.

```
# mmvdisk recoverygroup list --rg BB01L --pdisk

              declustered      paths                                AU
pdisk           array     active  total  capacity  free space  log size  state
------------  -----------  ------  -----  --------  ----------  --------  -----
n001v001      NVR              1      1    31 GiB      31 GiB   120 MiB  ok
n002v001      NVR              1      1    31 GiB      31 GiB   120 MiB  ok
e1s01ssd      SSD              2      4   745 GiB     744 GiB   120 MiB  ok
e1s02         DA1              2      4  9248 GiB    1592 GiB    40 MiB  ok
[...]
e1s19         DA1              2      4  9248 GiB    1592 GiB    40 MiB  simulatedDead/draining/replace
[...]
e1s66         DA1              2      4  9248 GiB    1592 GiB    40 MiB  simulatedDead/draining/replace
[...]
e1s85         DA1              2      4  9248 GiB    1592 GiB    40 MiB  ok
e2s01         DA1              2      4  9248 GiB    1608 GiB    40 MiB  ok
[...]
e2s84         DA1              2      4  9248 GiB    1608 GiB    40 MiB  ok
e2s85         DA1              2      4  9248 GiB    1608 GiB    40 MiB  ok
```

The preceding output shows that the following pdisks are marked for replacement:

- e1s19 in DA1
- e1s66 in DA1

The naming convention used during recovery group creation indicates that these disks are in Enclosure 1 Slot 19 and Enclosure 1 Slot 66. To confirm the physical locations of the failed disks, use the **mmvdisk pdisk list** command to list information about the pdisks in declustered array DA1 of recovery group BB01L that are marked for replacement:

```
# mmvdisk pdisk list --recovery-group BB01L --replace
recovery group  pdisk        priority       FRU (type)         location
--------------  ------------ --------       ---------------    --------
BB01L            e1s19        0.98         02PX531 Enclosure   5147-092-789A3B0 Drive 19
BB01L            e1s66        0.98         02PX531 Enclosure   5147-092-789A3B0 Drive 66
mmvdisk: A lower priority value means a higher need for replacement.
```

The physical locations of the failed disks are confirmed to be consistent with the pdisk naming convention and with the IBM Spectrum Scale RAID component database:

```
-------------------------------------------------------------------------------------
Disk                    Location               User Location
-------------------------------------------------------------------------------------
pdisk e1s19        789A3B0-19        Rack BB01 U01-05, Enclosure BB01ENC1 Slot 19
-------------------------------------------------------------------------------------
pdisk e1s66        789A3B0-66        Rack BB01 U01-05, Enclosure BB01ENC1 Slot 66
-------------------------------------------------------------------------------------
```

This shows how the component database provides an easier-to-use location reference for the affected physical disks. The pdisk name e1s19 means "Enclosure 1 Slot 19." Additionally, the location provides the serial number of enclosure 1, 789A3B0, with the slot number, -19. But the user location that has been defined in the component database can be used to precisely locate the disk in an equipment rack and a named disk enclosure. This is the disk enclosure that is labeled "BB01ENC1," found in compartments U01-U05 of the rack labeled "BB01," and the disk is in slot 19 of that enclosure.

The relationship between the enclosure serial number and the user location can be seen with the **mmlscomp** command:

```
# mmlscomp --serial-number 789A3B0

    Storage Enclosure Components
```

```
Comp ID  Part Number  Serial Number  Name              Node Number
-------  -----------  -------------  ----------------  ----------
      4  5147-092     789A3B0        5147-092-789A3B0     55
```

Replacing failed disks in an SL2 recovery group

Replacing each disk is a three-step process:

1. Using the **mmvdisk pdisk replace** command with the `--prepare` flag to inform IBM Spectrum Scale to locate the disk, suspend it, and allow it to be removed.
2. Locating and removing the failed disk, and replacing it with a new one.
3. Using the **mmvdisk pdisk replace** command to use the new disk.

Example:

Note: In this example, it is assumed that two new disks with the appropriate Field Replaceable Unit (FRU) code, as indicated by the fru attribute (`02PX531` in this case), have been obtained as replacements for the failed pdisks `e1s19` and `e1s66`.

1. Run the following command to release pdisk `e1s19` in recovery group `BB01L`:

```
# mmvdisk pdisk replace –prepare –recovery-group BB01L --pdisk e1s19
mmvdisk: Suspending pdisk e1s19 of RG BB01L in location 789A3B0-19.
mmvdisk: Location 789A3B0-19 is Enclosure 789A3B0 Drive 19.
mmvdisk: Carrier released.
mmvdisk:
mmvdisk:   - Remove carrier.
mmvdisk:   - Replace disk in location 789A3B0-19 with type '02PX531'.
mmvdisk:   - Reinsert carrier.
mmvdisk:   - Issue the following command:
mmvdisk:
mmvdisk:   mmvdisk pdisk replace --recovery-group BB01L --pdisk 'e1s19'
```

   IBM Spectrum Scale RAID issues instructions as to the physical actions that must be taken, and repeats the user-defined location to help find the disk.

2. To allow the enclosure `BB01ENC1` with serial number `789A3B0` to be located and identified, IBM Spectrum Scale RAID turns on the enclosure's amber "service required" LED. The enclosure's bezel must be removed. This will reveal that the amber service required LED has been turned on.

   Note: In this case the disk in slot 19 has its amber LED turned on.

   a. Unlatch and pull up the handle for the identified disk in slot 19. Lift out the failed disk and set it aside. The drive LEDs turn off when the slot is empty.

   b. A new disk with FRU `02PX531` should be lowered in place and have its handle pushed down and latched.

      Note: Since the second disk replacement in this example is also in the same enclosure, leave the enclosure bezel off. If the next replacement were in a different enclosure, the enclosure bezel would be replaced.

3. Run the following command to finish the replacement of pdisk `e1s19`:

```
# mmvdisk pdisk replace –recovery-group BB01L --pdisk e1s19
mmvdisk: 2020-07-17_10:18:02.800-0400: [I] Callback: /usr/lpp/mmfs/bin/tspreparenewpdiskforuse /dev/sdfy.
mmvdisk: Attempting to update firmware if necessary. Failure will not prevent drive replacement.
mmvdisk: Command: mmchfirmware --type drive --serial-number         JEJ6NHYN --new-pdisk
mmvdisk: Command: err 0: mmchfirmware --type drive --serial-number         JEJ6NHYN --new-pdisk
mmvdisk:
mmvdisk: The following pdisks will be formatted on node c145f08zn04.gpfs.net:
mmvdisk: //c145f08zn03/dev/sdgf,//c145f08zn03/dev/sdgg,//c145f08zn04/dev/sdgs,//c145f08zn04/dev/sdfy
mmvdisk: Pdisk e1s19 of RG BB01L successfully replaced.
mmvdisk: Resuming pdisk e1s19#0091 of RG BB01L.
mmvdisk: Carrier resumed.
mmvdisk:
mmvdisk: mmchcarrier : [I] Preparing a new pdisk for use may take many minutes.
mmvdisk:
```

When the **mmvdisk pdisk replace** command returns successfully, IBM Spectrum Scale RAID begins rebuilding and rebalancing IBM Spectrum Scale RAID strips onto the new disk, which assumes the pdisk name e1s19. The failed pdisk might remain in a temporary form, until all the data from it rebuilds, at which point it is deleted. The temporary form is indicated in this example by the name e1s19#0091.

Disk e1s66 is still marked for replacement, and DA1 of BB01L still needs service. This is because the IBM Spectrum Scale RAID replacement policy expects all the failed disks in the declustered array to be replaced after the replacement threshold is reached.

Repeat the same procedure to replace Pdisk e1s66:

1. Run the following command to release the pdisk e1s66 in recovery group BB01L:

```
# mmvdisk pdisk replace --prepare --rg BB01L --pdisk e1s66
mmvdisk: Suspending pdisk e1s66 of RG BB01L in location 789A3B0-66.
mmvdisk: Location 789A3B0-66is Enclosure 789A3B0 Drive 66.
mmvdisk: Carrier released.
mmvdisk:
mmvdisk:   - Remove carrier.
mmvdisk:   - Replace disk in location 789A3B0-66with type '02PX531'.
mmvdisk:   - Reinsert carrier.
mmvdisk:   - Issue the following command:
mmvdisk:
mmvdisk:   mmvdisk pdisk replace --recovery-group BB01L --pdisk 'e1s66'
```

2. Find the enclosure, unlatch and remove the enclosure's bezel, remove the disk in slot 66, place a new disk in slot 66, and replace the enclosure's bezel.
3. Run the following command to complete the replacement of pdisk e1s66:

```
# mmvdisk pdisk replace --rg BB01L --pdisk e1s66
mmvdisk: 2020-07-17_10:44:39.487-0400: [I] Callback: /usr/lpp/mmfs/bin/tspreparenewpdiskforuse /dev/sdgy.
mmvdisk: Attempting to update firmware if necessary. Failure will not prevent drive replacement.
mmvdisk: Command: mmchfirmware --type drive --serial-number         JEJ6E24N --new-pdisk
mmvdisk: Command: err 0: mmchfirmware --type drive --serial-number         JEJ6E24N --new-pdisk
mmvdisk:
mmvdisk: The following pdisks will be formatted on node c145f08zn04.gpfs.net:
mmvdisk: //c145f08zn04/dev/sdhe,//c145f08zn03/dev/sdgi,//c145f08zn03/dev/sdgh,//c145f08zn04/dev/sdgy
mmvdisk: Pdisk e1s66 of RG BB01L successfully replaced.
mmvdisk: Resuming pdisk e1s66#0089 of RG BB01L.
mmvdisk: Carrier resumed.
mmvdisk:
mmvdisk: mmchcarrier : [I] Preparing a new pdisk for use may take many minutes.
mmvdisk:
```

The disk replacements can be confirmed by using the **mmvdisk recoverygroup list --rg BB01L --pdisk** command as shown:

```
# mmvdisk rg list --rg BB01L --pdisk

              declustered      paths                                 AU
pdisk           array     active  total  capacity  free space  log size  state
------------  -----------  ------  -----  --------  ----------  --------  -----
n001v001      NVR               1      1    31 GiB      31 GiB   120 MiB  ok
n002v001      NVR               1      1    31 GiB      31 GiB   120 MiB  ok
e1s01ssd      SSD               2      4   745 GiB     744 GiB   120 MiB  ok
e1s02         DA1               2      4  9248 GiB    1592 GiB    40 MiB  ok
[...]
e1s19         DA1               2      4  9248 GiB    1592 GiB    40 MiB  ok
e1s19#0091    DA1                      0  9248 GiB    6312 GiB    40 MiB  simulatedDead/deleting/draining
[...]
e1s66         DA1               2      4  9248 GiB    1592 GiB    40 MiB  ok
e1s66#0089    DA1                      0  9248 GiB    6312 GiB    40 MiB  simulatedDead/deleting/draining
[...]
e2s85         DA1               2      4  9248 GiB    1608 GiB    40 MiB  ok
```

Notice that the temporary pdisks (The disk replacements can be confirmed by usinge1s19#0091 and e1s66#0089), representing the now-removed physical disks, are counted toward the total number of pdisks in the recovery group BB01L and the declustered array DA1. They exist until the IBM Spectrum Scale RAID rebuild completes the reconstruction of the data that they carried onto other disks including their replacements. When rebuild completes, the temporary pdisks disappear, and the number of disks in DA1 is 91 again.

Using the mmvdisk command to fix issues caused by improper disk removal for ESS

Pdisks are identified by the descriptors that are written onto the disks, not by their physical locations. If a pdisk is moved to a different enclosure slot, the system still correctly identifies the pdisk and continues to use it. In general, the system cannot prevent an operator from swapping disks between slots. Continuing to use a disk that is found in an unexpected location avoids risk of data unavailability.

The location code that is associated with a pdisk reflects the enclosure slot where the pdisk was last seen. Thus, if a pdisk is moved to a different slot, the system automatically updates the location code to reflect where it currently is.

There are only two ways a location code can be empty:

- The location is unknown since the time of installation.
- The pdisk was removed; another pdisk from the same GNR recovery group pair was inserted into the slot, and the new pdisk took over the location.

Devices such as logtip disks might not have location codes and can fall into the first case. But devices in external enclosures that automatically detect the location are not likely to be blank forever. Blank location codes on these disks, therefore, suggest that disks have been pulled out and other disks from the same recovery group pair have been placed into their slots.

The user location code comes from a table in the **mmcomp** database that maps location code to user location code. A blank user location might indicate a blank location code as mentioned above, or it may indicate a missing row in the table. Verify that the regular location code is also blank.

Test case of issues caused due to improper disk removal

Consider a situation where the pdisk has failed. The admin runs the **mmvdisk pdisk replace --prepare --recovery-group rg1 --pdisk e1s02** command, and removes the bad drive. The system is now expecting a new disk to be inserted. However, instead of inserting a new disk, the admin pulls pdisk `e1s03` from one slot over, inserts it into slot 2, then runs the **mmvdisk pdisk replace --recovery-group rg1 --pdisk e1s02** command. The replace command detects what happened and fails, and displays the following error message:

```
[E] Pdisk e1s03 of recovery group rg1 in location 78E00KW-2 cannot be used as a
replacement for pdisk e1s02 of recovery group rg1.
```

But because `e1s03` now occupies the slot, it has taken on the location code `78E00KW-2`, clearing it from pdisk `e2s02`. The system no longer knows the location `e1s03`; it just knows that the location is not `78E00KW-2`. Even, if the admin realizes the mistake and moves `e1s03` back into slot 3, `e1s03`'s location is updated to slot 3, but `e1s02`'s location remains blank.

Solution

You can put the disks back into the right slot and solve this issue in case the following criteria are met:

- You have all the drives.
- All the drives are functional and the system can read the descriptors from them.
- dd or other tools are not used to clear the descriptors.

When the system discovers the disks, it automatically updates the location codes. After the location codes are updated, replace any bad disks by using the **mmvdisk pdisk change** command. To pull a drive that is in the wrong slot, use the **mmvdisk pdisk change --recovery-group RGNAME --pdisk PDNAME --suspend** command to quiesce the disk before you pull it. Run the **mmvdisk pdisk change --recovery-group RGNAME --pdisk PDNAME --resume** command after you reinsert the disk. Suspending the disk before you pull it avoids unnecessary I/O errors and the risk of causing a recovery group resign.

If some of the disks are no longer available or the descriptors are unreadable, then you can use the `replace-at-location` script to replace them. This script is found in `/usr/lpp/mmfs/samples/vdisk` as shown:

1. Insert a new, blank disk into the empty slot 2 where the bad `e1s02` drive was.
2. Run replace-at-location `rg1 e1s02 78E00KW-2`.

Replacing failed ESS storage enclosure components: a sample scenario for ESS 5000

The scenario presented here shows how to detect and replace failed storage enclosure components in an ESS building block.

Detecting failed storage enclosure components

The **mmlsenclosure** command can be used to show you which enclosures need service along with the specific component. A best practice is to run this command every day to check for failures.

```
# mmlsenclosure all -L --not-ok

                  needs
 serial number    service  nodes
 -------------    -------  ------
 789A3AY          yes      c145f08zn03.gpfs.net

 component type  serial number     component id    failed value   unit   properties
 --------------  -------------     ------------    ------ -----   ----   ----------
 fan             789A3AY           1_BOT_LEFT      yes            RPM    FAILED
```

When you are ready to replace the failed component, use the **mmchenclosure** command to identify whether it is safe to complete the repair action, or whether IBM Spectrum Scale needs to be shut down first:

```
# mmchenclosure 789A3AY --component fan --component-id 1_BOT_LEFT

mmenclosure: Proceed with the replace operation.
```

The fan can now be replaced.

Special note about detecting failed enclosure components

In the following example, only the enclosure itself is being called out as having failed; the specific component that has actually failed is not identified.

```
mmlsenclosure all -L --not-ok

                   needs  nodes
 serial number     service
 -------------     ------- ------
 SV13306129        yes     c45f01n01-ib0.gpfs.net

 component type  serial number     component id    failed value   unit   properties
 --------------  -------------     ------------    ------ -----   ----   ----------
 enclosure       SV13306129        ONLY            yes                   NOT_IDENTIFYING,FAILED
```

This typically means that there are drive "Service Action Required (Fault)" LEDs that have been turned on in the drawers. In such a situation, the **mmvdisk pdisk list –-recovery-group all --not-ok** command can be used to check for dead or failing disks.

Replacing a failed ESS storage enclosure for ESS 5000

Enclosure replacement should be rare. This procedure assumes that the enclosure chassis is replaced, and the serial number of the replaced enclosure is moved to the replaced chassis. Contact IBM Service if the enclosure replacement changes the serial number of the enclosure.

Prerequisite information:

This procedure is intended to be done as a partnership between the storage administrator and a hardware service representative. The storage administrator is expected to understand the IBM Spectrum Scale

RAID concepts and the locations of the storage enclosures. The storage administrator is responsible for all the steps except those in which the hardware is actually being worked on.

To replace a failed storage enclosure, follow these steps:

1. Shut down IBM Spectrum Scale and perform the enclosure replacement as soon as possible.
2. Run the following enclosure replacement procedure:
   a. Replace the enclosure by running the following standard hardware procedures:
      - Remove the SAS connections in the rear of the enclosure.
      - Remove the enclosure.
      - Install the new enclosure.
   b. Replace the drives in the corresponding slots of the new enclosure.
   c. Connect the SAS connections in the rear of the new enclosure.
   d. Power up the enclosure.
   e. Verify the SAS topology on the servers to ensure that all drives from the new storage enclosure are present.
   f. Update the necessary firmware on the new storage enclosure as needed.

Other hardware service

This section contains information about hardware services for ESS systems.

While IBM Spectrum Scale RAID can easily tolerate a single disk fault with no significant impact, and failures of up to three disks with various levels of impact on performance and data availability, it still relies on a majority of all the disks functioning properly and reachable from the server. If a major equipment malfunction prevents both the primary and backup server from accessing more than that number of disks, or if those disks are destroyed, all vdisks in the recovery group become either unavailable or suffer permanent data loss. As IBM Spectrum Scale RAID cannot recover from such catastrophic problems, it also does not attempt to diagnose them or organize their maintenance.

Hardware service for ESS Legacy systems

In the case that a IBM Spectrum Scale RAID server becomes permanently disabled, a manual failover procedure exists that requires recabling to an alternative server. For more information, see the `mmchrecoverygroup` command in the *IBM Spectrum Scale RAID: Administration*. If both the primary and backup IBM Spectrum Scale RAID servers for a recovery group fail, the recovery group is unavailable until one of the servers is repaired.

Hardware service for ESS 3000, ESS 5000, and ESS 3200

Other hardware components of the ESS system such as boot drives, fans, and power supplies can be serviced by IBM authorized service personnel only. IBM service support representatives and lab based services personnel can access service information through the Service Guide.

Note: An IBM intranet connection is required to access these documents.

The status of many ESS components can be examined by using the **`mmlsenclosure`** command.

Directed maintenance procedures available in the GUI

The directed maintenance procedures (DMPs) assist you to repair a problem when you select the action **Run fix procedure** on a selected event from the **Monitoring** > **Events** page. DMPs are present for only a few events reported in the system.

The following table provides details of the available DMPs and the corresponding events.

Table 5. DMPs	
DMP	**Event ID**
Replace disks	gnr_pdisk_replaceable
Update enclosure firmware	enclosure_firmware_wrong
Update drive firmware	drive_firmware_wrong
Update host-adapter firmware	adapter_firmware_wrong
Start NSD	disk_down
Start GPFS daemon	gpfs_down
Increase fileset space	inode_error_high and inode_warn_high
Start performance monitoring collector service	pmcollector_down
Start performance monitoring sensor service	pmsensors_down
Activate AFM performance monitoring sensors	afm_sensors_inactive
Activate NFS performance monitoring sensors	nfs_sensors_inactive
Activate SMB performance monitoring sensors	smb_sensors_inactive
Configure NFS sensor	nfs_sensors_not_configured
Configure SMB sensor	smb_sensors_not_configured
Mount file systems	unmounted_fs_check
Start GUI service on remote node	gui_down
Repair a failed GUI refresh task	gui_refresh_task_failed

Replace disks

The replace disks DMP assists you to replace the disks.

The following are the corresponding event details and proposed solution:

- **Event name:** gnr_pdisk_replaceable
- **Problem:** The state of a physical disk is changed to “replaceable”.
- **Solution:** Replace the disk.

The ESS GUI detects if a disk is broken and whether it needs to be replaced. In this case, launch this DMP to get support to replace the broken disks. You can use this DMP either to replace one disk or multiple disks.

The DMP automatically launches in corresponding mode depending on situation. You can launch this DMP from the pages in the GUI and follow the wizard to release one or more disks:

- **Monitoring** > **Hardware** page: Select **Replace Broken Disks** from the **Actions** menu.
- **Monitoring** > **Hardware** page: Select the broken disk to be replaced in an enclosure and then select **Replace** from the **Actions** menu.
- **Monitoring** > **Events** page: Select the *gnr_pdisk_replaceable* event from the event listing and then select **Run Fix Procedure** from the **Actions** menu.
- **Storage** > **Physical Disks** page: Select **Replace Broken Disks** from the **Actions** menu.
- **Storage** > **Physical Disks** page: Select the disk to be replaced and then select **Replace Disk** from the **Actions** menu.

The system uses the following command on an *mmvdisk-enabled* environment to release and replace the disk:

```
mmvdisk pdisk replace [--prepare | --cancel] --recovery-group DiskRecoveryGroup --pdisk DiskName
```

Update enclosure firmware

The update enclosure firmware DMP assists to update the enclosure firmware to the latest level.

The following are the corresponding event details and the proposed solution:

- **Event name:** enclosure_firmware_wrong
- **Problem:** The reported firmware level of the environmental service module is not compliant with the recommendation.
- **Solution:** Update the firmware.

If more than one host-adapter is not running the newest version of the firmware, the system prompts to update the firmware. The system issues the **mmchfirmware** command to update firmware of the installed host-adapters. Consult the *IBM Spectrum Scale RAID: Administration* guide for the **mmchfirmware** command format.

Update drive firmware

The update drive firmware DMP assists to update the drive firmware to the latest level so that the physical disk becomes compliant.

The following are the corresponding event details and the proposed solution:

- **Event name:** drive_firmware_wrong
- **Problem:** The reported firmware level of the physical disk is not compliant with the recommendation.
- **Solution:** Update the firmware.

If more than one host-adapter is not running the newest version of the firmware, the system prompts to update the firmware. The system issues the **mmchfirmware** command to update firmware of the installed host-adapters. Consult the *IBM Spectrum Scale RAID: Administration* guide for the **mmchfirmware** command format.

Update host-adapter firmware

The Update host-adapter firmware DMP assists to update the host-adapter firmware to the latest level.

The following are the corresponding event details and the proposed solution:

- **Event name:** adapter_firmware_wrong
- **Problem:** The reported firmware level of the host adapter is not compliant with the recommendation.
- **Solution:** Update the firmware.

If more than one host-adapter is not running the newest version of the firmware, the system prompts to update the firmware. The system issues the **mmchfirmware** command to update firmware of the installed host-adapters. Consult the *IBM Spectrum Scale RAID: Administration* guide for the **mmchfirmware** command format.

Note: IBM Spectrum Scale RAID daemon must be down for host-adapter firmware upgrade.

Start NSD

The Start NSD DMP assists to start NSDs that are not working.

The following are the corresponding event details and the proposed solution:

- **Event ID:** disk_down

- **Problem:** The availability of an NSD is changed to “down”.
- **Solution:** Recover the NSD.

The DMP provides the option to start the NSDs that are not functioning. If multiple NSDs are down, you can select whether to recover only one NSD or all of them.

The system issues the **mmchdisk** command to recover NSDs as given in the following format:

```
/usr/lpp/mmfs/bin/mmchdisk <device> start -d <disk description>
```

For example: `/usr/lpp/mmfs/bin/mmchdisk r1_FS start -d G1_r1_FS_data_0`

Start GPFS daemon

When the GPFS daemon is down, GPFS functions do not work properly on the node.

The following are the corresponding event details and the proposed solution:

- **Event ID:** gpfs_down
- **Problem:** The GPFS daemon is down. GPFS is not operational on node.
- **Solution:** Start GPFS daemon.

The system issues the **mmstartup -N** command to restart GPFS daemon as given in the following format:

```
/usr/lpp/mmfs/bin/mmstartup -N <Node>
```

For example: `usr/lpp/mmfs/bin/mmstartup -N gss-05.localnet.com`

Increase fileset space

The system needs inodes to allow I/O on a fileset. If the inodes allocated to the fileset are exhausted, you need to either increase the number of maximum inodes or delete the existing data to free up space.

The procedure helps to increase the maximum number of inodes by a percentage of the already allocated inodes. The following are the corresponding event details and the proposed solution:

- **Event ID:** `inode_error_high` and `inode_warn_high`
- **Problem:** The inode usage in the fileset reached an exhausted level.
- **Solution:** Increase the maximum number of inodes.

The system issues the **mmchfileset** command to recover NSDs as given in the following format:

```
/usr/lpp/mmfs/bin/mmchfileset <Device> <Fileset> --inode-limit <inodesMaxNumber>
```

For example: `/usr/lpp/mmfs/bin/mmchfileset r1_FS testFileset --inode-limit 2048`

Synchronize node clocks

The time must be in sync with the time set on the GUI node. If the time is not in sync, the data that is displayed in the GUI might be wrong or it does not even display the details. For example, the GUI does not display the performance data if time is not in sync.

The procedure assists to fix timing issue on a single node or on all nodes that are out of sync. The following are the corresponding event details and the proposed solution:

- **Event ID:** time_not_in_sync
- **Limitation:** This DMP is not available in sudo wrapper clusters. In a sudo wrapper cluster, the user name is different from 'root'. The system detects the user name by finding the parameter GPFS_USER=<user name>, which is available in the file `/usr/lpp/mmfs/gui/conf/gpfsgui.properties`.

- **Problem:** The time on the node is not synchronous with the time on the GUI node. It differs more than 1 minute.
- **Solution:** Synchronize the time with the time on the GUI node.

The system issues the **sync_node_time** command as given in the following format to synchronize the time in the nodes:

```
usr/lpp/mmfs/gui/bin-sudo/sync_node_time <nodeName>
```

For example: `/usr/lpp/mmfs/gui/bin-sudo/sync_node_time c55f06n04.gpfs.net`

Start performance monitoring collector service

The collector services on the GUI node must be functioning properly to display the performance data in the IBM Spectrum Scale management GUI.

The following are the corresponding event details and the proposed solution:

- **Event ID:** pmcollector_down
- **Limitation:** This DMP is not available in sudo wrapper clusters when a remote *pmcollector* service is used by the GUI. A remote *pmcollector* service is detected in case a different value than localhost is specified in the ZIMonAddress in file, which is located at: `/usr/lpp/mmfs/gui/conf/gpfsgui.properties`. In a sudo wrapper cluster, the user name is different from 'root'. The system detects the user name by finding the parameter GPFS_USER=<user name>, which is available in the file `/usr/lpp/mmfs/gui/conf/gpfsgui.properties`.
- **Problem:** The performance monitoring collector service *pmcollector* is in inactive state.
- **Solution:** Issue the **systemctl status pmcollector** to check the status of the collector. If *pmcollector* service is inactive, issue **systemctl start pmcollector**.

The system restarts the performance monitoring services by issuing the **systemctl restart pmcollector** command.

The performance monitoring collector service might be on some other node of the current cluster. In this case, the DMP first connects to that node, then restarts the performance monitoring collector service.

```
ssh <nodeAddress> systemctl restart pmcollector
```

For example: `ssh 10.0.100.21 systemctl restart pmcollector`

In a sudo wrapper cluster, when collector on remote node is down, the DMP does not restart the collector services by itself. You need to do it manually.

Start performance monitoring sensor service

You need to start the sensor service to get the performance details in the collectors. If sensors and collectors are not started, the GUI and CLI do not display the performance data in the IBM Spectrum Scale management GUI.

The following are the corresponding event details and the proposed solution:

- **Event ID:** pmsensors_down
- **Limitation:** This DMP is not available in sudo wrapper clusters. In a sudo wrapper cluster, the user name is different from 'root'. The system detects the user name by finding the parameter GPFS_USER=<user name>, which is available in the file `/usr/lpp/mmfs/gui/conf/gpfsgui.properties`.
- **Problem:** The performance monitoring sensor service *pmsensor* is not sending any data. The service might be down or the difference between the time of the node and the node hosting the performance monitoring collector service *pmcollector* is more than 15 minutes.
- **Solution:** Issue **systemctl status pmsensors** to verify the status of the sensor service. If *pmsensor* service is inactive, issue **systemctl start pmsensors**.

The system restarts the sensors by issuing **systemctl restart pmsensors** command.

For example: `ssh gss-15.localnet.com systemctl restart pmsensors`

Activate AFM performance monitoring sensors

The activated AFM performance monitoring sensor's DMP assists to activate the inactive SMB sensors.

The following are the corresponding event details and the proposed solution:

- **Event ID:** afm_sensors_inactive
- **Problem:** The AFM performance cannot be monitored because one or more of the performance sensors like GPFSAFMFS, GPFSAFMFSET, and GPFSAFM are offline.
- **Solution:** Activate the AFM sensors.

The DMP provides the option to activate the AFM monitoring sensor and select a data collection interval that defines how frequently the sensors must collect data. It is recommended to select a value that is greater than or equal to 10 as the data collection frequency to reduce the impact on the system performance.

The system issues the **mmperfmon** command to activate AFM sensors as given in the following format:

```
/usr/lpp/mmfs/bin/mmperfmon config update <<sensor_name>>.restrict=<<afm_gateway_nodes>>
/usr/lpp/mmfs/bin/mmperfmon config update <<sensor_name>>.period=<<seconds>>
```

For example,

```
/usr/lpp/mmfs/bin/mmperfmon config update GPFSAFM.restrict=gss-41
/usr/lpp/mmfs/bin/mmperfmon config update GPFSAFM.period=30
```

Activate NFS performance monitoring sensors

The activate NFS performance monitoring sensors DMP assists to activate the inactive NFS sensors.

The following are the corresponding event details and the proposed solution:

- **Event ID:** nfs_sensors_inactive
- **Problem:** The NFS performance cannot be monitored because the performance monitoring sensor NFSIO is inactive.
- **Solution:** Activate the SMB sensors.

The DMP provides the option to activate the NFS monitoring sensor and select a data collection interval that defines how frequently the sensors must collect data. It is recommended to select a value that is greater than or equal to 10 as the data collection frequency to reduce the impact on the system performance.

The system issues the **mmperfmon** command to activate the sensors as given in the following format:

```
/usr/lpp/mmfs/bin/mmperfmon config update NFSIO.restrict=cesNodes NFSIO.period=<<seconds>>
```

For example: `/usr/lpp/mmfs/bin/mmperfmon config update NFSIO.restrict=cesNodes NFSIO.period=10`

Activate SMB performance monitoring sensors

The activate SMB performance monitoring sensors DMP assists to activate the inactive SMB sensors.

The following are the corresponding event details and the proposed solution:

- **Event ID:** smb_sensors_inactive
- **Problem:** The SMB performance cannot be monitored because either one or both the SMBStats and SMBGlobalStats sensors are inactive.
- **Solution:** Activate the SMB sensors.

The DMP provides the option to activate the SMB monitoring sensor and select a data collection interval that defines how frequently the sensors must collect data. It is recommended to select a value that is greater than or equal to 10 as the data collection frequency to reduce the impact on the system performance.

The system issues the **mmperfmon** command to activate the sensors as given in the following format:

```
/usr/lpp/mmfs/bin/mmperfmon config update SMBStats.restrict=cesNodes SMBStats.period=<<seconds>>
```

For example: `/usr/lpp/mmfs/bin/mmperfmon config update SMBStats.restrict=cesNodes SMBStats.period=10`

Configure NFS sensors

The configure NFS sensor DMP assists you to configure NFS sensors.

The following are the details of the corresponding event:

- **Event ID:** nfs_sensors_not_configured
- **Problem:** The configuration details of the NFS sensor is not available in the sensor configuration.
- **Solution:** The sensor configuration is stored in a temporary file that is located at: `/var/lib/mmfs/gui/tmp/sensorDMP.txt`. The DMP provides options to enter the following details in the `sensorDMP.txt` file and later add them to the configuration by using the **mmperfmon config add** command.

Table 6. NFS sensor configuration example

Sensor	Restrict to nodes	Intervals	Contents of the sensorDMP.txt file

NFSIO	Node class - cesNodes	1, 5, 10, 15, 30
		Default value is 10.
			`sensors={`
			`name = "sensorName"`
			`period = period`
			`proxyCmd = "/opt/IBM/zimon/GaneshaProxy"`
			`restrict = "cesNodes"`
			`type = "Generic"`
			`}`

Only users with *ProtocolAdministrator, SystemAdministrator, SecurityAdministrator,* and *Administrator* roles can use this DMP to configure NFS sensor.

After you complete the steps in the DMP, refresh the configuration by issuing the following command:

```
/usr/lpp/mmfs/bin/mmhealth node show nfs --refresh -N cesNodes
```

Issue the **mmperfmon config show** command to verify whether the NFS sensor is configured properly.

Configure SMB sensors

The configure SMB sensor DMP assists you to configure SMB sensors.

The following are the details of the corresponding event:

- **Event ID:** smb_sensors_not_configured
- **Problem:** The configuration details of the SMB sensor is not available in the sensor configuration.
- **Solution:** The sensor configuration is stored in a temporary file that is located at: `/var/lib/mmfs/gui/tmp/sensorDMP.txt`. The DMP provides options to enter the following details in the `sensorDMP.txt` file and later add them to the configuration by using the **mmperfmon config add** command.

Table 7. SMB sensor configuration example			
Sensor	**Restrict to nodes**	**Intervals**	**Contents of the sensorDMP.txt file**

SMBStats
SMBGlobalStats
	Node class - cesNodes	1, 5, 10, 15, 30
		Default value is 10.
			`sensors={`
			`  name = "sensorName"`
			`  period = period`
			`  restrict = "cesNodes"`
			`  type = "Generic"`
			`}`

Only users with *ProtocolAdministrator, SystemAdministrator, SecurityAdministrator,* and *Administrator* roles can use this DMP to configure SMB sensor.

After you complete the steps in the DMP, refresh the configuration by issuing the following command:

```
/usr/lpp/mmfs/bin/mmhealth node show SMB --refresh -N cesNodes
```

Issue the **mmperfmon config show** command to verify whether the SMB sensor is configured properly.

Mount file system if it must be mounted

The mount file system DMP assists you to mount the file systems that must be mounted.

The following are the details of the corresponding event:

- **Event ID:** unmounted_fs_check
- **Problem:** A file system is assumed to be mounted all the time because it is configured to mount automatically, but the file system is currently not mounted on all nodes.
- **Solution:** Mount the file system on the node where it is not mounted.

Only users with *ProtocolAdministrator, SystemAdministrator, SecurityAdministrator,* and *Administrator* roles can use this DMP to mount the file systems on the required nodes.

If there is more than one instance of *unmounted_fs_check* event for the file system, you can choose whether to mount the file system on all nodes where it is not mounted but supposed to be mounted.

The DMP issues the following command for mounting the file system on one node:

```
mmmount Filesystem -N Node
```

The DMP issues the following command for mounting the file system on several nodes if automatic mount is not included:

```
mmmount Filesystem -N all
```

The DMP issues the following command for mounting the file system on certain nodes if automatic mount is not included in those nodes:

```
mmmount Filesystem -N Nodes (comma-separated list)
```

Note: Nodes where the file `/var/mmfs/etc/ignoreStartupMount.filesystem` or `/var/mmfs/etc/ignoreStartupMount` exists are excluded from automatic mount of this file system.

After running the **mmmount** command, the DMP waits until the *unmounted_fs_check* event disappear from the event list. If the *unmounted_fs_check* event does not get removed from the event list after 120 seconds, a warning message is displayed.

Start the GUI service on the remote nodes

You can start the GUI service on the remote nodes by using this DMP.

The following are the details of the corresponding event:

- **Event ID:** gui_down
- **Problem:** A GUI service is supposed to be running but it is down.
- **Solution:** Start the GUI service.
- **Limitation:** This DMP can only be used if GUI service is down on the remote nodes.

Only users with *ProtocolAdministrator, SystemAdministrator, SecurityAdministrator,* and *Administrator* roles can use this DMP to mount the file systems on the required nodes.

The DMP issues the **`systemctl restart gpfsgui`** command to start the GUI service on the remote node.

After running the **`mmmount`** command, the DMP waits until the *gui_down* event disappears from the event list. If the *gui_down* event does not get removed from the event list after 120 seconds, a warning message is displayed.

Maintenance procedures for NVMe and PCIe issues for ESS 3000

This section details the maintenance procedures for NVMe and PCIe issues.

Verify the status of the all attached NVMe devices using the **`mmlsnvmestatus`** command:

```
# mmlsnvmestatus all
                                                    Optimal     Optimal      needs
 node              NVMe device    serial number     Link State  LBA Format   service
 ------            -----------    -------------     ----------  ----------   -------
 ess3k3a.gpfs.net  /dev/nvme0     S43RNE0KC00112    YES         YES          NO
 ess3k3a.gpfs.net  /dev/nvme1     S43RNE0KC00109    YES         YES          NO
 ess3k3a.gpfs.net  /dev/nvme10    S43RNE0KC00052    YES         YES          NO
 ess3k3a.gpfs.net  /dev/nvme11    S43RNE0KC00042    YES         YES          NO
 ess3k3a.gpfs.net  /dev/nvme12    S43RNE0KC00045    YES         YES          NO
 ess3k3a.gpfs.net  /dev/nvme13    S43RNE0KC00047    YES         YES          NO
 ess3k3a.gpfs.net  /dev/nvme14    S43RNE0KC00148    YES         YES          NO
 ess3k3a.gpfs.net  /dev/nvme15    S43RNE0KC00041    YES         YES          NO
```

NVMe drive listing is not verified

Follow these steps to verify that the expected number of NVMe drives are listed:

1. Run the **`nvme list`** Linux command to query NVMe drives.
2. Verify that the expected number of drives is reported.

NVMe drives are missing from one or both I/O nodes

Follow these steps if the NVMe listing is done, but the listing displays no drives:

1. Validate that the PERST service, `systemctl status ess3k_perst.service`, is enabled and has run after boot.
2. If the PERST service is not enabled or does not exist, then reinstall the `gpfs.ess.platform.ess3k` rpm.

Note: You must reboot the canister if you reinstall `gpfs.ess.platform.ess3k` rpm.

PCIe initialization settings are not validated

Various PCIe-related settings like error-reporting settings are set by `ess3k_initpcie.service`. Follow these steps to validate that the PCIe initialization settings are enabled:

1. Validate that the `systemctl status ess3k_initpcie.service` service is enabled and has run after boot.
2. If the service is not enabled or does not exist, then reinstall the `gpfs.ess.platform.ess3k` rpm.

Unexpected kernel crashes due to PCIe or NVMe activities:

PCIe or NVMe activities like reset, power off, power on, and so on might cause unexpected kernel crash if the system is not set up correctly. If NVMe drives encounter PCIe fabric-related errors or resets, those events produce a fabric error interrupt, that must be handled by the PCIe fabric. However, if the fabric-handling infrastructure does not exist, it might result in a kernel crash and reboot. To prevent such issues, verify that the Linux native PCIe interrupt handler is enabled. For more information, see "Linux native PCIe interrupt handler validation and enablement for ESS 3000" on page 66.

Downstream port containment (DPC) bits are not clearing

ESS 3000 I/O nodes are DPC-enabled to provide isolation and containment of the PCIe-related issues for the NVMe drive endpoints. When an NVMe drive is removed or powered off, the PCIe fabric handles the event by performing a DPC. If the NVMe drive is reinserted or the slot is powered back on, and the NVMe drive does not show up again, it might be because the Linux native PCIe interrupt handler is not enabled. For more information, see "Linux native PCIe interrupt handler validation and enablement for ESS 3000" on page 66.

Linux native PCIe interrupt handler validation and enablement for ESS 3000

For the ESS 3000 I/O nodes, the native PCIe interrupt handler is enabled during the manufacturing phase and validated during the deployment phase.

However, if for some reason the enablement was removed, this section helps determine how to validate and enable it again.

1. To validate the PCIe native error handler, run the following query:

   ```
   cat /proc/cmdline | grep pcie_ports=native
   ```

   If the query comes back empty, then the PCIe native error handler must be enabled:
2. To enable the PCIe native error handler, follow these steps:
   a. Open the `/etc/default/grub` file for editing.
   b. Find the `GRUB_CMDLINE_LINUX` line.
   c. Append the text `pcie_ports=native` to the end of the `GRUB_CMDLINE_LINUX` line as shown:

   ```
   GRUB_CMDLINE_LINUX="nvme.sgl_threshold=0 sshd=1 noht crashkernel=auto
   resume=UUID=f0cccb47-da43-404d-a8f3-578129d3b8f7 rd.md.uuid=53d2b2a3:0c7532dd:72ba276b:179d3b74
   rd.md.uuid=519c1d9a:68fa26be:755637c7:9db5d8e4 rhgb quiet pcie_ports=native"
   ```

   d. Save and close the file.
   e. Make a new configuration with the updated grub file by running the following command:

   ```
   grub2-mkconfig -o /boot/efi/EFI/redhat/grub.cfg
   ```

   f. Reboot the server node.
   g. When the server is back up, validate that the handler is enabled by running the following query:

   ```
   cat /proc/cmdline | grep pcie_ports=native
   ```

PCIe-related data collection and debug for ESS 3000

This section details the PCIe-related data collection and debug processes that can be done live on a node.

You can get more information about the active issues on the ESS 3000 for both the NVMe drive availability and the PCIe-related issues. Follow these steps to determine the possible steps towards resolving these issues:

1. Run the following script to show the NVMe-related PCIe fabric:

```
lspci -tv |sed -n '/ +-\[0000:\(85\|3a\)\]/,/8546/p'
```

2. Run the following script to show the PCIe device link status for the NVMe drives:

```
for u in 87 3c; do for i in $(seq 0 11); do d=$(printf "%02x" $i); lspci -vvs
$u:$d.0; done; done | grep -E "^[0-9a-f]|LnkSta:|Bus:" | sed "/^[0-9a-f]/{s/ .*//;N;s/, sec-
latency.*//;N;s/, TrErr.*//;s/\n//g;}"
```

3. Run the following script to show the Downstream Port Containment (DPC) status for the NVMe drives:

```
for u in 87 3c; do for i in $(seq 0 11); do d=$(printf "%02x" $i); echo -n "$u:$d.0:
"; lw1="0x"$(setpci -s $u:$d.0 0x1b4.l); lw2="0x"$(setpci -s $u:$d.0 0x1b8.l); echo "$lw1
$lw2";done; done
```

Note: If DPC is enabled for a particular PCIe port, observe a nonzero value in the rightmost column.

Detecting faulty DIMMs to solve canister boot issues for ESS 3000

A faulty DIMM in the ESS 3000 can prevent the server canister from booting up, and nothing is displayed on the VGA port. The BIOS is not shown via the VGA port.

In such cases, removing all but two DIMMs per CPU module might allow the system to function. If removing the DIMMs work, then the customer can replace the DIMMs to resolve the issue.

Follow these steps to find if the DIMMs are faulty:

1. Remove all of the DIMMs except A0/D0 slots of each CPU.

   Note: A minimum of 2 DIMMs is installed in A0/D0 slots of each CPU, therefore there are 4 DIMMs in total.

2. Power on the canister.

   a. If the VGA displays the BIOS within 4 minutes: The canister is operational. One or more of the DIMMs are defective, and need to be replaced. This procedure is complete.

   b. If the VGA does not display the BIOS within 4 minutes: Repeat steps 1 and 2 until the VGA displays the BIOS. If all of the DIMMs have been tried and the VGA never displayed the BIOS, then the DIMMs are not faulty, and do not need to be replaced. It is likely the canister FRU must be replaced.

Note: You require a crash-cart for this process.

DIMM locations and memory configurations

As the following image shows, each server canister contains two processors, which are identified as CPU1 (CPU_SrcID#0) CPU 1 and CPU2 (CPU_SrcID#1) CPU 2.

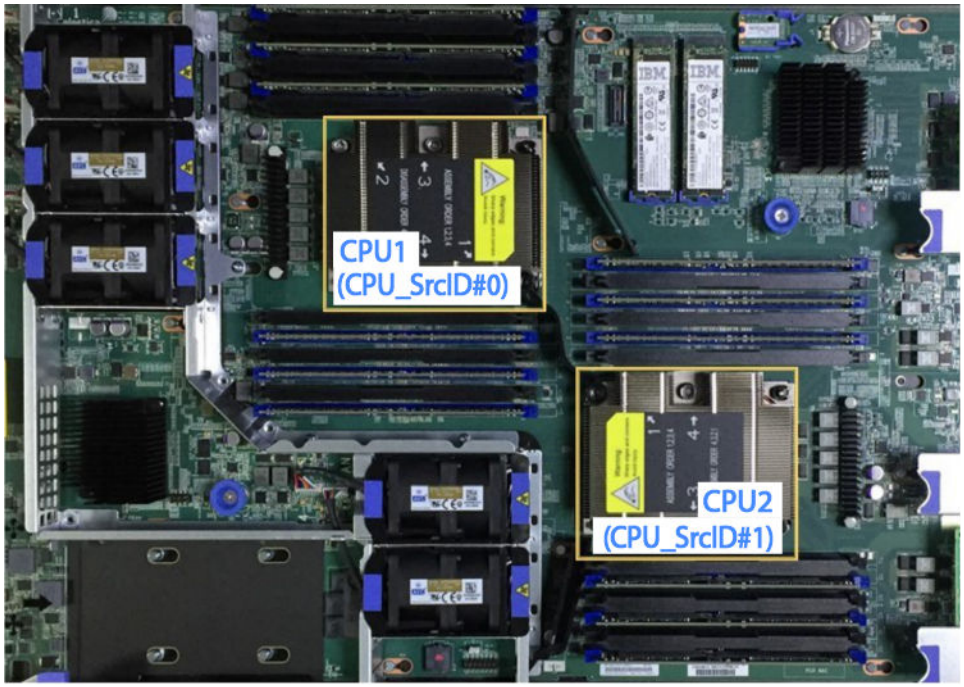

Figure 1. Location of CPUs and DIMM slots

A CPU processor has six memory channels, which are labeled A-F. Each memory channel has 2 DIMM slots, numbered 0-1. For example, DIMM slots A0 and A1 are in memory channel A. On the system board, the DIMM slots are labeled according to their memory channel and slot. They are associated with the CPU nearest to their DIMM slots.

The following table gives information about the DIMM locations and respective memory configuration:

Table 8. DIMM locations and memory configurations

DIMM slot	Slot index number	Configuration of DIMMs and blanks	
C0	5	32 GB	32 GB
C1	6		32 GB
B0	3	32 GB	32 GB
B1	4		32 GB
A0	1	32 GB	32 GB
A1	2		32 GB
CPU 1			
D1	8		32 GB
D0	7	32 GB	32 GB
E1	10		32 GB
E0	9	32 GB	32 GB
F1	12		32 GB
F0	11	32 GB	32 GB
F0	23	32 GB	32 GB
F1	24		32 GB

Table 8. DIMM locations and memory configurations (continued)

DIMM slot	Slot index number	Configuration of DIMMs and blanks	
E0	21	32 GB	32 GB
E1	22		32 GB
D0	19	32 GB	32 GB
D1	20		32 GB
CPU 2			
A1	14		32 GB
A0	13	32 GB	32 GB
B1	16		32 GB
B0	15	32 GB	32 GB
C1	18		32 GB
C0	17	32 GB	32 GB
Total memory			
Server canister		384 GB	768 GB
Control enclosure		768 GB	1536 GB

Chapter 9. References

The IBM Elastic Storage System displays a warning or error message when it encounters an issue that needs user attention. The message severity tags indicate the severity of the issue

Events

The recorded events are stored in the local database on each node. The user can get a list of recorded events by using the **mmhealth node eventlog** command. Users can use the **mmhealth node show** or **mmhealth cluster show** commands to display the active events in the node and cluster respectively.

The recorded events can also be displayed through the GUI.

When you upgrade to IBM Spectrum Scale 5.0.5.3 or a later version, the nodes where no `sqlite3` package is installed have their RAS event logs converted to a new database format to prevent known issues. The old RAS event log is emptied automatically. You can verify that the event log is emptied either by using the **mmhealth node eventlog** command or in the IBM Spectrum Scale GUI.

Note: The event logs are updated only the first time IBM Spectrum Scale is upgraded to version 5.0.5.3 or higher.

The following sections list the RAS events that are applicable to various components of the IBM Spectrum Scale system:

Array events

The following table lists the events that are created for the *Array* component.

Table 9. Events for the Array component

Event	Event Type	Severity	Message	Description	Cause	User Action
gnr_array_found	INFO_ADD_ENTITY	INFO	A GNR declustered array {0} was found.	A GNR declustered array that is listed in the IBM Spectrum Scale configuration was detected.	N/A	N/A
gnr_array_needs service	STATE_CHANGE	WARNING	A GNR declustered array {0} needs service.	The declustered array state needs service.	N/A	N/A
gnr_array_ok	STATE_CHANGE	INFO	A GNR declustered array {0} is OK.	The declustered array state is OK.	N/A	N/A
gnr_array_unknown	STATE_CHANGE	WARNING	A GNR declustered array {0} is in an unknown state.	The declustered array state is unknown.	N/A	N/A
gnr_array_vanished	INFO_DELETE_ENTITY	INFO	A GNR declustered array {0} vanished.	A GNR declustered array that is listed in the IBM Spectrum Scale configuration was not detected.	A GNR declustered array that is listed in the IBM Spectrum Scale configuration, which was mounted before, is not found. This condition can be a valid situation.	Run the **mmlsrecoverygroup** command to verify that all expected GNR declustered arrays exist.

Table 9. Events for the Array component (continued)						
Event	**Event Type**	**Severity**	**Message**	**Description**	**Cause**	**User Action**
gnr_da_out_of_space	STATE_CHANGE	ERROR	The GNR declustered array {0} is reporting zero disk space is available.	A GNR declustered array has zero free disk space remaining.	A GNR declustered array is being actively filled with data and now has zero free disk space remaining.	Inspect the available disk space of the GNR declustered array to remove unused files. Check the **mmfs.log.latest** log for any warning messages.
gnr_da_space_healthy	STATE_CHANGE	INFO	The GNR declustered array {0} now has no disk space issues.	A GNR declustered array that previously reported disk space issues is now reporting no disk space issues.	N/A	N/A
gnr_da_space_critical	STATE_CHANGE	WARNING	The GNR declustered array {0} has reached its critical free disk space threshold.	A GNR declustered array has reached its critical free disk space threshold.	A GNR declustered array is being actively filled with data and has reached its critical free disk space threshold.	Inspect the available disk space of the GNR declustered array to remove unused files. Check the **mmfs.log.latest** log for any warning messages.
gnr_da_space_low	STATE_CHANGE	WARNING	The GNR declustered array {0} has reached its low free disk space threshold.	A GNR declustered array has reached its low free disk space threshold.	A GNR declustered array is being actively filled with data and has reached its low free disk space threshold.	Inspect the available disk space of the GNR declustered array to remove unused files. Check the **mmfs.log.latest** log for any warning messages.

Enclosure events

The following table lists the events that are created for the *Enclosure* component.

Table 10. Events for the Enclosure component						
Event	**Event Type**	**Severity**	**Message**	**Description**	**Cause**	**User Action**
adapter_bios_notavail	STATE_CHANGE	WARNING	The BIOS level of adapter {0} is not available.	The BIOS level of the adapter is not available.	N/A	Check the installed bios level by using the **mmlsfirmware** command.
adapter_bios_ok	STATE_CHANGE	INFO	The BIOS level of adapter {0} is correct.	The BIOS level of the adapter is correct.	N/A	N/A
adapter_bios_wrong	STATE_CHANGE	WARNING	The BIOS level of adapter {0} is wrong.	The BIOS level of the adapter is wrong.	N/A	Check the installed bios level by using the **mmlsfirmware** command.
adapter_firmware_notavail	STATE_CHANGE	WARNING	The firmware level of adapter {0} is not available.	The firmware level of the adapter is not available.	N/A	Check the installed bios level by using the **mmlsfirmware** command.
adapter_firmware_ok	STATE_CHANGE	INFO	The firmware level of adapter {0} is correct.	The firmware level of the adapter is correct.	N/A	N/A
adapter_firmware_wrong	STATE_CHANGE	WARNING	The firmware level of adapter {0} is wrong.	The firmware level of the adapter is wrong.	N/A	Check the installed bios level by using the **mmlsfirmware** command.

Table 10. Events for the Enclosure component (continued)

Event	Event Type	Severity	Message	Description	Cause	User Action
current_failed	STATE_CHANGE	ERROR	The currentSensor {0} failed.	The current sensor state failed.	The current sensors measured wrong current.	Contact IBM support.
current_ok	STATE_CHANGE	INFO	The currentSensor {0} is OK.	The currentSensor state is OK.	N/A	N/A
current_warn	STATE_CHANGE	WARNING	The currentSensor {0} has degraded.	The currentSensor state has degraded.	N/A	N/A
dcm_drawer_open	STATE_CHANGE	WARNING	DCM {0} drawer is open.	The DCM drawer is open.	N/A	N/A
dcm_failed	STATE_CHANGE	WARNING	DCM {0} failed.	The DCM state failed.	N/A	N/A
dcm_not_available	STATE_CHANGE	WARNING	DCM {0} is not available.	The DCM is not installed or not responding.	N/A	N/A
dcm_ok	STATE_CHANGE	INFO	DCM {ID[1]} is OK.	The DCM state is OK.	N/A	N/A
drawer_failed	STATE_CHANGE	ERROR	The drawer {0} failed.	The drawer state failed.	N/A	N/A
drawer_ok	STATE_CHANGE	INFO	The drawer {0} is OK.	The drawer state is OK.	N/A	N/A
drive_firmware_notavail	STATE_CHANGE	WARNING	The firmware level of drive {0} is not available.	The firmware level of the drive is not available.	N/A	Check the installed firmware level by using the **mmlsfirmware** command.
drive_firmware_ok	STATE_CHANGE	INFO	The firmware level of drive {0} is correct.	The firmware level of the drive is correct.	N/A	N/A
drive_firmware_wrong	STATE_CHANGE	WARNING	The firmware level of drive {0} is wrong.	The firmware level of the drive is wrong.	N/A	Check the installed firmware level by using the **mmlsfirmware** command.
enclosure_data	STATE_CHANGE	INFO	Enclosure data is found.	Successfully queried the enclosure details.	The **mmlsenclosure all -L -Y** command reports enclosure data.	N/A
enclosure_firmware_notavail	STATE_CHANGE	WARNING	The firmware level of enclosure {0} is not available.	The firmware level of the enclosure is not available.	N/A	Check the installed firmware level by using the **mmlsfirmware** command.
enclosure_firmware_ok	STATE_CHANGE	INFO	The firmware level of enclosure {0} is correct.	The firmware level of the enclosure is correct.	N/A	N/A

Table 10. Events for the Enclosure component (continued)

Event	Event Type	Severity	Message	Description	Cause	User Action
enclosure_firmware_unknown	STATE_CHANGE	WARNING	The firmware level of enclosure {0} is unknown.	The SAS card is unable to read enclosure firmware.	The SAS card does not report the enclosure firmware.	Check the SAS connectivity from node to enclosure. Run the **mmlsrecoverygroup <rg_name> -L --pdisk** command to verify whether all the paths to pdisk are available. Check the SAS connectivity by using a combination of the `mmgetpdisktopology` and the **topsummary** command. If an issue is found with the SAS HBA or SAS cable, then restart the node to check whether this action resolves the issue. Otherwise, contact your IBM representative.
enclosure_firmware_wrong	STATE_CHANGE	WARNING	The firmware level of enclosure {0} is wrong.	The firmware level of the enclosure is wrong.	N/A	Check the installed firmware level by using the **mmlsfirmware** command.
enclosure_found	INFO_ADD_ENTITY	INFO	Enclosure {0} was found.	A GNR enclosure that is listed in the IBM Spectrum Scale configuration was detected.	N/A	N/A
enclosure_needsservice	STATE_CHANGE	WARNING	Enclosure {0} needs service.	The enclosure needs service.	The **mmlsenclosure all -L** command reports that enclosure needs service.	Contact IBM support.
enclosure_ok	STATE_CHANGE	INFO	Enclosure {0} is OK.	The enclosure state is OK.	N/A	N/A
enclosure_unknown	STATE_CHANGE	WARNING	Enclosure state {0} is unknown.	The enclosure state is unknown.	N/A	N/A
enclosure_vanished	INFO_DELETE_ENTITY	INFO	Enclosure {0} vanished.	A GNR enclosure that is listed in the IBM Spectrum Scale configuration was not detected.	A GNR enclosure, which is listed in the IBM Spectrum Scale configuration as mounted before, is not found. This condition can be a valid situation.	Run the **mmlsenclosure** command to verify that all expected enclosures exist.
esm_absent	STATE_CHANGE	WARNING	ESM {0} is absent.	The ESM state is not installed.	N/A	N/A
esm_failed	STATE_CHANGE	WARNING	ESM {0} failed.	The ESM state failed.	N/A	N/A
esm_ok	STATE_CHANGE	INFO	ESM {0} is OK.	The ESM state is OK.	N/A	N/A
expander_absent	STATE_CHANGE	WARNING	The expander {0} is absent.	The expander is absent.	N/A	N/A
expander_failed	STATE_CHANGE	ERROR	The expander {0} failed.	The expander state failed.	N/A	N/A
expander_ok	STATE_CHANGE	INFO	The expander {0} is OK.	The expander state is OK.	N/A	N/A

Table 10. Events for the Enclosure component (continued)

Event	Event Type	Severity	Message	Description	Cause	User Action
fan_failed	STATE_CHANGE	WARNING	Fan {0} failed.	The fan state failed.	The **mmlsensclosure all -L** command reports component fan as failed.	Contact IBM support for service action.
fan_ok	STATE_CHANGE	INFO	Fan {0} is OK.	The fan state is OK.	N/A	N/A
fan_speed_high	STATE_CHANGE	WARNING	Fan {0} speed is too high.	The fan speed is out of the tolerance range.	N/A	Check the enclosure cooling module LEDs for fan faults.
fan_speed_low	STATE_CHANGE	WARNING	Fan {0} speed is too low.	The fan speed is out of the tolerance range.	N/A	Check the enclosure cooling module LEDs for fan faults.
no_enclosure_data	STATE_CHANGE	WARNING	Enclosure data and state information cannot be queried.	Cannot query the enclosure details. State reporting for all enclosures and canisters are incorrect.	The **mmlsenclosure all -L -Y** command fails to report any enclosure data.	Run the **mmlsenclosure** command to check for errors. Run the **lsmod** command to verify that the pemsmod is loaded.
power_high_current	STATE_CHANGE	WARNING	Power supply {0} reports high current.	The DC power supply current is greater than the threshold.	N/A	N/A
power_high_voltage	STATE_CHANGE	WARNING	Power supply {0} reports high-voltage value.	The DC power supply voltage is greater than the threshold.	N/A	N/A
power_no_power	STATE_CHANGE	WARNING	Power supply {0} has no power.	Power supply has no input AC power. The power supply might be turned off or disconnected from the AC supply.	N/A	N/A
power_supply_absent	STATE_CHANGE	WARNING	Power supply {0} is missing.	The power supply is missing.	N/A	N/A
power_supply_failed	STATE_CHANGE	WARNING	Power supply {0} failed.	The power supply state failed.	The **mmlsenclosure all -L** command reported that a power supply failed.	For more information, see the *IBM Spectrum Scale: Problem Determination Guide*.
power_supply_off	STATE_CHANGE	WARNING	Power supply {0} is off.	The power supply is not providing power.	N/A	N/A
power_supply_ok	STATE_CHANGE	INFO	Power supply {0} is OK.	The power supply state is OK.	N/A	N/A
power_switched_off	STATE_CHANGE	WARNING	Power supply {0} is switched off.	The requested on bit is off, indicating that the power supply is not manually turned on or been requested to turn on by setting the requested on bit.	N/A	N/A
sideplane_failed	STATE_CHANGE	ERROR	The side plane {0} failed.	The side plane state failed.	N/A	N/A
sideplane_ok	STATE_CHANGE	INFO	The side plane {0} is OK.	The side plane state is OK.	N/A	N/A
temp_bus_failed	STATE_CHANGE	WARNING	Temperature sensor {0} I2C bus failed.	The temperature sensor I2C bus failed.	N/A	N/A

Table 10. Events for the Enclosure component (continued)

Event	Event Type	Severity	Message	Description	Cause	User Action
temp_high_critical	STATE_CHANGE	WARNING	Temperature sensor {0} measured a high temperature value.	The temperature exceeded the actual high critical threshold value for at least one sensor.	N/A	N/A
temp_high_warn	STATE_CHANGE	WARNING	Temperature sensor {0} measured a high temperature value.	The temperature exceeded the actual high warning threshold value for at least one sensor.	N/A	N/A
temp_low_critical	STATE_CHANGE	WARNING	Temperature sensor {0} measured a low temperature value.	The temperature value dropped less than the actual low critical threshold value for at least one sensor.	N/A	N/A
temp_low_warn	STATE_CHANGE	WARNING	Temperature sensor {0} measured a low temperature value.	The temperature value dropped less than the actual low warning threshold value for at least one sensor.	N/A	N/A
temp_sensor_failed	STATE_CHANGE	WARNING	Temperature sensor {0} failed.	The temperature sensor state failed.	N/A	N/A
temp_sensor_ok	STATE_CHANGE	INFO	Temperature sensor {0} is OK.	The temperature sensor state is OK.	N/A	N/A
voltage_bus_failed	STATE_CHANGE	WARNING	Voltage sensor {0} I2C bus failed.	The voltage sensor I2C bus failed.	N/A	N/A
voltage_high_critical	STATE_CHANGE	WARNING	Voltage sensor {0} measured a high-voltage value.	The voltage exceeded the actual high critical threshold value for at least one sensor.	N/A	N/A
voltage_high_warn	STATE_CHANGE	WARNING	Voltage sensor {0} measured a high-voltage value.	The voltage exceeded the actual high warning threshold value for at least one sensor.	N/A	N/A
voltage_low_critical	STATE_CHANGE	WARNING	Voltage sensor {0} measured a low voltage value.	The voltage dropped to less than the actual low critical threshold value for at least one sensor.	The voltage dropped less than the actual low critical threshold value for at least one sensor.	Check the power supply and cabling for connectivity and power.
voltage_low_warn	STATE_CHANGE	WARNING	Voltage sensor {0} measured a low voltage value.	The voltage dropped to less than the actual low warning threshold value for at least one sensor.	N/A	N/A
voltage_sensor_failed	STATE_CHANGE	WARNING	Voltage sensor {0} failed.	The voltage sensor state failed.	N/A	N/A
voltage_sensor_ok	STATE_CHANGE	INFO	Voltage sensor {0} is OK.	The voltage sensor state is OK.	N/A	N/A

Virtual disk events

The following table lists the events that are created for the *Virtual disk* component.

Table 11. Events for the virtual disk component

Event	Event Type	Severity	Message	Description	Cause	User Action
gnr_vdisk_critical	STATE_CHANGE	ERROR	GNR vdisk {0} is critically degraded.	The vdisk state is critically degraded.	N/A	N/A

Table 11. Events for the virtual disk component (continued)

Event	Event Type	Severity	Message	Description	Cause	User Action
gnr_vdisk_degraded	STATE_CHANGE	WARNING	GNR vdisk {0} is degraded.	The vdisk state is degraded.	N/A	N/A
gnr_vdisk_found	INFO_ADD _ENTITY	INFO	GNR vdisk {0} is found.	A GNR vdisk, which is listed in the IBM Spectrum Scale configuration, is detected.	N/A	N/A
gnr_vdisk_offline	STATE_CHANGE	ERROR	GNR vdisk {0} is offline.	The vdisk state is offline.	N/A	N/A
gnr_vdisk_ok	STATE_CHANGE	INFO	GNR vdisk {0} is OK.	The vdisk state is OK.	N/A	N/A
gnr_vdisk_unknown	STATE_CHANGE	WARNING	GNR vdisk {0} is unknown.	The vdisk state is unknown.	N/A	N/A
gnr_vdisk_vanished	INFO_DELETE _ENTITY	INFO	GNR vdisk {0} vanished.	A GNR vdisk listed in the IBM Spectrum Scale configuration was not detected.	A GNR vdisk, which is listed in the IBM Spectrum Scale configuration as mounted before, is not found. This condition can be a valid situation.	Run the **mmlsvdisk** command to verify that all expected GNR vdisk exist.

Physical disk events

The following table lists the events that are created for the *Physical disk* component.

Table 12. Events for the physical disk component

Event	Event Type	Severity	Message	Description	Cause	User Action
gnr_nvram_degraded	STATE_CHANGE	WARNING	The NVDIMM of the pdisk {0} is degraded.	The NVRAM drive of the disk is in degraded state.	The **tslsnvramstatus** command shows degraded state for the NVRAM drive of the disk.	N/A
gnr_nvram_disarmed	STATE_CHANGE	ERROR	The NVDIMM of the pdisk {0} is disarmed.	NVDIMM is unable to preserve future content.	The **tslsnvramstatus** command reports disarmed failure condition for the NVRAM drive of the disk.	Identify the NVDIMM cards or BPM, which encountered the errors from FSP log or call home data, and replace the faulty NVDIMM cards, BPM or both as soon as possible.
gnr_nvram_erased	STATE_CHANGE	ERROR	The NVDIMM of the pdisk {0} reports erased image.	Image erased. The NVDIMM contents not persisted.	The **tslsnvramstatus** command reports the erased-failure condition for the NVRAM drive of the disk.	Verify that any NVDIMM cards, BPM encountered any errors from FSP log or call home data. If any errors are found then replace the faulty NVDIMM cards, BPM or both as soon as possible. If no errors are found then try to add the drive back to RG.
gnr_nvram_error	STATE_CHANGE	ERROR	The NVDIMM of the pdisk {0} is failed.	The NVRAM drive of the disk is in error state.	The **tslsnvramstatus** command shows failed state for the NVRAM drive of the disk.	N/A
gnr_nvram_ok	STATE_CHANGE	INFO	The NVDIMM of the pdisk {0} is normal.	NVDIMM is in good condition.	N/A	N/A

Table 12. Events for the physical disk component (continued)

Event	Event Type	Severity	Message	Description	Cause	User Action
gnr_nvram_persist_error	STATE_CHANGE	ERROR	The NVDIMM of the pdisk {0} might not persist.	NVDIMM failed to save or restore memory contents.	The **tslsnvramstatus** command reports the fail-to-persist failure condition for the NVRAM drive of the disk.	Identify the NVDIMM cards or BPM, which encountered the errors from FSP log or call home data. Replace the faulty NVDIMM cards, BPM, or both as soon as possible.
gnr_nvram_unhealthy	STATE_CHANGE	WARNING	The NVDIMM of the pdisk {0} is unhealthy.	Error is detected but save or restore might still work for the NVRAM drive of the disk.	The **tslsnvramstatus** command reports unhealthy failure condition for the NVRAM drive of the disk.	Identify the NVDIMM cards or BPM, which has encountered the errors from FSP log or call home data. Replace the faulty NVDIMM cards, BPM or both as soon as possible.
gnr_pdisk_degraded	WARNING	WARNING	GNR pdisk {0} is degraded.	The pdisk state is degraded.	The **mmlspdisk** command reports degraded user condition for the disk.	The IOA cache battery might have failed. For more information, issue the **mmlspdisk** command or see the *Disk diagnosis* subsection under the *Maintenance procedure* section in the *IBM Spectrum Scale: Problem Determination Guide*.
gnr_pdisk_diagnosing	INFO	WARNING	GNR pdisk {0} diagnose runs into a timeout.	The system has started and pdisk is now in the diagnosing state.	A disk error or timeout (read / write) has occurred, as can be seen in the output of the **mmvdisk pdisk list --smart-data --recovery-group <recovery-group> -L** command.	For more information, see the *Disk diagnosis* subsection under the *Maintenance procedure* section in the *IBM Spectrum Scale: Problem Determination Guide*.
gnr_pdisk_draining	STATE_CHANGE	ERROR	GNR pdisk {0} is draining.	The pdisk state is draining. The data is being drained from the disk and moved to distributed spare space on other disks.	The **mmlspdisk** command shows draining user condition for the disk.	Wait for the draining process to finish.
gnr_pdisk_disks	STATE_CHANGE	INFO	Pdisks are detected on this node.	Pdisks found.	N/A	N/A
gnr_pdisk_found	INFO_ADD_ENTITY	INFO	GNR pdisk {0} was found.	A GNR pdisk, which is listed in the IBM Spectrum Scale configuration, was detected.	N/A	N/A
gnr_pdisk_maintenance	STATE_CHANGE	WARNING	GNR pdisk {0} is in maintenance.	The GNR pdisk is in maintenance because the state is suspended, serviceDrain, pathMaintenance, or deleting. This might be caused by some administration commands like the **mmdeldisk** command.	The **mmlspdisk** command shows maintenance user condition for the disk.	Complete the maintenance action. If the issue persists, then contact IBM support.

Table 12. Events for the physical disk component (continued)

Event	Event Type	Severity	Message	Description	Cause	User Action
gnr_pdisk_missing	STATE_CHANGE	WARNING	GNR pdisk {0} is missing.	The pdisk state is missing.	Native RAID has lost connectivity to the drive and further analysis is needed to find the root cause of the failure. There might be several technical reasons, which cannot be resolved by replacing the disk.	Check whether the drive is correctly seated in the socket and is getting power. Also, check whether there are SAS errors in the logs. If other drives are also missing, then check whether there is a common failure domain (such as enclosure powered off or disconnected cables). In rare cases the drive might be dead and must be replaced.
gnr_pdisk_ needanalysis	STATE_CHANGE	ERROR	GNR pdisk {0} needs analysis.	The GNR pdisk has a problem that has to be analyzed and solved by an expert.	The **mmlspdisk** command shows `attention` user condition for the disk.	If the issue persists, then contact IBM support.
gnr_pdisk_nodisks	STATE_CHANGE	INFO	No pdisks found on this node.	No pdisks found, but some pdisks are expected on recovery group nodes.	N/A	N/A
gnr_pdisk_ok	STATE_CHANGE	INFO	GNR pdisk {0} is OK.	The pdisk state is OK.	N/A	N/A
gnr_pdisk_ replaceable	STATE_CHANGE	ERROR	GNR pdisk {0} is replaceable.	The pdisk is ready to be replaced, which means that all the data is drained out of the disk.	The **mmlspdisk** command shows replaceable user condition for the disk.	Replace the pdisk.
gnr_pdisk_sedlocked	STATE_CHANGE	ERROR	GNR pdisk {0}, which is a self-encrypting drive, is locked.	A self-encrypting drive, which has encryption enabled, is locked. GNR does not have access to any data on the drive.	The **mmlspdisk** command shows that the pdisk state contains `sedLocked`.	The drive must be unlocked to be used by GNR.
gnr_pdisk_server_ down	STATE_CHANGE	ERROR	GNR server {0}, responsible for pdisk {1}, is unresponsive and hence causing the pdisk to be unavailable.	The recovery group server node, which is responsible for this pdisk, is reported as unresponsive. This causes this pdisk to be unavailable to the recovery group.	The recovery group server node, which is responsible for this pdisk, is down or unresponsive.	Determine the health of the recovery group server nodes and resolve any health issues that are found.
gnr_pdisk_server_ up	STATE_CHANGE	INFO	GNR server {0} responsible for pdisk {1} is active.	The recovery group server node, which is responsible for this pdisk and previously reported as unresponsive, is now active.	N/A	N/A
gnr_pdisk_unknown	STATE_CHANGE	WARNING	GNR pdisks are in unknown state.	The pdisk state is unknown.	The pdisk state, which was previously known, is now unknown.	Check the **mmlspdisk** command output for errors and verify that the pdisk states are correct.
gnr_pdisk_vanished	INFO_DELETE _ENTITY	INFO	GNR pdisk {0} has vanished.	A GNR pdisk, which was previously listed in the IBM Spectrum Scale configuration, was not detected.	N/A	N/A

Table 12. Events for the physical disk component (continued)

Event	Event Type	Severity	Message	Description	Cause	User Action
gnr_pdisk_vwce	STATE_CHANGE	ERROR	GNR pdisk {0} has volatile write cache enabled.	Volatile write cache is enabled on the drive. The writes, that are already committed, can be lost in case of a power loss. GNR only reads from this disk, does not write to it.	The **mmlspdisk** command shows that the pdisk state contains VWCE.	Check the reason behind enabling the volatile write cache. For example, a new drive was added with wrong defaults or wrong UDEV rules. Fix the modes by using the **sg_wr_modes** command.
gnr_pdisk_wcache_disabled	STATE_CHANGE	INFO	GNR pdisk {0} has write cache disabled.	Write cache is disabled for this pdisk, which is the recommended state.	N/A	N/A
gnr_pdisk_wcache_enabled	STATE_CHANGE	WARNING	GNR pdisk {0} has write cache enabled.	Write cache is enabled for this pdisk, a potential data loss is possible in case of power loss.	The **mmvdisk pdisk list --smart-data --recovery-group all -L -Y** command shows that write cache is enabled for the this disk.	Disable the write cache for the pdisk. For example, use the **sdparm --set WCE=0 -s <devicename>** command.
ssd_endurance_ok	STATE_CHANGE	INFO	The ssd-endurance-percentage of GNR pdisk {0} is OK.	The ssdEndurance Percentage value is OK.	N/A	N/A
ssd_endurance_warn	STATE_CHANGE	WARNING	The ssdEndurance Percentage of GNR pdisk {0} is on a warning value.	The ssdEndurance Percentage value is a warning value.	The ssdEndurance Percentage value of the pdisk is in the range between 95 and 100.	SSDs have a finite lifetime based on the number of drive writes per day. The ssd-endurance-percentage values are actually reported as a number between 0 and 255. This value indicates the percentage of life that is used by the drive. The value 0 indicates that full life remains, and 100 indicates that the drive is at or past its end of life. The drive must be replaced when the value exceeds 100.

Recovery group events

The following table lists the events that are created for the *Recovery group* component.

Table 13. Events for the Recovery group component

Event	Event Type	Severity	Message	Description	Cause	User Action
gnr_rg_failed	STATE_CHANGE	ERROR	GNR recovery group {0} is not active.	A configured recovery group is not listed as active.	The **mmlsrecoverygroup** command reports that a recovery group is configured but not listed.	Examine the health of the recovery group server node and resolve any found health issues. Issue the **mmlsrecoverygroup** command to verify your modifications.
gnr_rg_found	INFO_ADD_ENTITY	INFO	GNR recovery group {0} is found.	A GNR recovery group, which was previously listed in the IBM Spectrum Scale configuration, was detected.	N/A	N/A
gnr_rg_ok	STATE_CHANGE	INFO	GNR recovery group {0} is OK.	The recovery group is OK.	N/A	N/A
gnr_rg_server_down	STATE_CHANGE	WARNING	GNR recovery group server {0} in resource group {1} is unresponsive.	The server node in this recovery group is reported as unresponsive.	The server node in this recovery group is down or unresponsive.	Examine the health of the recovery group server node and resolve any found health issues.
gnr_rg_server_up	STATE_CHANGE	INFO	GNR recovery group server {0} in resource group {1} is active.	The recovery group server node, which was previously reported as unresponsive, is now active.	N/A	N/A

Table 13. Events for the Recovery group component (continued)						
Event	**Event Type**	**Severity**	**Message**	**Description**	**Cause**	**User Action**
gnr_rg_vanished	INFO_DELETE_ENTITY	INFO	GNR recovery group {0} vanished.	A GNR recovery group, which was previously listed in the IBM Spectrum Scale configuration, is not detected.	A GNR recovery group, which was previously listed in the IBM Spectrum Scale configuration, is no longer found. This can be a valid situation.	Run the **mmlsrecoverygroup** command to verify that all expected GNR recovery groups exist.

Server events

The following table lists the events that are created for the *Server* component.

Server events

Table 14. Server events						
Event	**Event Type**	**Severity**	**Message**	**Description**	**Cause**	**User Action**
cpu_peci_failed	STATE_CHANGE	ERROR	PECI state of CPU {0} failed.	The GUI checks the hardware state by using xCAT.	The hardware part failed.	N/A
cpu_peci_ok	STATE_CHANGE	INFO	PECI state of CPU {0} is OK.	The GUI checks the hardware state by using xCAT.	The hardware part is OK.	N/A
cpu_qpi_link_ok	STATE_CHANGE	INFO	QPI Link of CPU {0} is OK.	The GUI checks the hardware state by using xCAT.	The hardware part is OK.	N/A
cpu_qpi_link_failed	STATE_CHANGE	ERROR	QPI Link of CPU {0} is failed.	The GUI checks the hardware state by using xCAT.	The hardware part failed.	N/A
cpu_temperature_failed	STATE_CHANGE	ERROR	CPU {0} temperature is failed ({0}).	The GUI checks the hardware state by using xCAT.	The hardware part failed.	N/A
cpu_temperature_ok	STATE_CHANGE	INFO	CPU {0} temperature is normal ({1}).	The GUI checks the hardware state by using xCAT.	The hardware part is OK.	N/A

Table 14. Server events (continued)

Event	Event Type	Severity	Message	Description	Cause	User Action
dasd_backplane_failed	STATE_CHANGE	ERROR	DASD Backplane {0} failed.	The GUI checks the hardware state by using xCAT.	The hardware part failed.	N/A
dasd_backplane_ok	STATE_CHANGE	INFO	DASD Backplane {0} is OK.	The GUI checks the hardware state by using xCAT.	The hardware part is OK.	N/A
dimm_failed	STATE_CHANGE	ERROR	DIMM {0} failed.	The GUI checks the hardware state by using xCAT.	The hardware part failed.	N/A
dimm_ok	STATE_CHANGE	INFO	DIMM {0} is OK.	The GUI checks the hardware state by using xCAT.	The hardware part is OK.	N/A
drive_failed	STATE_CHANGE	ERROR	Drive {0} failed.	The GUI checks the hardware state by using xCAT.	The hardware part failed.	N/A
drive_ok	STATE_CHANGE	INFO	Drive {0} is OK.	The GUI checks the hardware state by using xCAT.	The hardware part is OK.	N/A
fan_zone_failed	STATE_CHANGE	ERROR	Fan Zone {0} failed.	The GUI checks the hardware state by using xCAT.	The hardware part failed.	N/A
fan_zone_ok	STATE_CHANGE	INFO	Fan Zone {0} is OK.	The GUI checks the hardware state by using xCAT.	The hardware part is OK.	N/A
hmc_event	STATE_CHANGE	INFO	HMC Event: {1}	The GUI collects events that are raised by the HMC.	An event from the HMC arrived.	N/A
pci_failed	STATE_CHANGE	ERROR	PCI {0} failed.	The GUI checks the hardware state by using xCAT.	The hardware part failed.	N/A
pci_ok	STATE_CHANGE	INFO	PCI {0} is OK.	The GUI checks the hardware state by using xCAT.	The hardware part is OK.	N/A
pci_riser_temp_failed	STATE_CHANGE	ERROR	The temperature of PCI Riser {0} is too high. ({1})	The GUI checks the hardware state by using xCAT.	The hardware part failed.	N/A
pci_riser_temp_ok	STATE_CHANGE	INFO	The temperature of PCI Riser {0} is OK. ({1})	The GUI checks the hardware state by using xCAT.	The hardware part is OK.	N/A

Table 14. Server events (continued)

Event	Event Type	Severity	Message	Description	Cause	User Action
server_boot_status_failed	STATE_CHANGE	ERROR	System Boot failed on server {0}.	The GUI checks the hardware state by using xCAT.	The hardware part failed.	N/A
server_boot_status_ok	STATE_CHANGE	INFO	The boot status of server {0} is normal.	The GUI checks the hardware state by using xCAT.	The hardware part is OK.	N/A
server_cpu_failed	STATE_CHANGE	ERROR	At least one CPU of server {0} failed.	The GUI checks the hardware state by using xCAT.	The hardware part failed.	N/A
server_cpu_ok	STATE_CHANGE	INFO	All CPUs of server {0} are fully available.	The GUI checks the hardware state by using xCAT.	The hardware part is OK.	N/A
server_dimm_failed	STATE_CHANGE	ERROR	At least one DIMM of server {0} failed.	The GUI checks the hardware state by using xCAT.	The hardware part failed.	N/A
server_dimm_ok	STATE_CHANGE	INFO	All DIMMs of server {0} are fully available.	The GUI checks the hardware state by using xCAT.	The hardware part is OK.	N/A
server_failed	STATE_CHANGE	ERROR	The server {0} failed.	The GUI checks the hardware state by using xCAT.	The hardware part failed.	N/A
server_fan_failed	STATE_CHANGE	ERROR	Fan {0} failed. ({1})	The GUI checks the hardware state by using xCAT.	The hardware part failed.	N/A
server_fan_ok	STATE_CHANGE	INFO	Fan {0} is OK. ({1})	The GUI checks the hardware state by using xCAT.	The hardware part is OK.	N/A
server_ok	STATE_CHANGE	INFO	The server {0} is healthy.	The GUI checks the hardware state by using xCAT.	The hardware part is OK.	N/A
server_pci_failed	STATE_CHANGE	ERROR	At least one PCI of server {0} failed.	The GUI checks the hardware state by using xCAT.	The hardware part failed.	N/A
server_pci_ok	STATE_CHANGE	INFO	All PCIs of server {0} are fully available.	The GUI checks the hardware state by using xCAT.	The hardware part is OK.	N/A
server_planar_failed	STATE_CHANGE	ERROR	Planar state of server {0} is unhealthy. The voltage is too low or too high ({1}).	The GUI checks the hardware state by using xCAT.	The hardware part failed.	N/A

Table 14. Server events (continued)

Event	Event Type	Severity	Message	Description	Cause	User Action
server_planar_ok	STATE_CHANGE	INFO	Planar state of server {0} is healthy. The voltage is normal ({1}).	The GUI checks the hardware state by using xCAT.	The hardware part is OK.	N/A
server_power_supply_aux_line _12V_failed	STATE_CHANGE	ERROR	AUX Line 12V of Power Supply {0} failed.	The GUI checks the hardware state by using xCAT.	The hardware part failed.	N/A
server_power_supply_aux_line_12V_ok	STATE_CHANGE	INFO	AUX Line 12V of Power Supply {0} is OK.	The GUI checks the hardware state by using xCAT.	The hardware part is OK.	N/A
server_power_supply_failed	STATE_CHANGE	ERROR	Power Supply {0} failed.	The GUI checks the hardware state by using xCAT.	The hardware part failed.	N/A
server_power_supply_ fan_failed	STATE_CHANGE	ERROR	Fan of Power Supply {0} failed.	The GUI checks the hardware state by using xCAT.	The hardware part failed.	N/A
server_power_supply_ fan_ok	STATE_CHANGE	INFO	Fan of Power Supply {0} is OK.	The GUI checks the hardware state by using xCAT.	The hardware part is OK.	N/A
server_power_supply_oc_line _12V_failed	STATE_CHANGE	ERROR	OC Line 12V of Power Supply {0} failed.	The GUI checks the hardware state by using xCAT.	The hardware part failed.	N/A
server_power_supply_oc_line_12V_ok	STATE_CHANGE	INFO	OC Line 12V of Power Supply {0} is OK.	The GUI checks the hardware state by using xCAT.	The hardware part is OK.	N/A
server_power_supply_ok	STATE_CHANGE	INFO	Power Supply {0} is OK.	The GUI checks the hardware state by using xCAT.	The hardware part is OK.	N/A
server_power_supply_ov_line _12V_failed	STATE_CHANGE	ERROR	OV Line 12V of Power Supply {0} failed.	The GUI checks the hardware state by using xCAT.	The hardware part failed.	N/A
server_power_supply_ov_line_12V_ok	STATE_CHANGE	INFO	OV Line 12V of Power Supply {0} is OK.	The GUI checks the hardware state by using xCAT.	The hardware part is OK.	N/A
server_power_supply_temp_failed	STATE_CHANGE	ERROR	Temperature of Power Supply {0} is too high. ({1})	The GUI checks the hardware state by using xCAT.	The hardware part failed.	N/A
server_power_supply_temp_ok	STATE_CHANGE	INFO	Temperature of Power Supply {0} is OK ({1}).	The GUI checks the hardware state by using xCAT.	The hardware part is OK.	N/A

Table 14. Server events (continued)

Event	Event Type	Severity	Message	Description	Cause	User Action
server_power_supply_uv_line _12V_failed	STATE_CHANGE	ERROR	UV Line 12V of Power Supply {0} failed.	The GUI checks the hardware state by using xCAT.	The hardware part failed.	N/A
server_power_supply_uv_line_12V_ok	STATE_CHANGE	INFO	UV Line 12V of Power Supply {0} is OK.	The GUI checks the hardware state by using xCAT.	The hardware part is OK.	N/A
server_power_supply_voltage_failed	STATE_CHANGE	ERROR	Voltage of Power Supply {0} is not OK.	The GUI checks the hardware state by using xCAT.	The hardware part failed.	N/A
server_power_supply_voltage_ok	STATE_CHANGE	INFO	Voltage of Power Supply {0} is OK.	The GUI checks the hardware state by using xCAT.	The hardware part is OK.	N/A
server_ps_ambient_failed	STATE_CHANGE	ERROR	At least one Power Supply ambient of server {0} is not OK.	The GUI checks the hardware state by using xCAT.	The hardware part failed.	N/A
server_ps_ambient_ok	STATE_CHANGE	INFO	Power Supply ambient of server {0} is OK.	The GUI checks the hardware state by using xCAT.	The hardware part is OK.	N/A
server_ps_conf_failed	STATE_CHANGE	ERROR	At least one Power Supply Configuration of server {0} is not OK.	The GUI checks the hardware state by using xCAT.	The hardware part failed.	N/A
server_ps_conf_ok	STATE_CHANGE	INFO	All Power Supply Configurations of server {0} are OK.	The GUI checks the hardware state by using xCAT.	The hardware part is OK.	N/A
server_ps_heavyload_failed	STATE_CHANGE	ERROR	At least one Power Supply of server {0} is under heavy load.	The GUI checks the hardware state by using xCAT.	The hardware part failed.	N/A
server_ps_heavyload_ok	STATE_CHANGE	INFO	No Power Supplies of server {0} are under heavy load.	The GUI checks the hardware state by using xCAT.	The hardware part is OK.	N/A
server_ps_resource_failed	STATE_CHANGE	ERROR	At least one Power Supply of server {0} has insufficient resources.	The GUI checks the hardware state by using xCAT.	The hardware part failed.	N/A
server_ps_resource_ok	STATE_CHANGE	INFO	Power Supply resources of server {0} are OK.	The GUI checks the hardware state by using xCAT.	The hardware part is OK.	N/A
server_ps_unit_failed	STATE_CHANGE	ERROR	At least one Power Supply unit of server {0} failed.	The GUI checks the hardware state by using xCAT.	The hardware part failed.	N/A

Table 14. Server events (continued)

Event	Event Type	Severity	Message	Description	Cause	User Action
server_ps_unit_ok	STATE_CHANGE	INFO	All Power Supply units of server {0} are fully available.	The GUI checks the hardware state by using xCAT.	The hardware part is OK.	N/A
server_sys_board_failed	STATE_CHANGE	ERROR	The system board of server {0} failed.	The GUI checks the hardware state by using xCAT.	The hardware part failed.	N/A
server_sys_board_ok	STATE_CHANGE	INFO	The system board of server {0} is healthy.	The GUI checks the hardware state by using xCAT.	The hardware part is OK.	N/A
server_system_event_log_full	STATE_CHANGE	ERROR	The system event log of server {0} is full.	The GUI checks the hardware state by using xCAT.	The hardware part failed.	N/A
server_system_event_log_ok	STATE_CHANGE	INFO	The system event log of server {0} operates normally.	The GUI checks the hardware state by using xCAT.	The hardware part is OK.	N/A

Canister events

The following table lists the events that are created for the *Canister* component.

Table 15. Events for the Canister component

Event	Event Type	Severity	Message	Description	Cause	User Action
bootdrive_mirror_degraded	STATE_CHANGE	WARNING	The bootdrive's mirroring is degraded.	The bootdrive's mirroring is degraded.	The **tsplatformstat -a** command returns a DEGRADED value for at least one partition.	N/A
bootdrive_endurance_ok	STATE_CHANGE	INFO	The bootdrive's endurance is OK.	The bootdrive's endurance is OK.	Hardware monitor returned bootdrive endurance is OK.	N/A
bootdrive_endurance _unknown	INFO	WARNING	The bootdrive's endurance is not known.	The bootdrive's endurance is not known.	Hardware monitor returned unknown bootdrive endurance.	N/A
bootdrive_endurance_warn	STATE_CHANGE	WARNING	The bootdrive's endurance reached its end.	The bootdrive's endurance reached its end.	Hardware monitor returned a warning for bootdrive endurance.	Replace the bootdrive.
bootdrive_installed	STATE_CHANGE	INFO	The bootdrive attached to port {0} is available.	The bootdrive is available.	The **tsplatformstat -a** command returns the bootdrives as expected.	N/A
bootdrive_mirror_failed	STATE_CHANGE	ERROR	The bootdrive's mirroring is failed.	The bootdrive's mirroring is failed.	The **tsplatformstat -a** command returns a FAILED value for at least one partition.	N/A
bootdrive_mirror_ok	STATE_CHANGE	INFO	The bootdrive's mirroring is OK.	The bootdrive's mirroring is OK.	The **tsplatformstat -a** command returns optimal for all partitions.	N/A

Table 15. Events for the Canister component (continued)

Event	Event Type	Severity	Message	Description	Cause	User Action
bootdrive_mirror _unconfigured	STATE_CHANGE	WARNING	The bootdrive's mirroring is unconfigured.	The bootdrive's mirroring is unconfigured.	The **tsplatformstat -a** command returns unconfigured for mirroring.	N/A
bootdrive_missing	STATE_CHANGE	ERROR	The bootdrive on port {0} is missing or dead.	One bootdrive is missing or dead. Redundancy is not given anymore.	The **tsplatformstat -a** command returns only one instead of two bootdrives. Two drives are expected to ensure redundancy.	Inspect that the drive is correctly installed on the referenced port. Else insert or replace the drive.
bootdrive_smart_failed	STATE_CHANGE	ERROR	The smart assessment of bootdrive {0} attached to port {1} does not return OK.	The bootdrive's smart assessment does not return OK.	The **tsplatformstat -a** command does not return a PASSED value in the **selfAssessment** field for the bootdrive.	Verify the smart status of the bootdrive using **tsplatformstat** command or `smartctl`.
bootdrive_smart_ok	STATE_CHANGE	INFO	The smart assessment of bootdrive {0} attached to port {1} returns OK.	The bootdrive's smart assessment returns OK.	The **tsplatformstat -a** command returns a PASSED in the **selfAssessment** field for the bootdrive.	N/A
can_fan_failed	STATE_CHANGE	WARNING	Fan {0} is failed.	The fan state is failed.	The **mmlsenclosure** command reports the fan as failed.	Check the fan status by using the **mmlsenclosure** command. Replace the fan module in the canister.
can_fan_ok	STATE_CHANGE	INFO	Fan {0} is OK.	The fan state is OK.	The **mmlsenclosure** command reports the fan as working.	N/A
can_temp_bus_failed	STATE_CHANGE	WARNING	Temperature sensor {0} I2C bus is failed.	The temperature sensor I2C bus failed.	The **mmlsenclosure** command reports the temperature sensor with a failure.	Check the temperature status by using the **mmlsenclosure** command.
can_temp_high_critical	STATE_CHANGE	WARNING	Temperature sensor {0} measured a high temperature value.	The temperature exceeded the actual high critical threshold value for at least one sensor.	The **mmlsenclosure** command reports the temperature sensor with a failure.	Check the temperature status by using the **mmlsenclosure** command.
can_temp_high_warn	STATE_CHANGE	WARNING	Temperature sensor {0} measured a high temperature value.	The temperature exceeded the actual high warning threshold value for at least one sensor.	The **mmlsenclosure** command reports the temperature sensor with a failure.	Check the temperature status by using the **mmlsenclosure** command.
can_temp_low_critical	STATE_CHANGE	WARNING	Temperature sensor {0} measured a low temperature value.	The temperature drops less than the actual low critical threshold value for at least one sensor.	The **mmlsenclosure** command reports the temperature sensor with a failure.	Check the temperature status by using the **mmlsenclosure** command.
can_temp_low_warn	STATE_CHANGE	WARNING	Temperature sensor {0} measured a low temperature value.	The temperature drops less than the actual low warning threshold value for at least one sensor.	The **mmlsenclosure** command reports the temperature sensor with a failure.	Check the temperature status by using the **mmlsenclosure** command.
can_temp_sensor_failed	STATE_CHANGE	WARNING	Temperature sensor {0} is failed.	The temperature sensor state is failed.	The **mmlsenclosure** command reports the temperature sensor with a failure.	Check the temperature status by using the **mmlsenclosure** command. Replace the canister.

Table 15. Events for the Canister component (continued)

Event	Event Type	Severity	Message	Description	Cause	User Action
can_temp_sensor_ok	STATE_CHANGE	INFO	Temperature sensor {0} is OK.	The temperature sensor state is OK.	N/A	N/A
canister_failed	STATE_CHANGE	ERROR	Canister {0} is failed.	The canister reports a failed hardware state, which might be caused by an underlying component (like fan) failure.	The **mmlsenclosure** command reports the canister as failed.	Check for detailed error events of canister components by using the **mmhealth** command. Inspect the output of **mmlsenclosure all -L** command for the referenced canister.
canister_ok	STATE_CHANGE	INFO	Canister {0} is OK.	The canister state is OK.	The **mmlsenclosure** command reports the canister as failed.	N/A
canister_thermal_shutdown	STATE_CHANGE	ERROR	Canister {ID} temperature value is more than the critical threshold value.	The canister temperature value is more than the critical threshold value.	Hardware monitored temperature sensors reached critical thresholds.	Check environmental conditions and system error logs for fan errors.
coin_battery_low	STATE_CHANGE	WARNING	The coin battery has low voltage.	The coin battery has low voltage.	Hardware monitor reports coin battery with low voltage.	Replace the coin battery.
coin_battery_missing	STATE_CHANGE	WARNING	The coin battery is absent.	The coin battery is absent.	Hardware monitor reports no coin battery.	Install a coin battery.
coin_battery_ok	STATE_CHANGE	INFO	The coin battery is OK.	The coin battery is OK.	Hardware monitor reports a healthy coin battery.	N/A
coin_battery_unknown	STATE_CHANGE	WARNING	The coin battery has low voltage.	The coin battery's voltage is unknown.	Hardware monitor reports unknown coin battery voltage.	Replace the coin battery.
cpu_inspection_failed	STATE_CHANGE	ERROR	The inspection of the CPU slots found a mismatch.	Number of populated CPU slots, number of enabled CPUs, number of CPU cores, number of CPU threads or CPU speed is not as expected.	The **/opt/ibm/gss/tools/bin/ess3kplt** command returned an `InspectionPassed unequal to True` value.	Check for specific events related to CPUs by using the **mmhealth** command. Inspect the output of the **ess3kplt** command for details.
cpu_inspection_passed	STATE_CHANGE	INFO	The CPUs of the canister are OK.	The CPU speed and number of populated CPU slots is as expected.	The **/opt/ibm/gss/tools/bin/ess3kplt** command returned an `InspectionPassed equal to True` value.	N/A
cpu_speed_ok	STATE_CHANGE	INFO	The CPU speed is OK.	The speed of all CPUs is as expected.	The **/opt/ibm/gss/tools/bin/ess3kplt** command returned no speed errors.	N/A
cpu_speed_wrong	STATE_CHANGE	ERROR	One or more CPUs have an unsupported speed.	The speed of one or more CPUs is not as expected. This configuration is not supported.	The **/opt/ibm/gss/tools/bin/ess3kplt** command returned one or more speed errors.	Inspect the output of the **ess3kplt** command to see which CPUs have an unsupported speed.
cpu_unit_missing	STATE_CHANGE	ERROR	CPU {ID} in canister {0} is missing.	A CPU is missing.	Hardware monitor returns a missing CPU unit.	N/A
cpu_unit_speed_ok	STATE_CHANGE	INFO	The CPU {ID} in canister {0} has correct speed.	The speed of this CPU unit is as expected.	Hardware monitor returned no speed error for this CPU unit.	N/A

Table 15. Events for the Canister component (continued)

Event	Event Type	Severity	Message	Description	Cause	User Action
cpu_unit_speed_unknown	INFO	WARNING	CPU {ID} in canister {0} has an unknown speed.	The speed of this CPU is not known. This configuration is not supported.	Hardware monitor cannot detect speed for this CPU.	This issue is similar to a transient state during the detection. If this event persists, then restart the node.
cpu_unit_speed_wrong	STATE_CHANGE	ERROR	CPU {ID} in canister {0} has an unsupported speed.	The speed of this CPU is not as expected. This configuration is not supported.	Hardware monitor returned a speed error for this CPU.	Replace this CPU.
dimm_inspection_failed	STATE_CHANGE	ERROR	The inspection of the memory dimm slots found a failure.	The capacity, speed, or number of populated dimm slots is not as expected.	The **/opt/ibm/gss/tools/bin/ess3kplt** command returned an InspectionPassed unequal to True value.	Check for specific events related to dimms by using the **mmhealth** command. Inspect the output of the **ess3kplt** command for details.
dimm_inspection_passed	STATE_CHANGE	INFO	The memory dimms of the canister is OK.	The capacity, speed, and number of populated dimm slots is as expected.	The **/opt/ibm/gss/tools/bin/ess3kplt** command returned an InspectionPassed equal to True value.	N/A
dimm_module_missing	STATE_CHANGE	ERROR	Memory dimm modules {ID} in canister {0} is missing.	A dimm module is either missing or not properly connected or broken.	Hardware monitor reported the dimm module as missing.	If slot is empty, insert a new dimm. Otherwise, replace or reinsert the current dimm.
dimm_module_size_ok	STATE_CHANGE	INFO	The installed memory dimm {ID} in canister {0} has the expected capacity.	The capacity of this dimm is as expected.	Hardware monitor detected no capacity error.	N/A
dimm_module_size_unknown	INFO	WARNING	Memory dimm module {ID} has in canister {0] has an unknown capacity.	The capacity of this dimm module is unknown.	Hardware monitor cannot detect the capacity for this dimm.	This issue is similar to a transient state during the detection. If this event persists, then restart the node.
dimm_module_size_wrong	STATE_CHANGE	ERROR	Memory dimm module {ID} in canister {0} has an unsupported capacity of {1}.	The capacity of this memory dimm is not as expected. This configuration is not supported.	Hardware monitor detected a capacity error for this dimm.	Replace this dimm module.
dimm_module_speed_ok	STATE_CHANGE	INFO	Memory dimm module {ID} in canister {0} has a supported speed.	The speed of this dimm module is as expected.	Hardware monitor returned no speed error for this dimm.	N/A
dimm_module_speed_wrong	STATE_CHANGE	ERROR	Memory dimm {ID} in canister {0} has an unsupported speed.	The speed of this memory dimm slot is not as expected. This configuration is not supported.	Hardware monitor returned a speed error for this dimm.	Replace this dimm.
dimm_size_ok	STATE_CHANGE	INFO	All installed memory dimms have the expected capacity.	The capacity of all populated memory dimm slots is as expected.	The **/opt/ibm/gss/tools/bin/ess3kplt** command returned no capacity errors.	N/A
dimm_size_wrong	STATE_CHANGE	ERROR	One or more memory dimm modules have an unsupported capacity.	The capacity of one or more memory dimm slots is not as expected. This configuration is not supported.	The **/opt/ibm/gss/tools/bin/ess3kplt** command returned some capacity errors.	Inspect the output of the **ess3kplt** command to see which memory dimm slots have an unsupported capacity and replace those dimm modules.

Table 15. Events for the Canister component (continued)

Event	Event Type	Severity	Message	Description	Cause	User Action
dimm_speed_ok	STATE_CHANGE	INFO	All installed memory dimms have the expected speed.	The speed of all populated memory dimm slots is as expected.	The **/opt/ibm/gss/tools/bin/ess3kplt** command returned no speed errors.	N/A
dimm_speed_wrong	STATE_CHANGE	ERROR	One or more memory dimm modules have an unsupported speed.	The speed of one or more memory dimm slots is not as expected. This configuration is not supported.	The **/opt/ibm/gss/tools/bin/ess3kplt** command returned some speed errors.	Inspect the output of the **ess3kplt** command to see which memory dimm slots have an unsupported speed and replace those dimm modules.
pair_canister_comm_error	STATE_CHANGE	WARNING	Pair canister {0} has communication error.	The internal communication between pair canisters has an error.	Internal communication, which uses interlink interface, is lost.	Refer to the *Enterprise Storage Server Problem Determination Guide*.
pair_canister_failed	STATE_CHANGE	ERROR	Pair canister {ID} failed.	The pair canister failed.	Hardware monitor reports that one canister failed.	Investigate the reason why a canister failed.
pair_canister_missing	STATE_CHANGE	WARNING	Pair canister {0} is missing or dead.	Cannot get the state of the pair canister. It might be missing or dead.	The **mmlsenclosure** command reports only one canister instead of two.	Check for detailed error events of the referenced canister node by using the **mmhealth** command. Inspect the output of the **mmlsenclosure all -L** command for the referenced canister.
pair_canister_power_off	STATE_CHANGE	INFO	Pair canister {ID} is powered off.	Pair canister power off detected.	Hardware monitor detected power off for the pair canister.	N/A
pair_canister_unknown	INFO	WARNING	Pair canister {ID} state is unknown.	An error occurred when a pair canister is detected.	Hardware monitor failed to detect the pair canister status.	This issue is similar to a transient state during the detection. If this event persists, then restart the node.
pair_canister_visible	STATE_CHANGE	INFO	Pair canister {0} is visible.	Successfully get the state of the pair canister.	The **mmlsenclosure** command reports both canisters.	N/A
rest_client_failed	INFO	WARNING	The hardware monitor client does not work: {0}.	The hardware monitor client does not work.	Hardware monitor client does not work correctly.	N/A
rest_client_ok	STATE_CHANGE	INFO	The hardware monitor client is OK.	The hardware monitor client is OK.	Hardware monitor client works correctly.	N/A
tpm_absent	STATE_CHANGE	TIP	The Trusted Platform Module is absent.	The Trusted Platform Module is absent.	Hardware monitor reports that the Trusted Platform Module is absent.	Add the Trusted Platform Module.
tpm_ok	STATE_CHANGE	INFO	The Trusted Platform Module is OK.	The Trusted Platform Module is OK.	Hardware monitor reports that the Trusted Platform Module works correctly.	N/A
tpm_unknown	INFO	WARNING	The Trusted Platform Module is absent.	The status of the Trusted Platform Module is unknown.	Hardware monitor cannot detect when the Trusted Platform Module is present or absent.	N/A

Messages

This topic contains explanations for IBM Spectrum Scale RAID and ESS GUI messages.

For information about IBM Spectrum Scale messages, see the *IBM Spectrum Scale: Problem Determination Guide*.

Message severity tags

IBM Spectrum Scale and ESS GUI messages include message severity tags.

A severity tag is a one-character alphabetic code (**A** through **Z**).

For IBM Spectrum Scale messages, the severity tag is optionally followed by a colon (**:**) and a number, and surrounded by an opening and closing bracket (**[ ]**). For example:

```
[E] or [E:nnn]
```

If more than one substring within a message matches this pattern (for example, **[A]** or [**A:***nnn*]), the severity tag is the first such matching string.

When the severity tag includes a numeric code (*nnn*), this is an error code associated with the message. If this were the only problem encountered by the command, the command return code would be *nnn*.

If a message does not have a severity tag, the message does not conform to this specification. You can determine the message severity by examining the text or any supplemental information provided in the message catalog, or by contacting the IBM Support Center.

Each message severity tag has an assigned priority.

For IBM Spectrum Scale messages, this priority can be used to filter the messages that are sent to the error log on Linux. Filtering is controlled with the `mmchconfig` attribute `systemLogLevel`. The default for `systemLogLevel` is `error`, which means that IBM Spectrum Scale will send all error **[E]**, critical **[X]**, and alert **[A]** messages to the error log. The values allowed for `systemLogLevel` are: `alert`, `critical`, `error`, `warning`, `notice`, `configuration`, `informational`, `detail`, or debug. Additionally, the value none can be specified so no messages are sent to the error log.

For IBM Spectrum Scale messages, alert **[A]** messages have the highest priority and debug **[B]** messages have the lowest priority. If the `systemLogLevel` default of `error` is changed, only messages with the specified severity and all those with a higher priority are sent to the error log.

The following table lists the IBM Spectrum Scale message severity tags in order of priority:

Table 16. IBM Spectrum Scale message severity tags ordered by priority

Severity tag	Type of message (`systemLogLevel` attribute)	Meaning
A	`alert`	Indicates a problem where action must be taken immediately. Notify the appropriate person to correct the problem.
X	`critical`	Indicates a critical condition that should be corrected immediately. The system discovered an internal inconsistency of some kind. Command execution might be halted or the system might attempt to continue despite the inconsistency. Report these errors to IBM.

Table 16. IBM Spectrum Scale message severity tags ordered by priority (continued)		
Severity tag	**Type of message (`systemLogLevel` attribute)**	**Meaning**
E	`error`	Indicates an error condition. Command execution might or might not continue, but this error was likely caused by a persistent condition and will remain until corrected by some other program or administrative action. For example, a command operating on a single file or other GPFS object might terminate upon encountering any condition of severity **E**. As another example, a command operating on a list of files, finding that one of the files has permission bits set that disallow the operation, might continue to operate on all other files within the specified list of files.
W	`warning`	Indicates a problem, but command execution continues. The problem can be a transient inconsistency. It can be that the command has skipped some operations on some objects, or is reporting an irregularity that could be of interest. For example, if a multipass command operating on many files discovers during its second pass that a file that was present during the first pass is no longer present, the file might have been removed by another command or program.
N	`notice`	Indicates a normal but significant condition. These events are unusual, but are not error conditions, and could be summarized in an email to developers or administrators for spotting potential problems. No immediate action is required.
C	`configuration`	Indicates a configuration change; such as, creating a file system or removing a node from the cluster.
I	`informational`	Indicates normal operation. This message by itself indicates that nothing is wrong; no action is required.
D	`detail`	Indicates verbose operational messages; no is action required.
B	`debug`	Indicates debug-level messages that are useful to application developers for debugging purposes. This information is not useful during operations.

For ESS GUI messages, error messages (**(E)**) have the highest priority and informational messages (**I**) have the lowest priority.

The following table lists the ESS GUI message severity tags in order of priority:

Table 17. ESS GUI message severity tags ordered by priority		
Severity tag	**Type of message**	**Meaning**
E	`Error`	Indicates a critical condition that should be corrected immediately. The system discovered an internal inconsistency of some kind. Command execution might be halted or the system might attempt to continue despite the inconsistency. Report these errors to IBM.
W	`warning`	Indicates a problem, but command execution continues. The problem can be a transient inconsistency. It can be that the command has skipped some operations on some objects, or is reporting an irregularity that could be of interest. For example, if a multipass command operating on many files discovers during its second pass that a file that was present during the first pass is no longer present, the file might have been removed by another command or program.

Table 17. ESS GUI message severity tags ordered by priority (continued)		
Severity tag	**Type of message**	**Meaning**
I	informational	Indicates normal operation. This message by itself indicates that nothing is wrong; no action is required.

IBM Spectrum Scale RAID messages

This section lists the IBM Spectrum Scale RAID messages.

For information about the severity designations of these messages, see "Message severity tags" on page 92.

6027-1850 [E] NSD-RAID services are not configured on node *nodeName*. Check the nsdRAIDTracks and nsdRAIDBufferPoolSizePct configuration attributes.

Explanation:
A IBM Spectrum Scale RAID command is being executed, but NSD-RAID services are not initialized either because the specified attributes have not been set or had invalid values.

User response:
Correct the attributes and restart the GPFS daemon.

6027-1851 [A] Cannot configure NSD-RAID services. The nsdRAIDBufferPoolSizePct of the pagepool must result in at least 128MiB of space.

Explanation:
The GPFS daemon is starting and cannot initialize the NSD-RAID services because of the memory consideration specified.

User response:
Correct the `nsdRAIDBufferPoolSizePct` attribute and restart the GPFS daemon.

6027-1852 [A] Cannot configure NSD-RAID services. nsdRAIDTracks is too large, the maximum on this node is *value*.

Explanation:
The GPFS daemon is starting and cannot initialize the NSD-RAID services because the `nsdRAIDTracks` attribute is too large.

User response:
Correct the `nsdRAIDTracks` attribute and restart the GPFS daemon.

6027-1853 [E] Recovery group *recoveryGroupName* does not exist or is not active.

Explanation:
A command was issued to a RAID recovery group that does not exist, or is not in the active state.

User response:
Retry the command with a valid RAID recovery group name or wait for the recovery group to become active.

6027-1854 [E] Cannot find declustered array *arrayName* in recovery group *recoveryGroupName*.

Explanation:
The specified declustered array name was not found in the RAID recovery group.

User response:
Specify a valid declustered array name within the RAID recovery group.

6027-1855 [E] Cannot find pdisk *pdiskName* in recovery group *recoveryGroupName*.

Explanation:
The specified pdisk was not found.

User response:
Retry the command with a valid pdisk name.

6027-1856 [E] Vdisk *vdiskName* not found.

Explanation:
The specified vdisk was not found.

User response:
Retry the command with a valid vdisk name.

6027-1857 [E] A recovery group must contain between *number* and *number* pdisks.

Explanation:
The number of pdisks specified is not valid.

User response:
Correct the input and retry the command.

6027-1858 [E] Cannot create declustered array *arrayName*; there can be at most *number* declustered arrays in a recovery group.

Explanation:
The number of declustered arrays allowed in a recovery group has been exceeded.

User response:
Reduce the number of declustered arrays in the input file and retry the command.

6027-1859 [E] **Sector size of pdisk *pdiskName* is invalid.**

Explanation:
All pdisks in a recovery group must have the same physical sector size.

User response:
Correct the input file to use a different disk and retry the command.

6027-1860 [E] **Pdisk *pdiskName* must have a capacity of at least *number* bytes.**

Explanation:
The pdisk must be at least as large as the indicated minimum size in order to be added to this declustered array.

User response:
Correct the input file and retry the command.

6027-1861 [W] **Size of pdisk *pdiskName* is too large for declustered array *arrayName*. Only *number* of *number* bytes of that capacity will be used.**

Explanation:
For optimal utilization of space, pdisks added to this declustered array should be no larger than the indicated maximum size. Only the indicated portion of the total capacity of the pdisk will be available for use.

User response:
Consider creating a new declustered array consisting of all larger pdisks.

6027-1862 [E] **Cannot add pdisk *pdiskName* to declustered array *arrayName*; there can be at most *number* pdisks in a declustered array.**

Explanation:
The maximum number of pdisks that can be added to a declustered array was exceeded.

User response:
None.

6027-1863 [E] **Pdisk sizes within a declustered array cannot vary by more than *number*.**

Explanation:
The disk sizes within each declustered array must be nearly the same.

User response:
Create separate declustered arrays for each disk size.

6027-1864 [E] **[E] At least one declustered array must contain *number* + vdisk configuration data spares or more pdisks and be eligible to hold vdisk configuration data.**

Explanation:
When creating a new RAID recovery group, at least one of the declustered arrays in the recovery group must contain at least 2T+1 pdisks, where T is the maximum number of disk failures that can be tolerated within a declustered array. This is necessary in order to store the on-disk vdisk configuration data safely. This declustered array cannot have canHoldVCD set to no.

User response:
Supply at least the indicated number of pdisks in at least one declustered array of the recovery group, or do not specify canHoldVCD=no for that declustered array.

6027-1866 [E] **Disk descriptor for *diskName* refers to an existing NSD.**

Explanation:
A disk being added to a recovery group appears to already be in-use as an NSD disk.

User response:
Carefully check the disks given to `tscrrecgroup`, `tsaddpdisk` or `tschcarrier`. If you are certain the disk is not actually in-use, override the check by specifying the `-v no` option.

6027-1867 [E] **Disk descriptor for *diskName* refers to an existing pdisk.**

Explanation:
A disk being added to a recovery group appears to already be in-use as a pdisk.

User response:
Carefully check the disks given to `tscrrecgroup`, `tsaddpdisk` or `tschcarrier`. If you are certain the disk is not actually in-use, override the check by specifying the `-v no` option.

6027-1869 [E] **Error updating the recovery group descriptor.**

Explanation:
Error occurred updating the RAID recovery group descriptor.

User response:
Retry the command.

6027-1870 [E] **Recovery group name *name* is already in use.**

Explanation:
The recovery group name already exists.

User response:
Choose a new recovery group name using the characters a-z, A-Z, 0-9, and underscore, at most 63 characters in length.

6027-1871 [E] There is only enough free space to allocate *number* spare(s) in declustered array *arrayName*.

Explanation:
Too many spares were specified.

User response:
Retry the command with a valid number of spares.

6027-1872 [E] Recovery group still contains vdisks.

Explanation:
RAID recovery groups that still contain vdisks cannot be deleted.

User response:
Delete any vdisks remaining in this RAID recovery group using the `tsdelvdisk` command before retrying this command.

6027-1873 [E] Pdisk creation failed for pdisk *pdiskName*: err=*errorNum*.

Explanation:
Pdisk creation failed because of the specified error.

User response:
None.

6027-1874 [E] Error adding pdisk to a recovery group.

Explanation:
`tsaddpdisk` failed to add new pdisks to a recovery group.

User response:
Check the list of pdisks in the `-d` or `-F` parameter of `tsaddpdisk`.

6027-1875 [E] Cannot delete the only declustered array.

Explanation:
Cannot delete the only remaining declustered array from a recovery group.

User response:
Instead, delete the entire recovery group.

6027-1876 [E] Cannot remove declustered array *arrayName* because it is the only remaining declustered array with at least *number* pdisks eligible to hold vdisk configuration data.

Explanation:
The command failed to remove a declustered array because no other declustered array in the recovery group has sufficient pdisks to store the on-disk recovery group descriptor at the required fault tolerance level.

User response:
Add pdisks to another declustered array in this recovery group before removing this one.

6027-1877 [E] Cannot remove declustered array *arrayName* because the array still contains vdisks.

Explanation:
Declustered arrays that still contain vdisks cannot be deleted.

User response:
Delete any vdisks remaining in this declustered array using the `tsdelvdisk` command before retrying this command.

6027-1878 [E] Cannot remove pdisk *pdiskName* because it is the last remaining pdisk in declustered array *arrayName*. Remove the declustered array instead.

Explanation:
The `tsdelpdisk` command can be used either to delete individual pdisks from a declustered array, or to delete a full declustered array from a recovery group. You cannot, however, delete a declustered array by deleting all of its pdisks -- at least one must remain.

User response:
Delete the declustered array instead of removing all of its pdisks.

6027-1879 [E] Cannot remove pdisk *pdiskName* because *arrayName* is the only remaining declustered array with at least *number* pdisks.

Explanation:
The command failed to remove a pdisk from a declustered array because no other declustered array in the recovery group has sufficient pdisks to store the on-disk recovery group descriptor at the required fault tolerance level.

User response:
Add pdisks to another declustered array in this recovery group before removing pdisks from this one.

6027-1880 [E] Cannot remove pdisk *pdiskName* because the number of pdisks in declustered array *arrayName* would fall below the code width of one or more of its vdisks.

Explanation:
The number of pdisks in a declustered array must be at least the maximum code width of any vdisk in the declustered array.

User response:
Either add pdisks or remove vdisks from the declustered array.

6027-1881 [E] Cannot remove pdisk *pdiskName* because of insufficient free space in declustered array *arrayName*.

Explanation:
The `tsdelpdisk` command could not delete a pdisk because there was not enough free space in the declustered array.

User response:
Either add pdisks or remove vdisks from the declustered array.

6027-1882 [E] Cannot remove pdisk *pdiskName*; unable to drain the data from the pdisk.

Explanation:
Pdisk deletion failed because the system could not find enough free space on other pdisks to drain all of the data from the disk.

User response:
Either add pdisks or remove vdisks from the declustered array.

6027-1883 [E] Pdisk *pdiskName* deletion failed: process interrupted.

Explanation:
Pdisk deletion failed because the deletion process was interrupted. This is most likely because of the recovery group failing over to a different server.

User response:
Retry the command.

6027-1884 [E] Missing or invalid vdisk name.

Explanation:
No vdisk name was given on the `tscrvdisk` command.

User response:
Specify a vdisk name using the characters a-z, A-Z, 0-9, and underscore of at most 63 characters in length.

6027-1885 [E] Vdisk block size must be a power of 2.

Explanation:
The `-B` or `--blockSize` parameter of `tscrvdisk` must be a power of 2.

User response:
Reissue the `tscrvdisk` command with a correct value for block size.

6027-1886 [E] Vdisk block size cannot exceed maxBlockSize (*number*).

Explanation:
The virtual block size of a vdisk cannot be larger than the value of the `maxblocksize` configuration attribute of the IBM Spectrum Scale `mmchconfig` command.

User response:
Use a smaller vdisk virtual block size, or increase the value of `maxBlockSize` using `mmchconfig maxblocksize=`*newSize*.

6027-1887 [E] Vdisk block size must be between *number* and *number* for the specified code.

Explanation:
An invalid vdisk block size was specified. The message lists the allowable range of block sizes.

User response:
Use a vdisk virtual block size within the range shown, or use a different vdisk RAID code.

6027-1888 [E] Recovery group already contains *number* vdisks.

Explanation:
The RAID recovery group already contains the maximum number of vdisks.

User response:
Create vdisks in another RAID recovery group, or delete one or more of the vdisks in the current RAID recovery group before retrying the `tscrvdisk` command.

6027-1889 [E] Vdisk name *vdiskName* is already in use.

Explanation:
The vdisk name given on the `tscrvdisk` command already exists.

User response:
Choose a new vdisk name less than 64 characters using the characters a-z, A-Z, 0-9, and underscore.

6027-1890 [E] A recovery group may only contain one log home vdisk.

Explanation:
A log vdisk already exists in the recovery group.

User response:
None.

6027-1891 [E] Cannot create vdisk before the log home vdisk is created.

Explanation:
The log vdisk must be the first vdisk created in a recovery group.

User response:
Retry the command after creating the log home vdisk.

6027-1892 [E] Log vdisks must use replication.

Explanation:
The log vdisk must use a RAID code that uses replication.

User response:
Retry the command with a valid RAID code.

6027-1893 [E] The declustered array must contain at least as many non-spare pdisks as the width of the code.

Explanation:
The RAID code specified requires a minimum number of disks larger than the size of the declustered array that was given.

User response:
Place the vdisk in a wider declustered array or use a narrower code.

6027-1894 [E] There is not enough space in the declustered array to create additional vdisks.

Explanation:
There is insufficient space in the declustered array to create even a minimum size vdisk with the given RAID code.

User response:
Add additional pdisks to the declustered array, reduce the number of spares or use a different RAID code.

6027-1895 [E] Unable to create vdisk *vdiskName* because there are too many failed pdisks in declustered array *declusteredArrayName*.

Explanation:
Cannot create the specified vdisk, because there are too many failed pdisks in the array.

User response:
Replace failed pdisks in the declustered array and allow time for rebalance operations to more evenly distribute the space.

6027-1896 [E] Insufficient memory for vdisk metadata.

Explanation:
There was not enough pinned memory for IBM Spectrum Scale to hold all of the metadata necessary to describe a vdisk.

User response:
Increase the size of the GPFS page pool.

6027-1897 [E] Error formatting vdisk.

Explanation:
An error occurred formatting the vdisk.

User response:
None.

6027-1898 [E] The log home vdisk cannot be destroyed if there are other vdisks.

Explanation:
The log home vdisk of a recovery group cannot be destroyed if vdisks other than the log tip vdisk still exist within the recovery group.

User response:
Remove the user vdisks and then retry the command.

6027-1899 [E] Vdisk *vdiskName* is still in use.

Explanation:
The vdisk named on the `tsdelvdisk` command is being used as an NSD disk.

User response:
Remove the vdisk with the `mmdelnsd` command before attempting to delete it.

6027-3000 [E] No disk enclosures were found on the target node.

Explanation:
IBM Spectrum Scale is unable to communicate with any disk enclosures on the node serving the specified pdisks. This might be because there are no disk enclosures attached to the node, or it might indicate a problem in communicating with the disk enclosures. While the problem persists, disk maintenance with the `mmchcarrier` command is not available.

User response:
Check disk enclosure connections and run the command again. Use `mmaddpdisk --replace` as an alternative method of replacing failed disks.

6027-3001 [E] Location of pdisk *pdiskName* of recovery group *recoveryGroupName* is not known.

Explanation:
IBM Spectrum Scale is unable to find the location of the given pdisk.

User response:
Check the disk enclosure hardware.

6027-3002 [E] Disk location code *locationCode* is not known.

Explanation:
A disk location code specified on the command line was not found.

User response:
Check the disk location code.

6027-3003 [E] Disk location code *locationCode* was specified more than once.

Explanation:

The same disk location code was specified more than once in the `tschcarrier` command.

User response:
Check the command usage and run again.

6027-3004 [E] Disk location codes *locationCode* and *locationCode* are not in the same disk carrier.

Explanation:
The `tschcarrier` command cannot be used to operate on more than one disk carrier at a time.

User response:
Check the command usage and rerun.

6027-3005 [W] Pdisk in location *locationCode* is controlled by recovery group *recoveryGroupName*.

Explanation:
The `tschcarrier` command detected that a pdisk in the indicated location is controlled by a different recovery group than the one specified.

User response:
Check the disk location code and recovery group name.

6027-3006 [W] Pdisk in location *locationCode* is controlled by recovery group id *idNumber*.

Explanation:
The `tschcarrier` command detected that a pdisk in the indicated location is controlled by a different recovery group than the one specified.

User response:
Check the disk location code and recovery group name.

6027-3007 [E] Carrier contains pdisks from more than one recovery group.

Explanation:
The `tschcarrier` command detected that a disk carrier contains pdisks controlled by more than one recovery group.

User response:
Use the `tschpdisk` command to bring the pdisks in each of the other recovery groups offline and then rerun the command using the `--force-RG` flag.

6027-3008 [E] Incorrect recovery group given for location.

Explanation:
The `mmchcarrier` command detected that the specified recovery group name given does not match that of the pdisk in the specified location.

User response:
Check the disk location code and recovery group name. If you are sure that the disks in the carrier are not being used by other recovery groups, it is possible to override the check using the `--force-RG` flag. Use this flag with caution as it can cause disk errors and potential data loss in other recovery groups.

6027-3009 [E] Pdisk *pdiskName* of recovery group *recoveryGroupName* is not currently scheduled for replacement.

Explanation:
A pdisk specified in a `tschcarrier` or `tsaddpdisk` command is not currently scheduled for replacement.

User response:
Make sure the correct disk location code or pdisk name was given. For the `mmchcarrier` command, the `--force-release` option can be used to override the check.

6027-3010 [E] Command interrupted.

Explanation:
The `mmchcarrier` command was interrupted by a conflicting operation, for example the `mmchpdisk --resume` command on the same pdisk.

User response:
Run the `mmchcarrier` command again.

6027-3011 [W] Disk location *locationCode* failed to power off.

Explanation:
The `mmchcarrier` command detected an error when trying to power off a disk.

User response:
Check the disk enclosure hardware. If the disk carrier has a lock and does not unlock, try running the command again or use the manual carrier release.

6027-3012 [E] Cannot find a pdisk in location *locationCode*.

Explanation:
The `tschcarrier` command cannot find a pdisk to replace in the given location.

User response:
Check the disk location code.

6027-3013 [W] Disk location *locationCode* failed to power on.

Explanation:
The `mmchcarrier` command detected an error when trying to power on a disk.

User response:
Make sure the disk is firmly seated and run the command again.

6027-3014 [E] **Pdisk *pdiskName* of recovery group *recoveryGroupName* was expected to be replaced with a new disk; instead, it was moved from location *locationCode* to location *locationCode*.**

Explanation:
The `mmchcarrier` command expected a pdisk to be removed and replaced with a new disk. But instead of being replaced, the old pdisk was moved into a different location.

User response:
Repeat the disk replacement procedure.

6027-3015 [E] **Pdisk *pdiskName* of recovery group *recoveryGroupName* in location *locationCode* cannot be used as a replacement for pdisk *pdiskName* of recovery group *recoveryGroupName*.**

Explanation:
The `tschcarrier` command expected a pdisk to be removed and replaced with a new disk. But instead of finding a new disk, the `mmchcarrier` command found that another pdisk was moved to the replacement location.

User response:
Repeat the disk replacement procedure, making sure to replace the failed pdisk with a new disk.

6027-3016 [E] **Replacement disk in location *locationCode* has an incorrect type *fruCode*; expected type code is *fruCode*.**

Explanation:
The replacement disk has a different field replaceable unit type code than that of the original disk.

User response:
Replace the pdisk with a disk of the same part number. If you are certain the new disk is a valid substitute, override this check by running the command again with the `--force-fru` option.

6027-3017 [E] **Error formatting replacement disk *diskName*.**

Explanation:
An error occurred when trying to format a replacement pdisk.

User response:
Check the replacement disk.

6027-3018 [E] **A replacement for pdisk *pdiskName* of recovery group *recoveryGroupName* was not found in location *locationCode*.**

Explanation:
The `tschcarrier` command expected a pdisk to be removed and replaced with a new disk, but no replacement disk was found.

User response:
Make sure a replacement disk was inserted into the correct slot.

6027-3019 [E] **Pdisk *pdiskName* of recovery group *recoveryGroupName* in location *locationCode* was not replaced.**

Explanation:
The `tschcarrier` command expected a pdisk to be removed and replaced with a new disk, but the original pdisk was still found in the replacement location.

User response:
Repeat the disk replacement, making sure to replace the pdisk with a new disk.

6027-3020 [E] **Invalid state change, *stateChangeName*, for pdisk *pdiskName*.**

Explanation:
The `tschpdisk` command received an state change request that is not permitted.

User response:
Correct the input and reissue the command.

6027-3021 [E] **Unable to change identify state to *identifyState* for pdisk *pdiskName*: err=*errorNum*.**

Explanation:
The `tschpdisk` command failed on an identify request.

User response:
Check the disk enclosure hardware.

6027-3022 [E] **Unable to create vdisk layout.**

Explanation:
The `tscrvdisk` command could not create the necessary layout for the specified vdisk.

User response:
Change the vdisk arguments and retry the command.

6027-3023 [E] **Error initializing vdisk.**

Explanation:
The `tscrvdisk` command could not initialize the vdisk.

User response:
Retry the command.

6027-3024 [E] **Error retrieving recovery group *recoveryGroupName* event log.**

Explanation:

Because of an error, the `tslsrecoverygroupevents` command was unable to retrieve the full event log.

User response:
None.

6027-3025 [E] Device *deviceName* does not exist or is not active on this node.

Explanation:
The specified device was not found on this node.

User response:
None.

6027-3026 [E] Recovery group *recoveryGroupName* does not have an active log home vdisk.

Explanation:
The indicated recovery group does not have an active log vdisk. This may be because the log home vdisk has not yet been created, because a previously existing log home vdisk has been deleted, or because the server is in the process of recovery.

User response:
Create a log home vdisk if none exists. Retry the command.

6027-3027 [E] Cannot configure NSD-RAID services on this node.

Explanation:
NSD-RAID services are not supported on this operating system or node hardware.

User response:
Configure a supported node type as the NSD RAID server and restart the GPFS daemon.

6027-3028 [E] There is not enough space in declustered array *declusteredArrayName* for the requested vdisk size. The maximum possible size for this vdisk is *size*.

Explanation:
There is not enough space in the declustered array for the requested vdisk size.

User response:
Create a smaller vdisk, remove existing vdisks or add additional pdisks to the declustered array.

6027-3029 [E] There must be at least *number* non-spare pdisks in declustered array *declusteredArrayName* to avoid falling below the code width of vdisk *vdiskName*.

Explanation:
A change of spares operation failed because the resulting number of non-spare pdisks would fall below the code width of the indicated vdisk.

User response:
Add additional pdisks to the declustered array.

6027-3030 [E] There must be at least *number* non-spare pdisks in declustered array *declusteredArrayName* for configuration data replicas.

Explanation:
A delete pdisk or change of spares operation failed because the resulting number of non-spare pdisks would fall below the number required to hold configuration data for the declustered array.

User response:
Add additional pdisks to the declustered array. If replacing a pdisk, use `mmchcarrier` or `mmaddpdisk --replace`.

6027-3031 [E] There is not enough available configuration data space in declustered array *declusteredArrayName* to complete this operation.

Explanation:
Creating a vdisk, deleting a pdisk, or changing the number of spares failed because there is not enough available space in the declustered array for configuration data.

User response:
Replace any failed pdisks in the declustered array and allow time for rebalance operations to more evenly distribute the available space. Add pdisks to the declustered array.

6027-3032 [E] Temporarily unable to create vdisk *vdiskName* because more time is required to rebalance the available space in declustered array *declusteredArrayName*.

Explanation:
Cannot create the specified vdisk until rebuild and rebalance processes are able to more evenly distribute the available space.

User response:
Replace any failed pdisks in the recovery group, allow time for rebuild and rebalance processes to more evenly distribute the spare space within the array, and retry the command.

6027-3034 [E] The input pdisk name (*pdiskName*) did not match the pdisk name found on disk (*pdiskName*).

Explanation:

Cannot add the specified pdisk, because the input *pdiskName* did not match the *pdiskName* that was written on the disk.

User response:
Verify the input file and retry the command.

6027-3035 [A] Cannot configure NSD-RAID services. maxblocksize must be at least *value*.

Explanation:
The GPFS daemon is starting and cannot initialize the NSD-RAID services because the `maxblocksize` attribute is too small.

User response:
Correct the `maxblocksize` attribute and restart the GPFS daemon.

6027-3036 [E] Partition size must be a power of 2.

Explanation:
The `partitionSize` parameter of some declustered array was invalid.

User response:
Correct the `partitionSize` parameter and reissue the command.

6027-3037 [E] Partition size must be between *number* and *number*.

Explanation:
The `partitionSize` parameter of some declustered array was invalid.

User response:
Correct the `partitionSize` parameter to a power of 2 within the specified range and reissue the command.

6027-3038 [E] AU log too small; must be at least *number* bytes.

Explanation:
The `auLogSize` parameter of a new declustered array was invalid.

User response:
Increase the `auLogSize` parameter and reissue the command.

6027-3039 [E] A vdisk with disk usage vdiskLogTip must be the first vdisk created in a recovery group.

Explanation:
The `--logTip` disk usage was specified for a vdisk other than the first one created in a recovery group.

User response:
Retry the command with a different disk usage.

6027-3040 [E] Declustered array configuration data does not fit.

Explanation:
There is not enough space in the pdisks of a new declustered array to hold the AU log area using the current partition size.

User response:
Increase the `partitionSize` parameter or decrease the `auLogSize` parameter and reissue the command.

6027-3041 [E] Declustered array attributes cannot be changed.

Explanation:
The `partitionSize`, `auLogSize`, and `canHoldVCD` attributes of a declustered array cannot be changed after the the declustered array has been created. They may only be set by a command that creates the declustered array.

User response:
Remove the `partitionSize`, `auLogSize`, and `canHoldVCD` attributes from the input file of the `mmaddpdisk` command and reissue the command.

6027-3042 [E] The log tip vdisk cannot be destroyed if there are other vdisks.

Explanation:
In recovery groups with versions prior to 3.5.0.11, the log tip vdisk cannot be destroyed if other vdisks still exist within the recovery group.

User response:
Remove the user vdisks or upgrade the version of the recovery group with `mmchrecoverygroup --version`, then retry the command to remove the log tip vdisk.

6027-3043 [E] Log vdisks cannot have multiple use specifications.

Explanation:
A vdisk can have usage `vdiskLog`, `vdiskLogTip`, or `vdiskLogReserved`, but not more than one.

User response:
Retry the command with only one of the `--log`, `--logTip`, or `--logReserved` attributes.

6027-3044 [E] Unable to determine resource requirements for all the recovery groups served by node *value*: to override this check reissue the command with the -v no flag.

Explanation:
A recovery group or vdisk is being created, but IBM Spectrum Scale can not determine if there are enough non-stealable buffer resources to allow the node to successfully serve all the recovery groups at the same time once the new object is created.

User response:

You can override this check by reissuing the command with the `-v flag`.

6027-3045 [W] Buffer request exceeds the non-stealable buffer limit. Check the configuration attributes of the recovery group servers: pagepool, nsdRAIDBufferPoolSizePct, nsdRAIDNonStealableBufPct.

Explanation:
The limit of non-stealable buffers has been exceeded. This is probably because the system is not configured correctly.

User response

Check the settings of the `pagepool`, `nsdRAIDBufferPoolSizePct`, and `nsdRAIDNonStealableBufPct` attributes and make sure the server has enough real memory to support the configured values.

Use the `mmchconfig` command to correct the configuration.

6027-3046 [E] The nonStealable buffer limit may be too low on server *serverName* or the pagepool is too small. Check the configuration attributes of the recovery group servers: pagepool, nsdRAIDBufferPoolSizePct, nsdRAIDNonStealableBufPct.

Explanation:
The limit of non-stealable buffers is too low on the specified recovery group server. This is probably because the system is not configured correctly.

User response

Check the settings of the `pagepool`, `nsdRAIDBufferPoolSizePct`, and `nsdRAIDNonStealableBufPct` attributes and make sure the server has sufficient real memory to support the configured values. The specified configuration variables should be the same for the recovery group servers.

Use the `mmchconfig` command to correct the configuration.

6027-3047 [E] Location of pdisk *pdiskName* is not known.

Explanation:
IBM Spectrum Scale is unable to find the location of the given pdisk.

User response:
Check the disk enclosure hardware.

6027-3048 [E] Pdisk *pdiskName* is not currently scheduled for replacement.

Explanation:
A pdisk specified in a `tschcarrier` or `tsaddpdisk` command is not currently scheduled for replacement.

User response:
Make sure the correct disk location code or pdisk name was given. For the `tschcarrier` command, the `--force-release` option can be used to override the check.

6027-3049 [E] The minimum size for vdisk *vdiskName* is *number*.

Explanation:
The vdisk size was too small.

User response:
Increase the size of the vdisk and retry the command.

6027-3050 [E] There are already *number* suspended pdisks in declustered array *arrayName*. You must resume pdisks in the array before suspending more.

Explanation:
The number of suspended pdisks in the declustered array has reached the maximum limit. Allowing more pdisks to be suspended in the array would put data availability at risk.

User response:
Resume one more suspended pdisks in the array by using the `mmchcarrier` or `mmchpdisk` commands then retry the command.

6027-3051 [E] Checksum granularity must be *number* or *number*.

Explanation:
The only allowable values for the `checksumGranularity` attribute of a data vdisk are 8K and 32K.

User response:
Change the `checksumGranularity` attribute of the vdisk, then retry the command.

6027-3052 [E] Checksum granularity cannot be specified for log vdisks.

Explanation:
The `checksumGranularity` attribute cannot be applied to a log vdisk.

User response:
Remove the `checksumGranularity` attribute of the log vdisk, then retry the command.

6027-3053 [E] Vdisk block size must be between *number* and *number* for the

specified code when checksum granularity *number* is used.

Explanation:
An invalid vdisk block size was specified. The message lists the allowable range of block sizes.

User response:
Use a vdisk virtual block size within the range shown, or use a different vdisk RAID code, or use a different checksum granularity.

6027-3054 [W] Disk in location *locationCode* failed to come online.

Explanation:
The `mmchcarrier` command detected an error when trying to bring a disk back online.

User response:
Make sure the disk is firmly seated and run the command again. Check the operating system error log.

6027-3055 [E] The fault tolerance of the code cannot be greater than the fault tolerance of the internal configuration data.

Explanation:
The RAID code specified for a new vdisk is more fault-tolerant than the configuration data that will describe the vdisk.

User response:
Use a code with a smaller fault tolerance.

6027-3056 [E] Long and short term event log size and fast write log percentage are only applicable to log home vdisk.

Explanation:
The `longTermEventLogSize`, `shortTermEventLogSize`, and `fastWriteLogPct` options are only applicable to log home vdisk.

User response:
Remove any of these options and retry vdisk creation.

6027-3057 [E] Disk enclosure is no longer reporting information on location *locationCode*.

Explanation:
The disk enclosure reported an error when IBM Spectrum Scale tried to obtain updated status on the disk location.

User response:
Try running the command again. Make sure that the disk enclosure firmware is current. Check for improperly-seated connectors within the disk enclosure.

6027-3058 [A] GSS license failure - IBM Spectrum Scale RAID services will not be configured on this node.

Explanation:
The Elastic Storage System has not been installed validly. Therefore, IBM Spectrum Scale RAID services will not be configured.

User response:
Install a licensed copy of the base IBM Spectrum Scale code and restart the GPFS daemon.

6027-3059 [E] The serviceDrain state is only permitted when all nodes in the cluster are running daemon version *version* or higher.

Explanation:
The `mmchpdisk` command option `--begin-service-drain` was issued, but there are backlevel nodes in the cluster that do not support this action.

User response:
Upgrade the nodes in the cluster to at least the specified version and run the command again.

6027-3060 [E] Block sizes of all log vdisks must be the same.

Explanation:
The block sizes of the log tip vdisk, the log tip backup vdisk, and the log home vdisk must all be the same.

User response:
Try running the command again after adjusting the block sizes of the log vdisks.

6027-3061 [E] Cannot delete path *pathName* because there would be no other working paths to pdisk *pdiskName* of RG *recoveryGroupName*.

Explanation:
When the `-v yes` option is specified on the `--delete-paths` subcommand of the `tschrecgroup` command, it is not allowed to delete the last working path to a pdisk.

User response:
Try running the command again after repairing other broken paths for the named pdisk, or reduce the list of paths being deleted, or run the command with `-v no`.

6027-3062 [E] Recovery group version *version* is not compatible with the current recovery group version.

Explanation:
The recovery group version specified with the `--version` option does not support all of the features currently supported by the recovery group.

User response:

Run the command with a new value for `--version`. The allowable values will be listed following this message.

6027-3063 [E] Unknown recovery group version *version*.

Explanation:
The recovery group version named by the argument of the `--version` option was not recognized.

User response:
Run the command with a new value for `--version`. The allowable values will be listed following this message.

6027-3064 [I] Allowable recovery group versions are:

Explanation:
Informational message listing allowable recovery group versions.

User response:
Run the command with one of the recovery group versions listed.

6027-3065 [E] The maximum size of a log tip vdisk is *size*.

Explanation:
Running `mmcrvdisk` for a log tip vdisk failed because the size is too large.

User response:
Correct the size parameter and run the command again.

6027-3066 [E] A recovery group may only contain one log tip vdisk.

Explanation:
A log tip vdisk already exists in the recovery group.

User response:
None.

6027-3067 [E] Log tip backup vdisks not supported by this recovery group version.

Explanation:
Vdisks with usage type `vdiskLogTipBackup` are not supported by all recovery group versions.

User response:
Upgrade the recovery group to a later version using the `--version` option of `mmchrecoverygroup`.

6027-3068 [E] The sizes of the log tip vdisk and the log tip backup vdisk must be the same.

Explanation:
The log tip vdisk must be the same size as the log tip backup vdisk.

User response:
Adjust the vdisk sizes and retry the `mmcrvdisk` command.

6027-3069 [E] Log vdisks cannot use code *codeName*.

Explanation:
Log vdisks must use a RAID code that uses replication, or be unreplicated. They cannot use parity-based codes such as 8+2P.

User response:
Retry the command with a valid RAID code.

6027-3070 [E] Log vdisk *vdiskName* cannot appear in the same declustered array as log vdisk *vdiskName*.

Explanation:
No two log vdisks may appear in the same declustered array.

User response:
Specify a different declustered array for the new log vdisk and retry the command.

6027-3071 [E] Device not found: *deviceName*.

Explanation:
A device name given in an `mmcrrecoverygroup` or `mmaddpdisk` command was not found.

User response:
Check the device name.

6027-3072 [E] Invalid device name: *deviceName*.

Explanation:
A device name given in an `mmcrrecoverygroup` or `mmaddpdisk` command is invalid.

User response:
Check the device name.

6027-3073 [E] Error formatting pdisk *pdiskName* on device *diskName*.

Explanation:
An error occurred when trying to format a new pdisk.

User response:
Check that the disk is working properly.

6027-3074 [E] Node *nodeName* not found in cluster configuration.

Explanation:
A node name specified in a command does not exist in the cluster configuration.

User response:
Check the command arguments.

6027-3075 [E] The --servers list must contain the current node, *nodeName*.

Explanation:

The `--servers` list of a tscrrecgroup command does not list the server on which the command is being run.

User response:
Check the `--servers` list. Make sure the tscrrecgroup command is run on a server that will actually server the recovery group.

6027-3076 [E] Remote pdisks are not supported by this recovery group version.

Explanation:
Pdisks that are not directly attached are not supported by all recovery group versions.

User response:
Upgrade the recovery group to a later version using the `--version` option of `mmchrecoverygroup`.

6027-3077 [E] There must be at least *number* pdisks in recovery group *recoveryGroupName* for configuration data replicas.

Explanation:
A change of pdisks failed because the resulting number of pdisks would fall below the needed replication factor for the recovery group descriptor.

User response:
Do not attempt to delete more pdisks.

6027-3078 [E] Replacement threshold for declustered array *declusteredArrayName* of recovery group *recoveryGroupName* cannot exceed *number*.

Explanation:
The replacement threshold cannot be larger than the maximum number of pdisks in a declustered array. The maximum number of pdisks in a declustered array depends on the version number of the recovery group. The current limit is given in this message.

User response:
Use a smaller replacement threshold or upgrade the recovery group version.

6027-3079 [E] Number of spares for declustered array *declusteredArrayName* of recovery group *recoveryGroupName* cannot exceed *number*.

Explanation:
The number of spares cannot be larger than the maximum number of pdisks in a declustered array. The maximum number of pdisks in a declustered array depends on the version number of the recovery group. The current limit is given in this message.

User response:
Use a smaller number of spares or upgrade the recovery group version.

6027-3080 [E] Cannot remove pdisk *pdiskName* because declustered array *declusteredArrayName* would have fewer disks than its replacement threshold.

Explanation:
The replacement threshold for a declustered array must not be larger than the number of pdisks in the declustered array.

User response:
Reduce the replacement threshold for the declustered array, then retry the `mmdelpdisk` command.

6027-3084 [E] VCD spares feature must be enabled before being changed. Upgrade recovery group version to at least *version* to enable it.

Explanation:
The vdisk configuration data (VCD) spares feature is not supported in the current recovery group version.

User response:
Apply the recovery group version that is recommended in the error message and retry the command.

6027-3085 [E] The number of VCD spares must be greater than or equal to the number of spares in declustered array *declusteredArrayName*.

Explanation:
Too many spares or too few vdisk configuration data (VCD) spares were specified.

User response:
Retry the command with a smaller number of spares or a larger number of VCD spares.

6027-3086 [E] There is only enough free space to allocate *n* VCD spare(s) in declustered array *declusteredArrayName*.

Explanation:
Too many vdisk configuration data (VCD) spares were specified.

User response:
Retry the command with a smaller number of VCD spares.

6027-3087 [E] Specifying Pdisk rotation rate not supported by this recovery group version.

Explanation:
Specifying the Pdisk rotation rate is not supported by all recovery group versions.

User response:
Upgrade the recovery group to a later version using the `--version` option of the `mmchrecoverygroup` command. Or, don't specify a rotation rate.

6027-3088 [E] Specifying Pdisk expected number of paths not supported by this recovery group version.

Explanation:
Specifying the expected number of active or total pdisk paths is not supported by all recovery group versions.

User response:
Upgrade the recovery group to a later version using the `--version` option of the `mmchrecoverygroup` command. Or, don't specify the expected number of paths.

6027-3089 [E] Pdisk *pdiskName* location *locationCode* is already in use.

Explanation:
The pdisk location that was specified in the command conflicts with another pdisk that is already in that location. No two pdisks can be in the same location.

User response:
Specify a unique location for this pdisk.

6027-3090 [E] Enclosure control command failed for pdisk *pdiskName* of RG *recoveryGroupName* in location *locationCode*: err *errorNum*. Examine mmfs log for tsctlenclslot, tsonosdisk and tsoffosdisk errors.

Explanation:
A command used to control a disk enclosure slot failed.

User response:
Examine the mmfs log files for more specific error messages from the **tsctlenclslot**, **tsonosdisk**, and **tsoffosdisk** commands.

6027-3091 [W] A command to control the disk enclosure failed with error code *errorNum*. As a result, enclosure indicator lights may not have changed to the correct states. Examine the `mmfs` log on nodes attached to the disk enclosure for messages from the `tsctlenclslot`, `tsonosdisk`, and `tsoffosdisk` commands for more detailed information.

Explanation:
A command used to control disk enclosure lights and carrier locks failed. This is not a fatal error.

User response:
Examine the mmfs log files on nodes attached to the disk enclosure for error messages from the **tsctlenclslot**, **tsonosdisk**, and **tsoffosdisk** commands for more detailed information. If the carrier failed to unlock, either retry the command or use the manual override.

6027-3092 [I] Recovery group *recoveryGroupName* assignment delay *delaySeconds* seconds for safe recovery.

Explanation:
The recovery group must wait before meta-data recovery. Prior disk lease for the failing manager must first expire.

User response:
None.

6027-3093 [E] Checksum granularity must be *number* or *number* for log vdisks.

Explanation:
The only allowable values for the checksumGranularity attribute of a log vdisk are 512 and 4K.

User response:
Change the checksumGranularity attribute of the vdisk, then retry the command.

6027-3094 [E] Due to the attributes of other log vdisks, the checksum granularity of this vdisk must be *number*.

Explanation:
The checksum granularities of the log tip vdisk, the log tip backup vdisk, and the log home vdisk must all be the same.

User response:
Change the checksumGranularity attribute of the new log vdisk to the indicated value, then retry the command.

6027-3095 [E] The specified declustered array name (*declusteredArrayName*) for the new pdisk *pdiskName* must be *declusteredArrayName*.

Explanation:
When replacing an existing pdisk with a new pdisk, the declustered array name for the new pdisk must match the declustered array name for the existing pdisk.

User response:
Change the specified declustered array name to the indicated value, then run the command again.

6027-3096 [E] Internal error encountered in NSD-RAID command: err=*errorNum*.

Explanation:

An unexpected GPFS NSD-RAID internal error occurred.

User response:
Contact the IBM Support Center.

6027-3097 [E] Missing or invalid pdisk name (*pdiskName*).

Explanation:
A pdisk name specified in an **mmcrrecoverygroup** or **mmaddpdisk** command is not valid.

User response:
Specify a pdisk name that is 63 characters or less. Valid characters are: a to z, A to Z, 0 to 9, and underscore (_).

6027-3098 [E] Pdisk name *pdiskName* is already in use in recovery group *recoveryGroupName*.

Explanation:
The pdisk name already exists in the specified recovery group.

User response:
Choose a pdisk name that is not already in use.

6027-3099 [E] Device with path(s) *pathName* is specified for both new pdisks *pdiskName* and *pdiskName*.

Explanation:
The same device is specified for more than one pdisk in the stanza file. The device can have multiple paths, which are shown in the error message.

User response:
Specify different devices for different new pdisks, respectively, and run the command again.

6027-3800 [E] Device with path(s) *pathName* for new pdisk *pdiskName* is already in use by pdisk *pdiskName* of recovery group *recoveryGroupName*.

Explanation:
The device specified for a new pdisk is already being used by an existing pdisk. The device can have multiple paths, which are shown in the error message.

User response:
Specify an unused device for the pdisk and run the command again.

6027-3801 [E] [E] The checksum granularity for log vdisks in declustered array *declusteredArrayName* of RG *recoveryGroupName* must be at least *number* bytes.

Explanation:
Use a checksum granularity that is not smaller than the minimum value given. You can use the mmlspdisk command to view the logical block sizes of the pdisks in this array to identify which pdisks are driving the limit.

User response:
Change the checksumGranularity attribute of the new log vdisk to the indicated value, and then retry the command.

6027-3802 [E] [E] Pdisk *pdiskName* of RG *recoveryGroupName* has a logical block size of *number* bytes; the maximum logical block size for pdisks in declustered array *declusteredArrayName* cannot exceed the log checksum granularity of *number* bytes.

Explanation:
Logical block size of pdisks added to this declustered array must not be larger than any log vdisk's checksum granularity.

User response:
Use pdisks with equal or smaller logical block size than the log vdisk's checksum granularity.

6027-3803 [E] [E] NSD format version 2 feature must be enabled before being changed. Upgrade recovery group version to at least *recoveryGroupVersion* to enable it.

Explanation:
NSD format version 2 feature is not supported in current recovery group version.

User response:
Apply the recovery group version recommended in the error message and retry the command.

6027-3804 [W] Skipping upgrade of pdisk *pdiskName* because the disk capacity of *number* bytes is less than the *number* bytes required for the new format.

Explanation:
The existing format of the indicated pdisk is not compatible with NSD V2 descriptors.

User response:
A complete format of the declustered array is required in order to upgrade to NSD V2.

6027-3805 [E] NSD format version 2 feature is not supported by the current recovery group version. A recovery group version of at least *rgVersion* is required for this feature.

Explanation:

NSD format version 2 feature is not supported in the current recovery group version.

User response:
Apply the recovery group version recommended in the error message and retry the command.

6027-3806 [E] **The device given for pdisk *pdiskName* has a logical block size of *logicalBlockSize* bytes, which is not supported by the recovery group version.**

Explanation:
The current recovery group version does not support disk drives with the indicated logical block size.

User response:
Use a different disk device or upgrade the recovery group version and retry the command.

6027-3807 [E] **NSD version 1 specified for pdisk *pdiskName* requires a disk with a logical block size of 512 bytes. The supplied disk has a block size of *logicalBlockSize* bytes. For this disk, you must use at least NSD version 2.**

Explanation:
Requested logical block size is not supported by NSD format version 1.

User response:
Correct the input file to use a different disk or specify a higher NSD format version.

6027-3808 [E] **Pdisk *pdiskName* must have a capacity of at least *number* bytes for NSD version 2.**

Explanation:
The pdisk must be at least as large as the indicated minimum size in order to be added to the declustered array.

User response:
Correct the input file and retry the command.

6027-3809 [I] **Pdisk *pdiskName* can be added as NSD version 1.**

Explanation:
The pdisk has enough space to be configured as NSD version 1.

User response:
Specify NSD version 1 for this disk.

6027-3810 [W] **[W] Skipping the upgrade of pdisk *pdiskName* because no I/O paths are currently available.**

Explanation:
There is no I/O path available to the indicated pdisk.

User response:
Try running the command again after repairing the broken I/O path to the specified pdisk.

6027-3811 [E] **Unable to *action* vdisk MDI.**

Explanation:
The **tscrvdisk** command could not create or write the necessary vdisk MDI.

User response:
Retry the command.

6027-3812 [I] **Log group *logGroupName* assignment delay *delaySeconds* seconds for safe recovery.**

Explanation:
The recovery group configuration manager must wait. Prior disk lease for the failing manager must expire before assigning a new worker to the log group.

User response:
None.

6027-3813 [A] **Recovery group *recoveryGroupName* could not be served by node *nodeName*.**

Explanation:
The recovery group configuration manager could not perform a node assignment to manage the recovery group.

User response:
Check whether there are sufficient nodes and whether errors are recorded in the recovery group event log.

6027-3814 [A] **Log group *logGroupName* could not be served by node *nodeName*.**

Explanation:
The recovery group configuration manager could not perform a node assignment to manage the log group.

User response:
Check whether there are sufficient nodes and whether errors are recorded in the recovery group event log.

6027-3815 [E] **Erasure code not supported by this recovery group version.**

Explanation:
Vdisks with 4+2P and 4+3P erasure codes are not supported by all recovery group versions.

User response:
Upgrade the recovery group to a later version using the **--version** option of the **mmchrecoverygroup** command.

6027-3816 [E] **Invalid declustered array name (*declusteredArrayName*).**

Explanation:

A declustered array name given in the **mmcrrecoverygroup** or **mmaddpdisk** command is invalid.

User response:
Use only the characters a-z, A-Z, 0-9, and underscore to specify a declustered array name and you can specify up to 63 characters.

6027-3817 [E] Invalid log group name (*logGroupName*).

Explanation:
A log group name given in the **mmcrrecoverygroup** or **mmaddpdisk** command is invalid.

User response:
Use only the characters a-z, A-Z, 0-9, and underscore to specify a declustered array name and you can specify up to 63 characters.

6027-3818 [E] Cannot create log group *logGroupName*; there can be at most *number* log groups in a recovery group.

Explanation:
The number of log groups allowed in a recovery group has been exceeded.

User response:
Reduce the number of log groups in the input file and retry the command.

6027-3819 [I] Recovery group *recoveryGroupName* delay *delaySeconds* seconds for assignment.

Explanation:
The recovery group configuration manager must wait before assigning a new manager to the recovery group.

User response:
None.

6027-3820 [E] Specifying canHoldVCD not supported by this recovery group version.

Explanation:
The ability to override the default decision of whether a declustered array is allowed to hold vdisk configuration data is not supported by all recovery group versions.

User response:
Upgrade the recovery group to a later version using the **--version** option of the **mmchrecoverygroup** command.

6027-3821 [E] Cannot set canHoldVCD=yes for small declustered arrays.

Explanation:
Declustered arrays with less than 9+vcdSpares disks cannot hold vdisk configuration data.

User response:
Add more disks to the declustered array or do not specify canHoldVCD=yes.

6027-3822 [I] Recovery group *recoveryGroupName* working index delay *delaySeconds* seconds for safe recovery.

Explanation:
Prior disk lease for the workers must expire before recovering the working index metadata.

User response:
None.

6027-3823 [E] Unknown node *nodeName* in the recovery group configuration.

Explanation:
A node name does not exist in the recovery group configuration manager.

User response:
Check for damage to the **mmsdrfs** file.

6027-3824 [E] The defined server *serverName* for recovery group *recoveryGroupName* could not be resolved.

Explanation:
The host name of recovery group server could not be resolved by gethostbyName().

User response:
Fix host name resolution.

6027-3825 [E] The defined server *serverName* for node class *nodeClassName* could not be resolved.

Explanation:
The host name of recovery group server could not be resolved by gethostbyName().

User response:
Fix host name resolution.

6027-3826 [A] Error reading volume identifier for recovery group *recoveryGroupName* from configuration file.

Explanation:
The volume identifier for the named recovery group could not be read from the **mmsdrfs** file. This should never occur.

User response:
Check for damage to the **mmsdrfs** file.

6027-3827 [A] Error reading volume identifier for vdisk *vdiskName* from configuration file.

Explanation:
The volume identifier for the named vdisk could not\ be read from the **mmsdrfs** file. This should never occur.

User response:
Check for damage to the **mmsdrfs** file.

6027-3828 [E] Vdisk *vdiskName* could not be associated with its recovery group *recoveryGroupName* and will be ignored.

Explanation:
The named vdisk cannot be associated with its recovery group.

User response:
Check for damage to the **mmsdrfs** file.

6027-3829 [E] A server list must be provided.

Explanation:
No server list is specified.

User response:
Specify a list of valid servers.

6027-3830 [E] Too many servers specified.

Explanation:
An input node list has too many nodes specified.

User response:
Verify the list of nodes and shorten the list to the supported number.

6027-3831 [E] A vdisk name must be provided.

Explanation:
A vdisk name is not specified.

User response:
Specify a vdisk name.

6027-3832 [E] A recovery group name must be provided.

Explanation:
A recovery group name is not specified.

User response:
Specify a recovery group name.

6027-3833 [E] Recovery group *recoveryGroupName* does not have an active root log group.

Explanation:
The root log group must be active before the operation is permitted.

User response:
Retry the command after the recovery group becomes fully active.

6027-3836 [I] Cannot retrieve MSID for device: *devFileName*.

Explanation:
Command usage message for **tsgetmsid**.

User response:
None.

6027-3837 [E] Error creating worker vdisk.

Explanation:
The **tscrvdisk** command could not initialize the vdisk at the worker node.

User response:
Retry the command.

6027-3838 [E] Unable to write new vdisk MDI.

Explanation:
The **tscrvdisk** command could not write the necessary vdisk MDI.

User response:
Retry the command.

6027-3839 [E] Unable to write update vdisk MDI.

Explanation:
The **tscrvdisk** command could not write the necessary vdisk MDI.

User response:
Retry the command.

6027-3840 [E] Unable to delete worker vdisk *vdiskName* err=*errorNum*.

Explanation:
The specified vdisk worker object could not be deleted.

User response:
Retry the command with a valid vdisk name.

6027-3841 [E] Unable to create new vdisk MDI.

Explanation:
The **tscrvdisk** command could not create the necessary vdisk MDI.

User response:
Retry the command.

6027-3843 [E] Error returned from node *nodeName* when preparing new pdisk *pdiskName* of RG *recoveryGroupName* for use: err *errorNum*

Explanation:
The system received an error from the given node when trying to prepare a new pdisk for use.

User response:
Retry the command.

6027-3844 [E] **Unable to prepare new pdisk *pdiskName* of RG *recoveryGroupName* for use: exit status *exitStatus*.**

Explanation:
The system received an error from the **tspreparenewpdiskforuse** script when trying to prepare a new pdisk for use.

User response:
Check the new disk and retry the command.

6027-3845 [E] **Unrecognized pdisk state: *pdiskState*.**

Explanation:
The given pdisk state name is invalid.

User response:
Use a valid pdisk state name.

6027-3846 [E] **Pdisk state change *pdiskState* is not permitted.**

Explanation:
An attempt was made to use the **mmchpdisk** command either to change an internal pdisk state, or to create an invalid combination of states.

User response:
Some internal pdisk state flags can be set indirectly by running other commands. For example, the *deleting* state can be set by using the **mmdelpdisk** command.

6027-3847 [E] **[E] The *serviceDrain* state feature must be enabled to use this command. Upgrade the recovery group version to at least *version* to enable it.**

Explanation:
The **mmchpdisk** command option **--begin-service-drain** was issued, but there are back-level nodes in the cluster that do not support this action.

User response:
Upgrade the nodes in the cluster to at least the specified version and run the command again.

6027-3848 [E] **The simulated dead and failing state feature must be enabled to use this command. Upgrade the recovery group version to at least *version* to enable it.**

Explanation:
The **mmchpdisk** command option **--begin-service-drain** was issued, but there are back-level nodes in the cluster that do not support this action.

User response:
Upgrade the nodes in the cluster to at least the specified version and run the command again.

6027-3849 [E] **The pdisk *pdiskName* of recovery group *recoveryGroupName* could not be revived. Pdisk state is *pdiskState*.**

Explanation:
An **mmchpdisk --revive** command was unable to bring a pdisk back online.

User response:
If the state is missing, restore connectivity to the disk. If the disk is in failed state replace the pdisk. A pdisk with the status dead, readOnly, failing, or slot is considered as failed.

6027-3850 [E] **Location *locationCode* contains multiple disk devices. You cannot use this command to replace disks in the specific location.**

Explanation:
The **mmvdisk pdisk replace** command or the **mmchcarrier** command was given a location that contains multiple disk devices. An example of a location with multiple disk devices is the situation where the operating system (OS) root disk and log tip devices share the same underlying storage.

User response:
If the problem PDisk is one of the log tip devices and it shares storage with other log tip devices or the OS root, first make sure that the device has failed. That is, it is in "dead", "readOnly" or "failing" state as opposed to being temporarily inaccessible because node is down. If the device is really down, delete the log tip VDisk and declustered array from the recovery group, then replace the failed hardware. Finally, re-create the log tip DA and VDisk. Refer to the product documentation for more detailed instructions.

6027-3851 [E] **Command interrupted by recovery group *recoveryGroupName* failover.**

Explanation:
A recovery group command failed because the recovery group stopped serving, probably because it failed over to another node.

User response:
Run the command again.

6027-3852 [A] **Cannot configure NSD-RAID services. The *nsdRAIDBufferPoolSizePct* attribute of the pagepool must result in at least *nsdRAIDMasterBufferPoolSize* (*number*) bytes + 128 MiB of space.**

Explanation:

The GPFS daemon is starting and cannot initialize the NSD-RAID services because of the memory consideration specified.

User response:
Correct the **nsdRAIDBufferPoolSizePct** attribute of the pagepool and restart the GPFS daemon.

6027-3853 [W] **Buffer request (*name*) exceeds the master reserved buffer limit (*number*). Check the configuration attributes of the recovery group servers: nsdRAIDMasterBufferPoolSize.**

Explanation:
The limit of master reserved buffers is exceeded. This is probably because of an improperly configured system. Check the setting of the **nsdRAIDMasterBufferPoolSize** parameter, and whether the server has sufficient memory to support the configured value.

User response:
Use the **mmchconfig** command to correct the configuration.

6027-3854 [E] **Recovery group configuration manager takeover failed: scheduled *scheduled* stopping *stopping***

Explanation:
The recovery group configuration manager takeover schedule failed.

User response:
Contact the IBM Support.

6027-3855 [E] ***rgcmRefreshConfig* error. Duplicated NID *nsdID* (*vdiskName*) found in *recoveryGroupName*.**

Explanation:
Duplicated ID found by RGCM during initialization.

User response:
Contact the IBM Support.

6027-3856 [E] **Recovery group configuration manager takeover failed: err *errorNum***

Explanation:
The recovery group configuration manager takeover failed with error.

User response:
Contact the IBM Support.

6027-3857 [E] **Log group *logGroupName* of recovery group *recoveryGroupName* could not be served.**

Explanation:
The recovery group configuration manager could not perform a node assignment to manage the log group.

User response:
Check whether there are sufficient nodes and whether errors are recorded in the recovery group event log.

6027-3858 [E] **Recovery group configuration manager failed to start. err *errorNum***

Explanation:
Recovery group configuration manager final takeover failed.

User response:
Contact IBM Support.

6027-3859 [E] **Trim to device not supported by this recovery group version. Upgrade the recovery group version to at least *version* to enable it.**

Explanation:
The ability to enable trim to device is not supported by the current recovery group version.

User response:
Upgrade the recovery group to a later version by using the **--version** option of the **mmchrecoverygroup** command.

6027-3860 [E] **Recovery group descriptor for PDisk *pdiskName* of recovery group *recoveryGroupName* could not be written because volatile write caching is enabled on this drive.**

Explanation:
GPFS Native RAID refused to write a recovery group descriptor to a drive because it detected that volatile write caching was enabled. This error can occur when creating a recovery group, adding a new PDisk to an existing recovery group, or when replacing a drive.

User response:
Disable volatile write caching and related settings to comply with supported configurations.

6027-3861 [E] **Recovery group descriptor for PDisk *pdiskName* of recovery group *recoveryGroupName* not be written err=*errNum*.**

Explanation:
GPFS Native RAID refused to write recovery group descriptor to a drive due to an internal error. This error can occur when creating a recovery group, adding a new PDisk to an existing recovery group, or when replacing a drive.

User response:
Contact IBM Support.

6027-3862 [E] Trim to declustered array *arrayName* of recovery group *recoveryGroupName* is not supported for hardware type *hardwareType*

Explanation:
Trim to device will not be enabled for a declustered array that contains drives of an unsupported hardware type.

User response:
Review hardware documentation for device trim capability or GNR trim documentation for a list of supported configurations.

6027-3863 [E] Recovery group *recoveryGroupName* stops serving after exceeding retry limit *nsdRAIDMaxRecoveryRetries*.

Explanation:
The recovery group could not start due to retry failure exceeding retry limit.

User response:
Check disk storage connection and run mmvdisk to restart the recovery group.

6027-3864 [E] [E] Slot location is missing from pdisk *pdiskName* device(s) *deviceName* of declustered array *arrayName* in recovery group *recoveryGroupName* with hardware type *hardwareType*.

Explanation:
Perform strict pdisk slot location checking to find empty slot location for this pdisk.

User response:
Review GNR documentation to make sure the disk drives are configured properly and the slot location mapping is set up correctly. Fix the problems for this pdisk and retry the command. Contact IBM support if it doesn't solve the problem.

6027-3865 [E] [E] *nFailures* empty slot location*string* found in recovery group *recoveryGroupName*. Make sure \ disk drives and slot location mapping are configured properly.

Explanation:
Perform strict pdisk slot location checking to find empty slot location for at least one pdisk. The pdisks that have problem are shown above this message.

User response:
Review GNR documentation to make sure the disk drives are configured properly and the slot location mapping is set up correctly. Fix the problems for all these pdisks and retry the command. Contact IBM support if it doesn't solve the problem.

6027-3866 [E] Log group *logGroupName* of recovery group *recoveryGroupName* not reachable. Verify recovery group health and try again.

Explanation:
Vdisk deletion fails if the log group is down. Log group and corresponding recovery group are shown in the message.

User response:
Verify the state of recovery group before performing vdisk deletion.

6027-3867 [E] Log vdisks cannot be exported as NVMe-oF target.

Explanation:
A log vdisk (loghome, logtip, logbackup) cannot be exported as an NVMe-oF target.

User response:
None

6027-3868 [E] [E] Log group affinity node feature must be enabled before creating a new log group with affinityNode option. Upgrade recovery group version to at least *rgVersion* to enable it.

Explanation:
Log group affinity node feature is not supported in the current recovery group version.

User response:
Apply the recovery group version that is recommended in the error message and retry the command.

6027-3869 [E] [E] Log group affinityNode option can only be used for creating log vdisk of a non-root log group.

Explanation:
Log group affinity node option can only be used when creating log vdisk of a non-root log group.

User response:
--affinityNode option can only be used in log vdisk stanza for a non-root log group.

6027-3870 [E] [E] Specifying daHospitalParmeters is supported starting in recovery group version *rgVersion*.

Explanation:
The ability to specify hospital parameters of a declustered array is not supported by all recovery group versions.

User response:
Upgrade the recovery group to a later version using the --version option of the mmchrecoverygroup command.

6027-3871 [E] **[E] Declustered array bit error rate settings are supported starting in recovery group version *rgVersion*.**

Explanation:
The ability to specify declustered array bit error rate settings is not supported by all recovery group versions.

User response:
Upgrade the recovery group to a later version using the --version option of the mmchrecoverygroup command.

6027-3872 [E] **[E] Invalid format for hospital stanza parameters *stanzaParameters* for declustered array *declusteredArrayName* of RG *recoveryGroupName*.**

Explanation:
A declustered array was supplied with the wrong hospital stanza parameters.

User response:
Verify the pdisk stanza file

6027-3873 [E] **[E] Bit error rate settings have been permanently disabled for declustered array *declusteredArrayName* of RG *recoveryGroupName*.**

Explanation:
The recovery group bit error rate declustered array settings were disabled during deployment and cannot be turned back on.

User response:
None.

6027-3874 [E] **[E] Cannot remove declustered array *arrayName* because it is the only remaining declustered array designated at recovery group creation time with at least *number* pdisks eligible to hold vdisk configuration data.**

Explanation:
The command failed to remove a declustered array because no other declustered array in the recovery group has sufficient pdisks to store the on-disk recovery group descriptor at the required fault tolerance level.

User response:
None

Chapter 10. Contacting IBM

Specific information about a problem such as: symptoms, traces, error logs, GPFS logs, and file system status is vital to IBM in order to resolve an IBM Spectrum Scale RAID problem.

Obtain this information as quickly as you can after a problem is detected, so that error logs will not wrap and system parameters that are always changing will be captured as close to the point of failure as possible. When a serious problem is detected, collect this information and then call IBM.

Information to collect before contacting the IBM Support Center

For effective communication with the IBM Support Center to help with problem diagnosis, you need to collect certain information.

Information to collect for all problems related to IBM Spectrum Scale RAID

Regardless of the problem encountered with IBM Spectrum Scale RAID, the following data should be available when you contact the IBM Support Center:

1. A description of the problem.
2. Output of the failing application, command, and so forth.

   Collect the output of the **gpfs.snap** and **essinstallcheck** commands from each I/O canister node.
3. A tar file generated by the `gpfs.snap` command that contains data from the nodes in the cluster. In large clusters, the `gpfs.snap` command can collect data from certain nodes (for example, the affected nodes, NSD servers, or manager nodes) using the `-N` option.

   For more information about gathering data using the `gpfs.snap` command, see the *IBM Spectrum Scale: Problem Determination Guide*.

   If the `gpfs.snap` command cannot be run, collect these items:

   a. Any error log entries that are related to the event:

      - On a Linux node, create a tar file of all the entries in the `/var/log/messages` file from all nodes in the cluster or the nodes that experienced the failure. For example, issue the following command to create a tar file that includes all nodes in the cluster:

        ```
        mmdsh -v -N all "cat /var/log/messages" > all.messages
        ```

      - On an AIX® node, issue this command:

        ```
        errpt -a
        ```

      For more information about the operating system error log facility, see the *IBM Spectrum Scale: Problem Determination Guide*.

   b. A master GPFS log file that is merged and chronologically sorted for the date of the failure. See the *IBM Spectrum Scale: Problem Determination Guide* for information about creating a master GPFS log file.

   c. If the cluster was configured to store dumps, collect any internal GPFS dumps written to that directory relating to the time of the failure. The default directory is `/tmp/mmfs`.

   d. On a failing Linux node, gather the installed software packages and the versions of each package by issuing this command:

      ```
      rpm -qa
      ```

e. On a failing AIX node, gather the name, most recent level, state, and description of all installed software packages by issuing this command:

```
lslpp -l
```

f. For the file system attributes for all of the failing file systems, issue:

```
mmlsfs Device
```

g. For the current configuration and state of the disks for all of the failing file systems, issue:

```
mmlsdisk Device
```

h. A copy of file `/var/mmfs/gen/mmsdrfs` from the primary cluster configuration server.

4. If you are experiencing one of the following problems, see the appropriate section before contacting the IBM Support Center:
   - For delay and deadlock issues, see "Additional information to collect for delays and deadlocks" on page 118.
   - For file system corruption or MMFS_FSSTRUCT errors, see "Additional information to collect for file system corruption or MMFS_FSSTRUCT errors" on page 118.
   - For GPFS daemon crashes, see "Additional information to collect for GPFS daemon crashes" on page 119.

Additional information to collect for delays and deadlocks

When a delay or deadlock situation is suspected, the IBM Support Center will need additional information to assist with problem diagnosis. If you have not done so already, make sure you have the following information available before contacting the IBM Support Center:

1. Everything that is listed in "Information to collect for all problems related to IBM Spectrum Scale RAID" on page 117.
2. The deadlock debug data collected automatically.
3. If the cluster size is relatively small and the `maxFilesToCache` setting is not high (less than 10,000), issue the following command:

   ```
   gpfs.snap --deadlock
   ```

   If the cluster size is large or the `maxFilesToCache` setting is high (greater than 1M), issue the following command:

   ```
   gpfs.snap --deadlock --quick
   ```

   For more information about the `--deadlock` and `--quick` options, see the *IBM Spectrum Scale: Problem Determination Guide*.

Additional information to collect for file system corruption or MMFS_FSSTRUCT errors

When file system corruption or `MMFS_FSSTRUCT` errors are encountered, the IBM Support Center will need additional information to assist with problem diagnosis. If you have not done so already, make sure you have the following information available before contacting the IBM Support Center:

1. Everything that is listed in "Information to collect for all problems related to IBM Spectrum Scale RAID" on page 117.
2. Unmount the file system everywhere, then run `mmfsck -n` in offline mode and redirect it to an output file.

The IBM Support Center will determine when and if you should run the `mmfsck -y` command.

Additional information to collect for GPFS daemon crashes

When the GPFS daemon is repeatedly crashing, the IBM Support Center will need additional information to assist with problem diagnosis. If you have not done so already, make sure you have the following information available before contacting the IBM Support Center:

1. Everything that is listed in "Information to collect for all problems related to IBM Spectrum Scale RAID" on page 117.
2. Make sure the `/tmp/mmfs` directory exists on all nodes. If this directory does not exist, the GPFS daemon will not generate internal dumps.
3. Set the traces on this cluster and *all* clusters that mount any file system from this cluster:

```
mmtracectl --set --trace=def --trace-recycle=global
```

4. Start the trace facility by issuing:

```
mmtracectl --start
```

5. Recreate the problem if possible or wait for the assert to be triggered again.
6. Once the assert is encountered on the node, turn off the trace facility by issuing:

```
mmtracectl --off
```

   If traces were started on multiple clusters, `mmtracectl --off` should be issued immediately on all clusters.

7. Collect `gpfs.snap` output:

```
gpfs.snap
```

How to contact the IBM Support Center

IBM support is available for various types of IBM hardware and software problems that IBM Spectrum Scale customers might encounter.

These problems include the following:

- IBM hardware failure
- Node halt or crash not related to a hardware failure
- Node hang or response problems
- Failure in other software supplied by IBM

If you have an IBM Software Maintenance service contract

If you have an IBM Software Maintenance service contract, contact IBM Support as follows:

Your location	Method of contacting IBM Support
In the United States	Call **1-800-IBM-SERV** for support.
Outside the United States	Contact your local IBM Support Center or see the Directory of worldwide contacts (www.ibm.com/planetwide).

When you contact IBM Support, the following will occur:

1. You will be asked for the information you collected in "Information to collect before contacting the IBM Support Center" on page 117.
2. You will be given a time period during which an IBM representative will return your call. Be sure that the person you identified as your contact can be reached at the phone number you provided in the PMR.

3. An online Problem Management Record (PMR) will be created to track the problem you are reporting, and you will be advised to record the PMR number for future reference.
4. You might be requested to send data related to the problem you are reporting, using the PMR number to identify it.
5. Should you need to make subsequent calls to discuss the problem, you will also use the PMR number to identify the problem.

If you do not have an IBM Software Maintenance service contract

If you do not have an IBM Software Maintenance service contract, contact your IBM sales representative to find out how to proceed. Be prepared to provide the information you collected in "Information to collect before contacting the IBM Support Center" on page 117.

For failures in non-IBM software, follow the problem-reporting procedures provided with that product.

Appendix A. Cleaning up ESS environments

This section contains information for users who want to securely erase or clean up their IBM Elastic Storage Server environments.

For more information to securely erase a drive, see https://www.cyberciti.biz/faq/how-do-i-permanently-erase-hard-disk/.

Complete the following steps to securely clean up your ESS environments:

1. Unmount the file system by issuing the **mmumount all -a** command.
2. Delete the file system by issuing the **mmdelfs <FS>** command.

   Where *<FS>* should be `fs3k`.

   Note: You can also get this information by issuing the **mmlsfs all** command.
3. Delete the virtual disk sets by issuing the **mmvdisk vdiskset delete --vdisk-set <vdisk set>** command.

   Note: You can get the virtual disk set by issuing the **mmvdisk vdiskset list** command.
4. Undefine virtual disk sets by issuing the **mmvdisk vdiskset undefine --vdisk-set <vdisk set>** command.
5. List recovery groups by issuing the **mmvdisk recoverygroup list** command.
6. Delete recovery groups by issuing the **mmvdisk recoverygroup delete --recovery-group <RG>** command.
7. List virtual disk servers by issuing the **mmvdisk server list** command.
8. Unconfigure the virtual disk servers by issuing the **mmvdisk server unconfigure --node-class <class>** command.
9. Delete node class by issuing the **mmvdisk nodeclass delete --node-class <class>** command.
10. Delete the cluster by issuing the **mmshutdown -a** and **mmdelnode -a** commands.
11. At this point, the recovery groups are deleted. However, to proceed further, overwrite the disk sectors as a form of secure erase by issuing the **esscheckdisks** command.
12. Overwrite the disk sectors by using shred, and then perform random and sequential write tests on the disks.

    ESSENV=INSTALL /opt/ibm/ess/tools/bin/esscheckdisks --enclosure-list all --iotest a --write-enable --local --ioengine s

    You can also add the following flags for longer scrubbing and bigger batch size desired:

    --batch-size BATCH-SIZE

    This provides the batch size of the test. Select *0* for all. Default batch size is 60.

    --duration TEST-DURATION

    This provides the run time per test in seconds. Default value is 30 seconds.

    Enter *0* to run to the end of the disk.

    For example:

```
top - 11:05:27 up 1 day, 3:08, 2 users, load average: 32.01, 10.81, 3.93
Tasks: 1420 total, 1 running, 1419 sleeping, 0 stopped, 0 zombie
 %Cpu(s): 1.4 us, 0.2 sy, 0.0 ni, 70.1 id, 28.3 wa, 0.0 hi, 0.0 si, 0.0 st
 KiB Mem : 12966195+total, 12140812+free, 3898368 used, 4355456 buff/cache
 KiB Swap: 4095936 total, 4095936 free, 0 used. 12401952+avail Mem
 PID USER PR NI VIRT RES SHR S %CPU %MEM TIME+ COMMAND
 15867 root 20 0 110208 2432 1792 D 17.1 0.0 0:02.59 shred
 15872 root 20 0 110208 2368 1792 D 16.8 0.0 0:02.57 shred
```

```
  15001 root 20 0 120256 10048 4480 R 1.3 0.0 0:00.60 top
  15834 root 20 0 110208 2368 1792 D 1.0 0.0 0:00.09 shred
  15796 root 20 0 110208 2368 1792 D 0.7 0.0 0:00.08 shred
  15798 root 20 0 110208 2432 1792 D 0.7 0.0 0:00.08 shred
  15802 root 20 0 110208 2368 1792 D 0.7 0.0 0:00.08 shred
```

Note: Run this command from one of the 3000 canisters. This is not needed for an ESS Management Server.

When this is done, you should be ready to tear down and ship. You might want to clean up any logs.

You can clear or delete the following directories from the ESS Management Server:

- `/var/log/xcat`
- `/var/log/consoles`
- `/var/adm/ras`
- `/var/mmfs/gen`
- `/root/.ssh`
- `/etc/yum.repos.d`
- `/etc/hosts`
- `/var/log/messages`
- `/var/log/anaconda`

I/O nodes clear the following directories:

- `/var/log/xcat`
- `/var/adm/ras`
- `/var/mmfs/gen`
- `/root/.ssh`
- `/etc/yum.repos.d`
- `/etc/hosts`
- `/var/log/messages`
- `/var/log/anaconda`

You can also delete the container image from the ESS Management Server, but it is not required.

podman stop <container>

podman rm <container name>

podman image rm <container id> -f

Accessibility features for the system

Accessibility features help users who have a disability, such as restricted mobility or limited vision, to use information technology products successfully.

Accessibility features

The following list includes the major accessibility features in IBM Spectrum Scale RAID:

- Keyboard-only operation
- Interfaces that are commonly used by screen readers
- Keys that are discernible by touch but do not activate just by touching them
- Industry-standard devices for ports and connectors
- The attachment of alternative input and output devices

IBM Documentation, and its related publications, are accessibility-enabled.

Keyboard navigation

This product uses standard Microsoft Windows navigation keys.

IBM and accessibility

See the IBM Human Ability and Accessibility Center (www.ibm.com/able) for more information about the commitment that IBM has to accessibility.

Notices

This information was developed for products and services offered in the U.S.A.

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service.

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not grant you any license to these patents. You can send license inquiries, in writing, to:

IBM Director of Licensing IBM Corporation North Castle Drive Armonk, NY 10504-1785 U.S.A.

For license inquiries regarding double-byte (DBCS) information, contact the IBM Intellectual Property Department in your country or send inquiries, in writing, to:

Intellectual Property Licensing Legal and Intellectual Property Law IBM Japan Ltd. 19-21,

Nihonbashi-Hakozakicho, Chuo-ku Tokyo 103-8510, Japan

The following paragraph does not apply to the United Kingdom or any other country where such provisions are inconsistent with local law:

INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some states do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this IBM product and use of those Web sites is at your own risk.

IBM may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation to you.

Licensees of this program who wish to have information about it for the purpose of enabling: (i) the exchange of information between independently created programs and other programs (including this one) and (ii) the mutual use of the information which has been exchanged, should contact:

IBM Corporation
Dept. 30ZA/Building 707
Mail Station P300
2455 South Road,
Poughkeepsie, NY 12601-5400
U.S.A.

Such information may be available, subject to appropriate terms and conditions, including in some cases, payment or a fee.

The licensed program described in this document and all licensed material available for it are provided by IBM under terms of the IBM Customer Agreement, IBM International Program License Agreement or any equivalent agreement between us.

Any performance data contained herein was determined in a controlled environment. Therefore, the results obtained in other operating environments may vary significantly. Some measurements may have been made on development-level systems and there is no guarantee that these measurements will be the same on generally available systems. Furthermore, some measurements may have been estimated through extrapolation. Actual results may vary. Users of this document should verify the applicable data for their specific environment.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to the names and addresses used by an actual business enterprise is entirely coincidental.

COPYRIGHT LICENSE:

This information contains sample application programs in source language, which illustrate programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to IBM, for the purposes of developing, using, marketing or distributing application programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. IBM, therefore, cannot guarantee or imply reliability, serviceability, or function of these programs. The sample programs are provided "AS IS", without warranty of any kind. IBM shall not be liable for any damages arising out of your use of the sample programs.

If you are viewing this information softcopy, the photographs and color illustrations may not appear.

Trademarks

IBM, the IBM logo, and ibm.com are trademarks or registered trademarks of International Business Machines Corp., registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the Web at "Copyright and trademark information" at www.ibm.com/legal/copytrade.shtml.

Intel is a trademark of Intel Corporation or its subsidiaries in the United States and other countries.

Java™ and all Java-based trademarks and logos are trademarks or registered trademarks of Oracle and/or its affiliates.

The registered trademark Linux is used pursuant to a sublicense from the Linux Foundation, the exclusive licensee of Linus Torvalds, owner of the mark on a worldwide basis.

Microsoft, Windows, and Windows NT are trademarks of Microsoft Corporation in the United States, other countries, or both.

Red Hat and Ansible are trademarks or registered trademarks of Red Hat, Inc. or its subsidiaries in the United States and other countries.

UNIX is a registered trademark of The Open Group in the United States and other countries.

Terms and conditions for product documentation

Permissions for the use of these publications are granted subject to the following terms and conditions.

IBM Privacy Policy

At IBM we recognize the importance of protecting your personal information and are committed to processing it responsibly and in compliance with applicable data protection laws in all countries in which IBM operates.

Visit the IBM Privacy Policy for additional information on this topic at https://www.ibm.com/privacy/details/us/en/.

Applicability

These terms and conditions are in addition to any terms of use for the IBM website.

Personal use

You can reproduce these publications for your personal, noncommercial use provided that all proprietary notices are preserved. You cannot distribute, display, or make derivative work of these publications, or any portion thereof, without the express consent of IBM.

Commercial use

You can reproduce, distribute, and display these publications solely within your enterprise provided that all proprietary notices are preserved. You cannot make derivative works of these publications, or reproduce, distribute, or display these publications or any portion thereof outside your enterprise, without the express consent of IBM.

Rights

Except as expressly granted in this permission, no other permissions, licenses, or rights are granted, either express or implied, to the Publications or any information, data, software or other intellectual property contained therein.

IBM reserves the right to withdraw the permissions that are granted herein whenever, in its discretion, the use of the publications is detrimental to its interest or as determined by IBM, the above instructions are not being properly followed.

You cannot download, export, or reexport this information except in full compliance with all applicable laws and regulations, including all United States export laws and regulations.

IBM MAKES NO GUARANTEE ABOUT THE CONTENT OF THESE PUBLICATIONS. THE PUBLICATIONS ARE PROVIDED "AS-IS" AND WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESSED OR IMPLIED, INCLUDING BUT NOT LIMITED TO IMPLIED WARRANTIES OF MERCHANTABILITY, NON-INFRINGEMENT, AND FITNESS FOR A PARTICULAR PURPOSE.

Glossary

This glossary provides terms and definitions for the IBM Elastic Storage System solution.

The following cross-references are used in this glossary:

- *See* refers you from a non-preferred term to the preferred term or from an abbreviation to the spelled-out form.
- *See also* refers you to a related or contrasting term.

For other terms and definitions, see the IBM Terminology website (opens in new window):

http://www.ibm.com/software/globalization/terminology

B

building block
A pair of servers with shared disk enclosures attached.

BOOTP
See *Bootstrap Protocol (BOOTP)*.

Bootstrap Protocol (BOOTP)
A computer networking protocol that is used in IP networks to automatically assign an IP address to network devices from a configuration server.

C

CEC
See *central processor complex (CPC)*.

central electronic complex (CEC)
See *central processor complex (CPC)*.

central processor complex (CPC)
A physical collection of hardware that consists of channels, timers, main storage, and one or more central processors.

cluster
A loosely-coupled collection of independent systems, or *nodes*, organized into a network for the purpose of sharing resources and communicating with each other. See also *GPFS cluster*.

cluster manager
The node that monitors node status using disk leases, detects failures, drives recovery, and selects file system managers. The cluster manager is the node with the lowest node number among the quorum nodes that are operating at a particular time.

compute node
A node with a mounted GPFS file system that is used specifically to run a customer job. ESS disks are not directly visible from and are not managed by this type of node.

CPC
See *central processor complex (CPC)*.

D

DA
See *declustered array (DA)*.

datagram
A basic transfer unit associated with a packet-switched network.

DCM
See *drawer control module (DCM)*.

declustered array (DA)
A disjoint subset of the pdisks in a recovery group.

dependent fileset
A fileset that shares the inode space of an existing independent fileset.

DFM
See *direct FSP management (DFM)*.

DHCP
See *Dynamic Host Configuration Protocol (DHCP)*.

drawer control module (DCM)
Essentially, a SAS expander on a storage enclosure drawer.

Dynamic Host Configuration Protocol (DHCP)
A standardized network protocol that is used on IP networks to dynamically distribute such network configuration parameters as IP addresses for interfaces and services.

E

Elastic Storage System (ESS)
A high-performance, GPFS NSD solution made up of one or more building blocks. The ESS software runs on ESS nodes - management server nodes and I/O server nodes.

encryption key
A mathematical value that allows components to verify that they are in communication with the expected server. Encryption keys are based on a public or private key pair that is created during the installation process. See also *file encryption key (FEK)*, *master encryption key (MEK)*.

ESS
See *Elastic Storage System (ESS)*.

environmental service module (ESM)
Essentially, a SAS expander that attaches to the storage enclosure drives. In the case of multiple drawers in a storage enclosure, the ESM attaches to drawer control modules.

ESM
See *environmental service module (ESM)*.

F

failback
Cluster recovery from failover following repair. See also *failover*.

failover
(1) The assumption of file system duties by another node when a node fails. (2) The process of transferring all control of the ESS to a single cluster in the ESS when the other clusters in the ESS fails. See also *cluster*. (3) The routing of all transactions to a second controller when the first controller fails. See also *cluster*.

failure group
A collection of disks that share common access paths or adapter connection, and could all become unavailable through a single hardware failure.

FEK
See *file encryption key (FEK)*.

file encryption key (FEK)
A key used to encrypt sectors of an individual file. See also *encryption key*.

file system
The methods and data structures used to control how data is stored and retrieved.

file system descriptor
A data structure containing key information about a file system. This information includes the disks assigned to the file system (*stripe group*), the current state of the file system, and pointers to key files such as quota files and log files.

file system descriptor quorum
The number of disks needed in order to write the file system descriptor correctly.

file system manager
The provider of services for all the nodes using a single file system. A file system manager processes changes to the state or description of the file system, controls the regions of disks that are allocated to each node, and controls token management and quota management.

fileset
A hierarchical grouping of files managed as a unit for balancing workload across a cluster. See also *dependent fileset*, *independent fileset*.

fileset snapshot
A snapshot of an independent fileset plus all dependent filesets.

flexible service processor (FSP)
Firmware that provides diagnosis, initialization, configuration, runtime error detection, and correction. Connects to the HMC.

FQDN
See *fully-qualified domain name (FQDN)*.

FSP
See *flexible service processor (FSP)*.

fully-qualified domain name (FQDN)
The complete domain name for a specific computer, or host, on the Internet. The FQDN consists of two parts: the hostname and the domain name.

G

GPFS cluster
A cluster of nodes defined as being available for use by GPFS file systems.

GPFS portability layer
The interface module that each installation must build for its specific hardware platform and Linux distribution.

GPFS Storage Server (GSS)
A high-performance, GPFS NSD solution made up of one or more building blocks that runs on System x servers.

GSS
See *GPFS Storage Server (GSS)*.

H

Hardware Management Console (HMC)
Standard interface for configuring and operating partitioned (LPAR) and SMP systems.

HMC
See *Hardware Management Console (HMC)*.

I

IBM Security Key Lifecycle Manager (ISKLM)
For GPFS encryption, the ISKLM is used as an RKM server to store MEKs.

independent fileset
A fileset that has its own inode space.

indirect block
A block that contains pointers to other blocks.

inode
The internal structure that describes the individual files in the file system. There is one inode for each file.

inode space
A collection of inode number ranges reserved for an independent fileset, which enables more efficient per-fileset functions.

Internet Protocol (IP)
The primary communication protocol for relaying datagrams across network boundaries. Its routing function enables internetworking and essentially establishes the Internet.

I/O server node
An ESS node that is attached to the ESS storage enclosures. It is the NSD server for the GPFS cluster.

IP
See *Internet Protocol (IP)*.

IP over InfiniBand (IPoIB)
Provides an IP network emulation layer on top of InfiniBand RDMA networks, which allows existing applications to run over InfiniBand networks unmodified.

IPoIB
See *IP over InfiniBand (IPoIB)*.

ISKLM
See *IBM Security Key Lifecycle Manager (ISKLM)*.

J

JBOD array
The total collection of disks and enclosures over which a recovery group pair is defined.

K

kernel
The part of an operating system that contains programs for such tasks as input/output, management and control of hardware, and the scheduling of user tasks.

L

LACP
See *Link Aggregation Control Protocol (LACP)*.

Link Aggregation Control Protocol (LACP)
Provides a way to control the bundling of several physical ports together to form a single logical channel.

logical partition (LPAR)
A subset of a server's hardware resources virtualized as a separate computer, each with its own operating system. See also *node*.

LPAR
See *logical partition (LPAR)*.

M

management network
A network that is primarily responsible for booting and installing the designated server and compute nodes from the management server.

management server (MS)
An ESS node that hosts the ESS GUI and is not connected to storage. It must be part of a GPFS cluster. From a system management perspective, it is the central coordinator of the cluster. It also serves as a client node in an ESS building block.

master encryption key (MEK)
A key that is used to encrypt other keys. See also *encryption key*.

maximum transmission unit (MTU)
The largest packet or frame, specified in octets (eight-bit bytes), that can be sent in a packet- or frame-based network, such as the Internet. The TCP uses the MTU to determine the maximum size of each packet in any transmission.

MEK
See *master encryption key (MEK)*.

metadata
A data structure that contains access information about file data. Such structures include inodes, indirect blocks, and directories. These data structures are not accessible to user applications.

MS
See *management server (MS)*.

MTU
See *maximum transmission unit (MTU)*.

N

Network File System (NFS)
A protocol (developed by Sun Microsystems, Incorporated) that allows any host in a network to gain access to another host or netgroup and their file directories.

Network Shared Disk (NSD)
A component for cluster-wide disk naming and access.

NSD volume ID
A unique 16-digit hexadecimal number that is used to identify and access all NSDs.

node
An individual operating-system image within a cluster. Depending on the way in which the computer system is partitioned, it can contain one or more nodes. In a Power Systems environment, synonymous with *logical partition*.

node descriptor
A definition that indicates how ESS uses a node. Possible functions include: manager node, client node, quorum node, and non-quorum node.

node number
A number that is generated and maintained by ESS as the cluster is created, and as nodes are added to or deleted from the cluster.

node quorum
The minimum number of nodes that must be running in order for the daemon to start.

node quorum with tiebreaker disks
A form of quorum that allows ESS to run with as little as one quorum node available, as long as there is access to a majority of the quorum disks.

non-quorum node
A node in a cluster that is not counted for the purposes of quorum determination.

O

OFED
See *OpenFabrics Enterprise Distribution (OFED)*.

OpenFabrics Enterprise Distribution (OFED)
An open-source software stack includes software drivers, core kernel code, middleware, and user-level interfaces.

P

pdisk
A physical disk.

PortFast
A Cisco network function that can be configured to resolve any problems that could be caused by the amount of time STP takes to transition ports to the Forwarding state.

R

RAID
See *redundant array of independent disks (RAID)*.

RDMA
See *remote direct memory access (RDMA)*.

redundant array of independent disks (RAID)
A collection of two or more disk physical drives that present to the host an image of one or more logical disk drives. In the event of a single physical device failure, the data can be read or regenerated from the other disk drives in the array due to data redundancy.

recovery
The process of restoring access to file system data when a failure has occurred. Recovery can involve reconstructing data or providing alternative routing through a different server.

recovery group (RG)
A collection of disks that is set up by ESS, in which each disk is connected physically to two servers: a primary server and a backup server.

remote direct memory access (RDMA)
A direct memory access from the memory of one computer into that of another without involving either one's operating system. This permits high-throughput, low-latency networking, which is especially useful in massively-parallel computer clusters.

RGD
See *recovery group data (RGD)*.

remote key management server (RKM server)
A server that is used to store master encryption keys.

RG
See *recovery group (RG)*.

recovery group data (RGD)
Data that is associated with a recovery group.

RKM server
See *remote key management server (RKM server)*.

S

SAS
See *Serial Attached SCSI (SAS)*.

secure shell (SSH)
A cryptographic (encrypted) network protocol for initiating text-based shell sessions securely on remote computers.

Serial Attached SCSI (SAS)
A point-to-point serial protocol that moves data to and from such computer storage devices as hard drives and tape drives.

service network
A private network that is dedicated to managing POWER8 servers. Provides Ethernet-based connectivity among the FSP, CPC, HMC, and management server.

SMP
See *symmetric multiprocessing (SMP)*.

Spanning Tree Protocol (STP)
A network protocol that ensures a loop-free topology for any bridged Ethernet local-area network. The basic function of STP is to prevent bridge loops and the broadcast radiation that results from them.

SSH
See *secure shell (SSH)*.

STP
See *Spanning Tree Protocol (STP)*.

symmetric multiprocessing (SMP)
A computer architecture that provides fast performance by making multiple processors available to complete individual processes simultaneously.

T

TCP
See *Transmission Control Protocol (TCP)*.

Transmission Control Protocol (TCP)
A core protocol of the Internet Protocol Suite that provides reliable, ordered, and error-checked delivery of a stream of octets between applications running on hosts communicating over an IP network.

V

VCD
See *vdisk configuration data (VCD)*.

vdisk
A virtual disk.

vdisk configuration data (VCD)
Configuration data that is associated with a virtual disk.

Index

Special Characters

A

B

C

D

E

F

G

H

I

L

M

N

O

P

R

S

T

U

V

W

Product Number: 5765-DME
5765-DAE

SC27-9875-00