

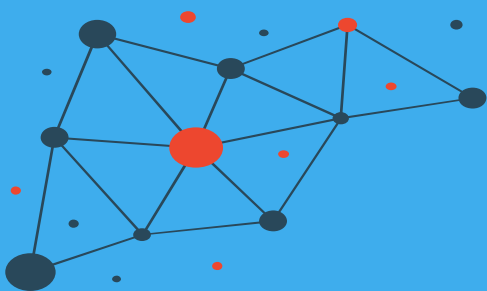
Spectrum Scale User Group, Australia

Spectrum Scale Erasure Code Edition Release Updates



Lin Feng Shen - shenlinf@cn.ibm.com
ECE Architect, Spectrum Scale Development

Join user group:
www.spectrumscaleug.org/join



About the user group

- Independent, work with IBM to develop events
- Not a replacement for PMR!
- Email and Slack community
- www.spectrumscaleug.org/join



#SSUG

Disclaimer

IBM's statements regarding its plans, directions, and intent are subject to change or withdrawal without notice at IBM's sole discretion. Information regarding potential future products is intended to outline our general product direction and it should not be relied on in making a purchasing decision. The information mentioned regarding potential future products is not a commitment, promise, or legal obligation to deliver any material, code, or functionality. The development, release, and timing of any future features or functionality described for our products remains at our sole discretion.

IBM reserves the right to change product specifications and offerings at any time without notice. This publication could include technical inaccuracies or typographical errors. References herein to IBM products and services do not imply that IBM intends to make them available in all countries.

Outline

- **Recap: ECE Overview**
- **ECE Release Updates (5.0.4 and 5.0.5)**
- **Hardware Requirements (5.0.5)**
- **References**



Spectrum Scale Erasure Code Edition Overview



A new Spectrum Scale offering that brings all of the benefits of “Data Management Edition” *plus* **Spectrum Scale RAID**

- Spectrum Scale running in storage rich servers connected to each other with a high-speed network infrastructure
- Bring your own hardware – select any hardware that meets minimum requirements
 - Provides Storage devices can be HDD, SSD, NVMe or a mixture
- Features of an Enterprise Storage Controller all in software
 - Enterprise ready storage software used in Spectrum Scale Elastic Storage Server (ESS)
 - Solved challenges with commodity server based distributed storage
- Restricted GA June 2019 *



* Required to verify supported hardware configuration

ECE Value Proposition



Delivers all the capability of Spectrum Scale Data Management Edition

- Enormous scalability with Software-based declustered RAID protection
- Very high performance no additional RAID hardware
- Enterprise manageability

Plus: Durable, robust, and storage-efficient

- Distributes data across nodes and drives for higher durability *without* the cost of replication
- End to end checksum identifies and corrects errors introduced by network or media
- Rapid recovery and rebuild after hardware failure

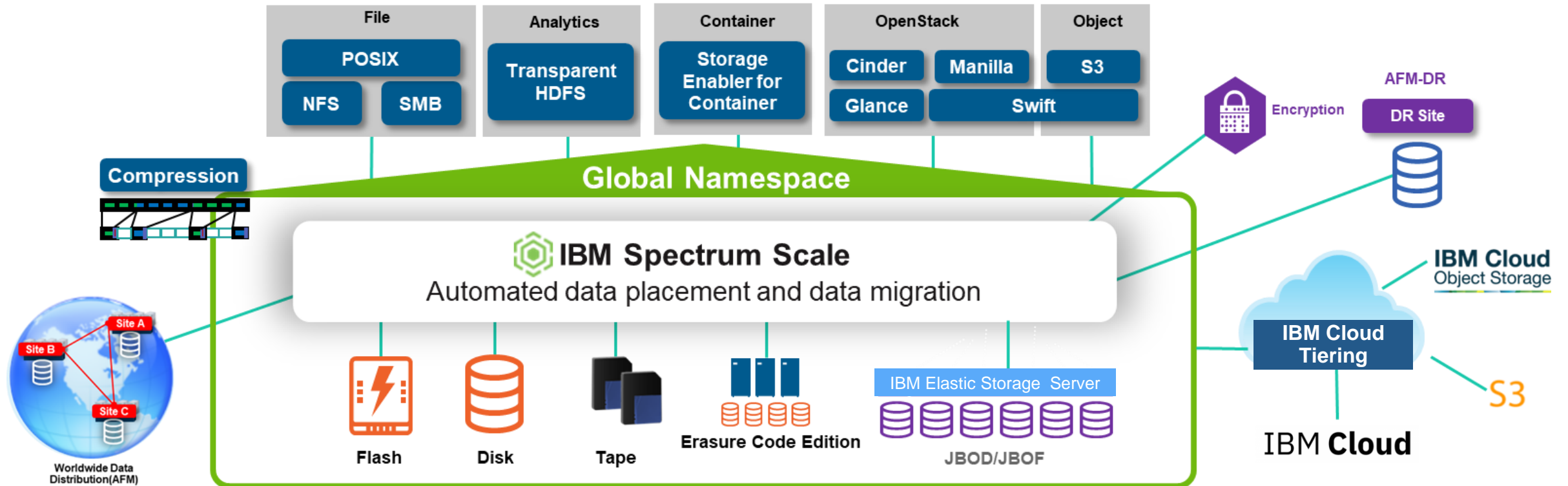
Plus: Delivered at hyperscale

- Hardware platform neutrality -Supports the user's choice of commodity servers and drives
- Disk Hospital manages drive issues before they become disasters
- Continuous background error correction supports deployment on very large numbers of drives



ECE Storage Pool Coexistence with IBM Spectrum Scale IBM Spectrum Scale

Unleash new storage economics on a global scale



Highest Performance Storage with Diverse Access Protocols using hardware that you select

Consolidate all your unstructured data storage on spectrum scale with unlimited and painless scaling of capacity and performance

ECE is Based on Proven IBM Spectrum Scale RAID Software Spectrum Scale



The software in ECE has been field-proven in over 1000 deployed ESS systems

ESS is the storage power behind the fastest supercomputers on the planet

- Summit and Sierra supercomputers at Oak Ridge National Laboratory and Lawrence Livermore National Laboratory are ranked the #1 and #2 fastest computers in the world
- They are helping to model supernovas, pioneer new materials, and explore cancer, genetics and the environment, using technologies available to all customers

ECE delivers the same capabilities on commodity compute, storage, and network components

Summary of changes for IBM Spectrum Scale Erasure Code Edition*

- Volatile write cache detection and data resiliency
 - Deferred Rebuilds for maintenance
 - New mmvdisk command option (spare node and suspend node)
 - Manual online upgrade procedure without install toolkit
 - RDMA network assessment tool
 - ECE based stretch cluster deployment (whole site down)
 - Support RHEL 7.7 and RHEL 8.1
 - IBM Spectrum Scale Erasure Code Edition Redpaper
-
- These are ECE specific updates only. All Spectrum Scale release updates are also applicable to ECE.

- Some drives are equipped with fast volatile memory to hide write latency.
- Not desired for GNR, where we require data to be persisted to disk after the kernel comes back from doing IO. Such volatile write caching violates this requirement.
- Two levels of checking in Hardware and OS Readiness Tool and the GPFS/GNR daemon.
- The readiness tool will check for SAS and NVMe Volatile Write Cache Enabled (VWCE).
 - Report “[FATAL] <hostname> <device> has Write Cache Enabled. This is not supported by ECE” when it’s enabled.
 - Tool halts at this point and halts the install. This should be sufficient to catch most of bad configs related to write caching.
 - Currently, the tool does not fix the drive settings for you. It’s the responsibility of the users.
- The daemon is the final line of defense to detect volatile write cache problem.
 - As part of disk discovery, queries mode sense pages to check write cache settings.
 - Put the disk into VWCE state, which makes it effectively read-only. We then drain the disk and treat it like missing for write purposes.
 - Can defense volatile write cache enablement unintendedly.

- Integrity Manager problems with node maintenance
 - Rebuild, Rebalance triggers automatically and unnecessarily even if the fault tolerance is sufficient
 - Waste of I/O and CPU bandwidth for planned maintenance
- Defer the rebuild and rebalance process when the fault tolerance is still good enough
 - Eliminate unnecessary data migration during maintenance
 - Services not stopped: GNR metadata rebuild, critical rebuild, readmit and scrub
 - Stop rebuild/rebalance in a specified period of timeout
 - Can specify a desired deferred rebuilds service level (the level to start rebuild)
 - Can be cancelled after timeout or in fly
- Use cases (in 5.0.5)
 - Support for online rolling upgrade with install toolkit
 - Mmvdisk suspend/resume node process for maintenance

Summary of changes for IBM Spectrum Scale Erasure Code Edition*

- Support for online rolling upgrade with install toolkit (start from 5.0.4.3 to 5.0.5 or later)
 - New disk performance precheck tool. Added disk Key Performance Indicators (KPIs).
 - Mmvdisk suspend/resume node process for maintenance
 - ECE based stretch cluster deployment
 - GNR TRIM support for NVMe devices
 - Support Mellanox ConnectX-6 (Ethernet or InfiniBand)
 - Support all 12 Gb/s LSI RAID Controller Cards
 - Support LSI Fusion-MPT Tri-Mode Host Bus Adapters, models SAS3008, SAS3408, and SAS3416
 - Procedure for setup and checking LSI cards disk location slots and remapping
 - Disk management trouble shooting improvements
 - Support RHEL 7.8 and RHEL 8.2
- These are ECE specific updates only. All Spectrum Scale release updates are also applicable to ECE.

IBM Spectrum Scale Erasure code Edition support is in 5.0.3.1 code and higher.

Ability to define a new setup type i.e. ECE.

Ability to define a scale-out node.

Ability to define a recovery group, vdiskset and filesystem.

Complete Install/deploy support with protocol support.

Ability to add new node into the existing ECE cluster.

Ability to add new recovery group into the existing ECE cluster.

Config populate support for ECE.

Offline Upgrade support.

ECE Install Toolkit New features: Online upgrade:

IBM Spectrum Scale Erasure Code Edition online upgrade support from version 5.0.4.3 or later to version 5.0.5 or later using toolkit.

Currency OS support:

RHEL8.1/8.2 support on x_86



Easy to run easy to read RAW disk performance tool:

https://github.com/IBM/SpectrumScale_ECE_STORAGE_READINESS

- Verification purpose: test and detect unexpected performance characteristics of the disk drives, including lower than expected performance or aggregate bottlenecks or big deviations
- Standard FIO benchmark based
- Only read by default – KPI as read only
- Must run and pass the KPI to get support from IBM (TDA process)
- It can guess drives when run on Python 3
- Logs per run saved on ./log directory

```
# ./nopeus.py -h

usage: nopeus.py [-h] [-b BS_CSV] [--guess-drives] [--i-want-to-lose-my-data]
                [-t FIO_RUNTIME] [--rpm_check_disabled] [-v]

optional arguments:
  -h, --help            show this help message and exit
  -b BS_CSV, --block-sizes BS_CSV
                        Block size for tests. The default and valid to certify
                        is 128k. The choices are: 4k 128k 256k 512k 1024k
  --guess-drives        It guesses the drives to test and adds them to the
                        drives.json file overwriting its content. You should
                        then manually review the file content before running
                        the tool again
  --i-want-to-lose-my-data
                        It makes the test a write test instead of read. This
                        will delete the data that is on the drives. So if you
                        care about the keeping the data on the drives you
                        really should not run with this parameter. Running
                        with this parameter will delete all data on the
                        drives
  -t FIO_RUNTIME, --time-per-test FIO_RUNTIME
                        The number of seconds to run each test. The value has
                        to be at least 30 seconds. The minimum required value
                        for certification is 300
  --rpm_check_disabled  Disables the RPM prerequisites check. Use only if you
                        are sure all required software is installed and no RPM
                        were used to install the required prerequisites.
                        Otherwise this tool will fail
  -v, --version          show program's version number and exit
```

ECE Storage Readiness Tool – Example 2/2



```
[root@mestor01 SpectrumScale_ECE_STORAGE_READINESS]# ./nopeus.py
```

Welcome to NOPEUS, version 1.6

JSON files versions:
 supported OS: 1.2
 packages: 1.0

Please use https://github.com/IBM/SpectrumScale_STORAGE_READINESS to get latest versions and report issues about this tool.

The purpose of NOPEUS is to obtain drive metrics, and compare them against KPIs

The FIO runtime per test of 300 seconds is sufficient to certify the environment

The FIO blocksize of 128k is valid to certify the environment

The FIO patterns of randread is valid to certify the environment

This test run estimation is 47 minutes

This software comes with absolutely no warranty of any kind. Use it at your own risk

NOTE: The bandwidth and latency numbers shown in this tool are for a very specific test. This is not a generic storage benchmark. The numbers do not reflect the numbers you would see with Spectrum Scale and your particular workload

We are going to test the following drives

Drive: sdd as SSD
Drive: sde as NVME
Drive: sdf as NVME
Drive: sda as HDD
Drive: sdb as HDD
Drive: sdc as SSD

Do you want to continue? (y/n): ☐

Mmvdisk Suspend/Resume Node Process for Maintenance

- The mmvdisk rg change --suspend command exploits deferred rebuild to avoid data movement during a longer maintenance window.
- It previously was dependent on the five-minute missing disk timeout to delay data rebuild from a suspended node's pdisks.
- Now suspend uses deferred rebuild with a default 20-minute window, which can be changed via the --window N option.
- Useful in various hardware/firmware/OS upgrade/maintenance tasks in storage nodes where node down is needed
- The mmvdisk rg change --resume command starts a suspended server and cancels deferred rebuild.

Example:

```
# mmvdisk rg change --rg RG --suspend -N server06 --window 30
mmvdisk: Suspended all pdisks from node 'server06'.
mmvdisk: Stopped serving log groups from node 'server06'.
mmvdisk: Shutting down GPFS on node 'server06'.
mmvdisk: Suspended node 'server06' in 'RG'.
mmvdisk: Non-critical rebuilds deferred for 30 minutes in recovery group 'RG'.
mmvdisk: After node maintenance is finished, run the command:
mmvdisk:      mmvdisk recoverygroup change --resume --recovery-group RG -N server06
#
```

<<< do your own maintenance tasks >>>

```
# mmvdisk rg change --rg RG --resume -N server06
mmvdisk: Starting GPFS on node 'server06'.
mmvdisk: Waiting up to 5 minutes for the GPFS daemon on node 'server06' to join the
cluster.
mmvdisk: Serving log groups is already enabled on node 'server06'.
mmvdisk: Resumed all pdisks from node 'server06'.
mmvdisk: Rebuilds are no longer deferred in recovery group 'RG'.
mmvdisk: Resumed node 'server06' in 'RG'.
#
```

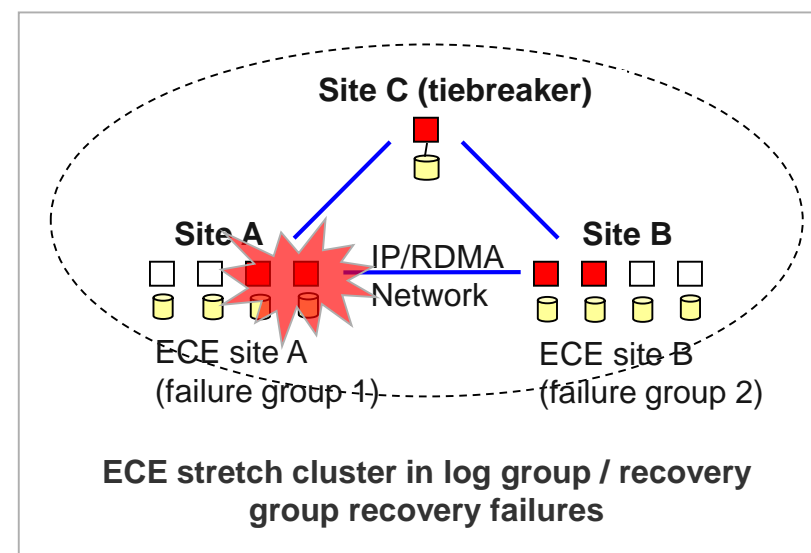
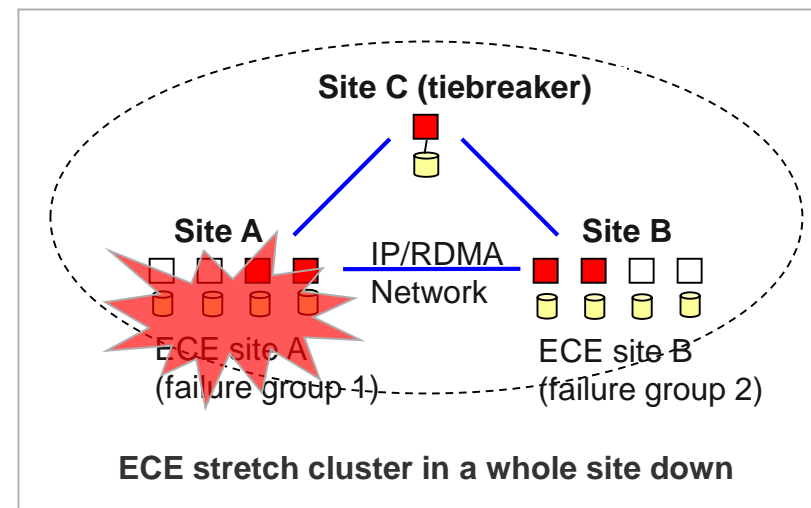
ECE based Stretch Cluster Deployment

- Single Scale cluster is defined across two geographic sites (A and B). The goal is to keep one of the sites operational should the other site fail.
- Both sites are ECE based serversets
- Same number of quorum nodes at each site
- The ECE nodes/disks are split into 2 failure groups (one at each site with all vdisks/NSDs in this site)

□ non-quorum node

■ quorum node

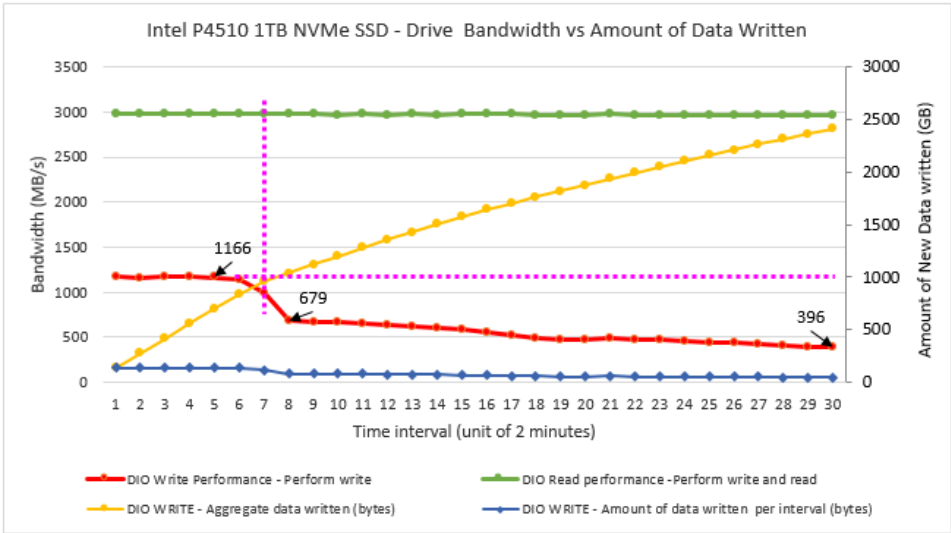
- Can tolerate various failure scenarios
 - Not matter network or node or disk failures
 - Whole site down
 - Partial site failures causing log group / recovery group recovery failures
 - Permanent log group / recovery group failures



GNR TRIM Support for NVMe Devices



- Solve tremendous write performance downgrade over time with NAND flash drives under certain workloads
- Support NVMe currently
- SAS SSD and HDD not supported yet (need tests)
- Use the same mmreclaimspace command and policy control interfaces for the same user experience across the whole Spectrum Scale family
- Additional mmvdisk interfaces for GNR recovery group TRIM functionality
 - Recovery groups can be created with TRIM enabled on declustered arrays of a specified disk hardware type
 - All DAs of the specified hardware type in the new RG will enable TRIM.
 - Can change the DA TRIM status of a recovery group
 - Only support file system vdisk NSD TRIM for vdisks from DAs with hardware TRIM enabled



```
# mmvdisk rg change --rg RG --da DA1 --trim-da yes
# mmvdisk rg list --rg RG --da --server
```

node number	server	active	remarks
1	ess3000a.gpfs.net	yes	serving RG: LG001, LG003
2	ess3000b.gpfs.net	yes	serving RG: root, LG002, LG004

declustered array	needs service	type	trim	vdisk user	log	pdisk total	spare	rt	capacity total	raw free	raw	background task
DA1	no	NVMe	yes	0	5	12	1	1	62 TiB	62 TiB		scrub 14d (10%)


```
#
# mmvdisk fs change --fs FS1 --trim auto
# mmvdisk fs list --fs FS1
```

vdisk set	recovery group	vdisk count	with trim	list of failure groups	holds metadata	holds data	storage pool
VS1	RG	4	4	1, 2	yes	yes	system

```
#
```

ECE Hardware/Architecture Requirements

ECE software is hardware platform neutral, but there are hardware requirements*



- An ECE storage system must have at least 4 servers, and up to 128 servers (128 is a test limitation).
 - Customers can create multiple ECE recovery groups. Each recovery group limits the number of servers to between 4 and 32.
 - Customers may scale out their ECE storage system with one server, multiple servers or a whole building block.
 - Every server in a recovery group must have the same configuration in terms of CPU, memory, network, storage, OS, etc.
 - For SSD and NVMe drives, it is recommended to use a file system block size of 4M or less with 8+2P or 8+3P erasure codes, and 2M file system block size or less for 4+2P or 4+3P erasure codes.
 - Minimum Declustered Array (DA) size DA is : At least one DA must contain 12 or more drives and every DA must have 6 or more drives
 - A DA is a subset of the physical disks within a recovery group that have matching 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. That means a 4 node RG must have one DA with 3 or more drives per node. A twelve node RG could have one drive per node, but that drive must be a “fast device”, either SSD or NVMe.
 - Each node must have at least one fast device (NVMe or SAS SSD)
 - All nodes/HBA's/drives in a Recovery Group must meet minimum firmware level requirements specified in Hardware Selection and Sizing Guide. (Tested number of drives per Recovery Group: 512)
 - To deliver the best performance, stability and functionality the next chart lists the minimal hardware requirements for each storage server. (This list will be expanded over time).
- For latest hardware requirements and readiness toolkit, see 'IBM Spectrum Scale Erasure Code Edition Hardware requirements' section in [IBM Knowledge Center](https://www.ibm.com/support/knowledgecenter/STXKQY_ECE_5.0.5/com.ibm.spectrum.scale.ece.v5r05.doc/b1lece_hwrequirements.htm) for a specific ECE release, e.g. ECE 5.0.5: https://www.ibm.com/support/knowledgecenter/STXKQY_ECE_5.0.5/com.ibm.spectrum.scale.ece.v5r05.doc/b1lece_hwrequirements.htm

ECE Hardware Requirements for each Storage Server 1/2

as of ECE 5.0.5 in August 2020*



CPU architecture	x86 64-bit processor with 8 or more processor cores per socket. Server should be dual socket with both sockets populated.
Memory	64 GB or more for configurations with up to 24 drives per node: <ul style="list-style-type: none"> 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 RecoveryGroup	A maximum of 512 drives per recovery group is supported.
Nodes per RecoveryGroup	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.

- For latest hardware requirements and readiness toolkit, see 'IBM Spectrum Scale Erasure Code Edition Hardware requirements' section in [IBM Knowledge Center](#) for a specific ECE release, e.g. ECE 5.0.5: https://www.ibm.com/support/knowledgecenter/STXKQY_ECE_5.0.5/com.ibm.spectrum.scale.ece.v5r05.doc/b1lece_hwrequirements.htm

ECE Hardware Requirements for each Storage Server 2/2

as of ECE 5.0.5 in August 2020*



Fast Drive Requirement	At least one SSD or NVMe drive is required in each server for IBM Spectrum Scale Erasure Code Edition logging.
SAS Storage Adapters/Controllers	<ul style="list-style-type: none">• 12 Gb/s LSI RAID Controller Cards, support JBOD mode, can be detected and managed by StorCLI utility. IBM verified cards types are recommended: SAS3008, SAS3108, SAS3408, SAS3508, or SAS3516.• 12 Gb/s LSI Fusion-MPT Tri-Mode Host Bus Adapters, models SAS3008, SAS3408, and SAS3416 can be detected and managed by StorCLI utility.• 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.
Network Adapter	Mellanox ConnectX-4, ConnectX-5 or ConnectX-6 (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.

- For latest hardware requirements and readiness toolkit, see 'IBM Spectrum Scale Erasure Code Edition Hardware requirements' section in [IBM Knowledge Center](https://www.ibm.com/support/knowledgecenter/STXKQY_ECE_5.0.5/com.ibm.spectrum.scale.ece.v5r05.doc/b1lece_hwrequirements.htm) for a specific ECE release, e.g. ECE 5.0.5: https://www.ibm.com/support/knowledgecenter/STXKQY_ECE_5.0.5/com.ibm.spectrum.scale.ece.v5r05.doc/b1lece_hwrequirements.htm

- **IBM Spectrum Scale Erasure Code Edition Redpaper:**

<http://www.redbooks.ibm.com/abstracts/redp5557.html>

- **IBM Spectrum Scale Erasure Code Edition Knowledge Center:**

https://www.ibm.com/support/knowledgecenter/STXKQY_ECE_5.0.5/ibmspectrumscaleece505_welcome.html

https://www.ibm.com/support/knowledgecenter/STXKQY_ECE_5.0.4/ibmspectrumscaleece504_welcome.html

https://www.ibm.com/support/knowledgecenter/STXKQY_ECE_5.0.3/ibmspectrumscaleece503_welcome.html

- **IBM Spectrum Scale FAQ with Erasure Code Edition:**

See section 'IBM Spectrum Scale Erasure Code Edition questions' in <https://www.ibm.com/support/knowledgecenter/STXKQY/gpfsclustersfaq.html>

- **ECE Readiness Toolkit:**

Hardware and OS Readiness Tool: https://github.com/IBM/SpectrumScale_ECE_OS_READINESS

Network Readiness Tool: https://github.com/IBM/SpectrumScale_NETWORK_READINESS

Storage Readiness Tool: https://github.com/IBM/SpectrumScale_ECE_STORAGE_READINESS

- **IBM Spectrum Scale Erasure Code Edition (ECE): Installation Demonstration**

<https://www.youtube.com/watch?v=6lf50EvgP-U&feature=youtu.be>

- **IBM Spectrum Scale Erasure Code Edition Blogs:**

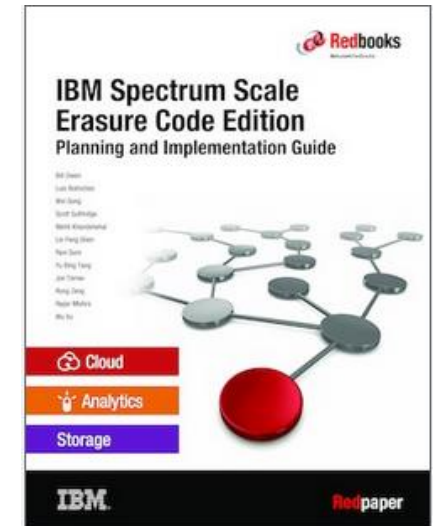
Introducing IBM Spectrum Scale Erasure Code Edition: <https://developer.ibm.com/storage/2019/07/07/introducing-ibm-spectrum-scale-erasure-code-edition/>

IBM Spectrum Scale Erasure Code Edition Fault Tolerance: <https://developer.ibm.com/storage/2019/05/30/ibm-spectrum-scale-erasure-code-edition-fault-tolerance/>

Installing IBM Spectrum Scale ECE using installation toolkit: <https://developer.ibm.com/storage/2019/06/09/installing-ibm-spectrum-scale-erasure-code-edition-using-installation-toolkit/>

Upgrading ECE using installation toolkit: <https://developer.ibm.com/storage/2019/06/09/upgrading-ibm-spectrum-scale-erasure-code-edition-using-installation-toolkit/>

IBM Spectrum Scale Erasure Code Edition in Stretched Cluster: <https://developer.ibm.com/storage/2020/07/10/ibm-spectrum-scale-erasure-code-edition-in-stretched-cluster/>




Thank you!



Please help us to improve Spectrum Scale with your feedback

- If you get a survey in email or a popup from the GUI, please respond
- We read every single reply

Provide Feedback

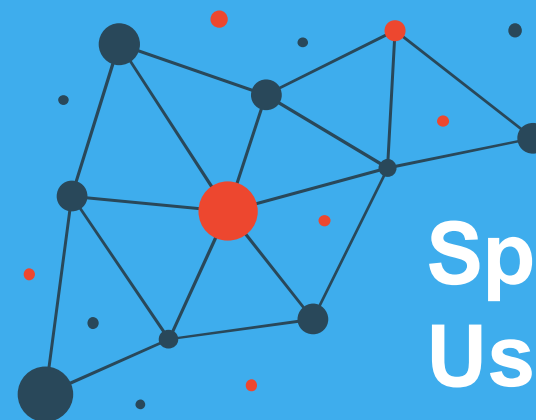


Tell IBM What You Think

Let us know what you think about IBM Spectrum Scale. It takes only a couple of minutes for you to help us improve our service. [IBM Privacy Policy](#)

Not Now

Provide Feedback



Spectrum Scale User Group

The Spectrum Scale (GPFS) User Group is free to join and open to all using, interested in using or integrating IBM Spectrum Scale.

The format of the group is as a web community with events held during the year, hosted by our members or by IBM.

See our web page for upcoming events and presentations of past events. Join our conversation via mail and Slack.

www.spectrumscaleug.org