IBM Spectrum Scale
Version 5.0.4

Erasure Code Edition Guide

SC27-9578-06

Note

Before using this information and the product it supports, read the information in "Notices" on page 73.

This edition applies to version 5 release 0 modification 4 of the following products, and to all subsequent releases and modifications until otherwise indicated in new editions:

- IBM Spectrum Scale Erasure Code Edition ordered through Passport Advantage® (product number 5737-J34)

Significant changes or additions to the text and illustrations are indicated by a vertical line (|) to the left of the change.

IBM® welcomes your comments; see the topic "How to send your comments" on page xxiv. 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

This edition applies to IBM Spectrum Scale version 5.0.4 for AIX®, Linux®, and Windows.

IBM Spectrum Scale is a file management infrastructure, based on IBM General Parallel File System (GPFS) technology, which provides unmatched performance and reliability with scalable access to critical file data.

To find out which version of IBM Spectrum Scale is running on a particular AIX node, enter:

```
lslpp -l gpfs\*
```

To find out which version of IBM Spectrum Scale is running on a particular Linux node, enter:

```
rpm -qa | grep gpfs     (for SLES and Red Hat Enterprise Linux)
```

```
dpkg -l | grep gpfs     (for Ubuntu Linux)
```

To find out which version of IBM Spectrum Scale is running on a particular Windows node, open **Programs and Features** in the control panel. The IBM Spectrum Scale installed program name includes the version number.

Which IBM Spectrum Scale information unit provides the information you need?

The IBM Spectrum Scale library consists of the information units listed in Table 1 on page viii.

To use these information units effectively, you must be familiar with IBM Spectrum Scale and the AIX, Linux, or Windows operating system, or all of them, depending on which operating systems are in use at your installation. Where necessary, these information units provide some background information relating to AIX, Linux, or Windows. However, more commonly they refer to the appropriate operating system documentation.

Note: Throughout this documentation, the term "Linux" refers to all supported distributions of Linux, unless otherwise specified.

Table 1. IBM Spectrum Scale library information units		
Information unit	**Type of information**	**Intended users**
IBM Spectrum Scale: Concepts, Planning, and Installation Guide	This guide provides the following information: **Product overview** • Overview of IBM Spectrum Scale • GPFS architecture • Protocols support overview: Integration of protocol access methods with GPFS • Active File Management • AFM-based Asynchronous Disaster Recovery (AFM DR) • Data protection and disaster recovery in IBM Spectrum Scale • Introduction to IBM Spectrum Scale GUI • IBM Spectrum Scale management API • Introduction to Cloud services • Introduction to file audit logging • Introduction to watch folder • Introduction to clustered watch • IBM Spectrum Scale in an OpenStack cloud deployment • IBM Spectrum Scale product editions • IBM Spectrum Scale license designation • Capacity based licensing • IBM Spectrum Storage™ Suite **Planning** • Planning for GPFS • Planning for protocols • Planning for Cloud services • Planning for AFM • Planning for AFM DR • Firewall recommendations • Considerations for GPFS applications	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

Table 1. IBM Spectrum Scale library information units (continued)		
Information unit	**Type of information**	**Intended users**
IBM Spectrum Scale: Concepts, Planning, and Installation Guide	**Installing** • Steps for establishing and starting your IBM Spectrum Scale cluster • Installing IBM Spectrum Scale on Linux nodes and deploying protocols • Installing IBM Spectrum Scale on AIX nodes • Installing IBM Spectrum Scale on Windows nodes • Installing Cloud services on IBM Spectrum Scale nodes • Installing and configuring IBM Spectrum Scale management API • Installation of Active File Management (AFM) • Installing and upgrading AFM-based Disaster Recovery • Installing call home • Installing file audit logging • Installing watch folder • Installing clustered watch • Steps to permanently uninstall GPFS **Upgrading** • IBM Spectrum Scale supported upgrade paths • Upgrading to IBM Spectrum Scale 5.0.x from IBM Spectrum Scale 4.2.y • Upgrading to IBM Spectrum Scale 4.2.y from IBM Spectrum Scale 4.1.x • Online upgrade support for protocols and performance monitoring	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

Table 1. IBM Spectrum Scale library information units (continued)		
Information unit	**Type of information**	**Intended users**
IBM Spectrum Scale: Concepts, Planning, and Installation Guide	• Upgrading IBM Spectrum® Scale non-protocol Linux nodes • Upgrading IBM Spectrum Scale protocol nodes • Upgrading AFM and AFM DR • Upgrading object packages • Upgrading SMB packages • Upgrading NFS packages • Upgrading call home • Manually upgrading the performance monitoring tool • Manually upgrading pmswift • Manually upgrading the IBM Spectrum Scale management GUI • Upgrading Cloud services • Upgrading to IBM Cloud Object Storage software level 3.7.2 and above • Upgrade paths and commands for file audit logging, watch folder API, and clustered watch folder • Upgrading with clustered watch folder enabled • Upgrading IBM Spectrum Scale components with the installation toolkit • Changing IBM Spectrum Scale product edition • Completing the upgrade to a new level of IBM Spectrum Scale • Reverting to the previous level of IBM Spectrum Scale	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

Table 1. IBM Spectrum Scale library information units (continued)		
Information unit	**Type of information**	**Intended users**
IBM Spectrum Scale: Concepts, Planning, and Installation Guide	• Coexistence considerations • Compatibility considerations • Considerations for IBM Spectrum Protect for Space Management • GUI user role considerations • Applying maintenance to your GPFS system • Guidance for upgrading the operating system on IBM Spectrum Scale nodes • Servicing IBM Spectrum Scale protocol nodes • Offline upgrade with complete cluster shutdown	

Table 1. IBM Spectrum Scale library information units (continued)		
Information unit	**Type of information**	**Intended users**
IBM Spectrum Scale: Administration Guide	This guide provides the following information: **Configuring** • Configuring the GPFS cluster • Configuring the CES and protocol configuration • Configuring and tuning your system for GPFS • Parameters for performance tuning and optimization • Ensuring high availability of the GUI service • Configuring and tuning your system for Cloud services • Configuring the message queue • Configuring file audit logging • Configuring clustered watch folder • Configuring Active File Management • Configuring AFM-based DR • Tuning for Kernel NFS backend on AFM and AFM DR **Administering** • Performing GPFS administration tasks • Verifying network operation with the mmnetverify command • Managing file systems • File system format changes between versions of IBM Spectrum Scale • Managing disks • Managing protocol services • Managing protocol user authentication • Managing protocol data exports • Managing object storage • Managing GPFS quotas • Managing GUI users • Managing GPFS access control lists	System administrators or programmers of IBM Spectrum Scale systems

Table 1. IBM Spectrum Scale library information units (continued)		
Information unit	**Type of information**	**Intended users**
IBM Spectrum Scale: Administration Guide	• Native NFS and GPFS • Considerations for GPFS applications • Accessing a remote GPFS file system • Information lifecycle management for IBM Spectrum Scale • Creating and maintaining snapshots of file systems • Creating and managing file clones • Scale Out Backup and Restore (SOBAR) • Data Mirroring and Replication • Implementing a clustered NFS environment on Linux • Implementing Cluster Export Services • Identity management on Windows / RFC 2307 Attributes • Protocols cluster disaster recovery • File Placement Optimizer • Encryption • Managing certificates to secure communications between GUI web server and web browsers • Securing protocol data • Cloud services: Transparent cloud tiering and Cloud data sharing • Managing file audit logging • Performing a watch with watch folder API • RDMA tuning • Administering AFM • Administering AFM DR • Highly-available write cache (HAWC) • Local read-only cache • Miscellaneous advanced administration • GUI limitations	System administrators or programmers of IBM Spectrum Scale systems

Table 1. IBM Spectrum Scale library information units (continued)		
Information unit	**Type of information**	**Intended users**
IBM Spectrum Scale: Problem Determination Guide	This guide provides the following information: **Monitoring** • Performance monitoring • Monitoring system health through the IBM Spectrum Scale GUI • Monitoring system health by using the mmhealth command • Monitoring events through callbacks • Monitoring capacity through GUI • Monitoring AFM and AFM DR • GPFS SNMP support • Monitoring the IBM Spectrum Scale system by using call home • Monitoring remote cluster through GUI • Monitoring the message queue • Monitoring file audit logging • Monitoring clustered watch **Troubleshooting** • Best practices for troubleshooting • Understanding the system limitations • Collecting details of the issues • Managing deadlocks • Installation and configuration issues • Upgrade issues • Network issues • File system issues • Disk issues • Security issues • Protocol issues • Disaster recovery issues • Performance issues	System administrators of GPFS systems who are experienced with the subsystems used to manage disks and who are familiar with the concepts presented in the *IBM Spectrum Scale: Concepts, Planning, and Installation Guide*

Table 1. IBM Spectrum Scale library information units (continued)		
Information unit	**Type of information**	**Intended users**
IBM Spectrum Scale: Problem Determination Guide	• GUI issues • AFM issues • AFM DR issues • Transparent cloud tiering issues • File audit logging issues • Troubleshooting watch folder API • Troubleshooting mmwatch • Message queue issues • Maintenance procedures • Recovery procedures • Support for troubleshooting • References	

Table 1. IBM Spectrum Scale library information units (continued)		
Information unit	**Type of information**	**Intended users**
IBM Spectrum Scale: Command and Programming Reference	This guide provides the following information: **Command reference** • gpfs.snap command • mmaddcallback command • mmadddisk command • mmaddnode command • mmadquery command • mmafmconfig command • mmafmctl command • mmafmlocal command • mmapplypolicy command • mmaudit command • mmauth command • mmbackup command • mmbackupconfig command • mmblock command • mmbuildgpl command • mmcachectl command • mmcallhome command • mmces command • mmcesdr command • mmchattr command • mmchcluster command • mmchconfig command • mmchdisk command • mmcheckquota command • mmchfileset command • mmchfs command • mmchlicense command • mmchmgr command • mmchnode command • mmchnodeclass command • mmchnsd command • mmchpolicy command • mmchpool command • mmchqos command • mmclidecode command	• 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

Table 1. IBM Spectrum Scale library information units (continued)		
Information unit	**Type of information**	**Intended users**
IBM Spectrum Scale: Command and Programming Reference	• mmclone command • mmcloudgateway command • mmcrcluster command • mmcrfileset command • mmcrfs command • mmcrnodeclass command • mmcrnsd command • mmcrsnapshot command • mmdefedquota command • mmdefquotaoff command • mmdefquotaon command • mmdefragfs command • mmdelacl command • mmdelcallback command • mmdeldisk command • mmdelfileset command • mmdelfs command • mmdelnode command • mmdelnodeclass command • mmdelnsd command • mmdelsnapshot command • mmdf command • mmdiag command • mmdsh command • mmeditacl command • mmedquota command • mmexportfs command • mmfsck command • mmfsctl command • mmgetacl command • mmgetstate command • mmhadoopctl command • mmhdfs command • mmhealth command • mmimgbackup command • mmimgrestore command • mmimportfs command • mmkeyserv command	• 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

Table 1. IBM Spectrum Scale library information units (continued)		
Information unit	**Type of information**	**Intended users**
IBM Spectrum Scale: Command and Programming Reference	• mmlinkfileset command • mmlsattr command • mmlscallback command • mmlscluster command • mmlsconfig command • mmlsdisk command • mmlsfileset command • mmlsfs command • mmlslicense command • mmlsmgr command • mmlsmount command • mmlsnodeclass command • mmlsnsd command • mmlspolicy command • mmlspool command • mmlsqos command • mmlsquota command • mmlssnapshot command • mmmigratefs command • mmmount command • mmmsgqueue command • mmnetverify command • mmnfs command • mmnsddiscover command • mmobj command • mmperfmon command • mmpmon command • mmprotocoltrace command • mmpsnap command • mmputacl command • mmquotaoff command • mmquotaon command • mmreclaimspace command • mmremotecluster command • mmremotefs command • mmrepquota command • mmrestoreconfig command • mmrestorefs command • mmrestripefile command	• 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

Table 1. IBM Spectrum Scale library information units (continued)		
Information unit	**Type of information**	**Intended users**
IBM Spectrum Scale: Command and Programming Reference	• mmrestripefs command • mmrpldisk command • mmsdrrestore command • mmsetquota command • mmshutdown command • mmsmb command • mmsnapdir command • mmstartup command • mmtracectl command • mmumount command • mmunlinkfileset command • mmuserauth command • mmwatch command • mmwinservctl command • spectrumscale command **Programming reference** • IBM Spectrum Scale Data Management API for GPFS information • GPFS programming interfaces • GPFS user exits • IBM Spectrum Scale management API commands • Watch folder API	• 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

Table 1. IBM Spectrum Scale library information units (continued)		
Information unit	**Type of information**	**Intended users**
IBM Spectrum Scale: Big Data and Analytics Guide	This guide provides the following information: Hadoop Scale Storage Architecture • Elastic Storage Server (ESS) • Erasure Code Edition • Share Storage (SAN-based storage) • File Placement Optimizer (FPO) • Deployment model • Additional supported features about storage IBM Spectrum Scale support for Hadoop • HDFS transparency • Supported IBM Spectrum Scale storage modes • Hadoop cluster planning • CES HDFS • Installation and configuration of HDFS transparency • Application interaction with HDFS transparency • Upgrading the HDFS Transparency cluster • Rolling upgrade for HDFS Transparency • Security • Advanced features • Hadoop distribution support • Limitations and differences from native HDFS • Problem determination IBM Spectrum Scale Hadoop performance tuning guide • Overview • Performance overview • Hadoop Performance Planning over IBM Spectrum Scale • Performance guide	• 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

Table 1. IBM Spectrum Scale library information units (continued)		
Information unit	**Type of information**	**Intended users**
IBM Spectrum Scale: Big Data and Analytics Guide	Hortonworks Data Platform 3.X • Planning • Installation • Upgrading and uninstallation • Configuration • Administration • Limitations • Problem determination Open Source Apache Hadoop • Open Source Apache Hadoop without CES HDFS • Open Source Apache Hadoop with CES HDFS BigInsights® 4.2.5 and Hortonworks Data Platform 2.6 • Installation • Upgrading software stack • Configuration • Administration • Troubleshooting • Limitations • FAQ	• 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

Table 1. IBM Spectrum Scale library information units (continued)		
Information unit	**Type of information**	**Intended users**
IBM Spectrum Scale Erasure Code Edition Guide	IBM Spectrum Scale Erasure Code Edition • Introduction to IBM Spectrum Scale Erasure Code Edition • Planning for IBM Spectrum Scale Erasure Code Edition • Installing IBM Spectrum Scale Erasure Code Edition • Incorporating IBM Spectrum Scale Erasure Code Edition in an Elastic Storage Server (ESS) cluster • Creating an IBM Spectrum Scale Erasure Code Edition storage environment • Upgrading IBM Spectrum Scale Erasure Code Edition • Administering IBM Spectrum Scale Erasure Code Edition • Troubleshooting • IBM Spectrum Scale RAID Administration [1] **Note:** [1] For PDF or EPUB format of IBM Spectrum Scale RAID Administration documentation, see Elastic Storage Server for Power® documentation on IBM Knowledge Center.	• 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

Table 1. IBM Spectrum Scale library information units (continued)		
Information unit	**Type of information**	**Intended users**
IBM Spectrum Scale Container Storage Interface Driver	This guide provides the following information: • Introduction to IBM Spectrum Scale Container Storage Interface Driver • Planning for IBM Spectrum Scale Container Storage Interface Driver • Installation of IBM Spectrum Scale Container Storage Interface Driver • Migrating from IBM Storage Enabler for Containers to IBM Spectrum Scale Container Storage Interface Driver • Configuring IBM Spectrum Scale Container Storage Interface Driver • Using IBM Spectrum Scale Container Storage Interface Driver • Managing IBM Spectrum Scale Container Storage Interface Driver • Limitations • Troubleshooting	• 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

Prerequisite and related information

For updates to this information, see IBM Spectrum Scale in IBM Knowledge Center (www.ibm.com/support/knowledgecenter/STXKQY/ibmspectrumscale_welcome.html).

For the latest support information, see the IBM Spectrum Scale FAQ in IBM Knowledge Center (www.ibm.com/support/knowledgecenter/STXKQY/gpfsclustersfaq.html).

Conventions used in this information

Table 2 on page xxiv describes the typographic conventions used in this information. UNIX file name conventions are used throughout this information.

Note: Users of IBM Spectrum Scale for Windows must be aware that on Windows, UNIX-style file names need to be converted appropriately. For example, the GPFS cluster configuration data is stored in the `/var/mmfs/gen/mmsdrfs` file. On Windows, the UNIX namespace starts under the `%SystemDrive%\cygwin64` directory, so the GPFS cluster configuration data is stored in the `C:\cygwin64\var\mmfs\gen\mmsdrfs` file.

Table 2. 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.
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.

Note: CLI options that accept a list of option values delimit with a comma and no space between values. As an example, to display the state on three nodes use `mmgetstate -N` *NodeA*,*NodeB*,*NodeC*. Exceptions to this syntax are listed specifically within the command.

How to send your comments

Your feedback is important in helping us to produce accurate, high-quality information. If you have any comments about this information or any other IBM Spectrum Scale documentation, send your comments to the following e-mail address:

`mhvrcfs@us.ibm.com`

Include the publication title and order number, and, if applicable, the specific location of the information about which you have comments (for example, a page number or a table number).

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

scale@us.ibm.com

Chapter 1. Summary of changes

This topic summarizes changes to the current version of the IBM Spectrum Scale Erasure Code Edition.

The following changes are made in the current release:

- Added hardware check list with respect to using KVM and VMware virtual machine as the storage node
- Support for manual online upgrade

Chapter 2. Introduction to IBM Spectrum Scale Erasure Code Edition

IBM Spectrum Scale Erasure Code Edition provides IBM Spectrum Scale RAID as software, allowing customers to create IBM Spectrum Scale clusters that use scale-out storage on any hardware that meets the minimum hardware requirements.

All of the benefits of IBM Spectrum Scale and IBM Spectrum Scale RAID can be realized using your own commodity hardware.

For example, IBM Spectrum Scale Erasure Code Edition provides:

- Reed-Solomon highly fault tolerant declustered Erasure Coding, protecting against individual drive failures as well as node failures.
- Disk Hospital to identify issues before they become disasters.
- End-to-end checksum to identify and correct errors introduced by network and/or media

IBM Spectrum Scale Erasure Code Edition uses the same software and most of the same concepts that are used in the Elastic Storage Server (ESS). Elastic Storage Server (ESS) is a solution consisting of two I/O (storage) servers and between one and several JBOD disk enclosures, with each storage device (pdisk) attached to both servers. In Elastic Storage Server (ESS), there are two recovery groups (RGs). Each RG takes half of each enclosure among all enclosures. Under normal conditions, each I/O server supports one of the two RGs. If either I/O server fails, the remaining I/O server takes over and supports both RGs.

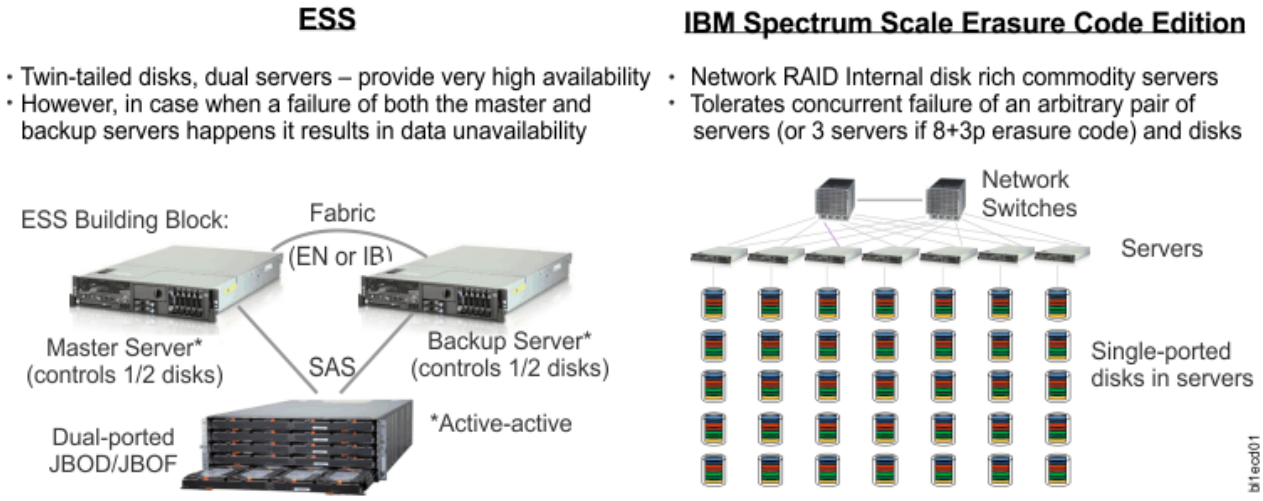

Figure 1. IBM Spectrum Scale Erasure Code Edition architecture

IBM Spectrum Scale Erasure Code Edition, in contrast, can have one or more recovery groups, but each RG is associated with between 4 and 32 storage servers, and each storage server belongs to only one RG. All of the storage servers in a recovery group must have a matching configuration, including identical CPU, memory, network, and storage device configurations. The storage devices (pdisks) are directly attached to only one storage server. Each storage server typically serves two log groups, each log group managing one half of the virtual disks (vdisk NSDs) assigned to a server. If a storage server fails, the log groups (and vdisk NSDs) it was serving are distributed to the remaining storage servers; any storage server failure will cause the remaining storage servers to serve at most one additional log group.

In both Elastic Storage Server (ESS) and IBM Spectrum Scale Erasure Code Edition, the placement of data is topology aware using a failure domain hierarchy of rack, node, enclosure, and storage device (pdisk). The RAID code makes placement decisions to maximize fault tolerance, depending on the RAID level you

choose. IBM Spectrum Scale Erasure Code Edition supports the following erasure codes and replication levels: 8+2p, 8+3p, 4+2p, 4+3p, 3WayReplication, and 4WayReplication.

With IBM Spectrum Scale Erasure Code Edition it is possible for either IBM Spectrum Scale Cluster Export Services with protocol software or customer applications to run directly on the storage servers if sufficient hardware resources are available. Customer applications must run in a constrained environment using Linux cgroups or Docker containers. For protocol workloads with high performance requirements, the Cluster Export Services should run on separate nodes.

In both Elastic Storage Server (ESS) and IBM Spectrum Scale Erasure Code Edition, the IBM Spectrum Scale file system, and file system features are independent of the storage configuration. A file system can be composed of NSDs provided by more than one recovery group, and the recovery groups can be from Elastic Storage Server (ESS) or IBM Spectrum Scale Erasure Code Edition or a combination of both. All of the IBM Spectrum Scale file system features can be used in a cluster with IBM Spectrum Scale Erasure Code Edition storage servers, but there are strict guidelines as to where the various components might run.

For an overview of IBM Spectrum Scale RAID, see the *Introducing IBM Spectrum Scale RAID* topic in the *IBM Spectrum Scale RAID: Administration*.

Minimum hardware requirements

At a high level, you must have between 4 and 32 storage servers per Recovery Group (RG), and each server must be an x86 server running Red Hat® Enterprise Linux version 7.5 or 7.6. The storage configuration must be identical for all storage servers. The supported storage types are SAS-attached HDD or SSD drives, using specified LSI adapters, or enterprise-class NVMe drives. Each storage server must have at least one SSD or NVMe drive, this is used for a fast write cache as well as user data storage. For more information on hardware requirement, see “Minimum hardware requirements and precheck” on page 9.

Maximum storage nodes in a cluster

There can be up to 128 IBM Spectrum Scale Erasure Code Edition storage nodes in a IBM Spectrum Scale cluster, for example 4 RGs with 32 nodes each or 8 RGs with 16 nodes each, or some other combination that results in no more than 128 total storage nodes.

Network configurations

The network can be either Ethernet or InfiniBand, and must be at least 25 Gbps bandwidth, with an average latency of 1.0 msec or less between any two storage nodes. It is recommended to have a dedicated network for storage server traffic. In most cases, the overall storage performance is dictated by network bandwidth and latency. Your performance requirements must be carefully considered when selecting the network hardware and the network architecture for your IBM Spectrum Scale Erasure Code Edition cluster. For more information on networking requirements, see “Network requirements and precheck” on page 14.

Administration and maintenance procedures

IBM Spectrum Scale Erasure Code Edition administration and maintenance procedures are similar to Elastic Storage Server (ESS), but not identical. With IBM Spectrum Scale Erasure Code Edition, the customer is responsible for managing the storage server hardware and software. For example, the customer is responsible for updating any firmware as well as the operating system, including security updates, when needed. The majority of the IBM Spectrum Scale RAID maintenance commands are accomplished using the **mmvdisk** command. For details of IBM Spectrum Scale Erasure Code Edition admin and maintenance procedures, see Chapter 8, “Administering IBM Spectrum Scale Erasure Code Edition,” on page 55.

Health monitoring and problem determination

IBM Spectrum Scale Erasure Code Edition health monitoring and problem determination procedures rely on IBM Spectrum Scale **mmhealth** capabilities, as well as IBM Spectrum Scale RAID troubleshooting guidelines. For more details, see Chapter 9, “Troubleshooting,” on page 65.

Understanding IBM Spectrum Scale Erasure Code Edition fault tolerance

When talking about fault tolerance, it is important to define what the potential faults could be. Some are obvious such as node failure and hard `pdisk` failure. But there can be other conditions as well. This topic explains those considerations.

The other conditions that affect the fault tolerance level are the following:

- Silent data corruption that can be detected in GNR scrub, but this may take up to 15 days, or whatever is defined to be the scrub duration.
- Transient `pdisk` problem, leading to stale strips.

When IBM Spectrum Scale RAID shows 2-node fault tolerance, it does mean you can tolerate two-node equivalent failures, but it does not mean you can always take down two nodes safely and unconditionally, given you may hit silent data corruption. Also, after you see a 2-node fault tolerance but before you take down the nodes, there may be stale strips generated. So, when considering taking down two nodes with 2-node fault tolerance, it is with the assumption that there is no silent data corruption and no stale strips.

The approach taken by IBM Spectrum Scale RAID is to display fault tolerance conservatively, so that you can expect no worse fault tolerance until there are new node or disk failures, and at that point fault tolerance is recalculated. But if the failure does not happen quickly, if you have the fault tolerance in hand, for example, 1 node fault tolerance, you will get better and better fault tolerance and no downgrade, so you can plan your maintenance operations, for example, take down one node safely and not have to worry about the downgrade. While this would seem too strict, there is a constraint. In theory, we can never get the exact fault tolerance. A fault-tolerance state that we obtain in time A could be changed very soon in time A+1.

The current behavior is due to some technical complexities:

- When we talk about fault, it is not only that the nodes are down or there is a failure in the disk hard drive, but also stale strips IBM Spectrum Scale RAID mark when a `pdisk` has transient problem. Given we still have enough fault tolerance, we do not have to wait for the disk hospital to complete its pdisk diagnosis and let I/O pause. We can mark the strip stale and let the I/O complete and respond back to the client. But with stale strips, the fault tolerance becomes more complex. If it is node down or pdisk hard failure, we can always look at the pdisk state and the partition group map to decide the fault tolerance. However, to consider the stale strips, every block/vtrack must be evaluated. Each partition group is divided into many vtracks, so a full system metadata scan is needed to calculate the fault tolerance.
- Close to the end of the rebuild, there is a swap procedure to get the best fault tolerance we can. But with very tight space at the end, it has some possibility that it cannot always succeed, and during the swap, there are multiple vtracks involved to move data around, and if the swap fails we need to move back to the original state. This means for some vtracks after moving with higher fault tolerance, we might need to revert the procedure and lower the fault tolerance to its original state. So, if we report fault tolerance before that and then have the backward downgrade, it might be misleading. For example, in this scenario the system could have displayed a 1-node fault tolerance, but due to the downgrade, it changes to 1 pdisk. To avoid confusion and to prevent maintenance operations, like a node down event from being planned, we show the more conservative fault tolerance value until the rebuild if completed.

The key points for dealing with fault tolerance are:

1. The possibility of the factors that can lower the fault tolerance soon after it is calculated.
2. The possible impact of a fault tolerance downgrade.
3. How to mitigate the risk of fault tolerance downgrade.

For 1, the possibility of multiple faults happening on the same block is usually low in an enterprise system.

For 2, we may look at the fault tolerance and see there is a 1-node fault tolerance and decide to do maintenance by taking down a node, but given the potential downgrade of fault tolerance, maybe there

are more faults than expected. Taking down a node for maintenance means missing some `pdisks`, the worst case here is to reassign the LG/RG without enough fault tolerance, which translates to out-of-service condition, rather than data loss. Maintenance is usually performed in a special time window without heavy workload or critical service time, so the impact of an unexpected short period of out-of-service can be minimized.

For 3, during the planning phase, we recommend the fault tolerance 1 node + 1 `pdisk` failure as the minimal setting. So regardless of a node failure or maintenance, we do not always drop into critical rebuild and can tolerate an additional unexpected fault. You could still use a 1-node fault tolerance or perform maintenance with only 1-node fault tolerance, but you should be aware of the risk and can accept it, especially for non-critical workloads.

Proper planning and better understanding of the fault tolerance is a good way to prepare. IBM Spectrum Scale Erasure Code Edition aims to protect the system from silent data corruption, and also to keep higher system performance even if when there are some transient disk errors.

The table below shows for various number of storage nodes and erasure codes, what are the number of strips per node and what is the fault tolerance level for that combination of nodes and erasure code. For example, with a 4+2P erasure code and 6 nodes, there are 6 strips (4 data and 2 parity) for each block, and they are distributed one on each node. This gives a fault tolerance of 2 nodes, one node and one disk or 2 disks. On the other hand with 8+2P erasure code on 6 nodes , there are 10 strips (8 data and 2 parity). There are 4 nodes with 2 strips each, and 2 nodes with one strip. This gives a fault tolerance of one node or 2 disks.

Table 3. Fault tolerances of nodes and disks for various RAID codes on different numbers of nodes

Nodes	**Code**	**Layout (strips per node)**	**Fault Tolerance (N Nodes, D Disks)**
4	4+2p	2,2,1,1	N, 2D
4	4+3p	2,2,2,1	N+D, 3D
4	8+2p	3,3,2,2	2D
4	8+3p	3,3,3,2	N, 3D
5	4+2p	2,1,1,1,1	N, 2D
5	4+3p	2,2,1,1,1	N+D, 3D
5	8+2p	2,2,2,2,2	N, 2D
5	8+3p	3,2,2,2,2	N, 3D
6	4+2p	1,1,1,1,1,1	2N, N+D, 2D
6	4+3p	2,1,1,1,1,1	2N, N+D, 3D
6	8+2p	2,2,2,2,1,1	N, 2D
6	8+3p	2,2,2,2,2,1	N+D, 3D
7	4+2p	1,1,1,1,1,1,0	2N, N+D, 2D
7	4+3p	1,1,1,1,1,1,1	2N+D, N+2D, 3D
7	8+2p	2,2,2,1,1,1,1	N, 2D
7	8+3p	2,2,2,2,1,1,1	N+D, 3D
8	4+2p	1,1,1,1,1,1,0,0	2N, N+D, 2D
8	4+3p	1,1,1,1,1,1,1,0	2N+D, N+2D, 3D
8	8+2p	2,2,1,1,1,1,1,1	N, 2D
8	8+3p	2,2,2,1,1,1,1,1	N+D, 3D

Table 3. Fault tolerances of nodes and disks for various RAID codes on different numbers of nodes (continued)

Nodes	Code	Layout (strips per node)	Fault Tolerance (N Nodes, D Disks)
9	4+3p	1,1,1,1,1,1,1,0,0	3N, 2N+D, N+2D, 3D
9	8+2p	2,1,1,1,1,1,1,1,1	N, 2D
9	8+3p	2,2,1,1,1,1,1,1,1	N+D, 3D
10	4+2p	1,1,1,1,1,1,0,0,0,0	2N, N+D, 2D
10	4+3p	1,1,1,1,1,1,1,0,0,0	3N, 2N+D, N+2D, 3D
10	8+2p	1,1,1,1,1,1,1,1,1,1	2N, N+D, 2D
10	8+3p	2,1,1,1,1,1,1,1,1,1	2N, N+D, 3D
11	4+2p	1,1,1,1,1,1,0,0,0,0,0	2N, N+D, 2D
11	4+3p	1,1,1,1,1,1,1,0,0,0,0	3N, 2N+D, N+2D, 3D
11	8+2p	1,1,1,1,1,1,1,1,1,1,0	2N, N+D, 2D
11	8+3p	1,1,1,1,1,1,1,1,1,1,1	3N, 2N+D, N+2D, 3D

IBM Spectrum Scale Erasure Code Edition limitations

This topic describes the known limitations of IBM Spectrum Scale Erasure Code Edition.

General limitations

- The installation toolkit does not support installing mixed Elastic Storage Server (ESS) and IBM Spectrum Scale Erasure Code Edition in the same cluster. If you need this configuration, see Chapter 5, "Incorporating IBM Spectrum Scale Erasure Code Edition in an Elastic Storage Server (ESS) cluster," on page 33. For information about installation toolkit limitations in an IBM Spectrum Scale Erasure Code Edition environment, see "Installation toolkit-related limitations" on page 24.
- Rack-level fault tolerance is not fully supported. This can be achieved by spreading servers evenly between racks, but it is recommended that there should be no more than 2 storage servers per rack for an N+3P erasure code, and 1 storage server per rack for N+2P Erasure Code.
- When using NVMe drives that are hot swappable from front panel, the customer must create an EDF file to specify what drives are in what slot. A tool is provided to assist with this mapping, but the user is responsible for defining the correct mapping.
- Configuration with single server per compute chassis is supported. Configurations with 2 or more servers packaged together in the same physical unit are not supported.
- Only disk drives that are attached to one server, and by a single path are supported.

Configuration limitations

- Supported range of nodes in a recovery group (RG) is 4 to 32 nodes
- All nodes in the RG must be configured the same (memory, drives, CPU, and network)
- Supported erasure codes are 4+2P, 4+3P, 8+2P, 8+3P, 3WayReplication, and 4WayReplication
- Minimum declustered array (DA) size: At least one DA must contain 12 or more drives and every DA must have 6 or more drives.

  Note: DA is a subset of the physical disks within a recovery group that all match in size and speed. A recovery group may contain multiple declustered arrays which are unique (that is, a pdisk must belong to exactly one declustered array). The minimum DA size is met by each node contributing a uniform number of disks.

- Each node must have at least one fast device (NVMe or SAS SSD)
- The maximum supported number of drives in an RG is 512
- All nodes/HBAs/drives in an RG must have consistent firmware levels, and be at a level that is supported by the hardware provided. For more information, see “Hardware checklist” on page 11.
- All limitations of IBM Spectrum Scale apply, notably:
  - There can be a maximum of 128 IBM Spectrum Scale Erasure Code Edition storage nodes in an IBM Spectrum Scale cluster.
- In this release, synchronous mirroring using GPFS replication is not supported for NSDs (vdisks) served by IBM Spectrum Scale Erasure Code Edition building blocks.

Chapter 3. Planning for IBM Spectrum Scale Erasure Code Edition

This topic describes information on various activities that must be planned for effective usage of IBM Spectrum Scale Erasure Code Edition in an enterprise.

IBM Spectrum Scale Erasure Code Edition Hardware requirements

This document describes the requirements for storage hardware, including network requirements that can be used with IBM Spectrum Scale Erasure Code Edition (ECE).

IBM Spectrum Scale Erasure Code Edition strives to provide the very best performance that a given hardware platform can provide. From this perspective, hardware requirements are also dictated by the performance requirements of the customer's use case. For example, the minimum network requirement of 25 Gbps may work for some use cases, but for high performance workloads, 100 Gbps Ethernet or InfiniBand may be required to achieve performance goals.

In the IBM Spectrum Scale Erasure Code Edition, it is the customer's responsibility to manage the operating system, firmware, and device driver software on each server. This guide is meant to be a starting point in system sizing, and not a substitute for performance engineering and tuning for each customer environment and use case.

Minimum hardware requirements and precheck

This topic describes the minimum requirements for IBM Spectrum Scale Erasure Code Edition.

These hardware requirements are for the base operating system and the IBM Spectrum Scale Erasure Code Edition storage functions. Additional resources are required when running IBM Spectrum Scale protocol software or other workloads on the IBM Spectrum Scale Erasure Code Edition storage servers, or to achieve specific performance goals.

Each IBM Spectrum Scale Erasure Code Edition recovery group must have at least 4 servers, but there is a limit on the number of IBM Spectrum Scale Erasure Code Edition storage nodes in a IBM Spectrum Scale cluster. In this release, there can be up to 128 storage nodes in the cluster. These nodes can be configured as 4 recovery groups of 32 nodes each, or 8 recovery groups of 16 nodes, or some other combination with 128 or fewer total storage nodes. Every server in a recovery group must have the same configuration in terms of CPU, memory, and storage.

Note:

- Drives with hardware compression enabled are not supported
- Drives with volatile cache enabled are not supported. For more information, see "Volatile write cache detection" on page 62.
- SED capable drives are not allowed if they have been enrolled, or if they require a key after power on to use.
- Disk drives in expansion enclosures are not allowed.
- For SSD and NVMe drives, it is recommended to use a file system block size of 4 M or less with 8+2P or 8+3P erasure codes, and 2M or less for 4+2P OR 4+3P erasure codes.

Table 4. IBM Spectrum Scale ECE hardware requirements for each storage server

CPU architecture	x86 64 bit processor with 8 or more processor cores per socket. Server should be dual socket with both sockets populated

Table 4. IBM Spectrum Scale ECE hardware requirements for each storage server (continued)	
Memory	64 GB or more for configurations with up to 24 drives per node: • For NVMe configurations, it is recommended to utilize all available memory DIMM sockets to get optimal performance. • For server configurations with more than 24 drives per node, contact IBM for memory requirements.
Server packaging	Single server per enclosure. Multi-node server packaging with common hardware components that provide a single point of failure across servers is not supported at this time.
Operating system	RHEL 7.5 or later for production deployments. See IBM Spectrum Scale FAQ for details of supported versions.
Drives per storage node	A maximum of 24 drives per storage node is supported.
Drives per Recovery Group	A maximum of 512 drives per recovery group is supported.
Nodes per Recovery Group	A maximum of 32 nodes per recovery group is supported.
Storage nodes per cluster	A maximum of 128 ECE storage nodes per cluster is supported.
System drive	A physical drive is required for each server's system disk. It is recommended to have this RAID1 protected and have a capacity of 100 GB or more.
SAS Data Drives	SAS or NL-SAS HDD or SSDs in JBOD mode and connected to the supported SAS host bus adapters. SATA drives and Shingled Magnetic Recording drives are not supported as data drives at this time.
NVMe Data Drives	Enterprise class NVMe drives with U.2 form factor and connected to PCIe buses directly or by PCIe switch. NVMe drives connected to SAS host bus adapters are not supported as data drives at this time.
Fast Drive Requirement	At least one SSD or NVMe drive is required in each server for IBM Spectrum Scale Erasure Code Edition logging.
Network Adapter	Mellanox ConnectX-4 or ConnectX-5, (Ethernet or InfiniBand)
Network Bandwidth	25 Gbps or more between storage nodes. Higher bandwidth may be required depending on your workload requirements.
Network Latency	Average latency must be less than 1 msec between any storage nodes.
Network Topology	To achieve the maximum performance for your workload, a dedicated storage network is recommended. For other workloads, a separate network is recommended but not required.
SAS Controller Card	12 Gb/s LSI RAID Controller Cards, models SAS3008, SAS3108, SAS3408, SAS3508, or SAS3516, support JBOD mode, can be detected and managed by StorCLI utility. **Note:** The StorCLI utility is a pre-requisite for managing these cards. Mixed card types in one IBM Spectrum Scale Erasure Code Edition recovery group is not suggested as it could introduce performance issues. The JBOD connection mode is required for the drives used for IBM Spectrum Scale Erasure Code Edition storage.

Note: You can use the *SpectrumScale_ECE_OS_READINESS* open source tool to check that your planned ECE servers meet the minimum hardware requirements. This tool is available on the IBM Spectrum Scale

Tools GitHub (https://github.com/IBM/SpectrumScale_ECE_OS_READINESS). Contact IBM® for further details.

Hardware checklist

This topic describes the hardware checklists that must be completed before installing IBM Spectrum Scale Erasure Code Edition at your site.

You can use the *SpectrumScale_ECE_OS_READINESS* open source tool to check the defined KPI. This tool is available on IBM Spectrum Scale Tools GitHub repository (https://github.com/IBM/SpectrumScale_ECE_OS_READINESS).

Disabling volatile write cache on IBM Spectrum Scale Erasure Code Edition drives

It is required that all drives that are managed by IBM Spectrum Scale Erasure Code Edition have their volatile write cache disabled. Not doing this could result in data loss on server failure. The procedure for this varies between drive types. Contact IBM if you need assistance with the checklist.

- Here is an example of how to disable volatile write cache on a SCSI drive:

```
sdparm --set WCE=0  --save <device>
```

- To verify the change:

```
sdparm --get WCE /dev/<device>
/dev/sda: HGST      HUH721010AL4204   C384
WCE         0  [cha: y, def:  1, sav:  0]  ----> sav is 0 for it persists across power cycles
```

Note: This example is for SCSI drives only.

- Here is an example of how to query WCE for NVMe devices:

  To show current/default/saved setting (it should be 0 IN ALL 3 cases for IBM Spectrum Scale Erasure Code Edition):

```
# nvme get-feature -f 0x6 /dev/nvme0 -n 0 -s 0
get-feature:0x6 (Volatile Write Cache), Current value:00000000

# nvme get-feature -f 0x6 /dev/nvme0 -n 0 -s 1
get-feature:0x6 (Volatile Write Cache), Default value:00000000

# nvme get-feature -f 0x6 /dev/nvme0 -n 0 -s 2
get-feature:0x6 (Volatile Write Cache), Saved value:00000000
```

  If your NVMe devices have *Volatile Write Cache* enabled, it can be disabled using the following command:

```
# nvme set-feature -f 0x6 /dev/nvme0 -v 0 -s 0
set-feature:06 (Volatile Write Cache), value:00000000
```

  Not every device supports saving this setting. If you see the following output when setting this feature, you need to disable write cache with a *udev* rule or some other mechanism that is automatically applied following a node reboot.

```
# nvme set-feature -f 0x6 /dev/nvme0 -v 0 -s
NVMe Status:FEATURE_NOT_SAVEABLE(210d)
```

  Contact IBM Support if you have questions about this procedure.

Verifying that SAS drives are in JBOD mode

- To verify that the disks are in JBOD mode, issue the following command:

```
/opt/MegaRAID/storcli/storcli64 /call show
```

The system displays an output similar to the following example:

```
PD LIST :
=======

-------------------------------------------------------------------------------------------
-------
EID:Slt DID State DG       Size Intf Med SED PI SeSz Model
Sp Type
-------------------------------------------------------------------------------------------
-------
134:0    23 JBOD  -  446.102 GB SATA SSD N   N  512B MTFDDAK480TCC-1AR1ZA 01GT749D7A09326LEN
U  -
134:1    19 JBOD  -  446.102 GB SATA SSD N   N  512B MTFDDAK480TCC-1AR1ZA 01GT749D7A09326LEN
U  -
134:2    21 JBOD  -  446.102 GB SATA SSD N   N  512B MTFDDAK480TCC-1AR1ZA 01GT749D7A09326LEN
U  -
134:3    22 JBOD  -  446.102 GB SATA SSD N   N  512B MTFDDAK480TCC-1AR1ZA 01GT749D7A09326LEN
U  -
134:4    20 Onln  0  557.861 GB SAS  HDD N   N  512B ST600MM0009
U  -
134:5    17 JBOD  -  557.861 GB SAS  HDD N   N  512B ST600MM0009
U  -
134:6    18 JBOD  -  557.861 GB SAS  HDD N   N  512B ST600MM0009
U  -
134:7    16 JBOD  -  557.861 GB SAS  HDD N   N  512B ST600MM0009
U  -
-------------------------------------------------------------------------------------------
-------
```

IBM Spectrum Scale Erasure Code Edition required NVMe drive format

NVMe drives used by IBM Spectrum Scale Erasure Code Edition must be formatted with metadata size of zero, and protection information disabled. All NVMe drives in the same de-clustered array should be formatted with same LBA size.

To see the format that is in use for NVMe drives, use the **`nvme list`** command. In this example, `nvme0n1` is formatted with 4KiB logical block size and 0 byte metadata, while `nvme1n1` is formatted with 8 bytes metadata size.

```
# nvme list
 Node             SN                   Model                                    Namespace
Usage                  Format           FW Rev
 ---------------- -------------------- ---------------------------------------- ---------
-------------------------- ---------------- --------
 /dev/nvme0n1     CVFT7155000D1P6NGN   INTEL SSDPEDMD016T4L                     1          1.60 TB / 1.60
TB     4 KiB + 0 B      8DV1LP13
 /dev/nvme1n1     CVFT7155500171P6NGN   INTEL SSDPEDMD016T4L                     1          1.60 TB / 1.60
TB     4 KiB + 8 B      8DV1LP13
```

To see the available formats for an NVMe drive (and all drives of that particular type), use the `nvme id-ns` command specifying the drive path:

```
# nvme id-ns /dev/nvme1n1
 NVME Identify Namespace 1:
 nsze    : 0x1749a956
 ncap    : 0x1749a956
 nuse    : 0x1749a956
 nsfeat  : 0
 nlbaf   : 6
 flbas   : 0x14
 mc      : 0x1
 dpc     : 0x11
 dps     : 0
 nmic    : 0
 rescap  : 0
 fpi     : 0
 dlfeat  : 0
 nawun   : 0
```

```
 nawupf  : 0
 nacwu   : 0
 nabsn   : 0
 nabo    : 0
 nabspf  : 0
 noiob   : 0
 nvmcap  : 0
 nvmsetid: 0
 endgid  : 0
 nguid   : 00000000000000000000000000000000
 eui64   : 0000000000000000
 lbaf 0  : ms:0 lbads:9 rp:0x2
 lbaf 1  : ms:8 lbads:9 rp:0x2
 lbaf 2  : ms:16 lbads:9 rp:0x2
 lbaf 3  : ms:0 lbads:12 rp:0
 lbaf 4  : ms:8 lbads:12 rp:0 (in use)
 lbaf 5  : ms:64 lbads:12 rp:0
 lbaf 6  : ms:128 lbads:12 rp:0
```

The entries at the bottom of the output indicate the available LBA formats (LBAF 0 - 6 in this example). For IBM Spectrum Scale Erasure Code Edition, use a format with metadata size of zero (ms:0). It is recommended to use a format with relative performance of 0 (rp:0) for best performance.

This example shows the nvme0n1 is formatted with a metadata size of 8, so it needs to be reformatted for use with IBM Spectrum Scale Erasure Code Edition. LBA format 3 has zero metadata size, and have *rp* of zero. To format the nvme drive with this format, use the following command:

```
# nvme format /dev/nvme1n1 --lbaf=3
         Success formatting namespace:1
```

Now, all the NVMe drives have metadata size of zero:

```
# nvme list
 Node             SN                   Model                                    Namespace
Usage                Format         FW Rev
 ---------------- -------------------- ---------------------------------------- ---------
-------------------------- ---------------- --------
 /dev/nvme0n1      CVFT7155000D1P6NGN    INTEL SSDPEDMD016T4L                     1          1.60 TB /
1.60 TB     4 KiB + 0 B    8DV1LP13
 /dev/nvme1n1      CVFT715500171P6NGN    INTEL SSDPEDMD016T4L                     1          1.60 TB /
1.60 TB     4 KiB + 0 B    8DV1LP13
```

Note: For all SCSI and NVMe drives that support volatile write cache, *udev* rules should be created that disable volatile cache for these drives. This simplifies disk replacement by ensuring that the write cache is disabled automatically before adding them into the recovery group. It also ensures drives are persistently in the correct state across storage node reboots.

Operating system and drive firmware levels

All servers should have the same level of operating system software installed, and should have the same levels of drive and adapter firmware. Some of this can be verified using **mmlsfirmware** command after your system is configured, but some of this is left to the customers to manage. Improved tools for monitoring software levels across a cluster are planned for future releases.

Using KVM and VMware virtual machine as the storage node

To use KVM and VMware virtual machine as the storage node, check the following:

- Disk drives must be presented as SCSI pass through device in virtual machine.
- Each drive used in Recovery Group must assign a WWID that is unique in the cluster. You can check this by **ls -l /dev/disk/by-id** or **lsscsi -i** command on the virtual machine.
- Run the hardware precheck tool to verify the virtual machine configuration. For systems planned to be used for test and evaluation, you can ignore error messages related to virtualized configuration.

Network requirements and precheck

This topic describes the networking requirements that must be met before using IBM Spectrum Scale Erasure Code Edition.

In the IBM Spectrum Scale Erasure Code Edition configuration, network bandwidth is consumed by the client workload as well as the backend erasure code traffic between nodes. For read I/O, every 1.0 Gbps of usable bandwidth requires 2.0 Gbps of total bandwidth. For write, the overhead depends on the selected erasure code. When writing with 8+3P, each 1.0 Gbps of usable bandwidth requires 2.4 Gbps of total bandwidth. This factor is 2.25 for 8+2P, 2.5 for 4+2P, and 2.75 for 4+3P.

Additional network considerations and requirements are as follows:

- Linux bonding is supported on mode 1 (active-backup) for Ethernet and RDMA and mode 4 (IEEE 802.3ad) on Ethernet only. For mode 4 any xmit_hash_policy is supported, however, it is recommended to use layer3+4.
- Jumbo frames of 9000 MTU (on Ethernet) or higher (on RDMA) is recommended.
- When using Cluster Export Services protocol software with IBM Spectrum Scale Erasure Code Edition, a dedicated network for CES protocol traffic is required.

Network Key Performance Indicators are listed as follows:

- The average ICMP latency between any two storage nodes should be 1 msec or less.
- The maximum ICMP latency between any two storage nodes should be 2 msec or less.
- The standard deviation should be 0.333 msec or less on the ICMP latency measurements.
- The minimum throughput test of 2000 MB/sec with 1 client and all the other nodes as server for read test. Note that this is a very specific test, not a performance estimator.
- The difference between the maximum and minimum throughput values cannot be more than 20%.
- The ICMP latency metrics should be collected over an extended period, at least 500 seconds for each measurement.
- The throughput metrics should be collected over an extended period, at least 1200 seconds for each measurement.

Note: You can use the *SpectrumScale_NETWORK_READINESS* open source tool to check the defined KPI. This tool is available on the IBM Spectrum Scale Tools GitHub (https://github.com/IBM/SpectrumScale_NETWORK_READINESS). Contact IBM for further details.

Planning for erasure code selection

This topic describes the various erasure codes and the factors that need to be considered while selecting an erasure code.

Minimizing the risk of data loss due to multiple failures and minimizing disk rebuilds can be done by using 4+3P or 8+3P encoding, at the expense of additional storage overhead.

IBM Spectrum Scale Erasure Code Edition supports 4 different erasure codes: 4+2P, 4+3P, 8+2P, and 8+3P in addition to 3 and 4 way replication. Choosing an erasure code involves considering several factors. We examine some of them below.

Data protection and storage utilization

Minimizing the risk of data loss due to multiple failures and minimizing disk rebuilds can be done by using 4+3P or 8+3P encoding, at the expense of additional storage overhead. The following table shows the approximate percentage of total capacity that is usable by the file system, excluding user-configurable spare space and IBM Spectrum Scale RAID metadata. Contact IBM Support if you require any more exact estimate of usable space for your selected configuration:

Table 5. Capacity usable by file system	
Protection Type	**Usable capacity**
4-way replication	25%
3-way replication	33%
4+3P	57%
4+2P	67%
8+3P	73%
8+2P	80%

RAID rebuild

IBM Spectrum Scale RAID performs intelligent rebuilds based on the number of failures to a vdisk. For example, with 8+2P protection if 1 failure occurs IBM Spectrum Scale RAID begins to rebuild the missing data or parity strip that was lost on the failed disk or node. Since data is still protected, this rebuild process occurs in the background and has little effect on the file system performance. If a second failure occurs, IBM Spectrum Scale RAID recognizes that another failure will result in data loss. It then begins a critical rebuild in order to restore data protection. This critical rebuild phase results in performance degradation until at least one level of protection can be restored.

Nodes in a recovery group

The number of nodes in a recovery group can also impact erasure code selection. A recovery group can contain between 4 and 32 nodes. If we consider a 4-node recovery group with 4+2P protection, each node contains 1 piece of data. In addition, for each stripe, 2 nodes contain 1 piece of parity data. A failure of a node that contains both parity and data results in a double-failure for that stripe of data, which causes that stripe to be critical and results in performance degradation during the critical rebuild phase. However, in a 6-node recovery group, with the same 4+2P protection, a single node failure only results in 1 failure to the RAID array.

Recommendations

This topic describes recommendations on what block sizes to be used with each erasure code and how many node failures can occur based on the recovery group size.

The following table shows how many node failures can occur based on a recovery group size, with different erasure code protections:

Number of Nodes	**4+2P**	**4+3P**	**8+2P**	**8+3P**
4	Not recommended 1 Node	1 Node + 1 Device	Not recommended 2 Devices	Not recommended 1 Node
5	Not recommended 1 Node	1 Node + 1 Device	Not recommended 1 Node	Not recommended 1 Node
6-8	2 Nodes	2 Nodes*	Not Recommended 1 Node	1 Node + 1 Device
9	2 Nodes	3 Nodes	Not Recommended 1 Node	1 Node + 1 Device
10	2 Nodes	3 Nodes	2 Nodes	2 Nodes

11+	2 Nodes	3 Nodes	2 Nodes	3 Nodes

Note: For 7 or 8 nodes, 4+3P is limited to 2 nodes by recovery group descriptors rather than by the erasure code.

There are limits on what block sizes can be used with each erasure code, depending on device media type. The following table provides information about the limits:

Block size	**4+2P**	**4+3P**	**8+2P**	**8+3P**
1 MiB	SSD or HDD	SSD or HDD	SSD or HDD	SSD or HDD
2 MiB	SSD or HDD	SSD or HDD	SSD or HDD	SSD or HDD
4 MiB	HDD	HDD	SSD or HDD	SSD or HDD
8 MiB	HDD	HDD	HDD	HDD
16 MiB	N/A	N/A	HDD	HDD

Key	
SSD or HDD	This combination of block size and erasure code may be used with SSD (NVMe or SAS) or HDD drives
HDD	This combination of block size and erasure code may be used with HDD drives only

Even though the number of failures that can be tolerated in a smaller recovery group is the same as the number of failures in a larger recovery group, the amount of data that is critical and must be rebuilt for each failure is less for a larger recovery group. For example, with an 8+3P array on an 11-node recovery group, 3 node failures would impact all of the data in the file system. On a 30-node recovery group, 3 node failures would impact only about 10% of the data on the file system (assuming all disks are the same size), and the critical rebuild will complete more quickly because the rebuild work is distributed across a larger number of remaining nodes.

When planning the erasure code type, also consider future expansion of the cluster and storage utilization. Erasure codes for a vdisks cannot be changed after the vdisk is created, and larger stripe widths have better storage utilization. A 4+3P code utilizes 57% of total capacity for usable data, while a 8+3P code uses 73% of total capacity for usable data. So, rather than creating a 9-node cluster with 4+3P and expanding it in the future, an 11-node cluster using 8+3P may be more cost-effective. In some cases, using a non-recommended erasure code may be tolerable if there are plans to increase the cluster size.

Planning for node roles

When configuring an IBM Spectrum Scale Erasure Code Edition system, it is important to account both for workload and roles of various nodes.

Each cluster requires manager nodes and quorum nodes. Each recovery group requires a recovery group master. The IBM Spectrum Scale installation toolkit helps configure the quorum and the manager node roles.

In addition, additional IBM Spectrum Scale features require additional node types:

- CES services require CES nodes, which can be also be part of an IBM Spectrum Scale Erasure Code Edition recovery group.
- AFM gateway nodes, which cannot be a part of a recovery group.
- Transparent cloud tiering (TCT) nodes, which cannot be a part of a recovery group.
- GUI nodes, which cannot be a part of a recovery group
- TSM backup nodes, which cannot be a part of a recovery group

- Other (non- IBM Spectrum Scale Erasure Code Edition) storage types, which cannot be a part of a recovery group

Before installing IBM Spectrum Scale Erasure Code Edition, a basic network test must be passed. We have provided a tool that is freely available, open sourced and with no warranty nor official support from IBM to help you achieve running the test. Any network that does not run or pass the test should be considered as not suited to install IBM Spectrum Scale Erasure Code Edition. For more information, see "Network requirements and precheck" on page 14

When planning a system, it is best to determine the minimum requirements for IBM Spectrum Scale RAID to get the performance and capacity required, then add additional hardware as needed to meet your functional requirements with hardware for the various node roles and applications.

As nodes take on more roles, the performance of applications running on that node may be affected by the operations of those roles. File system and CPU-intensive tasks may run slower on a node that is running as a recovery group master and file system manager than on other nodes in the cluster. There are two strategies to consider when distributing node roles and workload across a cluster:

- A small subset of these nodes may be used to act in several of these roles. For example, we may choose 3 nodes to act as file system managers, recovery group masters, and quorum. Other cluster applications can then avoid these 3 nodes entirely when determining when to run, as these nodes may be more heavily utilized.
- Distribute the roles of file system managers and recovery group masters to different nodes across the cluster. In this way, we can use any node in the cluster to run applications, with the expectation that they may only be slightly impacted.

The installation toolkit will assist with node role selection and configuration during system install.

Recovery group master

When a recovery group is defined in IBM Spectrum Scale RAID, a server is chosen to be the recovery group master. The node performing this role is automatically chosen by the system. The RG master can be used for other tasks in the cluster.

Quorum nodes

IBM Spectrum Scale uses a cluster mechanism called quorum to maintain data consistency in the event of a node failure.

Quorum operates on a simple majority rule, meaning that a majority of quorum nodes in the cluster must be accessible before any node in the cluster can access a file system. This keeps any nodes that are cut off from the cluster (by a network failure for example) from writing data to the file system. When nodes fail, quorum must be maintained in order for the cluster to remain online. If quorum is not maintained, IBM Spectrum Scale file systems unmount across the cluster until quorum is reestablished, at which point file system recovery occurs. For this reason, it is important that the set of quorum nodes be carefully considered.

IBM Spectrum Scale can use one of the following two methods for determining quorum:

- Node quorum

  Node quorum is the default quorum algorithm for IBM Spectrum Scale. Quorum is defined as one plus half of the explicitly defined quorum nodes in the IBM Spectrum Scale cluster. There are no default quorum nodes; you must specify which nodes have this role.

- Node quorum with tiebreaker disks

  Tiebreaker disks can be used in shared-storage configurations in order to preserve quorum. Because clusters running IBM Spectrum Scale Erasure Code Edition do not typically use shared storage, we normally use shared storage, quorum nodes are automatically configured based on number of recovery groups configured and the number of IBM Spectrum Scale Erasure Code Edition nodes in the cluster. It is best to configure an odd number of nodes, with 3, 5, or 7 nodes being the typical numbers used. If a cluster spans multiple failure domain; such as racks, power domains, or network domains, it is best to allocate quorum nodes from each failure domain in order to maintain availability. The number of

quorum nodes, along with the Erasure Code selection will determine the maximum number of nodes that can simultaneously fail in the cluster.

It is best to allocate quorum nodes as nodes that do not require frequent reboots or downtime. If possible, choose nodes that do not run intensive compute or network loads, as these may impact the quorum messages. This becomes more important as clusters grow larger in size, as the number of quorum messages increase. Finally, quorum nodes are used to maintain critical configuration data, which is stored on the operating system disk in the `/var` file system. In order to preserve access to this data, it is best to ensure that any workloads on the quorum node do not overly stress the disk that the `/var` file system resides on. Also note that `/var` file system must be on persistent local storage for each quorum node.

Manager nodes

When defining an IBM Spectrum Scale cluster, we define one or more manager nodes. Manager nodes are used for a variety of internal tasks.

For each file system, one manager node is designated as a file system manager. This node is responsible for providing certain tasks, such as file system configuration changes, quota management, and free space management. In addition, manager nodes are responsible for token management throughout the cluster. Due to the extra load on manager nodes, it is generally recommended to not run tasks on a manager node that are time sensitive, that require real-time response, or that may excessively use the system CPU or cluster network. Any tasks that may slow the IBM Spectrum Scale file system daemon affect the overall response of the file system throughout the cluster.

For large clusters of 100 or more nodes, or clusters where the `maxFilesToCache` parameter is modified from the default, it is necessary to consider the memory use on manager nodes for token management. Tokens are used in order to maintain locks and consistency when files are opened in the cluster. The number of tokens in use is dependent on the number of files that each node may have opened or cached and the number of nodes in the cluster. For very large clusters (generally 512 nodes or more), it may be beneficial to have dedicated nodes responsible for the manager role.

To determine the overall token memory used in a system, an approximation is to examine the `maxFilesToCache` (default 4000) and `maxStatCache` (default 1000) for all nodes. Each token uses approximately 512 bytes of memory on a token manager node. For example, a 20-node cluster using the default values use (4000 + 1000) tokens * 20 nodes * 512 bytes/token = approx. 49 MB of memory. This memory will be distributed across all manager nodes, as all manager nodes share the role of token management. If there are 4 manager nodes in the above example, each manager node is responsible for just over 12 MB of tokens. For fault tolerance, it is best to leave room for a manager node to go down, so we can assume just over 16 MB of memory required.

For default values, the token memory is not a consideration on small or mid-size clusters with default values. However, in some cases, it may be beneficial to increase the `maxFilesToCache` on nodes to 100's of thousands or even millions of files. In these cases, it is important to calculate the additional memory requirement, and to ensure that any nodes have enough memory beyond the IBM Spectrum Scale Erasure Code Edition requirements to perform token management tasks.

It is recommended to have uniform workload on each IBM Spectrum Scale Erasure Code Edition storage node, to the degree possible. For this reason, we recommend either all nodes in the recovery group be manager nodes or none of the nodes be manager nodes. In storage clusters that are composed of only IBM Spectrum Scale Erasure Code Edition storage nodes, all nodes would be manager nodes. In a large cluster or a cluster with more than one IBM Spectrum Scale Erasure Code Edition recovery group, the manager nodes could be on the nodes in one recovery group or on separate nodes altogether.

CES nodes

Cluster Export Services (CES) is used in order to provide SMB, NFS, or Object access to data in the IBM Spectrum Scale file system.

For environments with high performance requirements, separate CES nodes are required. In these environments, it is recommended that a CES node run no other workload other than the export services. For details of the memory and CPU requirements for CES nodes, see the *IBM Spectrum Scale FAQ*.

Finally, the network used for accessing the nodes via CES protocols should run on a different physical adapter and network than the network used for IBM Spectrum Scale Erasure Code Edition traffic. Typically, this means that a CES node have at least 2 adapters, one for node-to-node access for IBM Spectrum Scale, and one for CES protocol access. This recommendation helps ensure that CES protocol traffic does not interfere with the IBM Spectrum Scale traffic, which results in better overall performance as well as improved cluster stability.

NSD server nodes

In some cases, a cluster may contain both IBM Spectrum Scale RAID storage, as well as other storage subsystems, such as IBM V5000, V7000, or other storage arrays. This storage can be made available for separate file systems or to tier data from a single file system.

In this case, a number of servers, typically at least 2, are attached to the external storage system using Fibre Channel or a similar interconnect. These then serve NSDs to the rest of the cluster. It is mandatory that any servers providing NSDs to the rest of the cluster be dedicated servers, separate from the servers providing storage for IBM Spectrum Scale RAID. These servers typically should not run any applications. If applications are run on these servers, then they should not be time critical, as the demands of servicing disk requests may conflict with these applications. The connectivity of these servers should be sufficient to meet the requirements of the attached disk. Ensure that CPU and network bandwidth are capable of driving the attached disk system sufficiently.

Default helper node

Certain IBM Spectrum Scale commands that may generate a significant amount of IO, such as file system restripes, adding disks, or policy scans, use helper nodes in order to run faster.

These nodes can be specified using the '-N' flag to the command or using the *defaultHelperNode* configuration value. Some commands, such as **mmapplypolicy**, may use a lot of memory or CPU resources while running, in order to sort file lists. Other commands, such as **mmrestripefs**, or **mmdelsnapshot**, may generate a significant amount of IO in order to move data and update metadata structures. When specifying helper nodes, it is best to ensure that these nodes have sufficient memory, idle CPU, and network in order to handle these requests. It may be necessary to schedule these commands for a time when the nodes or cluster are not heavily utilized as well.

Commands that use helper nodes include: **mmadddisk**, **mmapplypolicy**,**mmbackup**, **mmchdisk**, **mmcheckquota**, **mmdefragfs**, **mmdeldisk**, **mmdelsnapshot**, **mmfileid**,**mmfsck**, **mmimgbackup**, **mmimgrestore**, **mmrestorefs**, **mmrestripefs**, and **mmrpldisk**. Helper nodes typically should be separated from the servers providing storage for IBM Spectrum Scale RAID.

AFM gateway node

On AFM cache clusters, AFM uses gateway nodes in order to connect to the home system. Each AFM-enabled fileset uses a designated primary gateway node in order to connect to home and fail over to other gateway nodes as required.

AFM gateway nodes may generate a large amount of network traffic between themselves and the home system in order to fetch and to synchronize files. The bandwidth and latency on this network can directly impact file operations on AFM-enabled filesets. In order to ensure the best performance and cluster stability, it is best to have AFM traffic use a different physical adapter than the IBM Spectrum Scale cluster network. It is best to use designated gateway nodes that are not used for other application workloads. AFM uses additional node memory and cache entries on gateway nodes, so applications running on these nodes compete for cache usage, which slows both the application and AFM operations. AFM gateway nodes are required to be separate from the servers providing storage for IBM Spectrum Scale RAID.

IBM Spectrum Protect backup node

This topic describes how IBM Spectrum Scale is integrated with IBM Spectrum Protect.

IBM Spectrum Scale can integrate with IBM Spectrum Protect in one of two ways. IBM Spectrum Scale can be used as a backup pool for IBM Spectrum Protect. In this use, external clients use IBM Spectrum Protect in order to back up their data to the file system. Alternatively, IBM Spectrum Protect can also be

used to back up the IBM Spectrum Scale file system itself. When using IBM Spectrum Scale as a backup target, one or more nodes will run the IBM Spectrum Protect server. This server is contacted by other clients in order to back up. The IBM Spectrum Scale server should communicate to external clients via a separate network used for internal cluster traffic, due to the bandwidth requirements on this server.

IBM Spectrum Scale can also integrate with IBM Spectrum Protect in order to back up the IBM Spectrum Scale file system. One or more nodes in the cluster can run the IBM Spectrum Protect agents, which transfer data to an IBM Spectrum Protect server. Other backup platforms also may utilize a similar agent to scan and migrate data on a file system.

Backup nodes can become very heavily utilized during the backup window, when data is scanned and transferred to the backup provider. It is best to use a separate network on these nodes for communication with the backup server. It is also best to not run any other applications on these nodes, especially during the backup window itself.

IBM Spectrum Protect uses the IBM Spectrum Scale policy engine to scan for changed files. This scan can run across multiple nodes in the cluster, other than just the node running the backup agent. See the *Default Helper Nodes* section for guidance on helper nodes during a policy scan.

Both nodes used to run the IBM Spectrum Protect server, as well as nodes running the client are required to be separate from the servers providing storage for IBM Spectrum Scale RAID.

Transparent cloud tiering nodes

Transparent cloud tiering may make use of 1-4 gateway nodes per file system in order to communicate to a cloud provider.

These nodes are used to transfer files to and from the cloud provider. During large file migrations, or if users need to recall files, these nodes may be used heavily for file transfer. It is best to communicate to the cloud provider on a different physical network than the network used for internal cluster communications. On heavily used clusters, Transparent cloud tiering may impact any other applications running on these nodes. Transparent cloud tiering gateway nodes are required to be separate from the servers providing storage for IBM Spectrum Scale RAID.

IBM Spectrum Scale Management Interface Node

IBM Spectrum Scale Management Interface supports both GUI and RESTful API access to an IBM Spectrum Scale cluster.

IBM Spectrum Scale Management Interface can run on 1 or more dedicated nodes within the cluster. These nodes run processes and databases to monitor the cluster. The GUI consumes extra memory as well as internal hard drive space for state databases. The GUI node may also run scheduled tasks to monitor the health and utilization of the cluster. It is best to not run any compute or memory-intensive applications on the GUI node, as the GUI may impact the performance of these applications. In many cases, the nodes running the management interface are also used as the call home server and the performance monitoring collector. Management interface nodes are required to be separate from the servers providing storage for IBM Spectrum Scale RAID.

IBM Spectrum Scale call home nodes

IBM Spectrum Scale call home is used to send diagnostic data to IBM.

Nodes are arranged into call home servers, which are responsible for collecting all of the data within a call home group and sending the data to IBM. Large clusters may consist of several groups. It is recommended to use call home whenever possible to assist in gathering data for support.

Call home servers are required to be separate from severs providing storage for IBM Spectrum Scale RAID. In the case of small clusters of 32 nodes or less, the call home server may be the same as the management interface node. In larger clusters, additional call home servers may be required. For additional information on sizing call home requirements, see the *Understanding call home* topic in the *IBM Spectrum Scale: Administration Guide*.

Performance monitoring

IBM Spectrum Scale performance monitoring divides nodes into collector and sensor nodes. Sensors run on all nodes that we wish to collect performance data from. Collectors run on a small number of nodes and are used to aggregate all of the sensor data into a single view. Sensors can run on all nodes, including nodes that provide storage for IBM Spectrum Scale Erasure Code Edition. Collectors should be run on nodes that do not provide IBM Spectrum Scale Erasure Code Edition storage. Typically, the same nodes used as management interface nodes will be used as collector nodes. On clusters with hundreds of nodes, multiple collectors may be required in order to aggregate data across the cluster. It is not recommended to run real-time or time-sensitive tasks on collector nodes.

File audit logging and watch folders

File Audit Logging (FAL) and Watch Folders use message queues in order to monitor file access on the cluster.

FAL producers create messages when certain file operations are performed (for example, file writes, reads, etc.). FAL consumers read these messages and perform required actions, such as writing to audit logs. All nodes, including the nodes providing IBM Spectrum Scale Erasure Code Edition storage may be producers, in order to provide complete access logging. Consumers must be on nodes that do not provide storage to the IBM Spectrum Scale Erasure Code Edition cluster, due to the additional load on the system caused by monitoring usage on the cluster. In addition, consumer nodes should not run real-time or time-sensitive applications.

Other IBM Spectrum Scale features

IBM Spectrum Scale offers caching features such as Local Read-Only Cache and High Availability Write Cache (LROC and HAWC), which can provide additional high-speed caching to speed up certain applications.

LROC and HAWC can be used on file systems that contain storage provided by IBM Spectrum Scale Erasure Code Edition. However, LROC and HAWC devices cannot be installed directly on nodes providing IBM Spectrum Scale Erasure Code Edition storage. Client nodes that are not part of the IBM Spectrum Scale Erasure Code Edition recovery group can use these devices.

Chapter 4. Installing IBM Spectrum Scale Erasure Code Edition

You can install IBM Spectrum Scale Erasure Code Edition by using the installation toolkit.

IBM Spectrum Scale Erasure Code Edition installation prerequisites

IBM Spectrum Scale Erasure Code Edition requires several software packages in addition to the base operating system.

Before installing IBM Spectrum Scale Erasure Code Edition, your network must pass the latency network KPIs for Ethernet networks to support RDMA network.

Note: In the IBM Spectrum Scale Erasure Code Edition, customers will be required to meet the following network KPI metrics before an installation is completed. For more information, see "Network requirements and precheck" on page 14. Also, you must verify that the hardware planned for ECE storage servers meets the minimum requirements. For more information, see "Minimum hardware requirements and precheck" on page 9. The installation toolkit will also verify that your hardware meets minimum requirements, but it is useful to execute this tool prior to beginning your installation.

The following rpms are required to be installed:

- `sg3_utils`
- `nvme-cli`
- `storcli` (if using SAS drives with LSI HBA)
- `dmidecode`
- `PyYAML`

Furthermore, it is important to ensure that you have the latest version of Mellanox OFED installed on each node. Likewise, the driver versions should be maintained at a consistent level across all nodes.

Note: All IBM Spectrum Scale cluster software and configuration prerequisites must also be satisfied. For more information, see *Installation prerequisites* in *IBM Spectrum Scale: Concepts, Planning, and Installation Guide*.

IBM Spectrum Scale Erasure Code Edition precheck

The IBM Spectrum Scale Erasure Code Edition precheck is integrated with the installation toolkit installation, deployment or upgrade precheck. For IBM Spectrum Scale Erasure Code Edition, the precheck includes the following on all scale-out nodes:

- Check whether the CPU requirements are met
- Check whether the memory requirements are met
- Check whether the OS is supported
- Check whether the networking requirements including the required NIC and SAS adapters are met
- Check whether the required syscall parameters are set correctly

Installation toolkit-related prerequisites

- Ensure that networking is set up in one of the following ways.
  - DNS is configured such that all host names, either short or long, are resolvable.
  - All host names are resolvable in the `/etc/hosts` file. The host entries in the `/etc/hosts` file must be in the following order:

```
<IP address> <Fully qualified domain name> <Short name>
```

- Passwordless SSH must be set up using the FQDN and the short name of the node.

For more information, see *Preparing to use the installation toolkit* in *IBM Spectrum Scale: Concepts, Planning, and Installation Guide*.

Installation toolkit-related limitations

- The installation toolkit is not supported in a sudo wrapper environment. Therefore, sudo wrappers cannot be used for installation, deployment, or upgrade of IBM Spectrum Scale Erasure Code Edition. After installation, deployment, or upgrade, you an use sudo wrappers for administration tasks in an IBM Spectrum Scale Erasure Code Edition environment.

  For more information, see *Limitations of the installation toolkit* in *IBM Spectrum Scale: Concepts, Planning, and Installation Guide*.
- The installation toolkit does not support advanced parameters of vdisk sets and file systems that can be specified by using the **mmvdisk** command. After recovery groups are created, you can use the **mmvdisk** command to create vdisk sets and file systems with the advanced configuration parameters. Thereafter, you can use the installation toolkit deployment operation for protocol deployment.
- The installation toolkit cannot accept multiple recovery groups as an argument while defining a vdisk set. If you want to specify more than one recovery group with the vdisk set, use the **mmvdisk** command after the installation phase is completed.
- The installation toolkit does not support declustered array as an argument while defining the vdisk set. If you want to specify one or more declustered arrays with the vdisk set, use the **mmvdisk** command after the installation phase is completed.
- The installation toolkit cannot accept multiple vdisk sets as an argument while defining the file system. If you want to specify multiple vdisk sets with the file system, use the **mmvdisk** command after the installation phase is completed.
- The installation toolkit does not support online upgrade of IBM Spectrum Scale Erasure Code Edition.
- The installation toolkit does not support the creation of hybrid clusters (IBM Spectrum Scale + ESS + IBM Spectrum Scale Erasure Code Edition).

IBM Spectrum Scale Erasure Code Edition installation overview

The installation of IBM Spectrum Scale Erasure Code Edition by using the installation toolkit occurs in these phases.

Phase 1: Network and hardware precheck

1. Download or clone the following two precheck tools on one of the nodes that are planned for your ECE storage configuration.
   - This tool is available on the IBM Spectrum Scale Tools GitHub, https://github.com/IBM/SpectrumScale_ECE_OS_READINESS
   - This tool is available on the IBM Spectrum Scale Tools GitHub), https://github.com/IBM/SpectrumScale_NETWORK_READINESS
2. Run the hardware precheck tool on at least one of your ECE storage nodes for each recovery group. Review the `README.md` file carefully for prerequisites and execution procedures.
3. Run the network precheck tool including each IBM Spectrum Scale Erasure Code Edition storage node. Review the `README.md` file carefully for prerequisites and execution procedures.

Phase 2: Cluster definition

By using the **./spectrumscale** command, the following steps are done.

1. Installer node is defined by the user.
2. Setup type is specified as `ece` by the user.

3. Scale-out nodes and other node designations are done by the user.

   Other types of nodes that can be designated include protocol, GUI, call home, and file audit logging. If you are planning to use GUI, call home, performance monitoring, or file audit logging, you must add a client node for each of these functions.

   Note: When you are adding a node in an existing cluster, the installation toolkit adds only the node in the existing cluster with the client or the server license. You must use the **mmvdisk** command to manually add the node into the existing node class.
4. Recovery group is defined by the user.
5. Vdisk set is defined by the user. [Vdisk set definition can be done after the installation phase]

   Note:

   - The installation toolkit cannot accept multiple recovery groups as an argument while defining a vdisk set. If you want to specify more than one recovery group with the vdisk set, use the **mmvdisk** command after the installation phase is completed.
   - The installation toolkit does not support declustered array as an argument while defining the vdisk set. If you want to specify one or more declustered arrays with the vdisk set, use the **mmvdisk** command after the installation phase is completed.
6. File system is defined by the user. [File system definition can be done after the installation phase]

   Note: The installation toolkit cannot accept multiple vdisk sets as an argument while defining the file system. If you want to specify multiple vdisk sets with the file system, use the **mmvdisk** command after the installation phase is completed.

Phase 3: Installation

This phase starts upon issuing the **./spectrumscale install** command.

1. IBM Spectrum Scale Erasure Code Edition packages including the IBM Spectrum Scale Erasure Code Edition license package are installed.
2. IBM Spectrum Scale Erasure Code Edition cluster is created.
3. Quorum and manager nodes are configured.
4. Server and client licenses are applied.
5. Node class is created.
6. Recovery group is created.

Note: During the installation, support packages are also installed. These support packages include supported disk topologies and starting udev rules for each node. There is a rule file that is placed here: `/etc/udev/rules.d/99-ibm-scaleout.rules`. These rules have these settings and they are meant to be a good starting point for a typical hardware configuration. You might need to adjust these settings for your hardware configuration:

```
#
# IBM Spectrum Scale RAID (GNR) block device attributes for
# Erasure Code Edition (ECE) storage-rich servers.
#
# These are least common denominator settings. It is likely
# that specific installations can increase especially the
# max_sectors_kb for GNR pdisks.
#
# After initial ECE installation and after any change to the
# contents of these rules, run
#       udevadm trigger --subsystem-match=block
# and inspect /var/log/messages for unexpected udev entries.
# Subsequent reboots and block device replacement will
# automatically invoke these rules as "add|change" events.
#
# ------------------------------------------------------------------------
#
# Identify the boot SCSI disk by the presence of a SWAP partition.
# Set boot disk nr_requests and queue_depth to reasonable values.
#
ACTION=="add|change", SUBSYSTEM=="block",
```

```
KERNEL=="sd*[^0-9]", PROGRAM="/usr/bin/lsblk -rno
FSTYPE,MOUNTPOINT,NAME /dev/%k", RESULT=="*SWAP*",
ATTR{queue/nr_requests}="128", ATTR{device/queue_depth}="64"
#
# Identify eligible GNR SCSI pdisks by the absence of a SWAP partition.
# Set preferred GNR attributes. The only attribute that should possibly
# be changed is max_sectors_kb, up to a value of 8192, depending on
# what the SCSI driver and disks support.
#
ACTION=="add|change", SUBSYSTEM=="block",
KERNEL=="sd*[^0-9]", PROGRAM="/usr/bin/lsblk -rno
FSTYPE,MOUNTPOINT,NAME /dev/%k",
RESULT!="*SWAP*", ATTR{queue/scheduler}="deadline",
ATTR{queue/nr_requests}="256", ATTR{device/queue_depth}="31",
ATTR{queue/max_sectors_kb}="1024", ATTR{queue/read_ahead_kb}="0",
ATTR{queue/rq_affinity}="2"
#
# Identify eligible GNR NVMe pdisks by the absence of a MOUNTPOINT.
# Set preferred GNR attributes. The only attribute that should possibly
# be changed is max_sectors_kb, up to a value of 8192, depending on
# what the NVMe driver and devices support.
#
ACTION=="add|change", SUBSYSTEM=="block",
KERNEL=="nvme*", KERNEL!="nvme*p[0-9]",
PROGRAM="/usr/bin/lsblk -rno
FSTYPE,MOUNTPOINT,NAME /dev/%k", RESULT!="*/*",
ATTR{queue/scheduler}="none", ATTR{queue/nr_requests}="256",
ATTR{queue/max_sectors_kb}="128",
ATTR{queue/read_ahead_kb}="0",
ATTR{queue/rq_affinity}="2"
```

Note: If you are planning to deploy protocols in the IBM Spectrum Scale Erasure Code Edition cluster, you must define a CES shared root file system before initiating the installation toolkit deployment phase by using the following command.

```
./spectrumscale config protocols -f FileSystem -m MountPoint
```

Phase 4: Deployment

This phase starts upon issuing the **`./spectrumscale deploy`** command.

1. Vdisk sets are created.
2. File systems are created.
3. Protocols are deployed, if applicable.

Additional IBM Spectrum Scale configuration items

It is recommended to add the following configuration settings for improved performance:

1. Set node class.

```
NC=Erasure Code Edition node class
```

2. Update tuning parameters for nodes in the IBM Spectrum Scale Erasure Code Edition node class.

```
mmchconfig nsdMaxWorkerThreads=3842     -N $NC
mmchconfig nsdMinWorkerThreads=3842     -N $NC
mmchconfig nsdRAIDThreadsPerQueue=16   -N $NC
mmchconfig nsdSmallThreadRatio=1 -N $NC
```

Quorum or manager node rules in IBM Spectrum Scale Erasure Code Edition

- In case of a single recovery group, the following quorum node rules apply.
  - When the number of scale-out nodes is 4, the number of quorum nodes is set to 3.
  - When the number of scale-out nodes is 5 or 6, the number of quorum nodes is set to 5.
  - When the number of scale-out nodes is 7 or more, the number of quorum nodes is set to 7.
- If the number of recovery groups is more than 1 and less than or equal to 7, 7 quorum nodes are distributed across recovery groups in a round robin manner.

- If the number of recovery groups is more than 7, 7 recovery groups are selected as quorum holders.
- If there is no recovery group or quorum node that is defined in the cluster configuration, the installation toolkit displays the following message.

```
"You have not defined any recovery group in the cluster configuration.
Installer will automatically define the quorum configuration. Do you want to continue"
```

  If you specify `yes` then quorum nodes are distributed according to the single recovery group rule.
- If you are adding a new recovery group in an existing cluster or if you want to add a new node into the existing node class, the existing quorum configuration is not modified by the installation toolkit.
- For an existing cluster, if you want to have quorum on a different node or a different recovery group then you must use an IBM Spectrum Scale command such as **mmchnode** to change this configuration.
- Every scale-out node has the manager mode designation. Scale-out nodes in a recovery group are equivalent so any of them can pick up the cluster manager or the file system manager role.

Installing IBM Spectrum Scale Erasure Code Edition by using the installation toolkit

IBM Spectrum Scale Erasure Code Edition is available in a separate installation package and you can install it by using the installation toolkit.

Use the following steps to install IBM Spectrum Scale Erasure Code Edition.

1. Download the IBM Spectrum Scale Erasure Code Edition self-extracting package from the IBM Spectrum Scale page on Fix Central.
2. Extract the installation package.

   ```
   # ./Spectrum_Scale_Erasure_Code-5.0.y.z-x86_64-Linux-install
   ```

   The installation toolkit gets extracted to the `/usr/lpp/mmfs/5.0.x.x/installer/` directory.
3. Change the directory to where the installation toolkit is extracted.

   ```
   cd /usr/lpp/mmfs/5.0.x.z/installer/
   ```

4. Specify the installer node and the setup type in the cluster definition file.

   The setup type must be `ece` for IBM Spectrum Scale Erasure Code Edition.

   ```
   ./spectrumscale setup -s InstallerNodeIP -st ece
   ```

5. Add scale-out nodes for IBM Spectrum Scale Erasure Code Edition in the cluster definition file.

   ```
   ./spectrumscale node add NodeName -so
   ```

   Specify any other node designations in the cluster definition file.

   Note: For environments with high-performance requirements, IBM Spectrum Scale Erasure Code Edition storage nodes must not be assigned file audit logging, call home, or protocol node roles.

   You can use the following command to display the list of nodes that are specified in the cluster definition file and the respective node designations.

   ```
   ./spectrumscale node list
   ```

   A sample output is as follows:

   ```
   [ INFO  ] List of nodes in current configuration:
   [ INFO  ] [Installer Node]
   [ INFO  ] 198.51.100.15
   [ INFO  ]
   [ INFO  ] [Cluster Details]
   [ INFO  ] Name: scalecluster.example.com
   [ INFO  ] Setup Type: Erasure Code Edition
   ```

```
[ INFO  ]
[ INFO  ] [Protocols]
[ INFO  ] Object : Disabled
[ INFO  ] SMB    : Disabled
[ INFO  ] NFS    : Enabled
[ INFO  ]
[ INFO  ] [Extended Features]
[ INFO  ] File Audit logging     : Enabled
[ INFO  ] Watch folder           : Disabled
[ INFO  ] Management GUI         : Disabled
[ INFO  ] Performance Monitoring : Disabled
[ INFO  ] Callhome               : Enabled
[ INFO  ]
[ INFO  ] GPFS                   Admin  Quorum  Manager  NSD   Protocol Callhome FAL/WF
Scale-out  OS   Arch
[ INFO  ] Node                   Node   Node     Node   Server   Node    Server  Broker
Node
[ INFO  ] node1.example.com        X       X       X
X     rhel7  x86_64
[ INFO  ] node2.example.com                X       X
X     rhel7  x86_64
[ INFO  ] node3.example.com                X       X
X     rhel7  x86_64
[ INFO  ] node4.example.com                X       X
X     rhel7  x86_64
[ INFO  ] node6.example.com
X                      rhel7  x86_64
[ INFO  ] node7.example.com                X       X
X                            rhel7  x86_64
[ INFO  ] node8.example.com                X       X
X                            rhel7  x86_64

[ INFO  ]
[ INFO  ] [Export IP address]
[ INFO  ] 198.51.100.11 (pool)
[ INFO  ] 198.51.100.12 (pool)
```

6. Define the recovery group for IBM Spectrum Scale Erasure Code Edition in the cluster definition file.

   ```
   ./spectrumscale recoverygroup define -N Node1,Node2,...,NodeN
   ```

   If you want to create more than one recovery groups, you must specify the recovery group name and the scale-out node class name. For example:

   ```
   ./spectrumscale recoverygroup define -rg RGName -nc ScaleOutNodeClassName -N
   Node1,Node2,...,NodeN
   ```

7. Perform environment prechecks before issuing the installation toolkit installation command.

   ```
   ./spectrumscale install --pre
   ```

8. Perform the installation toolkit installation procedure.

   ```
   ./spectrumscale install
   ```

9. Define vdisk sets for IBM Spectrum Scale Erasure Code Edition in the cluster definition file.

   ```
   ./spectrumscale vdiskset define -rg RgName -code RaidCode -bs BlockSize -ss VdiskSetSize
   ```

10. Define the file system for IBM Spectrum Scale Erasure Code Edition in the cluster definition file.

    ```
    ./spectrumscale filesystem define -fs FileSystem -vs VdiskSet
    ```

11. Perform environment prechecks before issuing the installation toolkit deployment procedure.

    ```
    ./spectrumscale deploy --pre
    ```

12. Perform the installation toolkit deployment procedure.

    ```
    ./spectrumscale deploy
    ```

Setting up IBM Spectrum Scale Erasure Code Edition for NVMe

IBM Spectrum Scale requires additional configuration for use with NVMe drives.

IBM Spectrum Scale Erasure Code Edition brings enclosure-like management services to direct attached storage disks, allowing users to identify and replace disks without compromising system availability or integrity. IBM Spectrum Scale Erasure Code Edition ships with support for NVMe disks with a U.2 form factor. The U.2 form factor allows system administrators to replace NVMe disks as if they were regular HDD or SSD drives. Drive LED control is not supported at this time, but replacement operations will work with their slot location. This means that NVMe drives may be replaced, but the replacement process will not trigger any identification or replace lights on the drive. For more information on disk replacement procedure, see “Physical disk procedures” on page 55.

To support disk replacement for NVMe drives on IBM Spectrum Scale Erasure Code Edition, users need to define a pseudo enclosure describing a server’s disk layout and capabilities.

Creating an Enclosure Descriptor File (EDF)

U.2 NVMe drives reside in a pseudo enclosure within their server node. This pseudo enclosure is defined using a plain-text EDF. The EDF describes the structure and layout of the storage components within the enclosure, as well as the capabilities of these components.

The EDF also contains a structure known as a “bay_map”, which describes a mapping from the server’s external drive slots to PCIe buses. The EDF refers to the PCIe buses as “ports”. A given server node’s slot to PCIe bus mapping may vary depending on its vendor and its internal cabling. This mapping is therefore crucial to ensure that disk replacement operations select the correct disk. It is recommended to use the same server node hardware across an IBM Spectrum Scale Erasure Code Edition recovery group, as this ensures a uniform NVMe drive mapping and allows a single EDF to be deployed on all nodes without additional configuration. Otherwise, a separate EDF has to be created on each node.

Note:

- NVMe drives might be organized into exclusive namespaces on a single controller or shared namespaces across multiple controllers. For use with IBM Spectrum Scale Erasure Code Edition, NVMe drives must be configured such that there is a single namespace on each controller.
- NVMe drive slot mapping must be done before the recovery group creation. IBM Spectrum Scale Erasure Code Edition supports doing the mapping and the re-mapping after the recovery group is created. The tool `dasEDFTool.py` only reads data from NVMe drives. Do not write data on NVMe drives after the recovery group created. If you want to do re-mapping, delete the `*.edf` files in `/usr/lpp/mmfs/data/gems` and do the procedures again.

Before starting, ensure the following:

- To define NVMe drive mapping, you must first select a server and populate all NVMe-capable slots with NVMe drives. After the mapping process is completed, the extra drives can be returned to the spare inventory or to other servers. This can be done once for each collection of servers with the same disk topology. IBM Spectrum Scale Erasure Code Edition does not support mapping additional NVMe server slots after this initial NVMe drive mapping is completed.

The creation of a properly formatted and named EDF with a correct bay_map may be produced using the sample script `/usr/lpp/mmfs/samples/vdisk/dasEDFTool.py`.

```
# /usr/lpp/mmfs/samples/vdisk/dasEDFTool.py [--slotrange [0-24] [0-24]] [--report] [--force]
```

For example, to create an EDF describing NVMe-capable server slots 16-18, issue this command:

```
# /usr/lpp/mmfs/samples/vdisk/dasEDFTool.py --slotrange 16 18
```

For each NVMe block device found in `/proc/partitions`, the tool will blink that block device’s activity light using a read workload and will prompt the user to enter the corresponding slot for the blinking disk.

```
Now blinking path /dev/nvme0n1
```

```
>>> Enter the slot number: 18
Now blinking path /dev/nvme1n1
>>> Enter the slot number: 17
Now blinking path /dev/nvme2n1
>>> Enter the slot number: 16
```

In this example, slots 16-18 represent all NVMe-capable drive slots on the server. The tool fails if it detects that you are trying to map more slots than the actual number of NVMe-capable drive slots. The EDF is written to `/usr/lpp/mmfs/data/gems/`, and it must be copied to all nodes with the same NVMe drive topology.

Verifying the Enclosure Descriptor File

Note: Ensure that you do not use any command to corrupt the data on the disk if the recovery group is already created.

It is recommended that you verify the EDF after running `dasEDFTool.py`. The tool can also be used to help verify your install by reporting the slot to bus mappings:

```
# /usr/lpp/mmfs/samples/vdisk/dasEDFTool.py --report

Summary Report:
  Slot 16 => X
  Slot 17 => Y
  Slot 18 => Z
```

where X, Y, and Z are PCIe bus numbers.

To check the slot to bus mapping from the report above, do the following steps:

1. Gather a list of your physical NVMe disk controllers. You can gather a list of physical controllers by running the following command:

   ```
   # lsblk | egrep -o "nvme[0-9][0-9]?" | uniq
   nvme0
   nvme1
   …
   ```

2. For each disk controller, use the sysfs file system to determine which PCIe bus connects to which controller. For disk controller “nvme0”:

   ```
   # find /sys/devices/ | egrep "nvme0$"
   /sys/devices/pciDOMAIN/DOMAIN:X:X.X/…/DOMAIN:BUS:DEVICE.FUNCTION/nvme/nvme0
   ```

   PCIe addresses are of the form “DOMAIN:BUS:DEVICE.FUNCTION”. The last bus in the path is the bus for the given disk controller. In this case, nvme0 will have bus BUS, where bus is a two-digit hexadecimal number. The information gathered in this step should match the generated report above.

3. For each disk controller, issue a **dd** read command to its corresponding block device:

   ```
   # dd if=/dev/nvmeXn1 of=/dev/null bs=1M count=10000 skip=1000
   ```

   Because the drives should be formatted with a single namespace, controller `nvmeX` corresponds to block device `/dev/nvmeXn1`. From the report above, if we ran this command on `/dev/nvme2n1`, we should expect to see the activity light in slot 16 light up.

Finally, the rest of the EDF can be checked with:

```
# tslsenclslot -a | mmyfields -s slot SlotHandle LocationCode | grep gems | awk '{print $2}'

SERIALNUMBER-SLOTX
SERIALNUMBER-SLOTY
...
```

This command should print out the location codes of the correctly configured slots. The serial number can be checked with:

```
# dmidecode -s system-serial-number
```

Chapter 5. Incorporating IBM Spectrum Scale Erasure Code Edition in an Elastic Storage Server (ESS) cluster

Use these procedures to incorporate IBM Spectrum Scale Erasure Code Edition in an Elastic Storage Server (ESS) cluster.

Ensure that the following prerequisites are met before you install IBM Spectrum Scale Erasure Code Edition in an Elastic Storage Server (ESS) cluster.

- ESS version is 5.3.4 or later.
- IBM Spectrum Scale Erasure Code Edition version is 5.0.3.1 or later.
- Hardware minimum requirements are met.
- All typical IBM Spectrum Scale and ESS prerequisites such as passwordless SSH, minimum OS levels, python, sg3_utils and pciutils software requirements are met.

  The IBM Spectrum Scale installation toolkit can help identify many missing prerequisites.
- Network performance minimum requirements are met.
- General understanding of how the IBM Spectrum Scale installation toolkit process works.
- Possible protocol architecture conflicts are mitigated.

The installation of IBM Spectrum Scale Erasure Code Edition in an Elastic Storage Server (ESS) cluster comprises 4 phases.

1. Phase 1: Convert ESS into mmvdisk management.
2. Phase 2: Add nodes to the ESS cluster using the installation toolkit.
3. Phase 3: Prepare the IBM Spectrum Scale Erasure Code Edition cluster using the installation toolkit.
4. Phase 4: Complete the configuration with mmvdisk commands.

Converting Elastic Storage Server (ESS) to mmvdisk management

In the 1st phase of incorporating IBM Spectrum Scale Erasure Code Edition in an ESS cluster, check if ESS is mmvdisk managed, and if required convert ESS to mmvdisk managed.

1. Check if ESS is mmvdisk managed.

   If the cluster is not mmvdisk managed, the `remarks` column in the output contains `non-mmvdisk`.

```
# mmvdisk server list

node
number  server                          node class  recovery groups           remarks
------  ------------------------------  ----------  ------------------------  -------
     1  gssio1-ib.example.com              -           rg_gssio1-ib, rg_gssio2-ib  non-
mmvdisk
     2  gssio2-ib.example.com              -           rg_gssio1-ib, rg_gssio2-ib  non-
mmvdisk

# mmvdisk rg list


                                                        needs    user
recovery group  active   current or master server      service  vdisks  remarks
--------------  -------  ------------------------------  -------  ------  -------
rg_gssio1-ib    yes      gssio1-ib.example.com             no           1  non-mmvdisk
rg_gssio2-ib    yes      gssio2-ib.example.com             no           1  non-mmvdisk
```

2. Convert ESS to mmvdisk managed.

```
# mmvdisk recoverygroup convert --recovery-group rg_gssio1-ib,rg_gssio2-ib --node-class
ess_nc1
```

```
mmvdisk: This command will permanently change the GNR configuration
mmvdisk: attributes and disable the legacy GNR command set for the
mmvdisk: servers and recovery groups involved, and their subsequent
mmvdisk: administration must be performed with the mmvdisk command.

mmvdisk: Do you wish to continue (yes or no)? yes

mmvdisk: Converting recovery groups 'rg_gssio1-ib' and 'rg_gssio2-ib'.
mmvdisk: Creating node class 'ess_nc1'.
mmvdisk: Adding 'gssio1-ib' to node class 'ess_nc1'.
mmvdisk: Adding 'gssio2-ib' to node class 'ess_nc1'.
mmvdisk: Associating recovery group 'rg_gssio1-ib' with node class 'ess_nc1'.
mmvdisk: Associating recovery group 'rg_gssio2-ib' with node class 'ess_nc1'.
mmvdisk: Recording pre-conversion cluster configuration
mmvdisk: in /var/mmfs/tmp/mmvdisk.convert.rg_gssio1-ib.rg_gssio2-ib.before.m07
mmvdisk: Updating server configuration attributes.
mmvdisk: Checking resources for specified nodes.
mmvdisk: Setting configuration for node class 'ess_nc1'.
mmvdisk: Defining vdisk set 'VS001_essFS' with recovery group
mmvdisk: 'rg_gssio1-ib' (vdisk 'rg_gssio1_ib_DA1_DataAndMetaData_16M_2p_1').
mmvdisk: Defining vdisk set 'VS002_essFS' with recovery group
mmvdisk: 'rg_gssio2-ib' (vdisk 'rg_gssio2_ib_DA1_DataAndMetaData_16M_2p_1').
mmvdisk: Committing cluster configuration changes.
mmvdisk: Recording post-conversion cluster configuration in
mmvdisk: /var/mmfs/tmp/mmvdisk.convert.rg_gssio1-ib.rg_gssio2-ib.after.m07
mmvdisk: For configuration changes to take effect, GPFS should be restarted
mmvdisk: on node class 'ess_nc1'.
```

3. Restart GPFS.

```
# mmshutdown -a
# mmstartup -a
```

4. Verify the GPFS state.

```
# mmgetstate -a
```

5. View the ESS cluster after it is converted to mmvdisk managed.

```
# mmvdisk server list


node
number  server                 node class  recovery groups            remarks
------  ---------------------  ----------  -------------------------  -------
   1    gssio1-ib.example.com  ess_nc1     rg_gssio1-ib, rg_gssio2-ib
   2    gssio2-ib.example.com  ess_nc1     rg_gssio1-ib, rg_gssio2-ib

# mmvdisk rg list


                                                        needs    user
recovery group  active   current or master server      service  vdisks  remarks
--------------  -------  -----------------------------  -------  ------  -------
rg_gssio1-ib    yes      gssio1-ib.example.com             no       1
rg_gssio2-ib    yes      gssio2-ib.example.com             no       1

# mmvdisk vdisk list --vdisk-set all


                                                                    declustered array,
RAID code,
vdisk               vdisk set            file system    recovery group          block
size           remarks
------------------  -------------------  -------------  --------------
--------------------------  -------
rg_gssio1_ib_DA1_DataAndMetaData_16M_2p_1  VS001_essFS  essFS  rg_gssio1-ib  DA1, 8+2p, 16
MiB
rg_gssio2_ib_DA1_DataAndMetaData_16M_2p_1  VS002_essFS  essFS  rg_gssio2-ib  DA1, 8+2p, 16
MiB
```

Adding nodes to the Elastic Storage Server (ESS) cluster using the installation toolkit

In the 2nd phase of incorporating IBM Spectrum Scale Erasure Code Edition in an ESS cluster, use the installation toolkit to create a generic cluster definition file that will be used to install and deploy Erasure Code Edition candidate nodes on the ESS cluster as generic IBM Spectrum Scale nodes.

1. From IBM FixCentral, download the IBM Spectrum Scale Advanced Edition 5.x.x.x installation package. You must download this package to the node that you plan to use as your installer node for the IBM Spectrum Scale Advanced Edition installation and the subsequent IBM Spectrum Scale Erasure Code Edition installation. Also, use a node that you plan to add in the existing ESS cluster.
2. Extract the IBM Spectrum Scale Advanced Edition 5.x.x.x installation package to the default directory or a directory of your choice on the node that you plan to use as the installer node.

   ```
   /DirectoryPathToDownloadedCode/Spectrum_Scale_Advanced-5.x.x.x-x86_64-Linux-install
   ```

3. Change the directory to the default directory for the installation toolkit.

   ```
   # cd /usr/lpp/mmfs/5.x.x.x/installer
   ```

4. Set up the installer node and the setup type as ess.

   In this command example, 198.51.100.1 is the IP address of the scale-out node that is planned to be designated as the installer node.

   ```
   # ./spectrumscale setup -s 198.51.100.1 -st ess


   [ INFO ] Installing prerequisites for install node
   [ INFO ] Existing Chef installation detected. Ensure the PATH is configured so that
   chefclient
   and knife commands can be run.
   [ INFO ] Your control node has been configured to use the IP 198.51.100.1 to communicate
   with other nodes.
   [ INFO ] Port 8889 will be used for chef communication.
   [ INFO ] Port 10080 will be used for package distribution.
   [ INFO ] Install Toolkit setup type is set to ESS. This mode will allow the EMS node to
   execute Install Toolkit commands.
   [ INFO ] SUCCESS
   [ INFO ] Tip : Designate an EMS node as admin node: ./spectrumscale node add <node> -a
   [ INFO ] Tip : After designating an EMS node, add nodes for the toolkit to act upon:
   ./spectrumscale node add <node> -p -n
   [ INFO ] Tip : After designating the EMS node, if you want to populate the cluster
   definition
   file with the current configuration, you can run: ./spectrumscale config populate -N
   <ems_node>
   ```

5. Add the existing EMS node to the cluster definition as admin, quorum, and EMS nodes.

   ```
   # ./spectrumscale node add ess.example.com -a -q -e


   [ INFO ] Adding node ess.example.com as a GPFS node.
   [ INFO ] Adding node ess.example.com as a quorum node.
   [ INFO ] Setting ess.example.com as an admin node.
   [ INFO ] Configuration updated.
   [ INFO ] Setting ess.example.com as an ESS node.
   [ INFO ] Configuration updated.

   # ./spectrumscale node list

   [ INFO  ] List of nodes in current configuration:
   [ INFO  ] [Installer Node]
   [ INFO  ] 198.51.100.1
   [ INFO  ]
   [ INFO  ] [Cluster Details]
   [ INFO  ] Name: scalecluster.example.com
   [ INFO  ] Setup Type: ESS
   [ INFO  ]
   [ INFO  ] [Extended Features]
   [ INFO  ] File Audit logging     : Disabled
   ```

```
[ INFO  ] Watch folder           : Disabled
[ INFO  ] Management GUI         : Enabled
[ INFO  ] Performance Monitoring : Disabled
[ INFO  ] Callhome               : Disabled
[ INFO  ]
[ INFO  ] GPFS                Admin  Quorum  Manager   NSD   Protocol   GUI    Perf Mon
EMS   OS   Arch
[ INFO  ] Node                Node   Node     Node   Server   Node    Server
Collector
[ INFO  ] ess.example.com     X       X                                X
X   rhel7  ppc64le
[ INFO  ]
[ INFO  ] [Export IP address]
[ INFO  ] No export IP addresses configured
```

6. Add the IBM Spectrum Scale Erasure Code Edition candidate nodes generically.

```
# ./spectrumscale node add 198.51.100.1
[ INFO ] Adding node node1.example.com as a GPFS node.
# ./spectrumscale node add 198.51.100.2
[ INFO ] Adding node node2.example.com as a GPFS node.
# ./spectrumscale node add 198.51.100.3
[ INFO ] Adding node node3.example.com as a GPFS node.
# ./spectrumscale node add 198.51.100.4
[ INFO ] Adding node node4.example.com as a GPFS node.
# ./spectrumscale node add 198.51.100.5
[ INFO ] Adding node node5.example.com as a GPFS node.
# ./spectrumscale node add 198.51.100.6
[ INFO ] Adding node node6.example.com as a GPFS node.
```

Verify the node details.

```
# ./spectrumscale node list

[ INFO  ] List of nodes in current configuration:
[ INFO  ] [Installer Node]
[ INFO  ] 198.51.100.1
[ INFO  ]
[ INFO  ] [Cluster Details]
[ INFO  ] Name: scalecluster.example.com
[ INFO  ] Setup Type: ESS
[ INFO  ]
[ INFO  ] [Extended Features]
[ INFO  ] File Audit logging     : Disabled
[ INFO  ] Watch folder           : Disabled
[ INFO  ] Management GUI         : Enabled
[ INFO  ] Performance Monitoring : Enabled
[ INFO  ] Callhome               : Disabled
[ INFO  ]
[ INFO  ] GPFS                Admin  Quorum  Manager   NSD   Protocol   GUI    Perf Mon
EMS   OS   Arch
[ INFO  ] Node                Node   Node     Node   Server   Node    Server
Collector
[ INFO  ] ess.example.com     X       X                                X
X  rhel7  ppc64le
[ INFO  ]
node1.example.com                                                                rhel7
x86_64
[ INFO  ]
node2.example.com                                                                rhel7
x86_64
[ INFO  ]
node3.example.com                                                                rhel7
x86_64
[ INFO  ]
node4.example.com                                                                rhel7
x86_64
[ INFO  ]
node5.example.com                                                                rhel7
x86_64
[ INFO  ]
node6.example.com                                                                rhel7
x86_64
[ INFO  ] [Export IP address]
[ INFO  ] No export IP addresses configured
```

7. Do an installation precheck by using the installation toolkit.

```
# ./spectrumscale install -pr

[ INFO ] Logging to file: /usr/lpp/mmfs/5.x.x.x/installer/logs/INSTALL-PRECHECK-06-08-
2019_13:17:42.log
[ INFO ] Validating configuration
[ INFO ] Performing Chef (deploy tool) checks.
[ WARN ] No NSD servers specified. The install toolkit will continue without creating any
NSDs. If you still want to continue, please ignore this warning. Otherwise, for information
on adding a node as an NSD server, see:
'http://www.ibm.com/support/knowledgecenter/STXKQY_5.0.3/com.ibm.spectrum.scale.v
5r03.doc/bl1ins_configuringgpfs.htm'
[ INFO ] Checking for knife bootstrap configuration...
[ INFO ] Performing GPFS checks.
[ INFO ] Running environment checks
[ WARN ] No manager nodes specified. Assuming managers already configured on ESS.gpfs.net
…
[ INFO ] Checking pre-requisites for portability layer.
[ INFO ] GPFS precheck OK
[ INFO ] Performing Performance Monitoring checks.
[ INFO ] Running environment checks for Performance Monitoring
[ INFO ] Performing FILE AUDIT LOGGING checks.
[ INFO ] Running environment checks for file Audit logging
[ INFO ] Network check from admin node node1.example.com to all other nodes
in the cluster passed
[ WARN ] Ephemeral port range is not set. Please set valid ephemeral port range using the
command ./spectrumscale config gpfs --ephemeral_port_range . You may set the default
values as 60000-61000
[ INFO ] The install toolkit will not configure call home as it is disabled. To enable call
home,
use the following CLI command: ./spectrumscale callhome enable
[ INFO ] Pre-check successful for install.
[ INFO ] Tip : ./spectrumscale install
```

8. Install the nodes defined in the cluster definition by using the installation toolkit.

```
# ./spectrumscale install

[ INFO ] Logging to file: /usr/lpp/mmfs/5.x.x.x/installer/logs/INSTALL-06-08-
2019_13:18:29.log
[ INFO ] Validating configuration
[ WARN ] No NSD servers specified. The install toolkit will continue without creating any
NSDs. If you still want to continue, please ignore this warning. Otherwise, for information
on adding a node as an NSD server, see:
'http://www.ibm.com/support/knowledgecenter/STXKQY_5.0.3/com.ibm.spectrum.scale.v
5r03.doc/bl1ins_configuringgpfs.htm'
[ INFO ] Checking for knife bootstrap configuration...
[ INFO ] Running pre-install checks
[ INFO ] Running environment checks
[ INFO ] The following nodes will be added to cluster scalecluster.example.com: node1-
.example.com, node2.example.com, node3.example.com, node4.example.com, node5.example.com,
node6.example.com, ess.example.com,
[ WARN ] No manager nodes specified. Assuming managers already configured on
ESS.gpfs.net.
…
…
…
…
[ INFO ] Checking for a successful install
[ INFO ] Checking state of Chef (deploy tool)
[ INFO ] Chef (deploy tool) ACTIVE
[ INFO ] Checking state of GPFS
[ INFO ] GPFS callhome has been successfully installed. To configure callhome run
'mmcallhome -h' on one of your nodes.
[ INFO ] Checking state of GPFS on all nodes
[ INFO ] GPFS active on all nodes
[ INFO ] GPFS ACTIVE
[ INFO ] Checking state of Performance Monitoring
[ INFO ] Running Performance Monitoring post-install checks
[ WARN ] Historical performance data is still kept on: node1.example.com in the
'/opt/IBM/zimon/data' directory. For documentation on migrating the data to the new
Performance Monitoring collectors: refer to the IBM Spectrum Scale Knowledge Center.
[ INFO ] pmcollector running on all nodes
[ INFO ] pmsensors running on all nodes
[ INFO ] Performance Monitoring ACTIVE
[ INFO ] SUCCESS
[ INFO ] All services running
[ INFO ] StanzaFile and NodeDesc file for NSD, filesystem, and cluster setup have been
saved to /usr/lpp/mmfs folder on node: ess.example.com
[ INFO ] Installation successful. 7 GPFS nodes active in cluster scalecluster.example.com.
```

```
Completed in 6
minutes 6 seconds.
[ INFO ] Tip :If all node designations and any required protocol configurations are
complete,
proceed to check the deploy configuration:./spectrumscale deploy --precheck
```

9. Verify that the installation completed successfully by issuing the following command.

```
# ./spectrumscale install -po

[ INFO ] Logging to file: /usr/lpp/mmfs/5.x.x.x/installer/logs/INSTALL-POSTCHECK-06-08-
2019_13:25:31.log
[ WARN ] NTP is not set to be configured with the install toolkit.See './spectrumscale
config
ntp -h' to setup.
[ WARN ] No NSD servers specified. The install toolkit will continue without creating any
NSDs. If you still want to continue, please ignore this warning. Otherwise, for information
on adding a node as an NSD server, see:
'http://www.ibm.com/support/knowledgecenter/STXKQY_5.0.3/com.ibm.spectrum.scale.v
5r03.doc/bl1ins_configuringgpfs.htm'
[ INFO ] Checking state of Chef (deploy tool)
[ INFO ] Chef (deploy tool) ACTIVE
[ INFO ] Checking state of GPFS
[ INFO ] GPFS callhome has been successfully installed. To configure callhome run
'mmcallhome -h' on one of your nodes.
[ INFO ] Checking state of GPFS on all nodes
[ INFO ] GPFS active on all nodes
[ INFO ] GPFS ACTIVE
[ INFO ] Checking state of Performance Monitoring
[ INFO ] Running Performance Monitoring post-install checks
[ WARN ] Historical performance data is still kept on: ess.example.com in the
'/opt/IBM/zimon/data' directory. For documentation on migrating the data to the new
Performance Monitoring collectors: refer to the IBM Spectrum Scale Knowledge Center.
[ INFO ] pmcollector running on all nodes
[ INFO ] pmsensors running on all nodes
```

10. Deploy the defined configuration by using the installation toolkit.

```
# ./spectrumscale deploy

INFO ] Logging to file: /usr/lpp/mmfs/5.x.x.x/installer/logs/DEPLOY-06-08-
2019_13:26:24.log
[ INFO ] Validating configuration
[ INFO ] Checking for knife bootstrap configuration...
[ INFO ] Running pre-install checks
[ INFO ] Running environment checks for Performance Monitoring
[ INFO ] Running environment checks for file Audit logging
[ INFO ] Network check from admin node node1.example.com to all other nodes
in the cluster passed
[ WARN ] Ephemeral port range is not set. Please set valid ephemeral port range using the
command ./spectrumscale config gpfs --ephemeral_port_range . You may set the default
values as 60000-61000
[ INFO ] The install toolkit will not configure call home as it is disabled. To enable call
home,
use the following CLI command: ./spectrumscale callhome enable
[ INFO ] Preparing nodes for install
…
…
…
…
[ INFO ] Checking for a successful install
[ INFO ] Checking state of Chef (deploy tool)
[ INFO ] Chef (deploy tool) ACTIVE
[ INFO ] Checking state of Performance Monitoring
[ INFO ] Running Performance Monitoring post-install checks
[ WARN ] Historical performance data is still kept on: node1.example.com in the
'/opt/IBM/zimon/data' directory. For documentation on migrating the data to the new
Performance Monitoring collectors: refer to the IBM Spectrum Scale Knowledge Center.
[ INFO ] pmcollector running on all nodes
[ INFO ] pmsensors running on all nodes
[ INFO ] Performance Monitoring ACTIVE
[ INFO ] SUCCESS
[ INFO ] All services running
[ INFO ] StanzaFile and NodeDesc file for NSD, filesystem, and cluster setup have been
saved to /usr/lpp/mmfs folder on node: ess.example.com
[ INFO ] Successfully installed protocol packages on 0 protocol nodes. Components
installed: Chef (deploy tool), Performance Monitoring, FILE AUDIT LOGGING. it took 2
minutes and 25 seconds.
```

Verify the node details.

```
# ./spectrumscale node list

[ INFO  ] List of nodes in current configuration:
[ INFO  ] [Installer Node]
[ INFO  ] 198.51.100.1
[ INFO  ]
[ INFO  ] [Cluster Details]
[ INFO  ] Name: scalecluster.example.com
[ INFO  ] Setup Type: ESS
[ INFO  ]
[ INFO  ] [Extended Features]
[ INFO  ] File Audit logging     : Disabled
[ INFO  ] Watch folder           : Disabled
[ INFO  ] Management GUI         : Enabled
[ INFO  ] Performance Monitoring : Disabled
[ INFO  ] Callhome               : Disabled
[ INFO  ]
[ INFO  ] GPFS                Admin  Quorum  Manager   NSD   Protocol   GUI    Perf Mon
EMS   OS   Arch
[ INFO  ] Node                Node   Node     Node    Server   Node    Server
Collector
[ INFO  ] ess.example.com     X       X                                 X
X  rhel7  ppc64le
[ INFO  ]
node1.example.com                                                                rhel7
x86_64
[ INFO  ]
node2.example.com                                                                rhel7
x86_64
[ INFO  ]
node3.example.com                                                                rhel7
x86_64
[ INFO  ]
node4.example.com                                                                rhel7
x86_64
[ INFO  ]
node5.example.com                                                                rhel7
x86_64
[ INFO  ]
node6.example.com                                                                rhel7
x86_64
[ INFO  ] [Export IP address]
[ INFO  ] No export IP addresses configured
```

Preparing the IBM Spectrum Scale Erasure Code Edition cluster using the installation toolkit

In the 3rd phase of incorporating IBM Spectrum Scale Erasure Code Edition in an ESS cluster, use the installation toolkit to create a new cluster definition file that will be used to create an unconfigured Erasure Code Edition cluster within the ESS cluster.

1. From IBM FixCentral, download the IBM Spectrum Scale Erasure Code Edition 5.x.x.x installation package on your installer node.
2. Extract the IBM Spectrum Scale Erasure Code Edition 5.x.x.x installation package to a directory on the installer node that is different from the installer directory that you used for the initial installation and deployment in phase 2. For example, `/usr/lpp/mmfs/5.x.x.x_ECE_New/`.

   ```
   /DirectoryPathToDownloadedCode/Spectrum_Scale_Erasure_Code-5.x.x.x-x86_64-Linux-install
   --dir /usr/lpp/mmfs/5.x.x.x_ECE_New/
   ```

3. Change the directory to the new directory in which the package was extracted.

   ```
   # cd /usr/lpp/mmfs/5.x.x.x_ECE_New/
   ```

4. Change the setup type of the installer node to `ece`.

In this command example, 198.51.100.1 is the IP address of the scale-out node that is designated as the installer node.

```
# ./spectrumscale setup -s 198.51.100.1 -st ece

[ INFO ] Installing prerequisites for install node
[ INFO ] Existing Chef installation detected. Ensure the PATH is configured so that
chefclient
and knife commands can be run.
[ INFO ] Your control node has been configured to use the IP 198.51.100.1 to communicate
with other nodes.
[ INFO ] Port 8889 will be used for chef communication.
[ INFO ] Port 10080 will be used for package distribution.
[ INFO ] Install Toolkit setup type is set to ECE (Erasure Code Edition).
[ INFO ] SUCCESS
[ INFO ] Tip : Designate scale out, protocol and admin nodes in your environment to use
during install:./spectrumscale node add <node> -p -a -so
```

Verify the node details.

```
# ./spectrumscale node list

[ INFO ] List of nodes in current configuration:
[ INFO ] [Installer Node]
[ INFO ] 198.51.100.1
[ INFO ]
[ INFO ] [Cluster Details]
[ INFO ] No cluster name configured
[ INFO ] Setup Type: Spectrum Scale
[ INFO ]
[ INFO ] [Extended Features]
[ INFO ] File Audit logging : Disabled
[ INFO ] Watch folder : Disabled
[ INFO ] Management GUI : Disabled
[ INFO ] Performance Monitoring : Disabled
[ INFO ] Callhome : Enabled
[ INFO ]
[ INFO ] No nodes configured. Use 'spectrumscale node add' to add nodes.
[ INFO ] If a cluster already exists use 'spectrumscale config populate -N node_in_cluster'
to
sync toolkit with existing cluster.
```

5. Add the same IBM Spectrum Scale Erasure Code Edition candidate nodes and any other nodes that you added previously for functions such as file audit logging to the cluster.

```
# ./spectrumscale node add 198.51.100.1 -a -so

[ INFO ] Setting node1.example.com as an admin node.
[ INFO  ] Setting node1.example.com as a scale-out node.
[ INFO  ] Configuration updated.

# ./spectrumscale node add 198.51.100.2 -so

[ INFO  ] Setting node2.example.com as a scale-out node.
[ INFO  ] Configuration updated

# ./spectrumscale node add 198.51.100.3 -so

[ INFO  ] Setting node3.example.com as a scale-out node.
[ INFO  ] Configuration updated.

# ./spectrumscale node add 198.51.100.4 -so

[ INFO  ] Setting node4.example.com as a scale-out node.
[ INFO  ] Configuration updated.


# ./spectrumscale node add 198.51.100.5 -so

[ INFO  ] Setting node5.example.com as a scale-out node.
[ INFO  ] Configuration updated.

# ./spectrumscale node add 198.51.100.6 -so

[ INFO  ] Setting node6.example.com as a scale-out node.
[ INFO  ] Configuration updated.
```

Verify the node details.

```
# ./spectrumscale node list

[ INFO ] List of nodes in current configuration:
[ INFO ] [Installer Node]
[ INFO ] 198.51.100.1
[ INFO ]
[ INFO ] [Cluster Details]
[ INFO ] No cluster name configured
[ INFO ] Setup Type: Erasure Code Edition
[ INFO ]
[ INFO ] [Extended Features]
[ INFO ] File Audit logging : Disabled
[ INFO ] Watch folder : Disabled
[ INFO ] Management GUI : Disabled
[ INFO ] Performance Monitoring : Disabled
[ INFO ] Callhome : Enabled
[ INFO ]
[ INFO  ] GPFS               Admin  Quorum  Manager   NSD   Protocol  Callhome   Scale-out
OS   Arch
[ INFO  ] Node                Node   Node     Node   Server   Node    Server
Node
[ INFO  ] node1.example.com    X                                                    X
rhel7  x86_64
[ INFO  ] node2.example.com                                                         X
rhel7  x86_64
[ INFO  ] node3.example.com                                                         X
rhel7  x86_64
[ INFO  ] node4.example.com                                                         X
rhel7  x86_64
[ INFO  ] node5.example.com                                                         X
rhel7  x86_64
[ INFO  ] node6.example.com                                                         X
rhel7  x86_64

[ INFO ]
[ INFO ] [Export IP address]
[ INFO ] No export IP addresses configured
```

6. Do a deployment precheck by using the installation toolkit.

```
# ./spectrumscale deploy --pr

[ INFO ] Logging to file: /usr/lpp/mmfs/5.x.x.x_ECE_New/installer/logs/DEPLOYPRECHECK-
06-08-2019_14:26:48.log
[ INFO ] Validating configuration
You have not defined any recovery group in the cluster configuration. Installer will
automatically define the quorum configuration. Do you want to continue [Y/n]: y
[ INFO ] Performing Chef (deploy tool) checks.
[ WARN ] No recoverygroup specified. The install toolkit will continue without creating any
recoverygroup. If you still want to continue, please ignore this warning. Otherwise, you can
use 'spectrumscale recoverygroup define' command to define recoverygroup configuration.
[ WARN ] Install toolkit will not reconfigure Performance Monitoring as it has been
disabled. See the IBM Spectrum Scale Knowledge center for documentation on manual
configuration.
[ WARN ] No GUI servers specified. The GUI will not be configured on any nodes.
[ INFO ] Install toolkit will not configure file audit logging as it has been disabled.
[ INFO ] Install toolkit will not configure watch folder as it has been disabled.
[ INFO ] Checking for knife bootstrap configuration...
[ INFO ] Performing FILE AUDIT LOGGING checks.
[ INFO ] Running environment checks for file Audit logging
[ INFO ] Performing Erasure Code checks.
[ INFO ] Running environment checks for Erasure Code Edition.
[ WARN ] No quorum nodes are configured. The Install Toolkit will assign quorum nodes.
[ INFO ] Erasure Code Edition precheck OK
[ WARN ] Ephemeral port range is not set. Please set valid ephemeral port range using the
command ./spectrumscale config gpfs --ephemeral_port_range . You may set the default
values as 60000-61000
[ INFO ] The install toolkit will not configure call home as it is disabled. To enable call
home,
use the following CLI command: ./spectrumscale callhome enable
[ INFO ] Pre-check successful for deploy.
[ INFO ] Tip : ./spectrumscale deploy
```

7. Deploy the defined IBM Spectrum Scale Erasure Code Edition configuration by using the installation toolkit.

```
# ./spectrumscale deploy

[ INFO ] Logging to file: /usr/lpp/mmfs/5.x.x.x_ECE_New/installer/logs/DEPLOY-06-08-
2019_15:04:27.log
[ INFO ] Validating configuration
[ WARN ] No recoverygroup specified. The install toolkit will continue without creating any
recoverygroup. If you still want to continue, please ignore this warning. Otherwise, you can
use 'spectrumscale recoverygroup define' command to define recoverygroup configuration.
[ WARN ] Install toolkit will not reconfigure Performance Monitoring as it has been
disabled. See the IBM Spectrum Scale Knowledge center for documentation on manual
configuration.
[ WARN ] No GUI servers specified. The GUI will not be configured on any nodes.
[ INFO ] Install toolkit will not configure file audit logging as it has been disabled.
[ INFO ] Install toolkit will not configure watch folder as it has been disabled.
[ INFO ] Checking for knife bootstrap configuration...
[ INFO ] Running pre-install checks
[ INFO ] Running environment checks for file Audit logging
[ INFO ] Running environment checks for Erasure Code Edition.
[ INFO ] Erasure Code Edition precheck OK
[ INFO ] Network check from admin node node1.example.com to all other
nodes in the cluster passed
[ WARN ] Ephemeral port range is not set. Please set valid ephemeral port range using the
command ./spectrumscale config gpfs --ephemeral_port_range . You may set the default
values as 60000-61000
[ INFO ] The install toolkit will not configure call home as it is disabled. To enable call
home,
use the following CLI command: ./spectrumscale callhome enable
[ INFO ] Preparing nodes for install
…
[ INFO ] Checking for a successful install
[ INFO ] Checking state of Chef (deploy tool)
[ INFO ] Chef (deploy tool) ACTIVE
[ INFO ] Checking state of Erasure Code
[ INFO ] Running Erasure Code Edition post-install checks
[ INFO ] Erasure Code ACTIVE
[ INFO ] SUCCESS
[ INFO ] All services running
[ INFO ] StanzaFile and NodeDesc file for NSD, filesystem, and cluster setup have been
saved to /usr/lpp/mmfs folder on node: node1.example.com
[ INFO ] Successfully installed protocol packages on 0 protocol nodes. Components
installed: Chef (deploy tool), FILE AUDIT LOGGING, Erasure Code. it took 1 minutes and 37
seconds.
```

You can verify that the deployment completed successfully by issuing the following command.

```
./spectrumscale deploy -po
```

Completing the IBM Spectrum Scale Erasure Code Edition configuration with mmvdisk commands

In the fourth phase of incorporating IBM Spectrum Scale Erasure Code Edition in an ESS cluster, use **mmvdisk** commands from any Erasure Code Edition mmvdisk enabled node in the cluster to complete the configuration of the IBM Spectrum Scale Erasure Code Edition cluster.

1. Create the Erasure Code Edition node class from the candidate scale-out nodes that you deployed earlier.

```
# mmvdisk nc create --node-class ece_nc1 -N node1,node2,node3,node4,node5,node6

mmvdisk: Node class 'ece_nc1' created.
```

2. Configure the Erasure Code Edition node class and restart GPFS.

```
# mmvdisk server configure --node-class ece_nc1 –recycle all

mmvdisk: Checking resources for specified nodes.
mmvdisk: Node 'node1' has a scale-out recovery group disk topology.
mmvdisk: Node 'node2' has a scale-out recovery group disk topology.
mmvdisk: Node 'node3' has a scale-out recovery group disk topology.
mmvdisk: Node 'node4' has a scale-out recovery group disk topology.
mmvdisk: Node 'node5' has a scale-out recovery group disk topology.
mmvdisk: Node 'node6' has a scale-out recovery group disk topology.
```

```
mmvdisk: Node class 'ece_nc1' has a scale-out recovery group disk topology.
mmvdisk: Setting configuration for node class 'nc1'.
mmvdisk: Node class 'ece_nc1' is now configured to be recovery group servers.
mmvdisk: Restarting GPFS on the following nodes:
mmvdisk:     node1.example.com
mmvdisk:     node2.example.com
mmvdisk:     node3.example.com
mmvdisk:     node4.example.com
mmvdisk:     node5.example.com
mmvdisk:     node6.example.com
```

Verify the node class details.

```
# mmvdisk nc list

node class recovery groups
---------- ----------------------
ece_nc1 -
ess_nc1 rg_gssio1-ib, rg_gssio2-ib
```

3. Configure and create the recovery group.

```
# mmvdisk rg create --rg ece_rg1 --nc ece_nc1

mmvdisk: Checking node class configuration.
mmvdisk: Checking daemon status on node 'node1.example.com'.
mmvdisk: Checking daemon status on node 'node4.example.com'.
mmvdisk: Checking daemon status on node 'node5.example.com'.
mmvdisk: Checking daemon status on node 'node6.example.com'.
mmvdisk: Checking daemon status on node 'node3.example.com'.
mmvdisk: Checking daemon status on node 'node2.example.com'.
mmvdisk: Analyzing disk topology for node 'node1.example.com'.
mmvdisk: Analyzing disk topology for node 'node4.example.com'.
mmvdisk: Analyzing disk topology for node 'node5.example.com'.
mmvdisk: Analyzing disk topology for node 'node6.example.com'.
mmvdisk: Analyzing disk topology for node 'node3.example.com'.
mmvdisk: Analyzing disk topology for node 'node2.example.com'.
mmvdisk: Creating recovery group 'ece_rg1'.
mmvdisk: Formatting log vdisks for recovery group.
mmvdisk: (mmcrvdisk) [I] Processing vdisk RG003ROOTLOGHOME
mmvdisk: (mmcrvdisk) [I] Processing vdisk RG003LG001LOGHOME
mmvdisk: (mmcrvdisk) [I] Processing vdisk RG003LG002LOGHOME
mmvdisk: (mmcrvdisk) [I] Processing vdisk RG003LG003LOGHOME
mmvdisk: (mmcrvdisk) [I] Processing vdisk RG003LG004LOGHOME
mmvdisk: (mmcrvdisk) [I] Processing vdisk RG003LG005LOGHOME
mmvdisk: (mmcrvdisk) [I] Processing vdisk RG003LG006LOGHOME
mmvdisk: (mmcrvdisk) [I] Processing vdisk RG003LG007LOGHOME
mmvdisk: (mmcrvdisk) [I] Processing vdisk RG003LG008LOGHOME
mmvdisk: (mmcrvdisk) [I] Processing vdisk RG003LG009LOGHOME
mmvdisk: (mmcrvdisk) [I] Processing vdisk RG003LG010LOGHOME
mmvdisk: (mmcrvdisk) [I] Processing vdisk RG003LG011LOGHOME
mmvdisk: (mmcrvdisk) [I] Processing vdisk RG003LG012LOGHOME
mmvdisk: Created recovery group 'ece_rg1'.
```

Verify the recovery group details.

```
# mmvdisk rg list


                                                 needs   user
recovery group active current or master server service vdisks remarks
-------------- ------ ----------------------- ------- ------ ------
ece_rg1         yes    node1.example.com       no       0
rg_gssio1-ib    yes    gssio1-ib.example.com    no       1
rg_gssio2-ib    yes    gssio2-ib.example.com    no       1
```

4. Define the vdisk set(s) with the desired parameters.

   - In this command example, the Erasure Code Edition vdisk set is defined as a dataOnly storage pool that is separate from the existing ESS pool. The ESS pool in this case is the system pool and it is defined as dataAndMetadata.

- Make sure you use the same block size (16M in this case) as the existing ESS file system if you are merging this vdisk set into that file system.

```
# mmvdisk vs define --vs ece_vs1 --rg ece_rg1 --code 8+2p --block-size 16M
  --set-size 80% --storage-pool ece_pool_1 --nsd-usage dataOnly

mmvdisk: Vdisk set 'ece_vs1' has been defined.
mmvdisk: Recovery group 'ece_rg1' has been defined in vdisk set 'ece_vs1'.

                    member vdisks
vdisk set       count   size   raw size  created  file system and attributes
--------------  -----  -------- --------  -------  -------------------------
ece_vs1          12   62 GiB   80 GiB    no       -, DA1, 8+2p, 16 MiB, dataOnly, ece_pool_1

                declustered                 capacity           all vdisk sets defined
recovery group     array     type  total raw  free raw  free%  in the declustered array
--------------  -----------  ----  ---------  --------  -----  ------------------------
ece_rg1             DA1      HDD    1213 GiB   253 GiB    20%  ece_vs1

              vdisk set map       memory per server
node class  available  required  required per vdisk set
----------  ---------  --------  ----------------------
ece_nc1      8996 MiB   390 MiB  ece_vs1 (2304 KiB)
```

5. Create vdisks, NSDs, and the vdisk set from the defined storage.

```
# mmvdisk vs create --vs ece_vs1

mmvdisk: 12 vdisks and 12 NSDs will be created in vdisk set 'ece_vs1'.
mmvdisk: (mmcrvdisk) [I] Processing vdisk RG003LG001VS003
mmvdisk: (mmcrvdisk) [I] Processing vdisk RG003LG002VS003
mmvdisk: (mmcrvdisk) [I] Processing vdisk RG003LG003VS003
mmvdisk: (mmcrvdisk) [I] Processing vdisk RG003LG004VS003
mmvdisk: (mmcrvdisk) [I] Processing vdisk RG003LG005VS003
mmvdisk: (mmcrvdisk) [I] Processing vdisk RG003LG006VS003
mmvdisk: (mmcrvdisk) [I] Processing vdisk RG003LG007VS003
mmvdisk: (mmcrvdisk) [I] Processing vdisk RG003LG008VS003
mmvdisk: (mmcrvdisk) [I] Processing vdisk RG003LG009VS003
mmvdisk: (mmcrvdisk) [I] Processing vdisk RG003LG010VS003
mmvdisk: (mmcrvdisk) [I] Processing vdisk RG003LG011VS003
mmvdisk: (mmcrvdisk) [I] Processing vdisk RG003LG012VS003
mmvdisk: Created all vdisks in vdisk set 'ece_vs1'.
mmvdisk: (mmcrnsd) Processing disk RG003LG001VS003
mmvdisk: (mmcrnsd) Processing disk RG003LG002VS003
mmvdisk: (mmcrnsd) Processing disk RG003LG003VS003
mmvdisk: (mmcrnsd) Processing disk RG003LG004VS003
mmvdisk: (mmcrnsd) Processing disk RG003LG005VS003
mmvdisk: (mmcrnsd) Processing disk RG003LG006VS003
mmvdisk: (mmcrnsd) Processing disk RG003LG007VS003
mmvdisk: (mmcrnsd) Processing disk RG003LG008VS003
mmvdisk: (mmcrnsd) Processing disk RG003LG009VS003
mmvdisk: (mmcrnsd) Processing disk RG003LG010VS003
mmvdisk: (mmcrnsd) Processing disk RG003LG011VS003
mmvdisk: (mmcrnsd) Processing disk RG003LG012VS003
mmvdisk: Created all NSDs in vdisk set 'ece_vs1'.
```

6. From any mmvdisk enabled node in the cluster, add the new vdisk set to the existing file system.

```
# mmvdisk fs add --fs ecefs1 --vs ece_vs1

mmvdisk: Creating file system 'ecefs1'.
mmvdisk: The following disks of ecefs1 will be formatted on node gssio2.example.com:
mmvdisk: RG003LG001VS003: size 64000 MB
mmvdisk: RG003LG002VS003: size 64000 MB
mmvdisk: RG003LG003VS003: size 64000 MB
mmvdisk: RG003LG004VS003: size 64000 MB
mmvdisk: RG003LG005VS003: size 64000 MB
mmvdisk: RG003LG006VS003: size 64000 MB
mmvdisk: RG003LG007VS003: size 64000 MB
mmvdisk: RG003LG008VS003: size 64000 MB
mmvdisk: RG003LG009VS003: size 64000 MB
mmvdisk: RG003LG010VS003: size 64000 MB
mmvdisk: RG003LG011VS003: size 64000 MB
mmvdisk: RG003LG012VS003: size 64000 MB
mmvdisk: Extending Allocation Map
mmvdisk: Creating Allocation Map for storage pool ece_pool_1
mmvdisk: Flushing Allocation Map for storage pool ece_pool_1
mmvdisk: Disks up to size 966.97 GB can be added to storage pool ece_pool_1.
```

```
mmvdisk: Checking Allocation Map for storage pool ece_pool_1
mmvdisk: Completed adding disks to file system ecefs1.
```

7. Verify the following entities from any mmvdisk enabled node.
   - File system details:

```
# mmvdisk fs list


file system     vdisk sets
-----------     ----------
ecefs1          VS001_essFS, VS002_essFS, ece_vs1
```

     Storage pools in the file system

```
# mmlspool ecefs1

Storage pools in file system at '/gpfs/ecefs1':
Name        Id     BlkSize Data Meta Total Data in (KB) Free Data in (KB) Total Meta in
(KB) Free Meta in (KB)
system      0      16 MB   yes yes 12501204992 12496994304 (100%) 12501204992 12497076224
(100%)
ece_pool_1  65537  16 MB   yes no  786432000 785252352 (100%) 0
0 ( 0%)
```

   - Recovery groups:

```
# mmvdisk rg list

                                                 needs    user
recovery group  active   current or master server           service  vdisks  remarks
--------------  -------  --------------------------------  -------  ------  -------
ece_rg1         yes      node1.example.com                   no           12
rg_gssio1-ib    yes      gssio1-ib.example.com               no            1
rg_gssio2-ib    yes      gssio2-ib.example.com               no            1
```

   - pdisks for the new recovery group ece_rg1:

```
# mmvdisk pdisk list --rg ece_rg1

                    declustered
recovery group  pdisk            array     paths  capacity  free space  FRU (type)
state
--------------  ------------  -----------  -----  --------  ----------  ---------------
-----
ece_rg1         n013p001      DA1          1      136 GiB   44 GiB      42D0623
ok
ece_rg1         n013p002      DA1          1      136 GiB   44 GiB      42D0422
ok
ece_rg1         n014p001      DA1          1      136 GiB   44 GiB      42D0623
ok
ece_rg1         n014p002      DA1          1      136 GiB   44 GiB      42D0422
ok
ece_rg1         n015p001      DA1          1      136 GiB   44 GiB      42D0623
ok
ece_rg1         n015p002      DA1          1      136 GiB   44 GiB      42D0422
ok
ece_rg1         n016p001      DA1          1      136 GiB   44 GiB      42D0623
ok
ece_rg1         n016p002      DA1          1      136 GiB   44 GiB      42D0422
ok
ece_rg1         n017p001      DA1          1      136 GiB   44 GiB      42D0623
ok
ece_rg1         n017p002      DA1          1      136 GiB   44 GiB      42D0422
ok
ece_rg1         n018p001      DA1          1      136 GiB   44 GiB      22R6802
ok
ece_rg1         n018p002      DA1          1      136 GiB   44 GiB      42D0422
ok
```

Chapter 6. Creating an IBM Spectrum Scale Erasure Code Edition storage environment

This topic describes the procedure for creating IBM Spectrum Scale Erasure Code Edition storage environment for your use.

Cluster creation

This topic describes the procedure for creating a IBM Spectrum Scale Erasure Code Edition cluster.

Install IBM Spectrum Scale Erasure Code Edition on all cluster nodes, and create an IBM Spectrum Scale cluster using either the IBM Spectrum Scale installation toolkit or manual procedures documented in the *Steps for establishing and starting your IBM Spectrum Scale cluster* topic in the *IBM Spectrum Scale: Concepts, Planning, and Installation Guide*.

Assign quorum nodes, cluster manager roles and other roles as described in the “Planning for node roles” on page 16.

Use the **mmnetverify connectivity all** option in the *mmnetverify* command in the *IBM Spectrum Scale: Command and Programming Reference* to ensure that your network is configured for use by IBM Spectrum Scale.

IBM Spectrum Scale Erasure Code Edition configurations

If you are using the installation toolkit, your initial recovery group, vdisksets and file systems are created. In that case, the commands shown here could be used to add additional recovery groups to your environment.

There are 6 steps to configuring IBM Spectrum Scale Erasure Code Edition. For details on each command and the supported arguments. see the *mmvdisk* topic in the *IBM Spectrum Scale RAID: Administration*.

1. Create a node class that contains a set of identical storage servers that belong to a single recovery group. There should be a minimum of 4 nodes and maximum of 32 nodes in a recovery group:

   ```
   mmvdisk nc create --nc <nodeclass-name>  -N <node-list>
   ```

2. To maintain quorum availability in the IBM Spectrum Scale cluster, exercise caution when you recycle nodes. The example below uses "--recycle one" so that nodes are recycled one at a time.

   ```
   mmvdisk server configure --nc <nodeclass-name>  --recycle one
   ```

3. Create a recovery group:

   ```
   mmvdisk rg create --rg <rg-name>  --nc <nodeclass-name>
   ```

4. Define one or more vdisk sets:

   ```
   mmvdisk vs define --vdisk-set <vs-name>  --rg <rg-name> --code <erasure-code>
    --block-size <bsize> --set-size <set-size>
   ```

5. Create the vdisk sets that you defined:

   ```
   mmvdisk vs create --vs <vs-name>
   ```

6. Create and mount the file system:

```
mmvdisk filesystem create --file-system <fs-name> --vs <vs-name>
mmmount <fs-name> -N <nodes-to-mount-on>
```

Chapter 7. Upgrading IBM Spectrum Scale Erasure Code Edition

You can upgrade to a newer available version of IBM Spectrum Scale Erasure Code Edition by using the installation toolkit or by using manual steps. In this release, only offline upgrade by using the installation toolkit is supported. For online upgrade, you must use the manual procedure. If you plan to use the installation toolkit for the upgrade, you must designate all nodes in the cluster as offline in the upgrade configuration.

Use one of the following available upgrade options depending on your requirements.

- Use the installation toolkit to do an offline upgrade of your IBM Spectrum Scale Erasure Code Edition cluster. For more information, see "Offline upgrade of IBM Spectrum Scale Erasure Code Edition by using the installation toolkit" on page 49.
- Use the manual procedure to do an online upgrade of your IBM Spectrum Scale Erasure Code Edition cluster. For more information, see "Manual online upgrade of IBM Spectrum Scale Erasure Code Edition" on page 50.

Offline upgrade of IBM Spectrum Scale Erasure Code Edition by using the installation toolkit

You can upgrade to a newer available version of IBM Spectrum Scale Erasure Code Edition by using the installation toolkit. In this release, you can use the installation toolkit only for offline upgrade.

1. Download the IBM Spectrum Scale Erasure Code Edition self-extracting package from the IBM Spectrum Scale page on Fix Central.

   The name of the IBM Spectrum Scale Erasure Code Edition self-extracting installation package is similar to `Spectrum_Scale_Erasure_Code-5.0.y.z-x86_64-Linux-install`.
2. Extract the installation package.

   The installation toolkit gets extracted to the `/usr/lpp/mmfs/5.0.y.z/installer/` directory.

   To verify that the extracted package is of IBM Spectrum Scale Erasure Code Edition, go to the `/usr/lpp/mmfs/5.0.y.z/gpfs_rpms` directory and check for `gpfs.gnr*` packages.
3. Change the directory to where the installation toolkit is extracted.

   ```
   cd /usr/lpp/mmfs/5.0.y.z/installer/
   ```
4. Specify the installer node and the setup type in the cluster definition file.

   The setup type must be `ece` for IBM Spectrum Scale Erasure Code Edition.

   ```
   ./spectrumscale setup -s InstallerNodeIP -st ece
   ```
5. Run the config populate command to populate the cluster definition file with the current cluster configuration.

   ```
   ./spectrumscale config populate -N ScaleOutNodeIP
   ```

   Note: If the config populate command does not work, you can still use the installation toolkit to populate the cluster configuration by using manual commands such as **`./spectrumscale node add`**.
6. Stop the workloads that are running on the nodes that you are upgrading.
7. If there are protocol nodes in the cluster, suspend Cluster Export Services (CES) on the protocol nodes and stop protocol services.

```
mmces node suspend -N ProtocolNodeList --stop
```

ProtocolNodeList is a list of all protocol nodes in the cluster.

8. Shut down GPFS on all nodes in the cluster.

```
mmshutdown -a
```

9. Designate all nodes in the cluster as offline in the installation toolkit upgrade configuration.

```
./spectrumscale upgrade config offline -N NodeList
```

NodeList is a list of all nodes in the cluster.

You can exclude nodes from the current upgrade process by using the following command:

```
./spectrumscale upgrade config exclude -N NodeName
```

Do the installation toolkit upgrade precheck and upgrade operations to upgrade the IBM Spectrum Scale Erasure Code Edition cluster after running the config populate command.

10. Do the installation toolkit upgrade precheck before the installation toolkit upgrade.

```
./spectrumscale upgrade precheck
```

11. Do the installation toolkit upgrade.

```
./spectrumscale upgrade run
```

You can access the installation toolkit upgrade logs from the `/usr/lpp/mmfs/5.0.x.x/installer/logs` directory.

12. If you are using customized udev rules on your storage nodes, you need to reapply those changes to the new udev rules. The previous rules are saved in the `/etc/udev/rules.d/` directory as part of the upgrade. After applying your changes, activate the changes with the **udevadm** command.
13. Start GPFS on all nodes in the cluster.

```
mmstartup -a
```

14. If there are protocol nodes in the cluster, resume CES on the protocol nodes and start protocol services.

```
mmces node resume -N ProtocolNodeList --start
```

ProtocolNodeList is a list of all protocol nodes in the cluster.

15. After the upgrade process is done, complete the upgrade to the new code level to take advantage of the new functions. For more information, see *Completing the upgrade to a new level of IBM Spectrum Scale* in *IBM Spectrum Scale: Concepts, Planning, and Installation Guide*.

After the upgrade is completed, you can use the installation toolkit for tasks such as adding new nodes, adding NSDs, creating more file systems, adding management GUI nodes, and adding protocol nodes. For more information, see *Performing additional tasks using the installation toolkit* in *IBM Spectrum Scale: Concepts, Planning, and Installation Guide*.

Manual online upgrade of IBM Spectrum Scale Erasure Code Edition

You can upgrade to a newer available version of IBM Spectrum Scale Erasure Code Edition by using the manual online upgrade procedure. In this release, online upgrade is supported only by using this manual procedure.

Before you begin:

- The versions of firmware drivers and operating system on each node must meet the requirements of the IBM Spectrum Scale Erasure Code Edition version that you are planning to upgrade to. For upgrading firmware or OS on the nodes, see the respective vendor documentation.
- It is recommended to plan the upgrade when the IBM Spectrum Scale Erasure Code Edition cluster is running a light workload.
- IBM Spectrum Scale Erasure Code Edition allows nodes in mixed old and new versions in the cluster. The administrator can divide the whole upgrade plan into several upgrade windows.

About fault tolerance: Fault tolerance is important in the entire upgrade progress and the administrator must monitor it from the beginning to the end of the upgrade. Fault tolerance can get affected due to an offline node or due to a pdisk failure and it automatically recovers after failures are repaired. The administrator must check the fault tolerance in each node during the upgrade because a node is offline from the IBM Spectrum Scale Erasure Code Edition cluster when it is being upgraded. It is recommended to recover the fault tolerance to the best possible configuration at the beginning of upgrade of each node. For more information, see "Understanding IBM Spectrum Scale Erasure Code Edition fault tolerance" on page 5.

Special scenarios: If the cluster has multiple recovery groups, the administrator can speed up the upgrading process by upgrading multiple nodes in different recovery groups at one time.

Typically, node quorum is sufficient for upgrading. However, there are some scenarios when there might be a risk of losing quorum. For example, if there are three recovery groups in the IBM Spectrum Scale Erasure Code Edition cluster and if you upgrade one quorum node in each recovery group, it results in three offline quorum nodes at a time. The administrator must be aware of the risk of losing quorum during each upgrade process. For more information, see *Node quorum* in *IBM Spectrum Scale: Concepts, Planning, and Installation Guide*.

1. Prepare for upgrading as follows.
   a) Download the IBM Spectrum Scale Erasure Code Edition self-extracting package from the IBM Spectrum Scale page on Fix Central.

      The name of the IBM Spectrum Scale Erasure Code Edition self-extracting installation package is similar to `Spectrum_Scale_Erasure_Code-5.0.y.z-x86_64-Linux-install`.

   b) Extract the installation package.

      ```
      ./Spectrum_Scale_Erasure_Code-5.0.y.z-x86_64-Linux-install --silent --text-only
      ```

      If no directory is specified, the self-extracting package extracts the GPFS RPMs to the `/usr/lpp/mmfs/5.0.y.z/gpfs_rpms` directory. Copy the `gpfs_rpms` directory to all nodes that are to be upgraded in the IBM Spectrum Scale Erasure Code Edition cluster.

   c) Do a health check of all nodes in the cluster.

      ```
      mmdsh -N all 'mmhealth node show'
      mmhealth cluster show
      ```

      If IBM Spectrum Scale Erasure Code Edition nodes in the cluster are not healthy, resolve the issues before you proceed with the upgrade. If you are unable to resolve the issues, Contact IBM Spectrum Scale support to assess the upgrade risk.

   d) Save the initial IBM Spectrum Scale Erasure Code Edition cluster configuration before doing any changes.

      ```
      mmfsadm dump config > ./cluster_config_before_upgrade.txt
      ```

   e) Save the initial mount map.

      ```
      mmlsmount all -L > mount_map_before_upgrade.txt
      ```

      Customer might mount several file systems with the auto-mount method. However, it is possible that some auto-mount points are manually unmounted. Or, some new mount points are mounted by mistake but they are not mounted before even if they are in the auto-mount list.

During the upgrade progress, the mount configuration must remain unchanged before and after upgrading.

Note: The following steps describe an example scenario wherein one node of a single recovery group is being upgraded. These steps must be done iteratively until all scale out nodes in the cluster are upgraded.

2. Check the current state of the cluster to ensure that it is ready for the upgrade.
   a) Monitor the pdisk status.

   ```
   mmvdisk pdisk list --recovery-group all --not-ok
   ```

   The expected output for this command is `mmvdisk: All pdisks are ok.` If the output is not similar, repair all pdisk failures before proceeding with the upgrade. If you are unable to resolve the issues, contact IBM Spectrum Scale support to assess the upgrade risk.

   b) View the recovery group fault tolerance information.

   ```
   mmvdisk recoverygroup list --recovery-group rgName --fault-tolerance
   ```

   The recovery group on which the upgrade procedure is being run must have a minimum fault tolerance of at least one node + one pdisk failure. However, a fault tolerance of 2-node is recommended. The administrator must ensure that the recovery group has the best possible fault tolerance. If the minimum fault tolerance cannot be satisfied, over a significant part of the upgrade window, stop upgrading and contact IBM Spectrum Scale support.

Online upgrade allows running a light workload on the cluster during the upgrade window.

3. Suspend or stop workloads on the node that is being upgraded.
   a) If the node that is being upgraded is also running protocols in the cluster, suspend Cluster Export Services (CES) on the protocol node and stop protocol services.

   ```
   mmces node suspend -N nodeName --stop
   ```

   For information on upgrading protocol nodes, see *Upgrading IBM Spectrum Scale protocol nodes* in *IBM Spectrum Scale: Concepts, Planning, and Installation Guide*.

   b) Stop or migrate the workloads off the node if they are running on the locally mounted GPFS file system.

   c) Unmount all the file systems including user file systems and the CES shared root file system.

   ```
   mmumount Device -N nodeName
   ```

   You can use the following command to view the mount information.

   ```
   mmlsmount all -L
   ```

To minimize upgrade time, the administrator must avoid upgrading nodes that are assigned with important roles. The administrator must migrate roles to different nodes that are not going to be upgraded or that are already upgraded.

4. Change the node roles as follows.
   a) Migrate the cluster manager.
      1) View the current cluster manager.

      ```
      mmlsmgr -c
      ```

      2) If the node that is to be upgraded is listed as a cluster manager, migrate the cluster manager role to another node.

      ```
      mmchmgr -c nodeName
      ```

   b) Migrate the file system manager for all file systems.
      1) For each file system, run the following command to view the current file system manager.

```
mmlsmgr filesystemName
```

2) If the node that is to be upgraded is listed as a file system manager, migrate the file system manager role to another node.

```
mmchmgr filesystemName nodeName
```

c) Use one of the following sets of steps depending on the version that you are upgrading from.

Note: If you are running an IBM Spectrum Scale Erasure Code Edition version earlier than 5.0.4.3, use the **tsrecgroupserver** command to migrate log groups to other nodes before shutting down GPFS on the node. If you are running version 5.0.4.3 or later, use the **mmvdisk** command to suspend the node.

- If upgrading from a version earlier than 5.0.4.3, migrate the log groups.

  1) View the balanced log groups.

  ```
  mmvdisk server list --recovery-group rgName
  ```

  Typically, log groups are balanced across each scale out node.

  2) For the node that is to be upgraded, move all the log groups that are residing on it to different nodes as follows.

  a) View the log groups that are residing on the node.

  ```
  mmvdisk server list --recovery-group rgName
  ```

  b) Move these log groups to different nodes.

  ```
  tsrecgroupserver rgName -f -l loggroupName nodeName
  ```

  For example, the output of the list command generates the following output:

  ```
  1  node1               yes      serving rg01: root, LG002, LG006
  ```

  According to the output, run the following commands to move all 3 log groups to three different nodes in the same recovery group (node2, node3, and node4 must be in the same recovery group).

  ```
  # tsrecgroupserver rgName -f -l root node2
  # tsrecgroupserver rgName -f -l LG002 node3
  # tsrecgroupserver rgName -f -l LG006 node4
  ```

- If upgrading from version 5.0.4.3 or later, suspend the node.

  1) Use the **mmvdisk** command to suspend the node.

  ```
  mmvdisk rg change --recovery-group rgName --suspend -N nodeName --window minutes
  ```

  This command enables defer rebuild in the node upgrading time. You need to estimate the upgrade process duration for this node and specify the time in minutes with the `--window` option. You only need to run this command for the node that is being upgraded which is an I/O server in the IBM Spectrum Scale Erasure Code Edition cluster.

  Note: If you are suspending a node, you must resume that node while starting up GPFS after the upgrade.

5. Shut down GPFS on the node.

```
mmshutdown -N nodeName
```

6. Upgrade the IBM Spectrum Scale Erasure Code Edition software.

a) Change the directory to where the RPMs are located.

```
cd /usr/lpp/mmfs/5.0.y.z/gpfs_rpms
```

b) Upgrade the RPMs.

```
rpm -Uvh --force --nodeps gpfs.base*.rpm gpfs.gpl*.rpm gpfs.crypto*.rpm
gpfs.adv*.rpm gpfs.gskit*.rpm gpfs.msg*.rpm gpfs.gnr*.rpm
gpfs.docs*.rpm gpfs.license.*.rpm
```

c) Check the version to ensure that the updated version of RPMs is installed on the node.

```
rpm -qa | grep gpfs
```

d) Rebuild the GPFS portability layer (GPL).

```
mmbuildgpl
```

e) If you are using customized udev rules on your storage nodes, you need to reapply those changes to the new udev rules. The previous rules are saved in the `/etc/udev/rules.d/` directory as part of the upgrade. After applying your changes, activate the changes with the **udevadm** command.

7. Start GPFS on the node.

```
mmstartup -N nodeName
```

After the **mmstartup** command completes, you can use the **mmgetstate -a** command to check the status of all nodes in the cluster.

If you used **mmvdisk suspend** command to suspend this node earlier in the procedure, use the following command to resume the node to disable defer rebuild.

```
mmvdisk rg change --recovery-group rgName --resume -N nodeName
```

You can use the **mmgetstate -a** and the **mmlsrecoverygroup rgName -L --pdisk** commands to check the status of all nodes in the cluster.

8. Resume or restart workloads on the node that is being upgraded.

a) Mount the file systems that were unmounted earlier.

Remount all file systems according to the original mount map. If the file systems are set to auto mount, check if those file systems are mounted as saved in the `mount_map_before_upgrade.text` file in an earlier step.

b) If protocol services were stopped on the node, resume CES on the protocol node and start protocol services.

```
mmces node resume -N nodeName --start
```

c) If the workloads were earlier running on a locally mounted file system, restart or migrate the workload back on the upgraded node.

Repeat steps 3 - 8 on all nodes one by one until all scale out nodes in the cluster are upgraded to the new IBM Spectrum Scale Erasure Code Edition version.

9. After the upgrade process is done, complete the upgrade to the new code level to take advantage of the new functions. For more information, see *Completing the upgrade to a new level of IBM Spectrum Scale* in *IBM Spectrum Scale: Concepts, Planning, and Installation Guide*.

Chapter 8. Administering IBM Spectrum Scale Erasure Code Edition

Physical disk procedures

This topic describes the various procedures that you can perform for the maintenance of disks.

Perform the following steps:

- To identify the pdisk to be replaced within all recovery groups:

```
mmvdisk pdisk list --rg all --replace
```

The system displays the following output:

```
recovery group  pdisk         priority  FRU (type)       location
--------------  ------------  --------  ---------------  --------
rg_1            n005p003         12.95  00YK014          Enclosure J1005744 Drive 6
rg_1            n005p004         12.95  00YK014          Enclosure J1005744 Drive 7

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

Note:

– If you replace a pdisk not on this list, you risk data loss. If the number of drained disks is below the replacement threshold for its member declustered array, then those disks will not show up in the list.

– It is recommended to set your replacement threshold to 1. This has the effect of listing all drives that are safely replaceable.

- To set your replacement threshold to 1:

```
mmvdisk rg change --rg RgName --da DaName --replace-threshold 1
```

- To replace hot swappable disk devices:

1. Issue this command:

```
mmvdisk pdisk replace --prepare --recovery-group RgName --pdisk PdiskName
```

The system displays an output similar to this:

```
mmvdisk: Suspending pdisk n005p003 of RG rg_1 in location J1005744-6.
 <<< replace the specific pdisk/rg names with persudo names
mmvdisk: Location J1005744-6 is Enclosure J1005744 Drive 6.
mmvdisk: Carrier released.
mmvdisk:
mmvdisk:   - Remove carrier.
mmvdisk:   - Replace disk in location J1005744-6 with type '00YK014'.
mmvdisk:   - Reinsert carrier.
mmvdisk:   - Issue the following command:
mmvdisk:
mmvdisk:   mmvdisk pdisk replace --recovery-group rg_1 --pdisk 'n005p003'
```

2. Go to the node to replace a new disk for the pdisk according to the slot location.
3. Issue this command:

```
mmvdisk pdisk replace --recovery-group RgName --pdisk PdiskName
```

The system displays an output similar to this:

```
mmvdisk:
mmvdisk: mmchcarrier : [I] Preparing a new pdisk for use may take many minutes.
mmvdisk:
mmvdisk: The following pdisks will be formatted on node HostName:
mmvdisk:     // HostName /dev/DevName
mmvdisk: Pdisk PdiskName of RG RgName successfully replaced.
mmvdisk: Resuming pdisk PdiskName#nnn of RG RgName.
mmvdisk: Carrier resumed.
```

Note: After adding a pdisk, ensure to check and disable the volatile write cache on the new pdisk. For more information, see "Volatile write cache detection" on page 62.

Virtual disk procedures

The **mmvdisk** command can be used to manage the IBM Spectrum Scale Erasure Code Edition storage. There are commands for listing individual or groups of virtual disks (`vdisks`), and for defining, creating and deleting groups of virtual disks (`vdisk sets`).

For more details, see the following topics:

- *mmvdisk* command in the *IBM Spectrum Scale RAID: Administration and Programming Reference (SA23-1354)*
- *mmvdisk vdisk* command in the *IBM Spectrum Scale RAID: Administration and Programming Reference (SA23-1354)*
- *mmvdisk vdiskset* command in the *IBM Spectrum Scale RAID: Administration and Programming Reference (SA23-1354)*

Node procedures

This topic describes various procedures that can be done on a node to accomplish various tasks.

When adding a new node or replacing a node, we need to prepare the following as the precondition for the new node:

- `edf` file for NVMe drive if there is: If the server is homogeneous with others including the drive mapping (which is what we recommend), the `edf` files (`/usr/lpp/mmfs/data/gems/*edf`) can be copied from the existing node to the new node. Otherwise, new `edf` files must be created. For more information, see "Setting up IBM Spectrum Scale Erasure Code Edition for NVMe" on page 29.
- Customer customized udev rules
- `systemctl` settings
- Run the precheck tools after preparing the node

Note: It is also recommended to run the precheck tool after preparing these preconditions.

Adding new I/O nodes

Adding a new node using the **mmvdisk** command:

1. Make sure the node is a member of the IBM Spectrum Scale cluster and the state is active (if not, issue **mmaddnode** and **mmstartup**). Also, make sure that the node has the server license (if not, run **mmchlicense**.)
2. Issue the **mmvdisk server list -N newnode --disk-topology** command to verify that this node has the same disk topology as the other nodes in the recovery group to which the node is added.
3. Issue the **mmvdisk server configure -N newnode --recycle one** command to configure it as IBM Spectrum Scale Erasure Code Edition server and restart the IBM Spectrum Scale daemon.
4. Issue the **mmvdisk rg add --rg rgname -N newnode** command to add this node to the current recovery group. After that, all DAs should be in rebalance state. In some cases, you may need to specify *--match* parameter if there are slight differences between your configuration and the standard

topology definitions, for example "--match 80". At this point, wait for all DAs to finish rebalance. While waiting, check DA status and rebalance progress by issuing this command:

```
mmvdisk recoverygroup list --recovery-group rg1 --declustered-array
```

The system displays a message similar to this:

```
Monitor rebalance is complete.
```

- Add new capacity to 1 or more `vdisk sets`:

  After all DAs finished rebalancing and are in scrub state, run this command to finish the node-add operation:

  ```
  mmvdisk recoverygroup add --recovery-group rg1 --complete-node-add
  ```

  This operation will create new log groups, create new vdisks for all existing vdisksets, create NSDs and add the free NSDs to file systems if the vdisk sets belong to some file system.

  Note: The add command will fail if you attempt to execute it before the rebalance completes. If that happens, continue monitoring state until the DAs are in scrub state and try again.

- Here are some examples and output for adding node to the current recovery group:
  - To verify the disk topology:

    ```
     mmvdisk server list  -N c72f4m5u09-ib0 --disk-topology -L
    GNR server: name c72f4m5u09-ib0 arch x86_64 model 7X06CTO1WW serial J1005746
    GNR enclosures found: internal
    Enclosure internal (internal, number 1):
    Enclosure internal sees 9 disks (6 SSDs, 3 HDDs)

    GNR server disk topology: C72 Mestor Cluster (match: 100/100)
    GNR configuration: 1 enclosure, 6 SSDs, 0 empty slots, 9 disks total, 0 NVRAM partitions
    ```

  - To add a node to the current recovery group:

    ```
    mmvdisk rg add --rg rg1 -N gpfstest10
    mmvdisk: Attempting to complete a previous add command.
    mmvdisk: Checking resources for specified nodes.
    mmvdisk: Analyzing disk topology for node 'gpfstest1'.
    mmvdisk: Analyzing disk topology for node 'gpfstest2'.
    mmvdisk: Analyzing disk topology for node 'gpfstest3'.
    mmvdisk: Analyzing disk topology for node 'gpfstest4'.
    mmvdisk: Analyzing disk topology for node 'gpfstest11'.
    mmvdisk: Analyzing disk topology for node 'gpfstest12'.
    mmvdisk: Analyzing disk topology for node 'gpfstest10'.
    mmvdisk: Updating server list for recovery group 'rg1'.
    mmvdisk: Updating pdisk list for recovery group 'rg1'.
    mmvdisk: The following pdisks will be formatted on node gpfstest1:
    mmvdisk:     //gpfstest10/dev/sdd
    mmvdisk:     //gpfstest10/dev/sdi
    mmvdisk:     //gpfstest10/dev/sdb
    mmvdisk:     //gpfstest10/dev/sdc
    mmvdisk:     //gpfstest10/dev/sdg
    mmvdisk:     //gpfstest10/dev/sdf
    mmvdisk:     //gpfstest10/dev/sde
    mmvdisk:     //gpfstest10/dev/sdh
    mmvdisk:     //gpfstest10/dev/sda
    mmvdisk: Updating parameters for declustered array 'DA1'.
    mmvdisk: Updating parameters for declustered array 'DA2'.
    mmvdisk: Node 'gpfstest10' added to recovery group 'rg1'.
    mmvdisk: Log group and vdisk set operations for recovery group 'rg1'
    mmvdisk: must be deferred until rebalance completes in all declustered arrays.
    mmvdisk: To monitor the progress of rebalance, use the command:
    mmvdisk:      mmvdisk recoverygroup list --recovery-group rg1 --declustered-array
    mmvdisk: When rebalance is completed, issue the command:
    mmvdisk:      mmvdisk recoverygroup add --recovery-group rg1 --complete-node-add
    ```

  - To verify the recovery group:

    ```
    mmvdisk recoverygroup list --recovery-group rg1 --declustered-array
    ```

```
declustered   needs    vdisks      pdisks     replace        capacity        scrub
   array     service  user log  total spare  threshold  total raw free raw  duration
background task
-----------  -------  ---- ---  ----- -----  ---------  --------- --------  --------
---------------
DA1          no         12  13     56     2          2     14 TiB   10 TiB   14 days
rebalance (0%)
DA2          no         24   0      7     0          1    620 GiB  253 GiB   14 days
rebalance (0%)

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

mmvdisk: Attention: Recovery group 'rg1' has an incomplete node addition (gpfstest10).
mmvdisk: Complete the node addition with the command:
mmvdisk:      mmvdisk recoverygroup add --recovery-group rg1 --complete-node-add
```

– To finish the node add operation:

```
 mmvdisk recoverygroup add --recovery-group rg1 --complete-node-add
mmvdisk: Verifying that the DAs in recovery group 'rg1' are idle.
mmvdisk: Updating log vdisks for recovery group 'rg1'.
mmvdisk: (mmcrvdisk) [I] Processing vdisk RG001LG013LOGHOME
mmvdisk: (mmcrvdisk) [I] Processing vdisk RG001LG014LOGHOME
mmvdisk: Updating vdisk sets for recovery group 'rg1.'
mmvdisk: 2 vdisks and 2 NSDs will be created in vdisk set 'vs_gpfs1'.
mmvdisk: 2 vdisks and 2 NSDs will be created in vdisk set 'vs_gpfs2'.
mmvdisk: 2 vdisks and 2 NSDs will be created in vdisk set 'vs_gpfs3'.
mmvdisk: (mmcrvdisk) [I] Processing vdisk RG001LG013VS001
mmvdisk: (mmcrvdisk) [I] Processing vdisk RG001LG014VS001
mmvdisk: (mmcrvdisk) [I] Processing vdisk RG001LG013VS002
mmvdisk: (mmcrvdisk) [I] Processing vdisk RG001LG014VS002
mmvdisk: (mmcrvdisk) [I] Processing vdisk RG001LG013VS003
mmvdisk: (mmcrvdisk) [I] Processing vdisk RG001LG014VS003
mmvdisk: Created all vdisks in vdisk set 'vs_gpfs1'.
mmvdisk: Created all vdisks in vdisk set 'vs_gpfs2'.
mmvdisk: Created all vdisks in vdisk set 'vs_gpfs3'.
mmvdisk: (mmcrnsd) Processing disk RG001LG013VS001
mmvdisk: (mmcrnsd) Processing disk RG001LG014VS001
mmvdisk: (mmcrnsd) Processing disk RG001LG013VS002
mmvdisk: (mmcrnsd) Processing disk RG001LG014VS002
mmvdisk: (mmcrnsd) Processing disk RG001LG013VS003
mmvdisk: (mmcrnsd) Processing disk RG001LG014VS003
mmvdisk: Created all NSDs in vdisk set 'vs_gpfs1'.
mmvdisk: Created all NSDs in vdisk set 'vs_gpfs2'.
mmvdisk: Created all NSDs in vdisk set 'vs_gpfs3'.
mmvdisk: Extending file system 'gpfs1'.
mmvdisk: The following disks of gpfs1 will be formatted on node gpfstest6:
mmvdisk:     RG001LG013VS001: size 208888 MB
mmvdisk:     RG001LG014VS001: size 208888 MB
mmvdisk: Extending Allocation Map
mmvdisk: Checking Allocation Map for storage pool system
mmvdisk: Completed adding disks to file system gpfs1.
mmvdisk: Extending file system 'gpfs2'.
mmvdisk: The following disks of gpfs2 will be formatted on node gpfstest6:
mmvdisk:     RG001LG013VS002: size 8692 MB
mmvdisk:     RG001LG014VS002: size 8692 MB
mmvdisk:     RG001LG013VS003: size 8692 MB
mmvdisk:     RG001LG014VS003: size 8692 MB
mmvdisk: Extending Allocation Map
mmvdisk: Checking Allocation Map for storage pool system
mmvdisk: Completed adding disks to file system gpfs2.
```

Replacing an I/O node with a new node and disks

In this scenario, a failed server is to be replaced with an entirely new server, including new drives.

1. Prepare a new node with the same disk topology as the node needs to be replaced. The server type, memory, disks, should all be the same.
2. Issue the **mmaddnode** command to add this node into the IBM Spectrum Scale, accept the license as the server, and issue the **mmstartup -N** command to bring up the IBM Spectrum Scale daemon.
3. Define the node as the same role as the old server, such as quorum, fsmgr, and so on.
4. Run the **mmvdisk server configure -N nodename** command to configure the node, then restart the daemon on this node.

5. Run the **mmvdisk rg replace** command to replace the existing node with a new node. In some cases, you may need to specify "--match" parameter if there are slight differences between your configuration and the standard topology definitions, for example "--match 90". For example,

```
mmvdisk rg replace --rg rg1 -N c72f4m5u01-ib0 --new-node c72f4m5u07-ib0
mmvdisk: Attempting to complete a previous replace command.
mmvdisk: Analyzing disk topology for node 'c72f4m5u01-ib0'.
mmvdisk: Analyzing disk topology for node 'c72f4m5u03-ib0'.
mmvdisk: Analyzing disk topology for node 'c72f4m5u05-ib0'.
mmvdisk: Analyzing disk topology for node 'c72f4m5u11-ib0'.
mmvdisk: Analyzing disk topology for node 'c72f4m5u09-ib0'.
mmvdisk: Analyzing disk topology for node 'c72f4m5u15-ib0'.
mmvdisk: Analyzing disk topology for node 'c72f4m5u13-ib0'.
mmvdisk: Analyzing disk topology for node 'c72f4m5u07-ib0'.
mmvdisk: Updating server list for recovery group 'rg1'.
mmvdisk: Updating pdisk list for recovery group 'rg1'.
mmvdisk: This could take a long time.
mmvdisk: The following pdisks will be formatted on node c72f4m5u01.gpfs.net:
mmvdisk:     //c72f4m5u07-ib0/dev/nvme1n1
mmvdisk:     //c72f4m5u07-ib0/dev/nvme0n1
mmvdisk:     //c72f4m5u07-ib0/dev/sda
mmvdisk:     //c72f4m5u07-ib0/dev/sdc
mmvdisk:     //c72f4m5u07-ib0/dev/sdb
mmvdisk:     //c72f4m5u07-ib0/dev/sde
mmvdisk:     //c72f4m5u07-ib0/dev/sdg
mmvdisk:     //c72f4m5u07-ib0/dev/sdf
mmvdisk:     //c72f4m5u07-ib0/dev/sdd
mmvdisk: Removing node 'c72f4m5u01-ib0' from node class 'r1'.
mmvdisk: Updating server list for recovery group 'rg1'.
```

6. Run the **mmvdisk rg list** command to make sure the new node joins the node class, and that all related pdisks work fine. Also make sure that the replaced node and the related pdisks are not in the RG anymore. Then wait for some time to make sure all DAs into scrub state.
7. Now we have replaced the node from RG successfully. Run **mmshutdown -N and mmdelnode -N** to delete the replaced node from the cluster (if we do not need the node in the cluster anymore).

Replacing broken I/O nodes with moving disks to new nodes

1. Make sure the node is totally broken, not pingable, or cannot be logged in. We can pull the network cable on this broken node if we can physically access the node.
2. Prepare a new node that is of the same hardware as that of the broken node, install the same OS on it, check the time to sync with all other nodes in the IBM Spectrum Scale Erasure Code Edition cluster, and then install the same IBM Spectrum Scale build on the new node.
3. Connect the new node to the switch, change the host name and IP address of the new node as that of the old node.
4. Pull the pdisks that the old node was using and insert them into the new node.
5. Make sure that all disks are visible on the new node and that none of the pdisks are broken. If the pdisks are broken, data in this disk never gets restored.
6. Make sure that the *ssh* and *scp* commands work on the new node. We should configure passwordless ssh and scp for root users.
7. Make sure that *ssh/scp* works between ALL nodes and the new node.
8. Issue the **mmsdrrestore -p <node name> -R /usr/bin/scp** command on the new node, where <node name > is one of the active nodes in the node class.

Manually online upgrade IBM Spectrum Scale Erasure Code Edition software

To upgrade IBM Spectrum Scale software on IBM Spectrum Scale Erasure Code Edition storage nodes, one node at a time, do the following steps:

1. Precheck: Before any node upgrade, we should double-check the ability of this cluster to tolerate this node failure. For IBM Spectrum Scale Erasure Code Edition, you can run **mmvdisk rg list** to

confirm we have at least 1-node fault tolerance and all the DAs in this recovery group are in scrub state.

```
mmvdisk rg list --rg rg1 –-fault-tolerance
```

2. Move out all "server functions" off this node to another active node. For example, fsmgr, ccmgr, LG server, etc. This can be done by **mmchmgr** command.
3. Move out all serving log groups (LGs) to another active node, this could be done by **tschrecgroup**. Then run the **mmshutdown** command to shut down the IBM Spectrum Scale service on the node. For details to simplify/expedite fail-overs, see the *Migrating resources off a IBM Spectrum Scale Erasure Code Edition node*.
4. Update the IBM Spectrum Scale packages, then make GPL using the **mmbuildgpl** command.
5. Run the **mmstartup** command to start the IBM Spectrum Scale daemon on this node. Wait for some period of time, approximately 3 minutes, and then perform the following check to make sure the DA is in scrub state, and all pdisks are in OK state and all the LGs are balanced in the node class.

```
# mmvdisk rg list –-rg rg1 –-da --lg
# mmvdisk pdisk list –-rg rg1 –-not-ok
```

6. Repeat the steps 1 to 5 on all the other nodes.

Manually online upgrade OS/driver for IBM Spectrum Scale Erasure Code Edition node

1. Precheck: Before any node upgrade, we should double check the ability of this cluster to tolerate this node failure, since the IBM Spectrum Scale update needs to stop the service on this node. We must also check the IBM Spectrum Scale level. For IBM Spectrum Scale Erasure Code Edition, we can run **mmvdisk RG list** to confirm if we have at least 1 node fault tolerance and all the DAs in this RG are in scrub state.
2. Move out all "server functions" off this node to another active node. For example, fsmgr, ccmgr, LG server, etc. this can be done by **mmchmgr** command.
3. Move out all serving LGs to another active node. This could be done by **tschrecgroup**. Then issue the **mmshutdown** command to shut down the IBM Spectrum Scale service on the node.
4. Upgrade the node from old OS level to new OS level (yum update). After that, double check if the network driver needs upgrade also. (We see that when upgrading RH7.5 to RH7.6, IB driver for Mellanox also needs to be reinstalled.). Then make GPL using the **mmbuildgpl** command.
5. Run the **mmstartup** command to start the IBM Spectrum Scale daemon on this node. Wait for about 3 minutes, and then perform the following check:

   All pdisks are in OK state and all the LGs are balanced in the node class.
6. Repeat Step 1 to Step 5 on all the other nodes.

To migrate resources off IBM Spectrum Scale Erasure Code Edition node to simplify/expedite fail-overs

It is recommended that you use **tsrecgroupserver** to move an LG off node prior to shut down. It is the background mechanism used in the **mmchrecoverygroup RecoveryGroupName --active ServerName** name. By doing this, we reduce the time of node failure detection and lease recovery wait. The failover time is only the LG recovery time, and during LG recovery, we will not see pdisk missing for this node.

Firmware updates

In the IBM Spectrum Scale Erasure Code Edition, it is the customer's responsibility to ensure that the firmware and operating system software are kept current. The procedures below are meant as a model, but exact procedures may vary depending on your hardware configuration.

- **HBA firmware update**:

1. Make sure that the RG have at least 1-node fault tolerance and all DAs are in scrub state. Move out all LGs that are served by this node, and then move out all other server functions such as fsmgr, ccmgr, and so on. For more information, see the *mmchmgr command* in the *IBM Spectrum Scale: Command and Programming Reference*.

   Here is an example on how to move out LGs, which will be taken over by another node:

```
mmvdisk server list --nc r1
node  server    active   remarks
number
 2  gpfstest1    yes      serving rg1: LG002, LG003
 9  gpfstest10   yes      serving rg1: LG006, LG013
 6  gpfstest11   yes      serving rg1: LG005, LG012
 7  gpfstest12   yes      serving rg1: root, LG007, LG014
 3  gpfstest2    yes      serving rg1: LG001, LG009
 4  gpfstest3    yes      serving rg1: LG008, LG010
 5  gpfstest4    yes      serving rg1: LG004, LG011

[root@gpfstest2]# tsrecgroupserver  rg1 -f -l LG001 gpfstest1
Node (192.168.10.1 (gpfstest1)) was asked to take over log group (LG001). (err 0)
[root@gpfstest2]# tsrecgroupserver  rg1 -f -l LG009 gpfstest4
Node (192.168.10.4 (gpfstest4)) was asked to take over log group (LG009). (err 0)
[root@gpfstest2]# mmvdisk server list --nc r1
```

2. Issue the **mmshutdown** command on the node where we update the HBA firmware.
3. Start to upgrade firmware for HBA
4. Start to upgrade firmware for HBA.
5. After the node boot-up, issue the **mmstartup -N** command to start IBM Spectrum Scale on this node.
6. After IBM Spectrum Scale has started up and the node is serving LGs, issue this command to verify that all LGs are balanced well:

```
mmvdisk server list --nc nodeclass
```

- **To update disk firmware on one node**:
  1. Check the firmware level using **mmlsfirmware** command.
  2. Suspend the drives using the **mmvdisk pdisk change --suspend** command.
  3. Update the drive firmware.
  4. Resume the drives using **mmvdisk pdisk change --resume** command.
  5. Check the firmware level to verify the update:

     The procedure depends on the way that disk firmware gets upgraded. There are 2 kinds of upgrades:

     – Case 1: Upgrade disk firmware one by one

       If you have a firmware image, which allows you to upgrade disk firmware one by one:

       a. Issue the **mmvdisk pdisk change --rg rgname --pdisk pdiskname –suspend** command to suspend one pdisk.
       b. Issue the external tool to update disk firmware.
       c. Run the **mmvdisk pdisk change --rg rgname --pdisk pdiskname –resume** to resume the pdisk.
       d. Repeat steps a to step c to make sure all pdisks firmware gets updated.
       e. On the RG master node, issue the **tschrecgroup --rg ALL --path-discovery enable** command to trigger GNR load new firmware level for all pdisks.

     – Case 2: Upgrade disk firmware in batch

       If the firmware upgrade tool only supports update all pdisk firmware together instead of upgrading firmware one by one, we need to take the node out of service, run the tools, and then bring the node back to service.

a. Make sure IBM Spectrum Scale Erasure Code Edition has at least 1-node fault tolerance and all DAs are in scrub state, then move the LGs served by the node where the disks need to be updated. Issue the **tsrecgroupserver rg -f -l LGname newnode**. See the steps in HBA firmware update.

b. Issue **mmchmgr** to move out all the server services to another node. For example, if this node is fsmgr, run **mmchmgr fsname –newnode**.

c. Issue the **mmshutdown** command to shut down the IBM Spectrum Scale daemon on this node.

d. Issue the external tool to update the disk firmware. Sometimes node reboot is required, this depends on disks and the tools that are used.

e. After all the disk firmware is updated, issue the **mmstartup** command to start IBM Spectrum Scale daemon on this node.

f. On RG master node, issue the **tschrecgroup --rg ALL --path-discovery enable** command to trigger IBM Spectrum Scale RAID load new firmware level for all pdisk.

g. Issue the **mmvdisk server list --nc nodeclass** command to verify all the LGs are balanced well.

h. Issue the **mmvdisk pdisk list –L** command to verify all the pdisk firmware was updated as expected.

Volatile write cache detection

IBM Spectrum Scale Erasure Code Edition now has the ability to test if volatile write caching mode is enabled on the physical disks.

Many SCSI and NVMe drives support a volatile write caching mode in which a drive reports success back from write operations as soon as data has been received into the drive's internal cache memory. IBM Spectrum Scale Erasure Code Edition cannot be used with drives operating in this mode because on power failure, the cached data is lost, causing already committed data to revert to an older version. This can lead to corruption of both the RAID and file system metadata, resulting in data integrity issues. If IBM Spectrum Scale Erasure Code Edition detects a drive with volatile write caching mode enabled, it puts the pdisk into a new volatile write cache enabled (VWCE) state and drains all data from the drive. If IBM Spectrum Scale Erasure Code Edition detects a large number of drives with volatile write caching enabled, it stops service of the recovery group and waits for volatile write caching mode to be disabled on the drives.

The volatile write cache detection feature is enabled for all new IBM Spectrum Scale Erasure Code Edition installations starting from version 5.0.4. On previous installations, the feature is disabled by default and must be manually enabled in order to take advantage of the check.

Check IBM Spectrum Scale Erasure Code Edition cluster configuration for VWCE

IBM Spectrum Scale Erasure Code Edition supports volatile write cache detection from version 5.0.4, upgrade from previous versions need to enable it.

Use the following commands to check the current IBM Spectrum Scale Erasure Code Edition configuration for VWCE detection:

1. `# mmdiag --config|grep nsdRAIDDiskCheckVWCE`

   If nsdRAIDDiskCheckVWCE is 1, it means enabled. If nsdRAIDDiskCheckVWCE is 0, it means disabled. Check all physical disks volatile write cache state before enabling it.

2. After making sure that all disks have disabled volatile write cache, use this command to enable it:

   ```
   # mmchconfig nsdRAIDDiskCheckVWCE=yes -i
   ```

3. Rediscover the disk state with this command:

   ```
   # mmvdisk rg change --recovery-group rg_name --refresh-pdisk-info
   ```

Creation of recovery group will fail if volatile write cache mode is enabled on disk

Before you install IBM Spectrum Scale Erasure Code Edition and create a recovery group, run the `SpectrumScale_ECE_OS_READINESS` tool first, it will detect volatile write cache of disks and give you warning messages. When disks have volatile write cache mode enabled, creation of recovery group will fail with error messages in the `/var/adm/ras/mmfs.log.latest` file.

Failure of disk replacement

For replacing failure disk, check and disable volatile write cache mode for the new physical disk. If volatile write cache mode is not disabled, replace command will fail.

Scale out IBM Spectrum Scale Erasure Code Edition by adding new node

Run the `SpectrumScale_ECE_OS_READINESS` tool first on new node, and disable volatile write cache mode for each disks if needed before adding a node into an IBM Spectrum Scale Erasure Code Edition recovery group.

What to do if volatile write cache is detected

For instructions on how to disable volatile writer caching on SCSI and NVMe disks, see “Hardware checklist” on page 11.

Chapter 9. Troubleshooting

This topic describes the known issues and workarounds of IBM Spectrum Scale Erasure Code Edition.

Monitoring the overall health

This topic describes different methods to monitor and troubleshoot IBM Spectrum Scale Erasure Code Edition.

For monitoring:

- From GUI, see the *Monitoring system health using IBM Spectrum Scale GUI* topic in the *IBM Spectrum Scale: Administration Guide*.
- From command line, see the *Monitoring system health by using the mmhealth command* topic in the *IBM Spectrum Scale: Administration Guide*.
- For general IBM Spectrum Scale troubleshooting, see the Troubleshooting topic in the *IBM Spectrum Scale: Problem Determination Guide*.
- For IBM Spectrum Scale RAID troubleshooting best practices, see the *Best practices for troubleshooting* topic in the *IBM Spectrum Scale RAID: Administration*.

What to do if you see degraded performance over NSD protocol

This topic describes the issues relating to degraded performance over NSD protocol.

Compared degraded performance to what? Is there a repeatable test and a baseline to compare to? "It is slow" is not a valid measurable metric. You should have a baseline to compare to.

There are multiple tools to create that. The product includes `nsdperf`, but you can choose other available tools in the market such as `ior`, `iozone`, `bonnie++`.

Things to check:

- First and foremost, check the network end to end
- Review any changes done to either the clients or servers (sysctl, software updates, ...)
- Check OS resources on the client system (CPU, memory, swap in and out, ...)
- Check OS resources on the server system
- Look for **`mmhealth`** events
- Look for SMART events (if applicable)
- Reboot the client

If you still see degraded performance compared to your baseline with the repeatable test, it is time to gather some information and contact IBM, as follows:

- Generate an IBM Spectrum Scale snap on the IBM Spectrum Scale Erasure Code Edition cluster.
- Generate an IBM Spectrum Scale snap on the client cluster.

You can already contact IBM support with the above snaps. If you suspect any issues at the disks level, you should engage with the disk vendor tools. In addition, you could gather the following information and attach to the IBM case.

ICT (intercompletion time) data is a full I/O trace giving size, seek distance, LBA, queue depth at time of completion, overall response time of the I/O and the completion time of this I/O relative to the previous or relative to the start of the I/O, whichever is later, for each `pdisk` I/O request. Things to look for would be the distribution of the ICT times, comparison of the response time to ICT time, etc. and checking if

anomalies are specific to hardware domains or to particular ranges of time. This data can be very useful to IBM support to help determine many different types of issues.

When contacting IBM support, compile the following data in addition to your baseline and the results that you obtain that differ from the baseline. Also include an overview of the environment and the tools as versions used to create the baseline:

- Gather ICT debug data:
  - Create a directory to host the debugs. You can use NFS or separate disk, as it can generate a fair amount of data. In our example, we are going to use `/tmp/mmfs/ict` but is just an example:

    ```
    # mkdir /tmp/mmfs/ict
    ```

  - Enable the gather of ICT data on the IBM Spectrum Scale Erasure Code Edition node:

    ```
    #mmchconfig nsdRAIDICTLogDir=/tmp/mmfs/ict,nsdRAIDDetailedICTLogging=all -N NODE i
    ```

  - Once you have recreated the performance degradation against the baseline, set the login back to default and tar the information to be sent to IBM:

    ```
    #  mmchconfig nsdRAIDICTLogDir=default,nsdRAIDDetailedICTLogging=default -N NODE –i
    ```

    ```
    # tar -czf ict.tgz -C /tmp/mmfs ict
    ```

  - Attach the compressed file to the IBM case.
- Unbalance of vdisk partition distribution:
  - Add the output of the following command from the Erasure Code Edition nodes to a text file and add it to the IBM case.

    ```
    # /usr/lpp/mmfs/bin/mmfsadm test vdisk vdDist 1
    ```

What to do if you see degraded performance over CES with NFS and/or SMB

This topic describes the procedure to troubleshoot any issues relating to degraded performance over CES with NFS or SMB.

Compared degraded performance to what? Is there a repeatable test and a baseline to compare to? "It is slow" is not a valid measurable metric. You should have a baseline to compare to. It is also important to identify if the issue is only reproducible on CES-served protocols instead of on NSD protocol. If it is also reproducible on NSD protocol, see “What to do if you see degraded performance over NSD protocol” on page 65.

There are multiple tools to create that. The product includes `nsdperf`, but you can choose to use other tools available in the market such as `ior`, `iozone`, `bonnie++`.

Whichever tool that you choose when you deploy the system, use the same tool to compare against baseline. Mention the tool you used and the results of baseline and the current results when you contact IBM support.

Things to check:

- First and foremost, check the network end to end
- Review any changes done to either the clients or servers (sysctl)
- Check OS resources on the client system (CPU, memory, swap in and out, ...)
- Check OS resources on the server system
- Look for **`mmhealth`** events
- Look for SMART events (if applicable)
- Reboot the client

If you still see degraded performance compared to your baseline with the repeatable test, it is time to gather some information and contact IBM support with the following data.

For detailed information, see the *CES tracing and debug data collection* topic in the *IBM Spectrum Scale: Problem Determination Guide*.

- Generate an IBM Spectrum Scale snap on the CES cluster. Use `--performance` and `--protocol` with the protocol of interest (nfs or smb or nfs, smb, ...).
- Gather protocol traces:
  - For SMB:
    - Start the traces from a CES node that serves SMB

      ```
      #  mmprotocoltrace start smb -c <clientIP>
      ```

    - You can check the status of the trace as well as the output files with:

      ```
      # mmprotocoltrace status smb
      ```

    - Once the problem has been reproduced from the client, stop the traces and send the files to IBM:

      ```
      # mmprotocoltrace stop smb
      ```

  - For NFS:
    - NFS traces are obtained by changing the log level to `FULL_DEBUG`. Be aware that the change of log level will do a restart of the CES NFS daemons on all nodes and that generates a vast amount of data that might impact performance.

      ```
      # mmnfs config change LOG_LEVEL=FULL_DEBUG
      ```

    - When the issue has been reproduced from the client, gather a snap and restore to default (EVENT) log level. Be aware that the restore of the log level will trigger a restart of all CES NFS daemons.

      ```
      # gpfs.snap --protocol nfs
      ```

      ```
      # mmnfs config change LOG_LEVEL=EVENT
      ```

Monitoring NVMe Devices

You can monitor the health of any NVMe drives in your system using the **mmlsnvmestatus** command. You can monitor the status of all devices or a specific device, specified by serial number.

For each NVMe device, the **mmlsnvmestatus** command will identify any devices where the link status does not match the link capabilities (speed and width). Additionally, it will identify any devices where the device LBA format is not one of the designated “best” formats for that device.

This example shows the output of the command on a 4-server system:

```
mmlsnvmestatus all
                                                 Optimal     Optimal      needs
 node       NVMe device    serial number        Link State  LBA Formats service
 ------     -----------    -------------        ----------  -----------  ------
 node1      /dev/nvme0     57L0A03LTZ5D         NO          YES          NO
 node1      /dev/nvme1     57L0A03KTZ5D         YES         YES          NO
 node2      /dev/nvme0     57M0A01GTZ5D         YES         NO           NO
 node2      /dev/nvme1     57M0A01JTZ5D         YES         YES          NO
 node3      /dev/nvme0     57M0A00UTZ5D         YES         YES          NO
 node3      /dev/nvme1     57M0A00KTZ5D         YES         YES          NO
 node4      /dev/nvme0     57M0A019TZ5D         YES         YES          NO
 node4      /dev/nvme1     57M0A00QTZ5D         YES         YES          NO
```

You can pass the *--not-ok* flag example to only return devices with Link State or LBA Format that is not optimal. For example:

```
mmlsnvmestatus all --not-ok
                                          Optimal     Optimal      needs
 node       NVMe device    serial number    Link State  LBA Formats service
 ------     -----------    -------------    ----------  -----------  ------
 node1      /dev/nvme0     57L0A03LTZ5D     NO          YES          NO
 node2      /dev/nvme0     57M0A01GTZ5D     YES         NO           NO
```

In this example, the NVMe device on node1 is shown to have "Optimal Link State" value of "NO". This is likely due to device not being seated properly in PCIe slot. You can see more details by comparing at the *LnkCap* and *LnkSta* output of **lspci** command for this device. The NVMe device on node1 is shown to have "Optimal LBA Formats" value of "NO". You can view the available format values and the current in use value with the **nvme id-ns** command for the NVMe device.

Monitoring the endurance of SSD Devices

You can monitor the endurance of the SSD drives in your system by using the **mmhealth** command.

An SSD or physical disk has a finite lifetime based on the number of drive writes per day. The SSD endurance is a number between 0 and 255. The **ssd-endurance-percentage** 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. When the endurance number exceeds this threshold, the **mmhealth** command displays a `ssd_endurance_warn` warning with the specific physical disk name and the recovery group name information. The drive must be replaced when the value exceeds 100, and the state of its health is reported as DEGRADED by the **mmhealth** command.

Issue the following command to display the health status of the **NATIVE_RAID** component:

```
[root@client21 ~]# mmhealth node show NATIVE_RAID
```

If the endurance number exceeds 100, the system gives an output similar to the following:

```
Node name:      client21.sonasad.almaden.ibm.com

Component             Status        Status Change     Reasons
-------------------------------------------------------------------------------------------------
----------------
NATIVE_RAID           DEGRADED      Now               ssd_endurance_warn(rg1/n001p013)
  ARRAY               HEALTHY       Now               -
  NVME                HEALTHY       1 hour ago        -
  PHYSICALDISK        DEGRADED      Now               ssd_endurance_warn(rg1/n001p013)
  RECOVERYGROUP       HEALTHY       Now               -
  VIRTUALDISK         HEALTHY       Now               -
```

You can replace the SSD physical disk to resolve this warning message. After the SSD is replaced, issue the **mmhealth** command as shown to check the health status of the SSD:

```
[root@client21 ~]# mmhealth node show NATIVE_RAID
```

After the issue is resolved the system gives an output similar to the following:

```
Node name:      client21.sonasad.almaden.ibm.com

Component             Status        Status Change     Reasons
-----------------------------------------------------------------
NATIVE_RAID           HEALTHY       Now               -
  ARRAY               HEALTHY       Now               -
  NVME                HEALTHY       1 hour ago        -
  PHYSICALDISK        HEALTHY       Now               -
```

```
  RECOVERYGROUP       HEALTHY       Now                 -
  VIRTUALDISK         HEALTHY       Now                 -
```

Detecting unsupported firmware in a IBM Spectrum Scale Erasure Code Edition network

You can detect unsupported firmware in a recovery group by using the **mmhealth** command.

Issue the following command to display the health status of the **NETWORK** component:

```
mmhealth node show NETWORK
```

If any of the firmware is unsupported, the system displays an output similar to the following:

```
Node name:      c941f3n08-ib0

Component     Status        Status Change     Reasons
------------------------------------------------------------------------------------------------------------
-
NETWORK       DEGRADED      11 hours ago      nic_firmware_unexpected(00W0038YK50200006EP,00W0038YK50200006EL)
  ib0         HEALTHY       11 hours ago      -
  mlx4_0/1    HEALTHY       11 hours ago      -


Event                       Parameter     Severity    Active Since      Event Message
------------------------------------------------------------------------------------------------------------
-
nic_firmware_unexpected     NETWORK       WARNING     11 hours ago      The adapter 00W0038YK50200006EP
                                                                        has firmware level 2.10.0700
                                                                        and not the expected
                                                                        firmware level 12.24.1000.
nic_firmware_unexpected     NETWORK       WARNING     11 hours ago      The adapter 00W0038YK50200006EL
                                                                        has firmware level 2.10.0700
                                                                        and not the expected
                                                                        firmware level 12.24.1000.
```

Note: The command raises a warning for any unsupported firmware attached to the IB network, but not for the Ethernet cluster.

You can replace the upgrade or change the firmware to resolve this warning message. After the firmware is replaced, issue the **mmhealth** command as shown:

```
mmhealth node show NETWORK
```

After the issue is resolved the system gives an output similar to the following:

```
Node name:      c941f3n08-ib0

Component     Status        Status Change     Reasons
-------------------------------------------------------------
NETWORK       HEALTHY       1 day ago         -
  ib0         HEALTHY       1 day ago         -
  mlx4_0/1    HEALTHY       1 day ago         -

There are no active error events for the component NETWORK on this node (c941f3n08-ib0)
```

Accessibility features for IBM Spectrum Scale

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:

- 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 Knowledge Center, and its related publications, are accessibility-enabled. The accessibility features are described in IBM Knowledge Center (www.ibm.com/support/knowledgecenter).

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 US. This material might be available from IBM in other languages. However, you may be required to own a copy of the product or product version in that language in order to access it.

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, MD-NC119 Armonk, NY 10504-1785 US

For license inquiries regarding double-byte character set (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

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 jurisdictions 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 websites are provided for convenience only and do not in any manner serve as an endorsement of those websites. The materials at those websites are not part of the materials for this IBM product and use of those websites is at your own risk.

IBM may use or distribute any of the information you provide 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 Director of Licensing IBM Corporation North Castle Drive, MD-NC119 Armonk, NY 10504-1785 US

Such information may be available, subject to appropriate terms and conditions, including in some cases, payment of 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.

The performance data discussed herein is presented as derived under specific operating conditions. Actual results may vary.

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.

Statements regarding IBM's future direction or intent are subject to change or withdrawal without notice, and represent goals and objectives only.

All IBM prices shown are IBM's suggested retail prices, are current and are subject to change without notice. Dealer prices may vary.

This information is for planning purposes only. The information herein is subject to change before the products described become available.

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 actual people or business enterprises 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.

Each copy or any portion of these sample programs or any derivative work must include a copyright notice as follows:

© (your company name) (year).
Portions of this code are derived from IBM Corp.
Sample Programs. © Copyright IBM Corp. _enter the year or years_.

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.

Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.

Microsoft and Windows are trademarks of Microsoft Corporation in the United States, other countries, or both.

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.

Applicability

These terms and conditions are in addition to any terms of use for the IBM website.

Personal use

You may reproduce these publications for your personal, noncommercial use provided that all proprietary notices are preserved. You may not distribute, display or make derivative work of these publications, or any portion thereof, without the express consent of IBM.

Commercial use

You may reproduce, distribute and display these publications solely within your enterprise provided that all proprietary notices are preserved. You may not 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 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 may not download, export or re-export 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.

IBM Online Privacy Statement

IBM Software products, including software as a service solutions, (“Software Offerings”) may use cookies or other technologies to collect product usage information, to help improve the end user experience, to tailor interactions with the end user or for other purposes. In many cases no personally identifiable information is collected by the Software Offerings. Some of our Software Offerings can help enable you to collect personally identifiable information. If this Software Offering uses cookies to collect personally identifiable information, specific information about this offering’s use of cookies is set forth below.

This Software Offering does not use cookies or other technologies to collect personally identifiable information.

If the configurations deployed for this Software Offering provide you as customer the ability to collect personally identifiable information from end users via cookies and other technologies, you should seek your own legal advice about any laws applicable to such data collection, including any requirements for notice and consent.

For more information about the use of various technologies, including cookies, for these purposes, See IBM’s Privacy Policy at http://www.ibm.com/privacy and IBM’s Online Privacy Statement at http://www.ibm.com/privacy/details the section entitled “Cookies, Web Beacons and Other Technologies” and the “IBM Software Products and Software-as-a-Service Privacy Statement” at http://www.ibm.com/software/info/product-privacy.

Glossary

This glossary provides terms and definitions for IBM Spectrum Scale.

The following cross-references are used in this glossary:

- *See* refers you from a nonpreferred 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 (www.ibm.com/software/globalization/terminology) (opens in new window).

B

block utilization
The measurement of the percentage of used subblocks per allocated blocks.

C

cluster
A loosely-coupled collection of independent systems (nodes) organized into a network for the purpose of sharing resources and communicating with each other. See also *GPFS cluster*.

cluster configuration data
The configuration data that is stored on the cluster configuration servers.

Cluster Export Services (CES) nodes
A subset of nodes configured within a cluster to provide a solution for exporting GPFS file systems by using the Network File System (NFS), Server Message Block (SMB), and Object protocols.

cluster manager
The node that monitors node status using disk leases, detects failures, drives recovery, and selects file system managers. The cluster manager must be a quorum node. The selection of the cluster manager node favors the quorum-manager node with the lowest node number among the nodes that are operating at that particular time.

Note: The cluster manager role is not moved to another node when a node with a lower node number becomes active.

clustered watch folder
Provides a scalable and fault-tolerant method for file system activity within an IBM Spectrum Scale file system. A clustered watch folder can watch file system activity on a fileset, inode space, or an entire file system. Events are streamed to an external Kafka sink cluster in an easy-to-parse JSON format. For more information, see the *mmwatch command* in the *IBM Spectrum Scale: Command and Programming Reference*.

control data structures
Data structures needed to manage file data and metadata cached in memory. Control data structures include hash tables and link pointers for finding cached data; lock states and tokens to implement distributed locking; and various flags and sequence numbers to keep track of updates to the cached data.

D

Data Management Application Program Interface (DMAPI)
The interface defined by the Open Group's XDSM standard as described in the publication *System Management: Data Storage Management (XDSM) API Common Application Environment (CAE) Specification C429*, The Open Group ISBN 1-85912-190-X.

deadman switch timer
A kernel timer that works on a node that has lost its disk lease and has outstanding I/O requests. This timer ensures that the node cannot complete the outstanding I/O requests (which would risk causing file system corruption), by causing a panic in the kernel.

dependent fileset
A fileset that shares the inode space of an existing independent fileset.

disk descriptor
A definition of the type of data that the disk contains and the failure group to which this disk belongs. See also *failure group*.

disk leasing
A method for controlling access to storage devices from multiple host systems. Any host that wants to access a storage device configured to use disk leasing registers for a lease; in the event of a perceived failure, a host system can deny access, preventing I/O operations with the storage device until the preempted system has reregistered.

disposition
The session to which a data management event is delivered. An individual disposition is set for each type of event from each file system.

domain
A logical grouping of resources in a network for the purpose of common management and administration.

E

ECKD
See *extended count key data (ECKD)*.

ECKD device
See *extended count key data device (ECKD device)*.

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*, *master encryption key*.

extended count key data (ECKD)
An extension of the count-key-data (CKD) architecture. It includes additional commands that can be used to improve performance.

extended count key data device (ECKD device)
A disk storage device that has a data transfer rate faster than some processors can utilize and that is connected to the processor through use of a speed matching buffer. A specialized channel program is needed to communicate with such a device. See also *fixed-block architecture disk device*.

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*.

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.

file audit logging
Provides the ability to monitor user activity of IBM Spectrum Scale file systems and store events related to the user activity in a security-enhanced fileset. Events are stored in an easy-to-parse JSON format. For more information, see the *mmaudit command* in the *IBM Spectrum Scale: Command and Programming Reference*.

file clone
A writable snapshot of an individual file.

file encryption key (FEK)
A key used to encrypt sectors of an individual file. See also *encryption key*.

file-management policy
A set of rules defined in a policy file that GPFS uses to manage file migration and file deletion. See also *policy*.

file-placement policy
A set of rules defined in a policy file that GPFS uses to manage the initial placement of a newly created file. See also *policy*.

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.

fixed-block architecture disk device (FBA disk device)
A disk device that stores data in blocks of fixed size. These blocks are addressed by block number relative to the beginning of the file. See also *extended count key data device*.

fragment
The space allocated for an amount of data too small to require a full block. A fragment consists of one or more subblocks.

G

global snapshot
A snapshot of an entire GPFS file system.

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 recovery log
A file that contains a record of metadata activity, and exists for each node of a cluster. In the event of a node failure, the recovery log for the failed node is replayed, restoring the file system to a consistent state and allowing other nodes to continue working.

I

ill-placed file
A file assigned to one storage pool, but having some or all of its data in a different storage pool.

ill-replicated file
A file with contents that are not correctly replicated according to the desired setting for that file. This situation occurs in the interval between a change in the file's replication settings or suspending one of its disks, and the restripe of the file.

independent fileset
A fileset that has its own inode space.

indirect block
A block containing 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.

ISKLM
IBM Security Key Lifecycle Manager. For GPFS encryption, the ISKLM is used as an RKM server to store MEKs.

J

journaled file system (JFS)
A technology designed for high-throughput server environments, which are important for running intranet and other high-performance e-business file servers.

junction
A special directory entry that connects a name in a directory of one fileset to the root directory of another fileset.

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.

M

master encryption key (MEK)
A key used to encrypt other keys. See also *encryption key*.

MEK
See *master encryption key*.

metadata
Data structures that contain information that is needed to access file data. Metadata includes inodes, indirect blocks, and directories. Metadata is not accessible to user applications.

metanode
The one node per open file that is responsible for maintaining file metadata integrity. In most cases, the node that has had the file open for the longest period of continuous time is the metanode.

mirroring
The process of writing the same data to multiple disks at the same time. The mirroring of data protects it against data loss within the database or within the recovery log.

Microsoft Management Console (MMC)

A Windows tool that can be used to do basic configuration tasks on an SMB server. These tasks include administrative tasks such as listing or closing the connected users and open files, and creating and manipulating SMB shares.

multi-tailed
A disk connected to multiple nodes.

N

namespace
Space reserved by a file system to contain the names of its objects.

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 hex 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 may contain one or more nodes.

node descriptor
A definition that indicates how GPFS uses a node. Possible functions include: manager node, client node, quorum node, and nonquorum node.

node number
A number that is generated and maintained by GPFS 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 GPFS 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.

Non-Volatile Memory Express (NVMe)
An interface specification that allows host software to communicate with non-volatile memory storage media.

P

policy
A list of file-placement, service-class, and encryption rules that define characteristics and placement of files. Several policies can be defined within the configuration, but only one policy set is active at one time.

policy rule
A programming statement within a policy that defines a specific action to be performed.

pool
A group of resources with similar characteristics and attributes.

portability
The ability of a programming language to compile successfully on different operating systems without requiring changes to the source code.

primary GPFS cluster configuration server
In a GPFS cluster, the node chosen to maintain the GPFS cluster configuration data.

private IP address
A IP address used to communicate on a private network.

public IP address
A IP address used to communicate on a public network.

Q

quorum node
A node in the cluster that is counted to determine whether a quorum exists.

quota
The amount of disk space and number of inodes assigned as upper limits for a specified user, group of users, or fileset.

quota management
The allocation of disk blocks to the other nodes writing to the file system, and comparison of the allocated space to quota limits at regular intervals.

R

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.

remote key management server (RKM server)
A server that is used to store master encryption keys.

replication
The process of maintaining a defined set of data in more than one location. Replication involves copying designated changes for one location (a source) to another (a target), and synchronizing the data in both locations.

RKM server
See *remote key management server*.

rule
A list of conditions and actions that are triggered when certain conditions are met. Conditions include attributes about an object (file name, type or extension, dates, owner, and groups), the requesting client, and the container name associated with the object.

S

SAN-attached
Disks that are physically attached to all nodes in the cluster using Serial Storage Architecture (SSA) connections or using Fibre Channel switches.

Scale Out Backup and Restore (SOBAR)
A specialized mechanism for data protection against disaster only for GPFS file systems that are managed by IBM Spectrum Protect Hierarchical Storage Management (HSM).

secondary GPFS cluster configuration server
In a GPFS cluster, the node chosen to maintain the GPFS cluster configuration data in the event that the primary GPFS cluster configuration server fails or becomes unavailable.

Secure Hash Algorithm digest (SHA digest)
A character string used to identify a GPFS security key.

session failure
The loss of all resources of a data management session due to the failure of the daemon on the session node.

session node
The node on which a data management session was created.

Small Computer System Interface (SCSI)
An ANSI-standard electronic interface that allows personal computers to communicate with peripheral hardware, such as disk drives, tape drives, CD-ROM drives, printers, and scanners faster and more flexibly than previous interfaces.

snapshot
An exact copy of changed data in the active files and directories of a file system or fileset at a single point in time. See also *fileset snapshot*, *global snapshot*.

source node
The node on which a data management event is generated.

stand-alone client
The node in a one-node cluster.

storage area network (SAN)
A dedicated storage network tailored to a specific environment, combining servers, storage products, networking products, software, and services.

storage pool
A grouping of storage space consisting of volumes, logical unit numbers (LUNs), or addresses that share a common set of administrative characteristics.

stripe group
The set of disks comprising the storage assigned to a file system.

striping
A storage process in which information is split into blocks (a fixed amount of data) and the blocks are written to (or read from) a series of disks in parallel.

subblock
The smallest unit of data accessible in an I/O operation, equal to one thirty-second of a data block.

system storage pool
A storage pool containing file system control structures, reserved files, directories, symbolic links, special devices, as well as the metadata associated with regular files, including indirect blocks and extended attributes The `system storage pool` can also contain user data.

T

token management
A system for controlling file access in which each application performing a read or write operation is granted some form of access to a specific block of file data. Token management provides data consistency and controls conflicts. Token management has two components: the token management server, and the token management function.

token management function
A component of token management that requests tokens from the token management server. The token management function is located on each cluster node.

token management server
A component of token management that controls tokens relating to the operation of the file system. The token management server is located at the file system manager node.

transparent cloud tiering (TCT)
A separately installable add-on feature of IBM Spectrum Scale that provides a native cloud storage tier. It allows data center administrators to free up on-premise storage capacity, by moving out cooler data to the cloud storage, thereby reducing capital and operational expenditures. .

twin-tailed
A disk connected to two nodes.

U

user storage pool
A storage pool containing the blocks of data that make up user files.

V

VFS
See *virtual file system*.

virtual file system (VFS)
A remote file system that has been mounted so that it is accessible to the local user.

virtual node (vnode)
The structure that contains information about a file system object in a virtual file system (VFS).

W

watch folder API
Provides a programming interface where a custom C program can be written that incorporates the ability to monitor inode spaces, filesets, or directories for specific user activity-related events within IBM Spectrum Scale file systems. For more information, a sample program is provided in the following directory on IBM Spectrum Scale nodes: `/usr/lpp/mmfs/samples/util` called tswf that can be modified according to the user's needs.

Index

L

M

N

O

P

R

S

U

V

Product Number: 5737-J34

SC27-9578-06