SC2019 - Spectrum Scale User Group

Spectrum Scale & ESS update

Christopher D. Maestas
cdmaestas@us.ibm.com

Please Note

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. Information about potential future products may not be incorporated into any contract. The development, release, and timing of any future features or functionality described for our products remains at our sole discretion.

Notices and disclaimers

Notices and disclaimers continued

- 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 about this publication 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. IBM does not warrant the quality of any third-party products, or the ability of any such third-party products to interoperate with IBM's products. **IBM expressly disclaims all warranties, expressed or implied, including but not limited to, the implied warranties of merchantability and fitness for a purpose.**
- The provision of the information contained herein is not intended to, and does not, grant any right or license under any IBM patents, copyrights, trademarks or other intellectual property right.
- IBM, the IBM logo, ibm.com and [names of other referenced IBM products and services used in the presentation] are trademarks of International Business Machines Corporation, 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.

Poll – what are you running?

IBM Spectrum Scale	Minimum Recommended Level	Field proven level	Latest Level
IBM Spectrum Scale	4.2.3.14 [1] [Mar 21, 2019]	4.x stream 4.2.3.14 [Mar 21, 2019] [1] 5.x stream 5.0.2.2 [Dec 13, 2018]	4.x stream 4.2.3.18 [Oct 2019] 5.x stream 5.0.4 [Oct 2019]
IBM Spectrum Scale for ESS	ESS 5.2.6 [Apr 2019]	4.x stream: ESS 5.2.6 [Apr 2019] 5.x stream: 5.3.4 [Jun 2019]	4.x stream ESS 5.2.8 [Oct 2019] [2] 5.x stream: ESS 5.3.4.2 [Oct 2019]

[1] Clients with a file system created on GPFS 2.2 or earlier releases should read the Flash before upgrading.

[2] Clients are recommended to upgrade to 4.2.3.14 and ESS 5.2.6. For more information, please read the Flash.

Spectrum Scale Early Programs

Types of Programs:

Alpha

Influence the development of new technology by gaining before market access to product code. Alpha programs are typically confidential and the first opportunity for you to interact with a feature or function.

Beta

Try out a new offering with the team who owns the product and influence its usability and design. A Beta program gives you the ability to evaluate and provide feedback on IBM products before the products general availability. Beta programs are typically confidential and run prior to GA.

Early Support Program (ESP)

Be one of the few selected participants to validate new Software or Hardware and potentially give your enterprise an edge over the competition. The IBM early support programs give you and IBM the opportunity to develop, evaluate, and gain experience with a product or a set of products in your enterprise environment.

Customer Success

- ❑ Evaluate new IBM HW or SW in your environment.
- ❑ Validate procedures and interoperability with other products in your enterprise.
- ❑ Opportunity to Influence Product Design
- ❑ Early Enablement and education
- ❑ Strengthen Partnership with IBM

IBM Early Programs website:
https://ibm.biz/NTIPrograms

NTI Program Interest Form:
www-355.ibm.com/technologyconnect/cna/epInterestForm.xhtml

Spectrum Scale Updates

Core enhancements – Thin Provisioning and TRIM support

Thin provisioning support (RPQ only)

- Add the ability to use thinly provisioned and compressed volumes for both file system data and metadata
- Contact your sales or account team (RPQ / SCORE process) for assessment on the use of thin provisioning
- New CLI command (**mmreclaimspace**) and NSD configuration

TRIM support for NVMe devices

- NSD configuration and the new **mmreclaimspace** command to enable TRIM support, which reduces write amplification on solid-state devices under certain workloads

QOS improvements in large clusters

QoS node collects stats X time and sends to QoS manager node Y time

Report via **mmlsqos**

Large clusters => more communication, performance degradation

Dynamically based on client mounts

Allow changes

stat-slot-time : QoS collects

stat-poll-interval : QoS -> QoS manager

Table 1. Default intervals for collecting and sending statistics

Number of nodes that have mounted the file system	Interval between collecting statistics, in milliseconds	Interval between sending statistics to the QoS manager, in seconds
< 32	1000	5
< 64	2000	10
< 128	3000	15
< 256	4000	20
< 512	6000	30
< 1024	8000	40
< 2048	10000	50
< 4096	12000	60
< 8192	12000	60
< 16384	12000	60
16384 or more	24000	120

mmchqos *Device --enable [--stat-poll-interval Seconds] [--stat-slot-time Milliseconds]*

Network enhancements

daemon startup service waits for active RDMA port (configurable)

adjust the length of the timeout period. see

verbsPortsWaitTimeout
verbsRdmaFailBackTCPIfNotAvailable

nsdperf for "stress" testing

now opensource!

https://github.com/IBM/SpectrumScale_NETWORK_READINESS/blob/master/nsdperf.C

Attempt to reconnect socket before expel

- **DEFAULT ON! (in Linux)**

mmchconfig *proactiveReconnect=yes*

Raise network reconnects to **mmhealth**

```
# mmhealth node eventlog
Timestamp        Event Name       Severity        Details
2019-TIME TZ     reconnect_start  WARNING         Attempting to …
2019-TIME TZ     reconnect_done   INFO            Reconnected to …
```

```
# mmhealth node eventlog
Timestamp        Event Name        Severity        Details
2019-TIME TZ     reconnect_start   WARNING         Attempting to …
2019-TIME TZ     reconnect_failed  ERROR           Reconnect … failed
2019-TIME TZ     reconnect_aborted INFO            Reconnect … aborted
```

Spectrum Scale misc.

Support **mmsdrrestore --ccr-repair** option with

- sudo wrapper
- Windows environments

mmsdrrestore -N in CCR clusters with *adminMode=central*

gpfs.snap - Displays an error message if the output directory (-d) is in a file system managed by the same cluster you are running against.

Monitoring critical threads in mmfsd for stuck or overloaded critical threads

mmhealth other updates

New health events for

- ssd wear level (new monitor)
- firmware level of NICs (ECE only)
- Nameserver issues related to AD authentication
- ESS3000 (officially with 5.0.4-1):
  - in CANISTER/SERVER (new component)
  - in ENCLOSURE (some events moved to CANISTER/SERVER)

Improvements in the usability of mmprotocoltrace

SMB updates

vfs_fruit module: enhances the support of Mac OS SMB2 clients.
Enabling results in changes how Apple particular metadata is handled that improves file browsing.

Support for RHEL version 8

Enhancements for using immutable files from SMB clients: Files in an immutable fileset can now be set immutable from SMB clients by setting the READONLY attribute. The retention time can be set by modifying the LastAccessTime from a SMB client. After the retention time expires, the READONLY attribute can be cleared from an SMB client and the file can be deleted.

Spectrum Scale Release	General Availability	Samba Version	Platform Support (accum.)
4.1.1	2Q15	4.2	x86_64/RHEL7
4.2.0	4Q15	4.3	ppc64/RHEL7
4.2.1	2Q16	4.3	x86_64/SLES12
4.2.2	4Q16	4.4	ppc64le, ppc64, x86_64 / RHEL7.2
4.2.3.0 - 4.2.3.8	2Q17	4.5	x86_64, ppc64, ppc64le / RHEL 7.3, 7.4
5.0.0	4Q17	4.6	x86_64/Ubuntu 16.04.2
5.0.1	1Q18	4.6	RHEL 7.5 (5.0.1.1)
5.0.2 >= 4.2.3.9	3Q18	4.6	+ Ubuntu 18.04
5.0.3	2Q19	4.9	RHEL 7.6 (bringing mutex fixes)
5.0.4	4Q2019	4.9	RHEL8

PROTOCOLS – NFS – Ganesha 2.7.5

Ganesha grace period is changed from 60 seconds to 90 seconds

New configuration keyword: RPC_IOQ_THRDMAX

- **Deprecates several existing configuration parameters for simplified tuning** (NB_WORKER, Dispatch_Max_Reqs, Dispatch_max_Reqs_Xprt)

Enhancements to Ganesha stats (ganesha_stats) command

- Reset & Duration field added for statistics
- Authentication related statistics added

Enhancements to Ganesha Mgr (ganesha_mgr) command

- Memory trim options
- File system cache display

Enhanced Memory Management methods

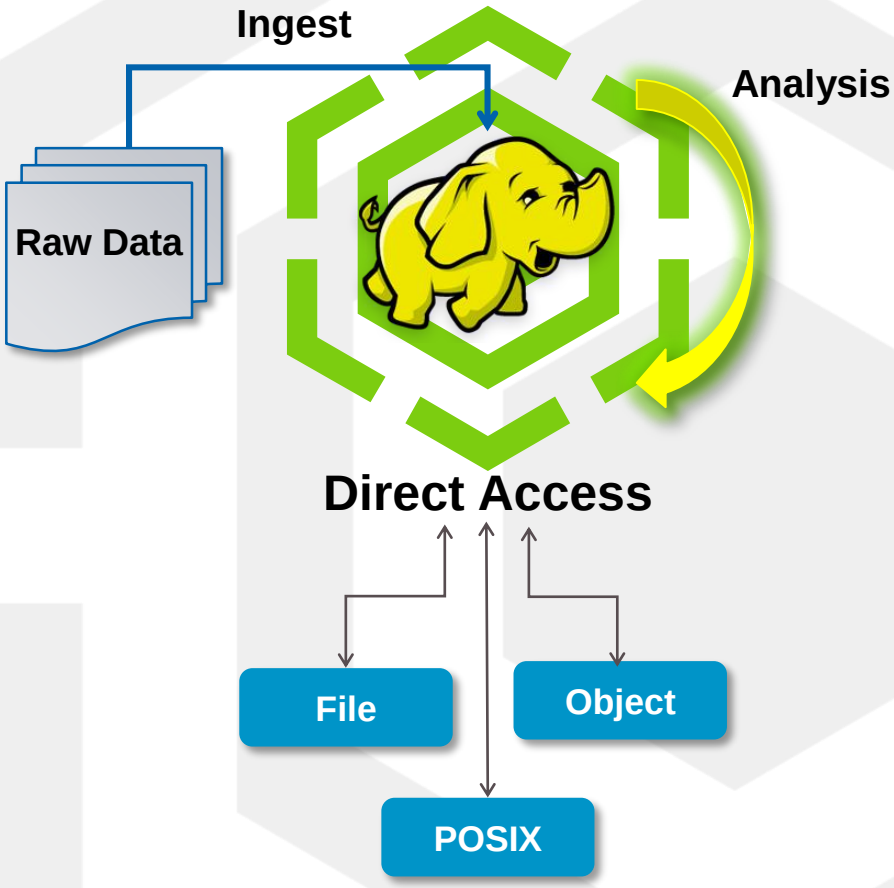
Big Data and Analytics Enhancements

Support for Hortonworks Data Platform (HDP) 3.1.4

Issue fixed when a map reduce task fails after running for one hour when Ranger is enabled.

Issue fixed when Hadoop permission settings do not work properly in a kerberized environment.

Open Source Apache Hadoop version 3.1.1 is now supported



Scale and Containers

Container Storage Interface (CSI) version 1.0

Allow a Spectrum Scale file system to surface into a pod/container running inside OpenShift 4.2

Open Source Beta driver in out:

https://github.com/IBM/ibm-spectrum-scale-csi-driver

Official GA support later in Q4

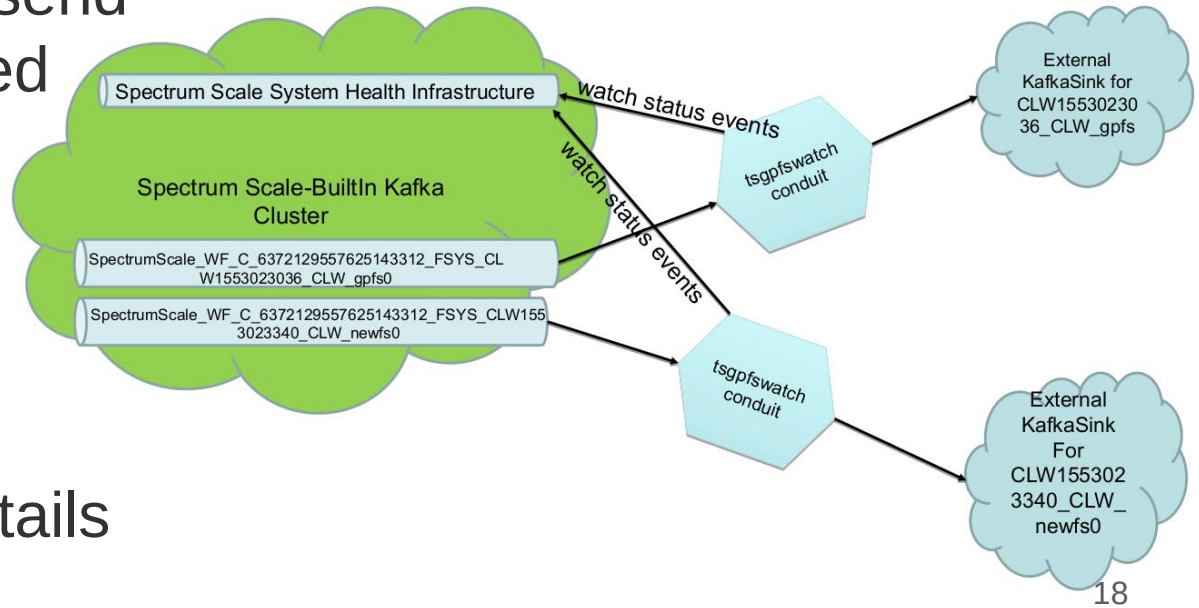
Security - Support for sending events to a secondary sink when a watch is suspended

When a watch is suspended, events are written to a secondary sink, which is a configurable fileset

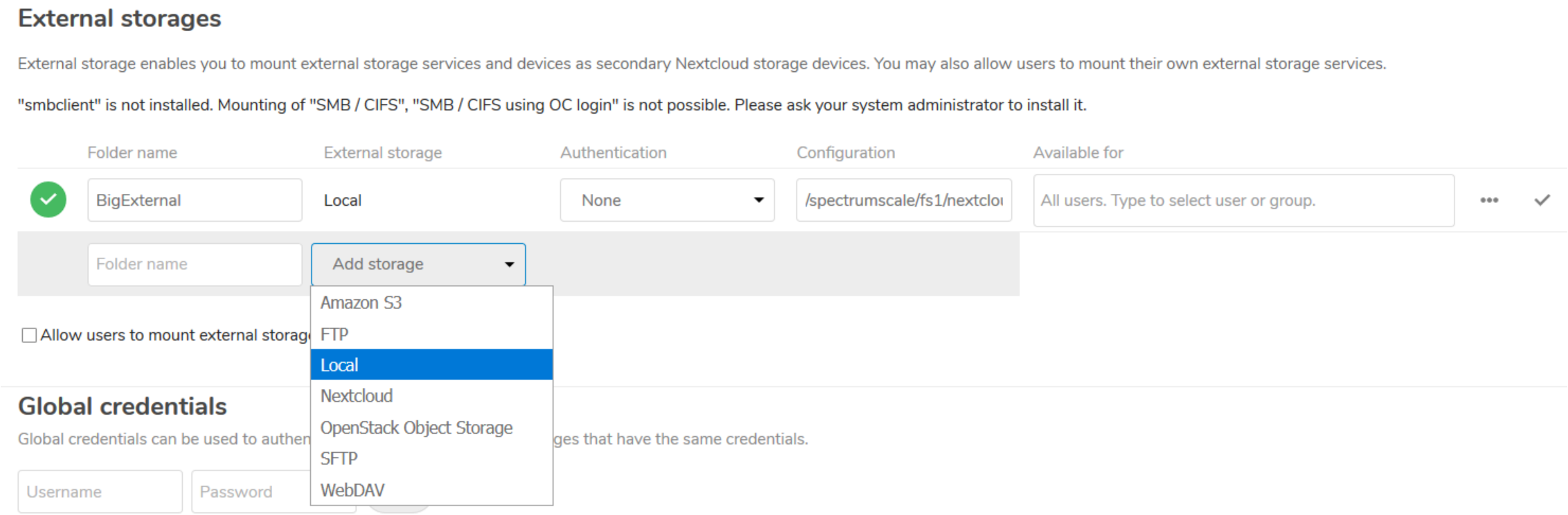
When the watch is resumed, replay conduits read events from the secondary sink and send them to the external queue to be processed

Ensures no events are lost during maintenance

See the **mmwatch** man page for more details on how to configure a secondary sink

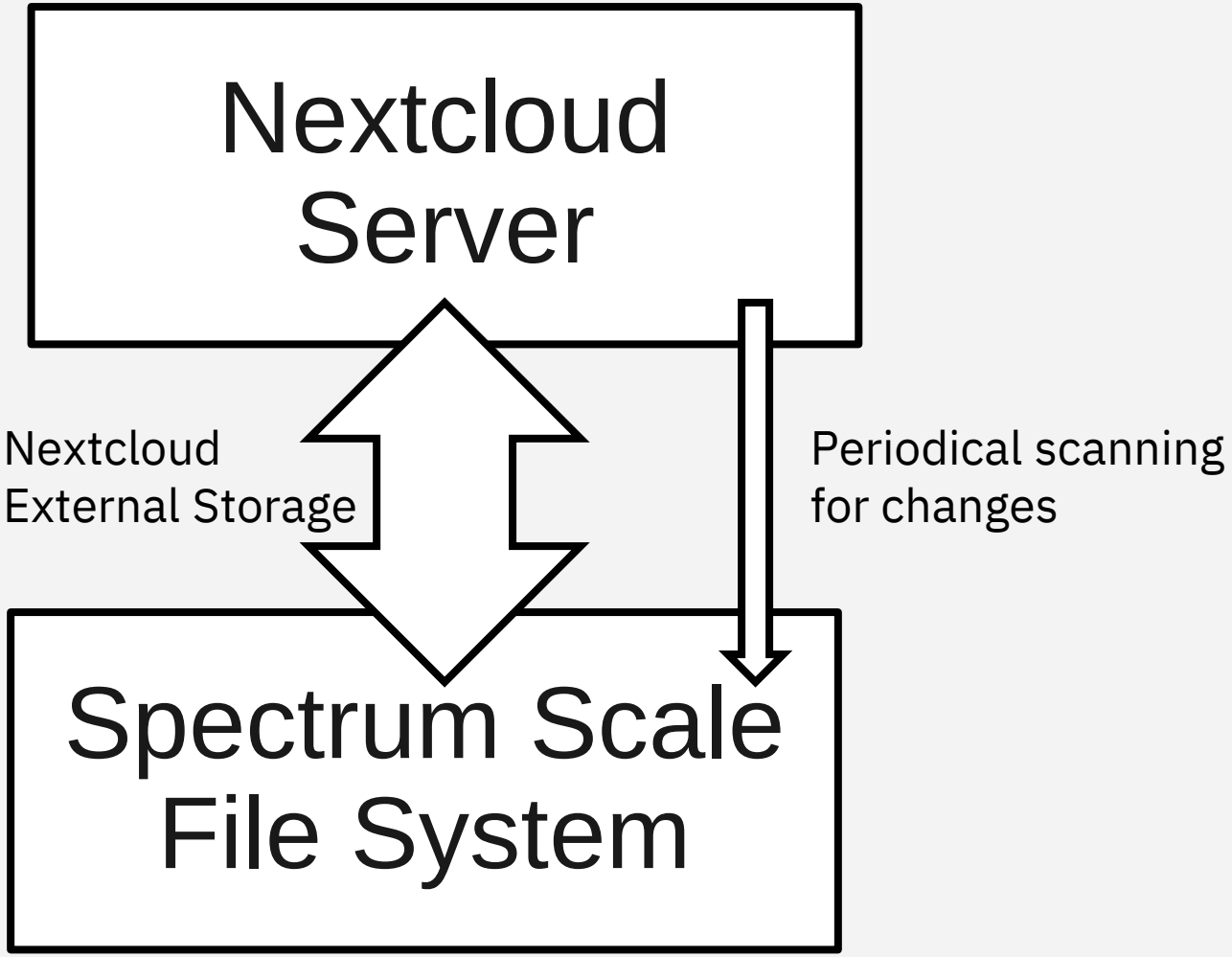


Auditing – an example Nextcloud External Storage

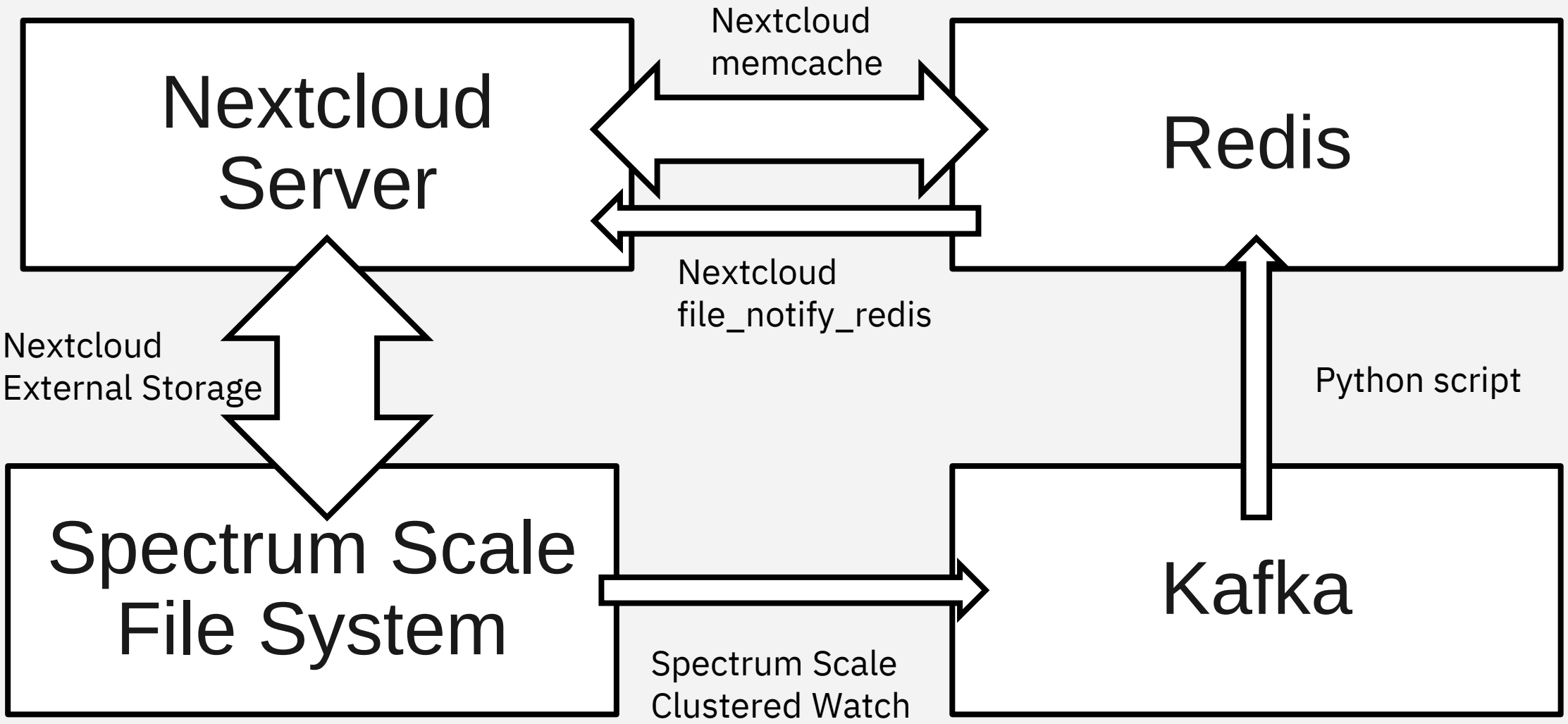


- Configure Spectrum Scale as local file system
- Past: Periodically scan file system to update Nextcloud file system view
- New: Real-time update form Spectrum Scale to Nextcloud using Spectrum Scale Clustered Watch

Integration Architecture (Past)



Integration Architecture (New)



Active File Management (AFM)

afmParallelMounts -
for parallel data transfers by using multiple remote mounts

Fixes for AFM DR over remote cluster mounts

Support **dependent filesets** inside an AFM or AFM-DR fileset.

AFM and AFM DR is supported when SELinux is enabled.

Recovery fixes, resync is not required in all the cases

5.0.4 – Stabilized, Deprecated, and Discontinued

Certain functions within Spectrum Scale may be stabilized, deprecated or discontinued

Stabilized function

- There is **no** plan to remove this functionality
- Continue to use this functionality, as it will remain supported
- Expect currency updates and fixes only. No significant new function or enhancements are planned

Deprecated function

- The function is supported in this release
- The function may be discontinued in a future release
- In some cases, it may be best to begin planning for alternatives to this function for long-term support

Discontinued function

- This function is no longer available in this release. Stay on previous releases of code if you are using this function.

Stabilized

Category	Recommend Action
Transparent Cloud Tiering (TCT)	TCT can still be used, no plans to extend purpose. It talks Swift, S3 and Azure. YMMV.
File Placement Optimizer (FPO)	FPO and SNC are available. FPO limit to 32 nodes. Direction to use Erasure Code Edition (ECE)
cNFS	Investing in user space solutions for NFS. Deprecation when performance and functionality sufficient for Ganesha.

Deprecated

Category	Deprecated functionality	Recommend Action
iSCSI target (**mmblock**)	Use of IBM Spectrum Scale as an iSCSI target for remote boot of servers.	Plan to replace the use of Scale as a target through iSCSI with other block storage providers.
The watch API and tswf sample program	The watch API and sample program was for creating a single node watch using the API. We now provide more resilient and fully integrated cluster watch with the mmwatch command.	Plan to use the improved **mmwatch** command to start clustered watches.
Kafka	IBM Spectrum Scale will no longer support gpfs.kafka on IBM Spectrum Scale clusters. This means that there will be no concept of brokers or zookeepers. Although, we will still provide gpfs.librdkafka.	No actions needed. IBM Spectrum Scale will provide a single command to do this conversion at the time it is removed.
Message Queue	The message queue will no longer be needed since kafka will be removed. The **mmmsgqueue** command will be removed entirely and no longer needed to run **mmwatch** or **mmaudit** commands.	No actions needed. IBM Spectrum Scale will provide a single command to do this conversion at the time it is removed.
Audit fileset residing on separate filesystem	IBM Spectrum Scale will no longer support creating the audit fileset on a filesystem that is not the one being audited. This means that the audit fileset has to belong to the audited filesystem.	Reconfigure audit with the mmaudit command to change this configuration. mmaudit Device enable [--log-fileset FilesetName]

Discontinued

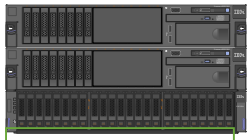
Category	Deprecated functionality	Recommend Action
OpenStack	Support for OpenStack other than support for Swift interfaces to Scale Objects. IBM Spectrum Scale will not be certified in additional releases beyond the OpenStack Rocky release. The Train release of OpenStack will not support the Cinder driver for IBM Spectrum Scale.	Plan to move Scale deployments under OpenStack to a new deployment environment. Plan to replace use of IBM Spectrum Scale through the Cinder driver with other block storage providers.
GPFS version 2.2 file system format	File systems originally created under GPFS version 2.2 or earlier are not supported with IBM Spectrum Scale 5.0. This includes file systems that were originally created in 2.2 or earlier and were subsequently upgraded to a later file system version.	Create a new file system in IBM Spectrum Scale Version 4.2 or Version 5.0 and migrate the data from the old file system.

ESS Updates

IBM Elastic Storage Server: building blocks small and large

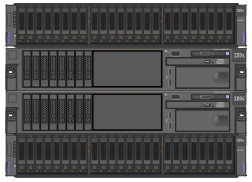
Speed ← → **Capacity**

Model GS1S
24 SSD



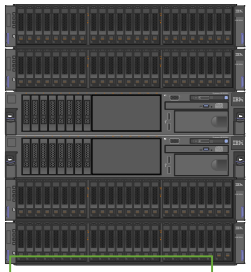
10+ GB/s

Model GS2S
48 SSD



20+ GB/s

Model GS4S
96 SSD



40 GB/s

Model GH22:
2 2U24 Enclosure SSD
2 5U84 Enclosure HDD
166 NL-SAS, 48 SSD

Model GH12:
1 2U24 Enclosure SSD
2 5U84 Enclosure HDD
166 NL-SAS, 24 SSD

20+ GB/s*

Model GH14:
1 2U24 Enclosure SSD
4 5U84 Enclosure HDD
334 NL-SAS, 24 SSD

30+ GB/s*

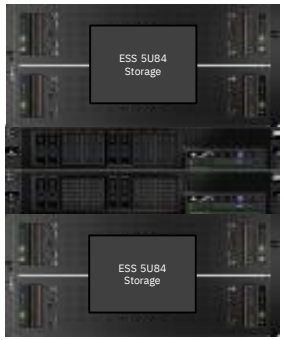
Model GH24:
2 2U24 Enclosure SSD
4 5U84 Enclosure HDD
334 NL-SAS, 48 SSD

40+ GB/s*

Model GL1S:
1 Enclosures, 9U
82 NL-SAS, 2 SSD

5+ GB/s

Model GL2S:
2 Enclosures, 14U
166 NL-SAS, 2 SSD

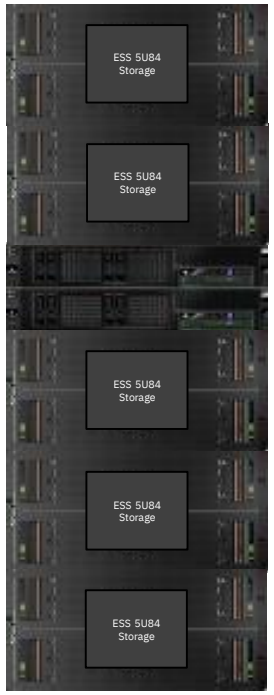


10+ GB/s

Model GL4S:
4 Enclosures, 24U
334 NL-SAS, 2 SSD

20+ GB/s

Model GL5S:
5 Enclosures, 29U
418 NL-SAS, 2 SSD



25+ GB/s

Model GL6S:
6 Enclosures, 34U
502 NL-SAS, 2 SSD

30+ GB/s

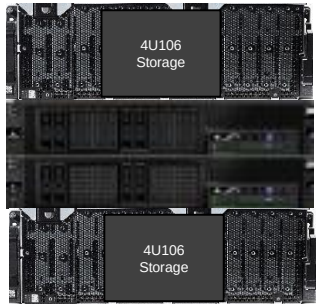
* Estimate of performance aggregated across SSD and HDD. NOTE: All estimates assume EDR Infiniband connections and are read performance

IBM Elastic Storage Server GLxC models

Model GL1C:
1 Enclosure, 8U
104 NL-SAS, 2 SSD

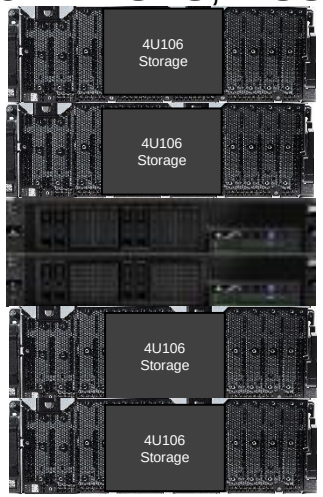
1.46 PB raw

Model GL2C:
2 Enclosures, 12U
210 NL-SAS, 2 SSD



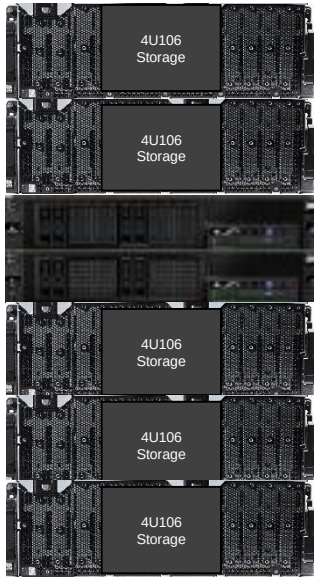
2.9 PB

Model GL4C
4 Enclosures, 16U
432 NL-SAS, 2 SSD



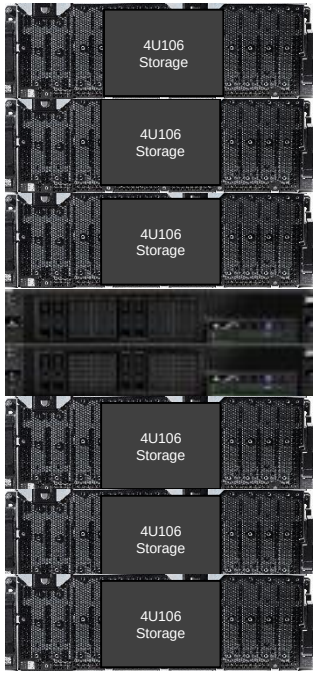
5.9 PB

Model GL5C
5 Enclosures, 28U
528 NL-SAS, 2 SSD



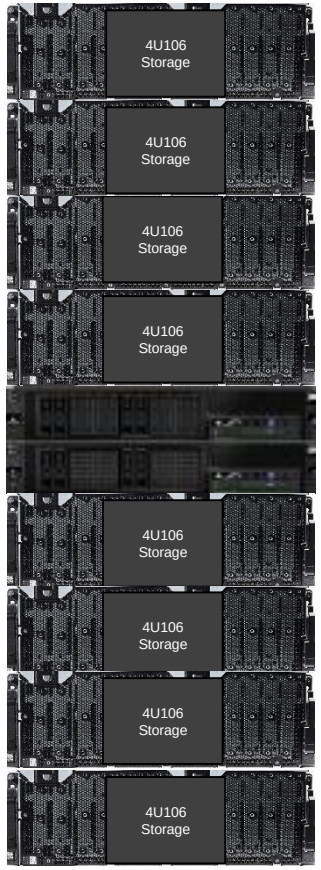
7.3 PB

Model GL6C
6 Enclosures, 28U
634 NL-SAS, 2 SSD



8.8 PB

Model GL8C
8 Enclosures, 36U
846 NL-SAS, 2 SSD



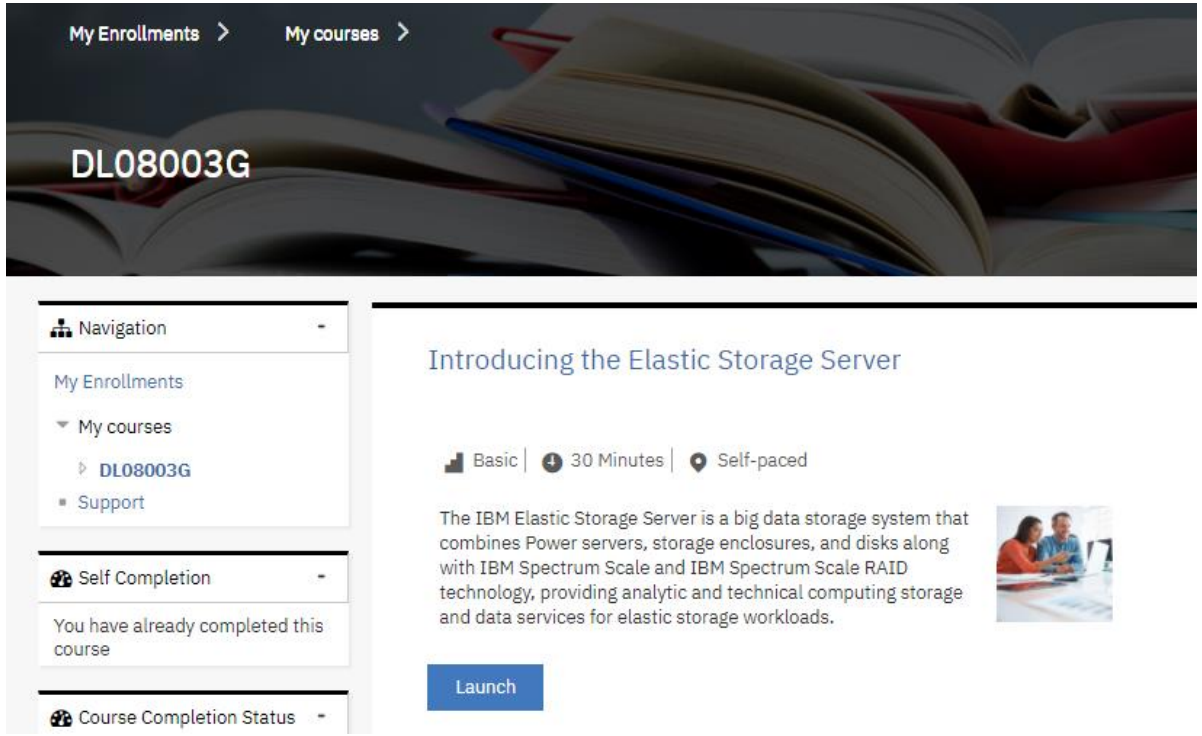
11.8 PB raw

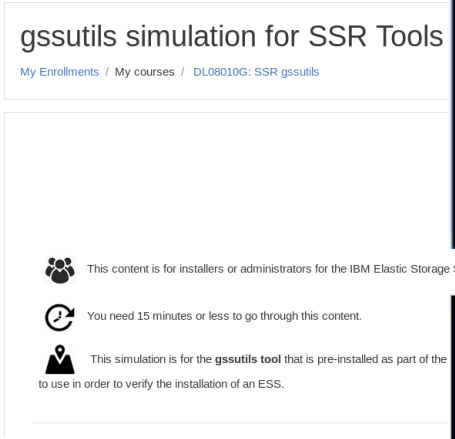
Free! Introduction to IBM Elastic Storage Server and Spectrum Scale RAID and gssutils

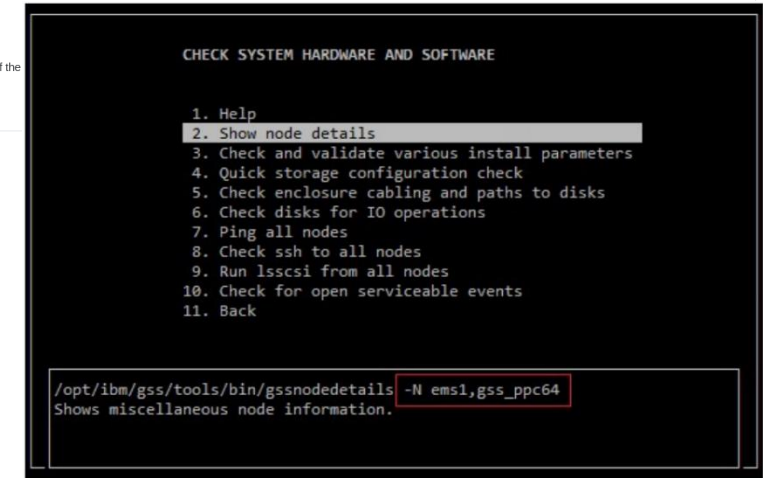
(Log on with your IBM ID)

https://www.onlinedigitallearning.com/course/view.php?id=2173

https://www.onlinedigitallearning.com/course/view.php?id=3570







Non-disruptive upgrades

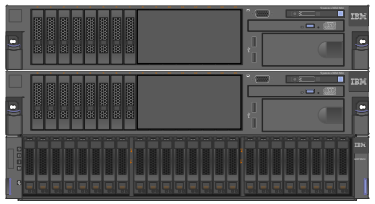
Simple expansion of Storage Capacity

- Spectrum Scale will automatically rebalance data in the background
- System automatically puts the new capacity to use
- No need to Archive & Restore data
- No System disruption*

Non Disruptive Upgrades	
From	To
GS1S	GS2S
GS2S	GS4S
GL1S	GL2S
GL2S	GL4S
GL4S	GL6S
GL1C	**GL2C**
GL2C	**GL4C**
GL4C	**GL5C**

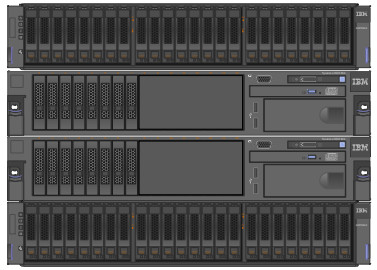
Example

Model GS1S
With 24 SSDs



Install additional drawer with 24 SSDs

Model GS2S
With 48 SSDs



*Requires space available in the rack

Software Changes

Software Name	Previous Version 5.3.4.1	Current Version 5.3.4.2
Spectrum Scale	5.0.3.2 efix4	**5.0.3.3**
HMC (For classic only)	860 SP3	860 SP3
xCAT	2.14.6	2.14.6
System Firmware	FW860.60 (SV860_180)	**FW860.70 (SV860_205)**
Red Hat Enterprise Linux (PPC64BE and PPC64LE)	7.6	7.6
Kernel Systemd Network Manager	3.10.0-957.21.3 219-62.el7_6.6 1.12.0-10.el7_6	**3.10.0-957.27.4** **219-67.el7_7.1** **1.18.0-5**
Open Fabrics Enterprise Distribution (Mellanox, Infiniband, some Ethernet)	MOFED 4.6-3.1.9.1	MOFED 4.6-3.1.9.1
IPR (for boot drives)	19512200	19512200
ES AGENT	4.5.1	**4.5.1-1**

IBM Elastic Storage System 3000

NVMe Flash for AI & Big Data Workflows

NEW

All new storage solution

- Integrated scale-out advanced data management with end-to-end NVMe storage
- Containerized software for ease of install and update
- Hours, not days, for initial configuration
- Fast and easy update and scale-out expansion
- Performance, capacity, & ease of integration for AI and Big Data workflows

IBM Spectrum Scale

IBM Elastic Storage System 3000

Overview

Scalable high-performance unified storage for files and objects

File management	IBM Spectrum Scale Version 5
Data protection	IBM Spectrum Scale erasure coding
Internal operating system	Red Hat Enterprise Linux 8.x
Protocols and interfaces	POSIX with Spectrum Scale client, NFS v4.0, SMB v3.0, Hadoop MapReduce, OpenStack Swift (object), S3 (object), CSI (Container Storage Interface)
Controllers	Highly available dual active-active controllers
Storage	NVMe flash drives (1.92TB, 3.84TB, 7.68TB or 15.4TB)
Number of drives	12 or 24 drives per 2U enclosure
Memory	384 GB or 768 GB memory per controller
Network adapters	Up to two PCIe network interface cards per controller (2 controllers per ESS 3000) Mellanox Connect X5 with Infiniband EDR and 100GBps Ethernet with RoCE support

ESS 3000 versus GS4S

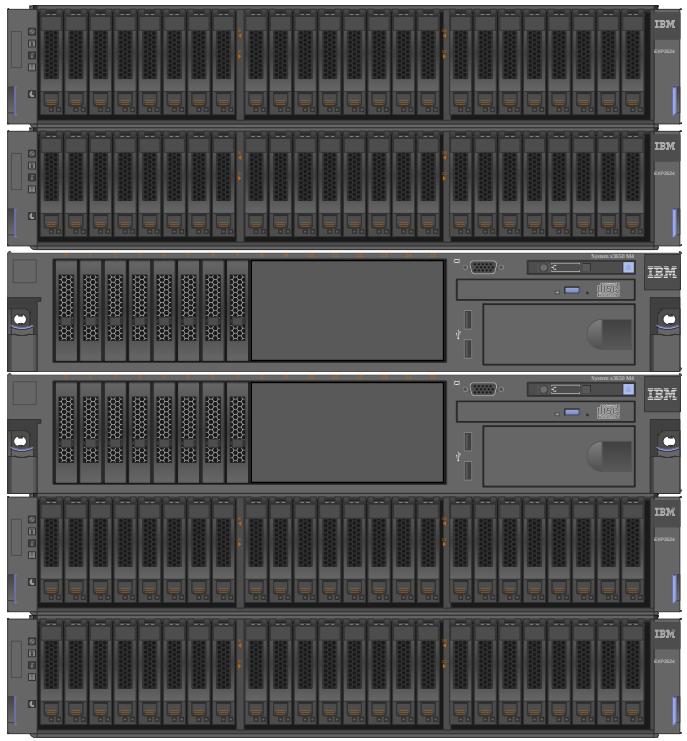
Similar Performance

Bandwidth and IOPS

Running Spectrum Scale and Spectrum Scale RAID

2U space

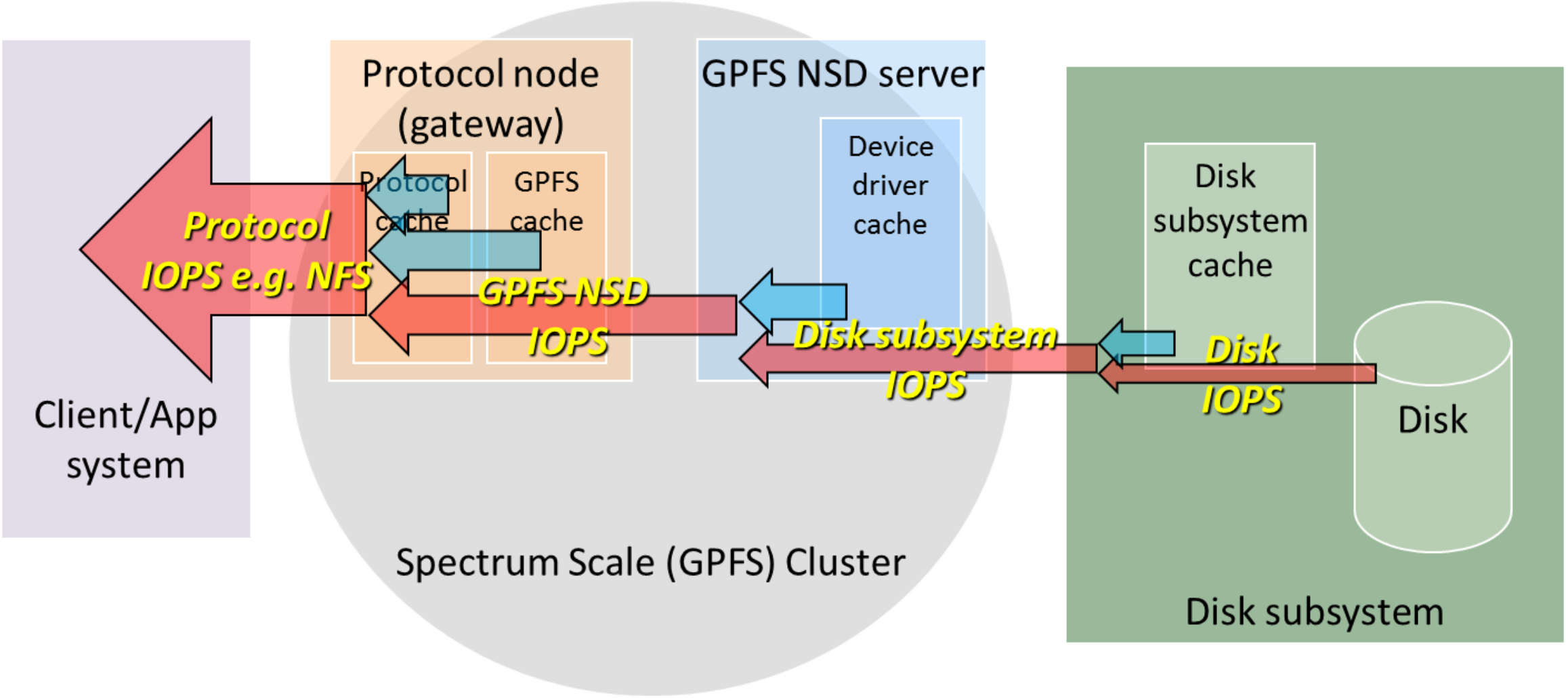
I/O servers integrated



12U space

~~IOPS~~ POSIX Transactions per second!

The many meanings of IOPS



POSIX Transactions per Second
Random 4k reads (think meta data searching)

In 3.5 was about **60k** per NSD server

Changed in a PTF to about **120k** per NSD server

ESS with (Scale 4.2.X.Y) - recorded 185k per ESS

ESS 5.3.0/1 code (Scale 5.0.1.1) – Increased to 450k per ESS

- **225k per NSD Server**
- Measured with IOR different options for
  - Oil and Gas and Government

Today's testing with ESS3000 showing about

330k per NSD server
660k per ESS3000

Thank You.

IBM Storage & SDI