Spectrum Scale 5.0.2 Updates

Christopher D. Maestas

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

-

- **U.S. Government Users Restricted Rights — use, duplication or disclosure restricted by GSA ADP Schedule Contract with IBM.**

- Information in these presentations (including information relating to products that have not yet been announced by IBM) has been reviewed for accuracy as of the date of initial publication and could include unintentional technical or typographical errors. IBM shall have no responsibility to update this information. **This document is distributed "as is" without any warranty, either express or implied. In no event, shall IBM be liable for any damage arising from the use of this information, including but not limited to, loss of data, business interruption, loss of profit or loss of opportunity.** IBM products and services are warranted per the terms and conditions of the agreements under which they are provided.

- IBM products are manufactured from new parts or new and used parts.
In some cases, a product may not be new and may have been previously installed. Regardless, our warranty terms apply."

- **Any statements regarding IBM's future direction, intent or product plans are subject to change or withdrawal without notice.**

- Performance data contained herein was generally obtained in a controlled, isolated environments. Customer examples are presented as illustrations of how those

- customers have used IBM products and the results they may have achieved. Actual performance, cost, savings or other results in other operating environments may vary.

- References in this document to IBM products, programs, or services does not imply that IBM intends to make such products, programs or services available in all countries in which IBM operates or does business.

- Workshops, sessions and associated materials may have been prepared by independent session speakers, and do not necessarily reflect the views of IBM. All materials and discussions are provided for informational purposes only, and are neither intended to, nor shall constitute legal or other guidance or advice to any individual participant or their specific situation.

- It is the customer's responsibility to insure its own compliance with legal requirements and to obtain advice of competent legal counsel as to the identification and interpretation of any relevant laws and regulatory requirements that may affect the customer's business and any actions the customer may need to take to comply with such laws. IBM does not provide legal advice or represent or warrant that its services or products will ensure that the customer follows any law.

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.

- .

IBM User Group Meetings

Sun	11/11	12:30-17:30	IBM Spectrum Scale User Group Meeting
Mon	11/12	13:00-16:00	Open Compute HPC Project Meeting
Tue	11/13	15:00-17:00	IBM Spectrum LSF User Group Meeting
Thu	11/15	8:30-12:30	IBM HPC & AI University User Meeting

IBM Seminars

Tue	11/13	10:00-11:00	MC01: PowerAI Enterprise: Elastic Distributed Training and High Performance Inference
Tue	11/13	13:00-14:00	MC02: PowerAI Vision: Data Labeling to Inference at the Edge, Made Easy For All
Tue	11/13	14:30-15:30	MC03: IBM Spectrum Metadata Solutions Deep Dive and Demo
Wed	11/14	10:00-11:00	MC05: High Performance and Capacity: Options for Spectrum Scale and Object Storage
Wed	11/14	13:00-14:00	MC06: H2O Driverless AI on Power: AI to do AI
Wed	11/14	14:30-15:30	MC07: Machine Learning and Deep Learning at Scale

1:1 Meetings

Carl Zetie	Offering Manager for Spectrum Scale
Sam Werner	Offering Executive for Spectrum Scale

IBM Spectrum Scale

Summary!

Use Cases for Spectrum Scale and the Elastic Storage Server (ESS)

1. Back-up / Restore
2. Archive
3. Information Life Cycle Management
4. Unified Storage view in your “Data Ocean”
5. Big Data and Analytics
6. Data-intensive Technical Computing
7. AI
8. Selected Solutions
   - Industry Solutions
   - ISV Solutions and Offerings

Spectrum Scale Parallel Architecture

No Hot Spots

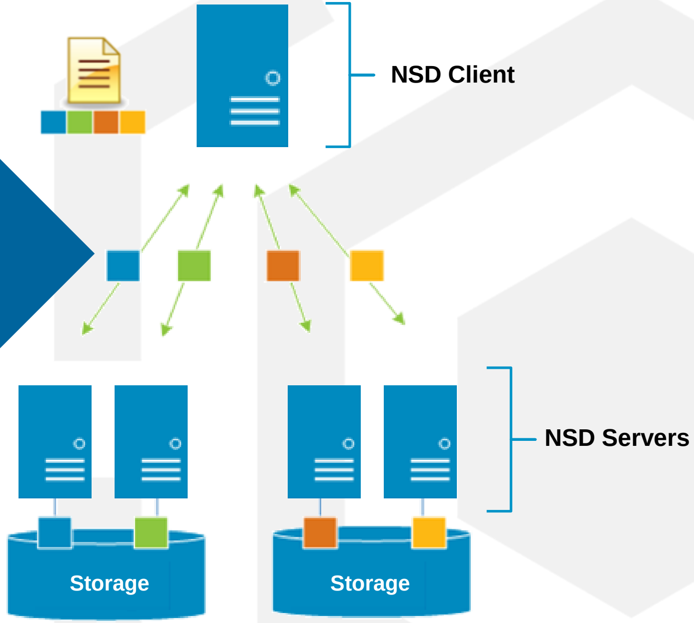
All NSD servers export to all clients in active-active mode

Spectrum Scale stripes files across NSD servers and NSDs in units of file-system block-size

File-system load spread evenly

Easy to scale file-system capacity and performance while keeping the architecture balanced

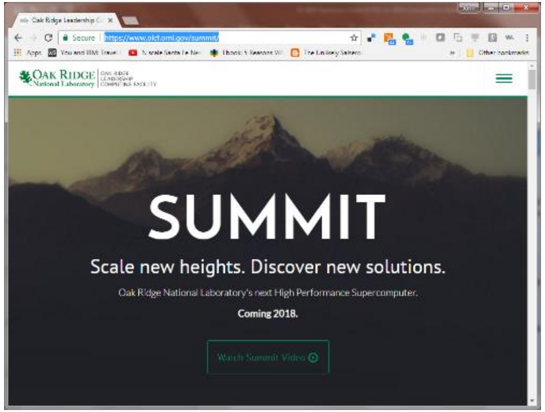
Primary Protocol



NSD Client does real-time parallel I/O
to all the NSD servers and storage volumes/NSDs

Performance engineering matters

CORAL
COLLABORATION
OAK RIDGE • ARGONNE • LIVERMORE


Oak Ridge Leadership C
Secure
https://www.olcf.ornl.gov/summit/
Apps
Other bookmarks
OAK RIDGE National Laboratory
OAK RIDGE LEADERSHIP COMPUTING FACILITY
Watch Summit Video

SUMMIT

Scale new heights. Discover new solutions.

Oak Ridge National Laboratory's next High Performance Supercomputer.

Coming 2018.

https://www.olcf.ornl.gov/summit/

Imagine you need to deliver:

- 2.5 TB/sec single stream IOR
  as requested from ORNL
- 1 TB/sec 1MB sequential read/write
  as stated in CORAL RFP
- Single Node 16 GB/sec sequential read/write
  as requested from ORNL
- 50K creates/sec per shared directory
  as stated in CORAL RFP
- 2.6 Million 32K file creates/sec
  as requested from ORNL

What innovations in storage would this require?

Performance - Benchmark efforts

What have we done and where can we go?

IBM has a benchmark center in Poughkeepsie

- On the truck code of Scale and ESS with EDR
- IOR runs on ESS GL6S and GS4S with **Scale 5.0.1.1**
- Now preparing to upgrade ESS systems with next scale release

Take what research and performance teams do and replicate

How about a external place to submit?

1. **ELK stack (Elastic Search, Kibana, Logstack)**
2. **tick from influxdata ... www.influxdata.com/time-series-platform**

Chris’s performance summary from this year

Number of runs	Benchmark request
1	bonnie
2	fio
2	gpfsperf
19	IOR
13	mdtest
1	Single node

Create 10 different filesystems on each ESS (GL6S and GS4S)

Run IOR via LSF

- 1 job at a time
- Total of 1074 jobs
- 12 nodes with 1 process per node
- smpi 10.1.1.0 - now testing newer version

same results regardless of the benchmark

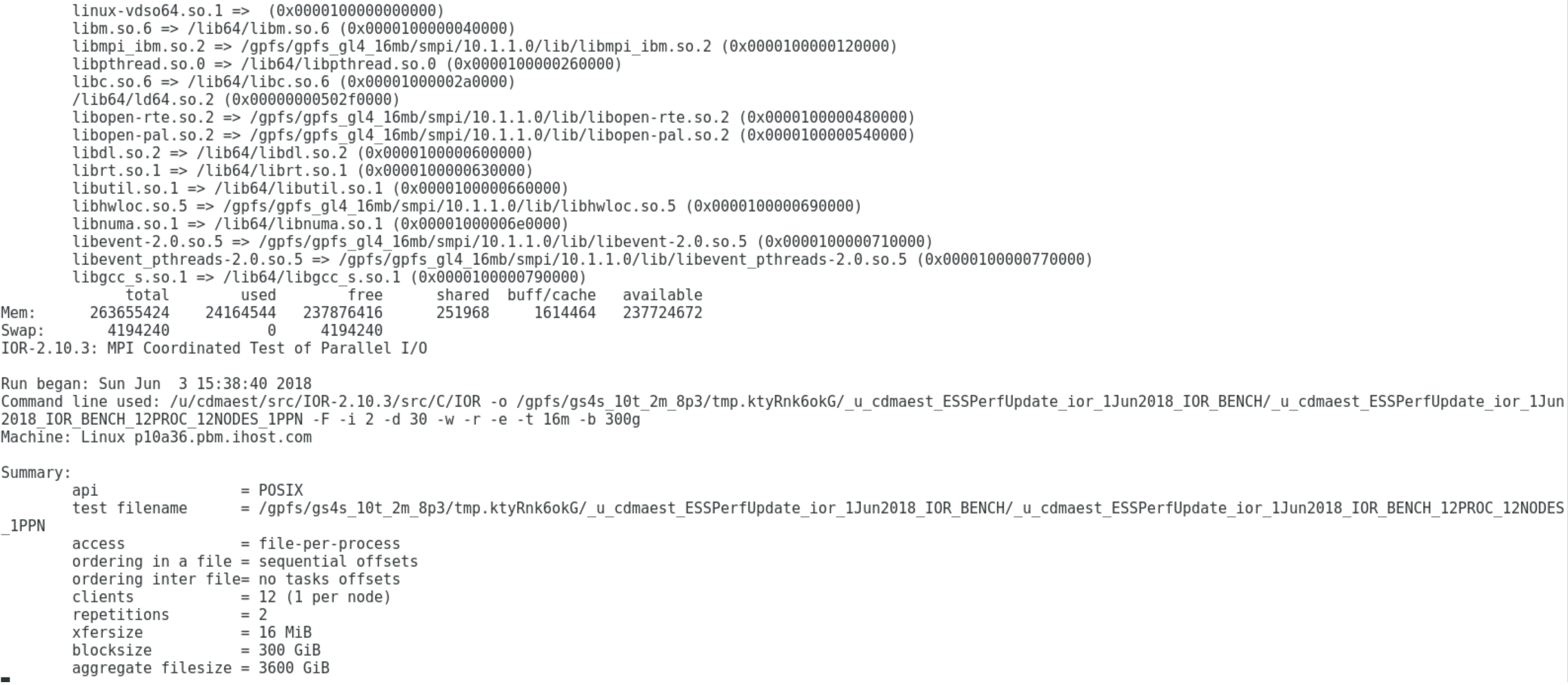
~11GB/s each EDR port (client 2 EDR cards 1 ports)

1. **IOR**_gpfs_gl6s_16mb_bench_12PROC_1NODES_12PPN.stdout.173216: aggregate filesize = 1536 GiB
   IOR_gpfs_gl6s_16mb_bench_12PROC_1NODES_12PPN.stdout.173216:Max Read: **20763.40 MiB/sec (21772.00 MB/sec)**

2. **gpfsperf**_gpfs_gl6s_16mb_bench_8PROC_1NODES_8PPN.stdout.173229: Data rate was **20342078.24 Kbytes/sec**, Op Rate was 1212.48 Ops/sec, Avg Latency was 6.512 milliseconds, thread utilization 0.987, bytesTransferred 322122547200

3. **iozone**_gpfs_gl6s_16mb_bench_8PROC_1NODES_8PPN.stdout.173267: Parent sees throughput for 8 readers = **20938448.48 kB/sec**

IOR run parameters



```
        linux-vdso64.so.1 =>  (0x0000100000000000)
        libm.so.6 => /lib64/libm.so.6 (0x0000100000040000)
        libmpi_ibm.so.2 => /gpfs/gpfs_gl4_16mb/smpi/10.1.1.0/lib/libmpi_ibm.so.2 (0x0000100000120000)
        libpthread.so.0 => /lib64/libpthread.so.0 (0x0000100000260000)
        libc.so.6 => /lib64/libc.so.6 (0x00001000002a0000)
        /lib64/ld64.so.2 (0x00000000502f0000)
        libopen-rte.so.2 => /gpfs/gpfs_gl4_16mb/smpi/10.1.1.0/lib/libopen-rte.so.2 (0x0000100000480000)
        libopen-pal.so.2 => /gpfs/gpfs_gl4_16mb/smpi/10.1.1.0/lib/libopen-pal.so.2 (0x0000100000540000)
        libdl.so.2 => /lib64/libdl.so.2 (0x0000100000600000)
        librt.so.1 => /lib64/librt.so.1 (0x0000100000630000)
        libutil.so.1 => /lib64/libutil.so.1 (0x0000100000660000)
        libhwloc.so.5 => /gpfs/gpfs_gl4_16mb/smpi/10.1.1.0/lib/libhwloc.so.5 (0x0000100000690000)
        libnuma.so.1 => /lib64/libnuma.so.1 (0x00001000006e0000)
        libevent-2.0.so.5 => /gpfs/gpfs_gl4_16mb/smpi/10.1.1.0/lib/libevent-2.0.so.5 (0x0000100000710000)
        libevent_pthreads-2.0.so.5 => /gpfs/gpfs_gl4_16mb/smpi/10.1.1.0/lib/libevent_pthreads-2.0.so.5 (0x0000100000770000)
        libgcc_s.so.1 => /lib64/libgcc_s.so.1 (0x0000100000790000)
              total        used        free      shared  buff/cache   available
Mem:      263655424    24164544   237876416      251968     1614464   237724672
Swap:       4194240           0     4194240
IOR-2.10.3: MPI Coordinated Test of Parallel I/O

Run began: Sun Jun  3 15:38:40 2018
Command line used: /u/cdmaest/src/IOR-2.10.3/src/C/IOR -o /gpfs/gs4s_10t_2m_8p3/tmp.ktyRnk6okG/_u_cdmaest_ESSPerfUpdate_ior_1Jun2018_IOR_BENCH/_u_cdmaest_ESSPerfUpdate_ior_1Jun
2018_IOR_BENCH_12PROC_12NODES_1PPN -F -i 2 -d 30 -w -r -e -t 16m -b 300g
Machine: Linux p10a36.pbm.ihost.com

Summary:
        api                = POSIX
        test filename      = /gpfs/gs4s_10t_2m_8p3/tmp.ktyRnk6okG/_u_cdmaest_ESSPerfUpdate_ior_1Jun2018_IOR_BENCH/_u_cdmaest_ESSPerfUpdate_ior_1Jun2018_IOR_BENCH_12PROC_12NODES
_1PPN
        access             = file-per-process
        ordering in a file = sequential offsets
        ordering inter file= no tasks offsets
        clients            = 12 (1 per node)
        repetitions        = 2
        xfersize           = 16 MiB
        blocksize          = 300 GiB
        aggregate filesize = 3600 GiB
```

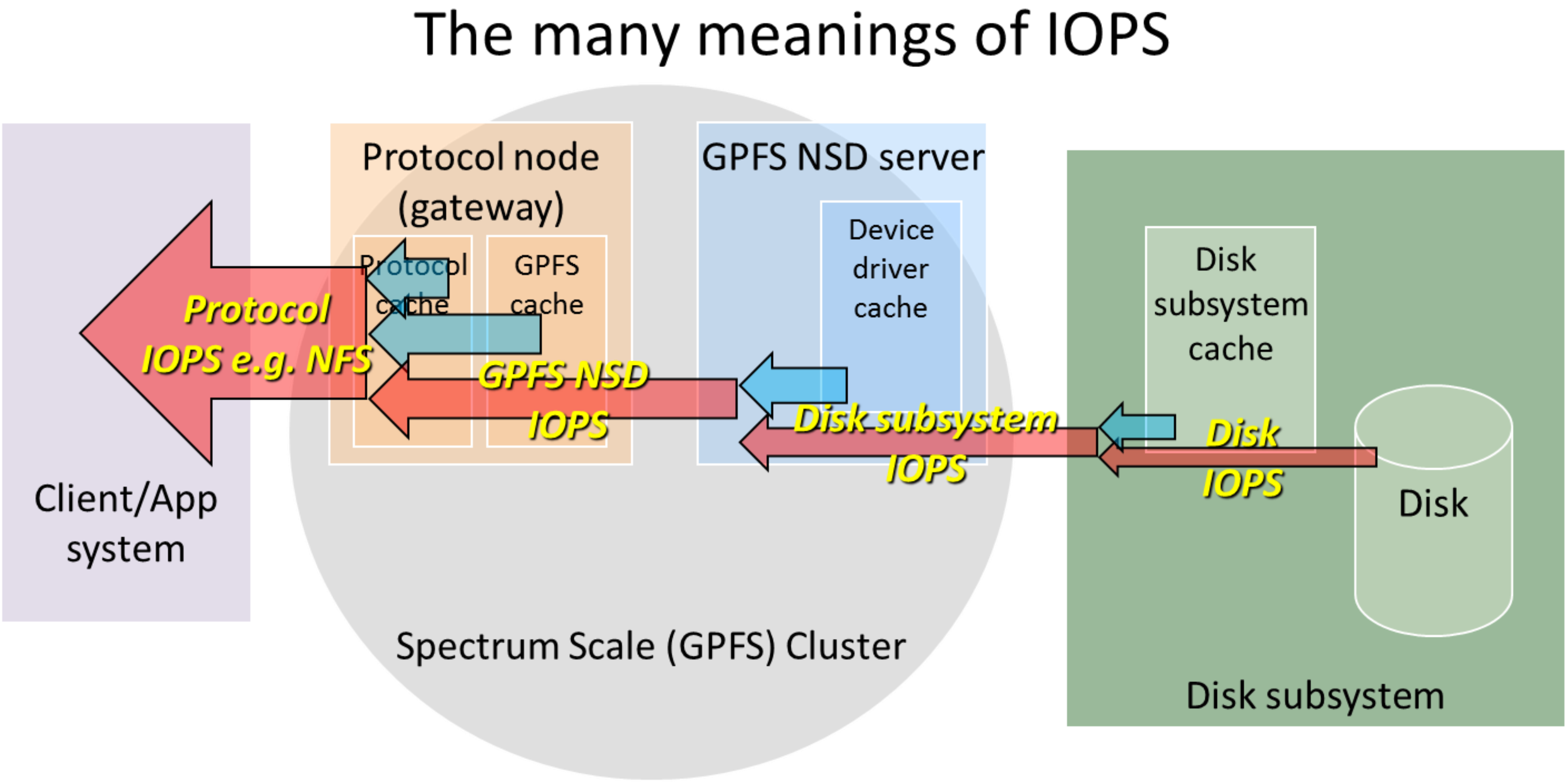
GS4S Bandwidth Summary (GB/sec)
YMMV and remember charts 2-4

Block Size/ Erasure Encoding	1M	2M	4M	8M	16M
GS4S 8+2p READ	35.04427	42.70552	42.56804	39.88963	34.35266
GS4S 8+3p READ	35.81005	43.42365	41.62348	40.15347	38.22962
GS4S 8+2p WRITE	27.98365	30.82226	30.509.48	30.34373	33.19305
GS4S 8+3p WRITE	25.64657	28.17133	29.40512	29.12085	28.25616

GL6S Bandwidth Summary (GB/sec)
YMMV and remember charts 2-4

Block Size/ Erasure Encoding	2M	4M	8M	16M
GL6S 8+2p READ	19.36236	29.67862	36.02717	36.53436
GL6S 8+3p READ	18.97629	28.88162	37.28137	35.66792
GL6S 8+2p WRITE	12.94642	19.77895	26.75490	30.97978
GL6S 8+3p WRITE	11.78215	18.38796	25.78975	29.67814

~~IOPS~~ POSIX Transactions per second!



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

- Measured with IOR different options for
  - Oil and Gas
  - Government

Gathering data to focus on future improvements

Future

Upgrade to new ESS/Scale release

Re-run benchmarks for bandwidth

Cadence with performance team measurements for mdtest

Publish here or somewhere global?

New in IBM Spectrum Scale 5.0.2

Performance!

maxActiveIallocSegs enhancement

A single node has created a large number of files in multiple directories

Processes and threads on multiple nodes are now concurrently attempting to delete or unlink files in those directories.

Configuration parameter– *maxActiveIallocSegs*

Specifies the number of active inode allocation segments maintained

The default value is 8 on file systems that are created at file system format version 5.0.2 or later, otherwise it is 1

- change of this attribute is not effective until after the file system is remounted.
- Not dependent on fs version format

If equal nodes creating and deleting, no BIG difference between 5.0.1 and 5.0.2

maxStatCache enhancement

Spectrum Scale < 5.0.2, the stat cache is not effective on the Linux platform

maxStatCache=0 || LROC (man mmchconfig)

Spectrum Scale >= 5.0.2 stat cache is effective on the Linux platform for all configurations

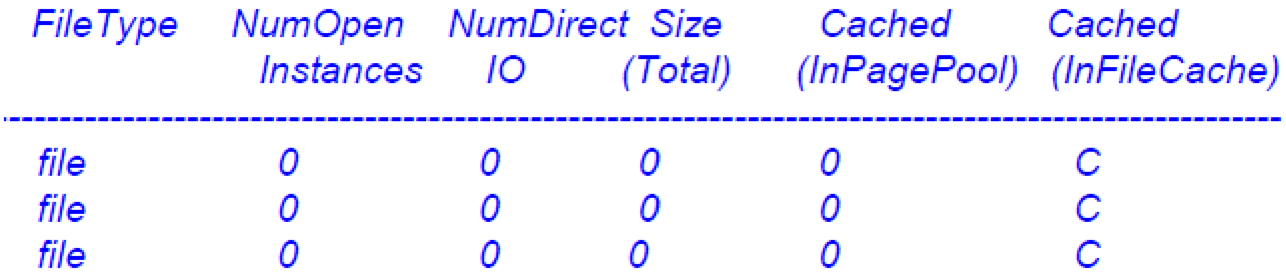
Configuration parameter – maxStatCache
maintains only enough inode information to perform a query on the file system.

file and dir stat operation performance may be improved when the inode is in the stat cache.

If not set, maxStatCache = 4 * maxFilesToCache

“*mmcachectl show*” can be used to verify if file inode is in the stat cache

Commands: ls -l and mdtest
have shown improvement.

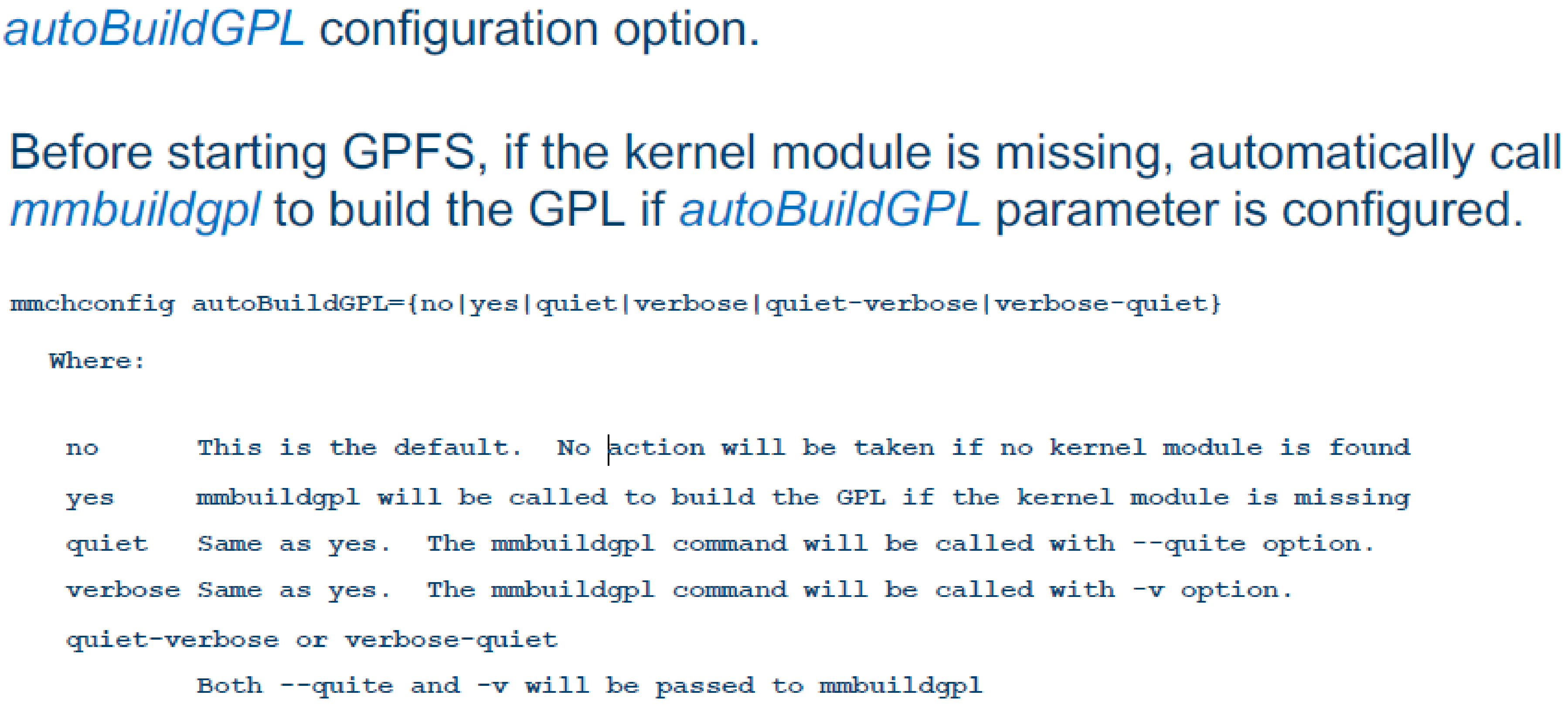


```
FileType    NumOpen     NumDirect  Size      Cached        Cached
             Instances    IO        (Total)   (InPagePool)  (InFileCache)
-----------------------------------------------------------------------------
file          0            0          0         0              C
file          0            0          0         0              C
file          0            0         0          0              C
```

IBM Spectrum Scale 5.0.2

Operational Efficiencies

Rebuild GPL module if new kernel detected

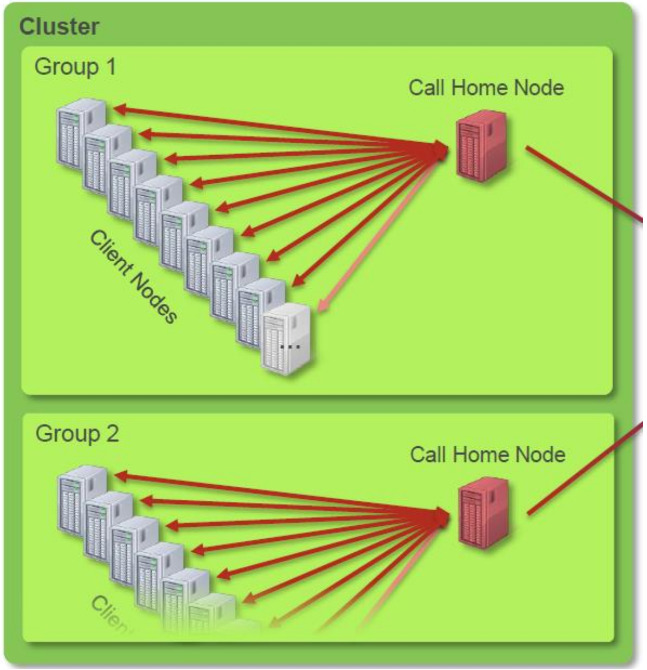


autoBuildGPL configuration option.

Before starting GPFS, if the kernel module is missing, automatically call *mmbuildgpl* to build the GPL if *autoBuildGPL* parameter is configured.

```
mmchconfig autoBuildGPL={no|yes|quiet|verbose|quiet-verbose|verbose-quiet}

  Where:

   no       This is the default.  No action will be taken if no kernel module is found
   yes      mmbuildgpl will be called to build the GPL if the kernel module is missing
   quiet    Same as yes.  The mmbuildgpl command will be called with --quite option.
   verbose  Same as yes.  The mmbuildgpl command will be called with -v option.
   quiet-verbose or verbose-quiet
            Both --quite and -v will be passed to mmbuildgpl
```

Proactive Services - callhome



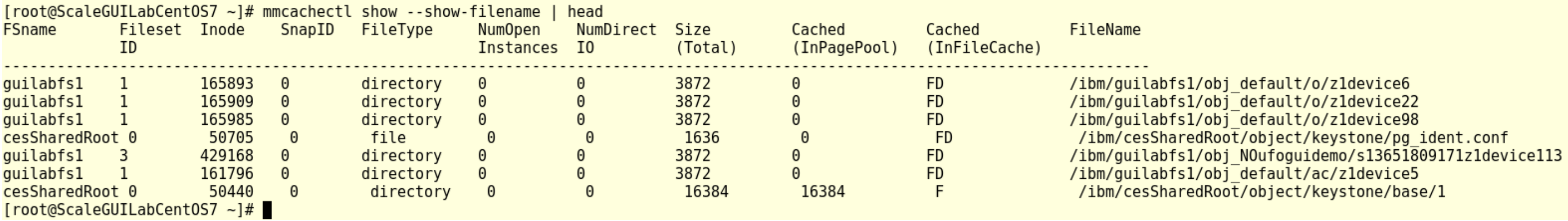
Can group nodes by class

Can find configuration challenges and recommend changes

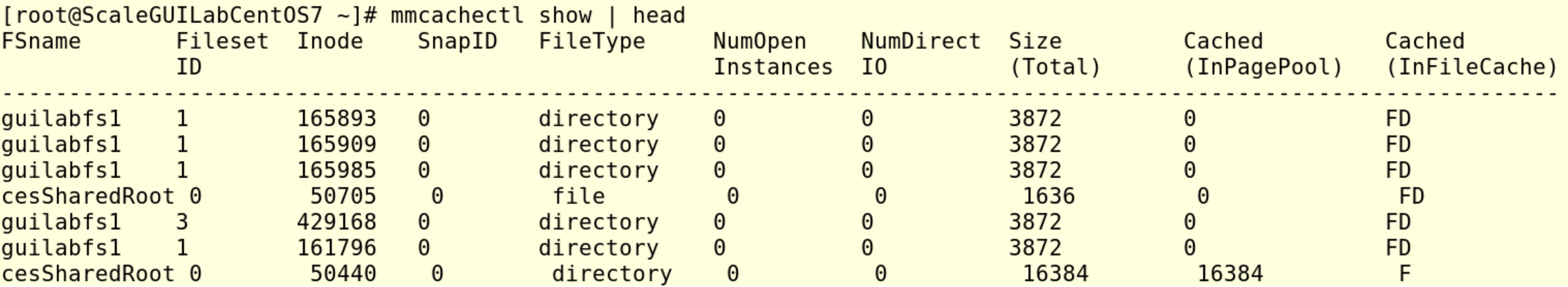
Uses the –Y flag for mm commands

Looking at pagepool - mmcachectl

mmcachectl show –show-filename



```
[root@ScaleGUILabCentOS7 ~]# mmcachectl show --show-filename | head
FSname       Fileset  Inode    SnapID   FileType     NumOpen    NumDirect  Size         Cached         Cached          FileName
             ID                                      Instances  IO         (Total)      (InPagePool)   (InFileCache)
--------------------------------------------------------------------------------------------------------------------------------
guilabfs1    1        165893   0        directory    0          0          3872         0              FD              /ibm/guilabfs1/obj_default/o/z1device6
guilabfs1    1        165909   0        directory    0          0          3872         0              FD              /ibm/guilabfs1/obj_default/o/z1device22
guilabfs1    1        165985   0        directory    0          0          3872         0              FD              /ibm/guilabfs1/obj_default/o/z1device98
cesSharedRoot 0        50705    0        file         0          0          1636         0              FD              /ibm/cesSharedRoot/object/keystone/pg_ident.conf
guilabfs1    3        429168   0        directory    0          0          3872         0              FD              /ibm/guilabfs1/obj_NOufoguidemo/s13651809171z1device113
guilabfs1    1        161796   0        directory    0          0          3872         0              FD              /ibm/guilabfs1/obj_default/ac/z1device5
cesSharedRoot 0        50440    0        directory    0          0          16384        16384          F               /ibm/cesSharedRoot/object/keystone/base/1
[root@ScaleGUILabCentOS7 ~]# █
```

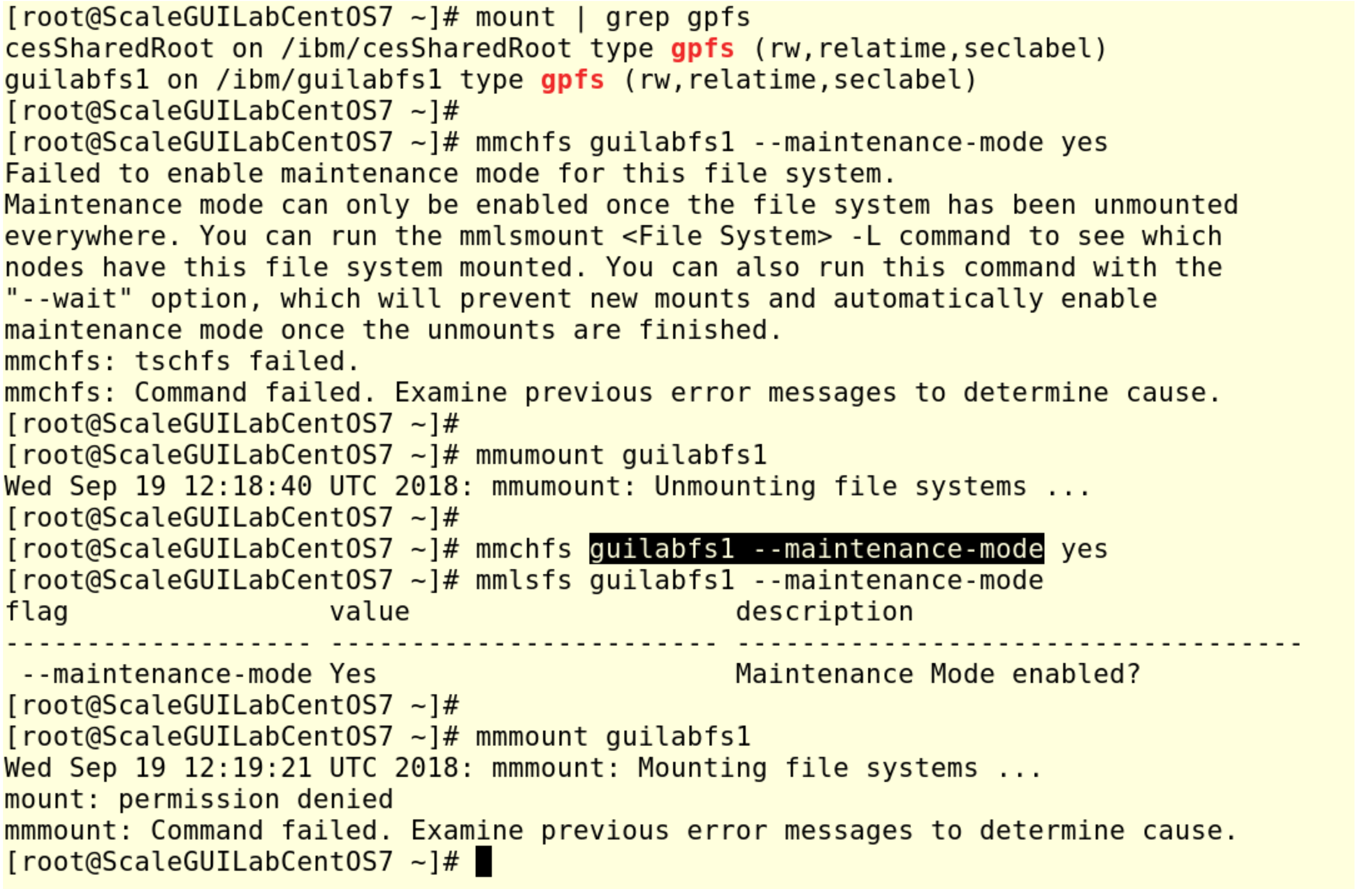


```
[root@ScaleGUILabCentOS7 ~]# mmcachectl show | head
FSname       Fileset  Inode    SnapID   FileType     NumOpen    NumDirect  Size         Cached         Cached
             ID                                      Instances  IO         (Total)      (InPagePool)   (InFileCache)
--------------------------------------------------------------------------------------------------------------------
guilabfs1    1        165893   0        directory    0          0          3872         0              FD
guilabfs1    1        165909   0        directory    0          0          3872         0              FD
guilabfs1    1        165985   0        directory    0          0          3872         0              FD
cesSharedRoot 0        50705    0        file         0          0          1636         0              FD
guilabfs1    3        429168   0        directory    0          0          3872         0              FD
guilabfs1    1        161796   0        directory    0          0          3872         0              FD
cesSharedRoot 0        50440    0        directory    0          0          16384        16384          F
```

File System Maintenance

Create maintenance period on NSD disks, server or entire cluster

Users can still by pass this, but disks may be marked as down and *mmchdisk* to start these down disks could take a long time



```
[root@ScaleGUILabCentOS7 ~]# mount | grep gpfs
cesSharedRoot on /ibm/cesSharedRoot type gpfs (rw,relatime,seclabel)
guilabfs1 on /ibm/guilabfs1 type gpfs (rw,relatime,seclabel)
[root@ScaleGUILabCentOS7 ~]#
[root@ScaleGUILabCentOS7 ~]# mmchfs guilabfs1 --maintenance-mode yes
Failed to enable maintenance mode for this file system.
Maintenance mode can only be enabled once the file system has been unmounted
everywhere. You can run the mmlsmount <File System> -L command to see which
nodes have this file system mounted. You can also run this command with the
"--wait" option, which will prevent new mounts and automatically enable
maintenance mode once the unmounts are finished.
mmchfs: tschfs failed.
mmchfs: Command failed. Examine previous error messages to determine cause.
[root@ScaleGUILabCentOS7 ~]#
[root@ScaleGUILabCentOS7 ~]# mmumount guilabfs1
Wed Sep 19 12:18:40 UTC 2018: mmumount: Unmounting file systems ...
[root@ScaleGUILabCentOS7 ~]#
[root@ScaleGUILabCentOS7 ~]# mmchfs guilabfs1 --maintenance-mode yes
[root@ScaleGUILabCentOS7 ~]# mmlsfs guilabfs1 --maintenance-mode
flag                value                    description
------------------- ------------------------ ----------------------------------
 --maintenance-mode Yes                      Maintenance Mode enabled?
[root@ScaleGUILabCentOS7 ~]#
[root@ScaleGUILabCentOS7 ~]# mmmount guilabfs1
Wed Sep 19 12:19:21 UTC 2018: mmmount: Mounting file systems ...
mount: permission denied
mmmount: Command failed. Examine previous error messages to determine cause.
[root@ScaleGUILabCentOS7 ~]#
```

More network checks and long I/O waits

Check remote clusters

mmnetverify now can check remote clusters for host-name resolution, network connectivity via ping, and GPFS daemon connectivity.

mmnetverify –cluster NAME

nsdperf - bandwidth verification is still recommended

Callbacks for IO hangs (man mmaddcallback)

diskIOHang callback add notification and datacollection scripts to analyze a local I/O request pending for more than 5 minutes.

panicOnIOHang panics the node kernel when a local I/O request pends for more than five minutes.

Remember deadlocks? Don't do it like that right out of the gate!

Estimate an offline mmfsck

- New mmfsck option: --estimate-only
- Displays estimation of offline fsck run time for given mmfsck options
- Does not scan the file system
- Can be run when file system is online or offline
- Works for offline fsck only
- Participating nodes must be at 5.0.2 or later
- The estimate is based on mmfsck command line options, configuration of the target file system and average disk and network I/O throughput of the participating nodes

```
mmfsck fs1 -nv --estimate-only
Checking "fs1"
  FsckFlags                  0x2000009
  ...
Estimating fsck run time
Measuring disk stat...
Measuring RPC stat...
Estimating bytes to scan...
Fsck will complete in 0 hours 0 minutes 58 seconds
(+/- 4 seconds)
Note that this estimate does not factor in any CPU
processing overhead and assumes balanced scan
workload across all threads and nodes
...
File system is clean.
Fsck completed in 0 hours 0 minutes 0 seconds
```

Network Improvements in Spectrum Scale 5.0.2

Network PD improvement – dump the TCP_INFO when disk lease overdue occurs (Linux only)

- Is it a GPFS problem or network problem by looking at fields of TCP_INFO

Network resiliency enhancement - prioritize commMsgCheckMessages RPC to avoid RPC time-out node requested expels

- When sending commMsgCheckMessages RPC could be blocked because of heavy TCP connection (lots of NSD read and write RPC), and if the wait time of exclusive use exceeds 300s, this could cause expel even if the network is good though it's just slow.

Network resiliency enhancement - when CM pings a node near to being expelled, due to a lease timeout, ensure take into account the subnets configuration if set.

- When doing ping check, such as disk lease overdue, current design is to do ping check on the primary address, then cannot detect network problems on the subnets IP address, so check subnets IP address

5.0.2 Spectrum Scale GUI –What's new

- Remote Cluster Capacity data for Filesets and File Systems
- Remote Cluster Quota info
- Node Class Management
- CES IP Health status, Preferred CES nodes and non-hostable nodes exposed
- File Audit Log enable/disable

- Extended Legend in Dashboard views
- More lines in charts (up to 20)
- Cluster Name in banner
- Filtered views by health state
- Enhanced event filtering

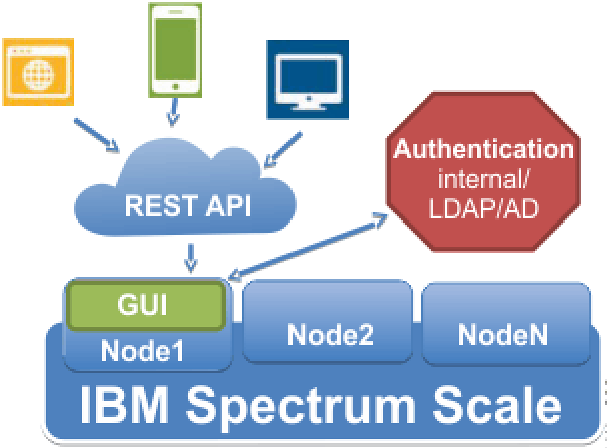
GUI and the REST API

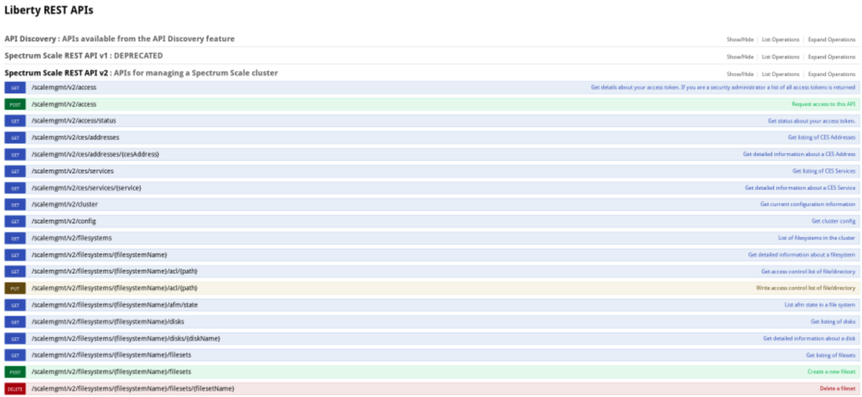
Driven by same WebSphere server

Authentication shared between GUI and REST API

THE strategic interface for integrating with 3rd party customer applications, automation or monitoring

https://[GUI_NODE]:443/ibm/api/explorer/#!/Spectrum_Scale_REST_API_v2/





Liberty REST APIs

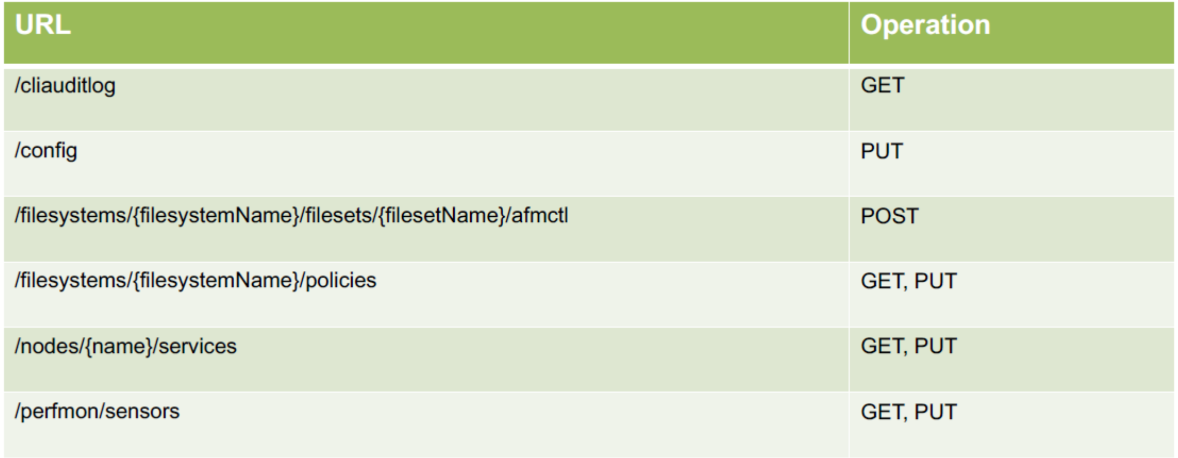
API Discovery : APIs available from the API Discovery feature — Show/Hide | List Operations | Expand Operations

Spectrum Scale REST API v1 : DEPRECATED — Show/Hide | List Operations | Expand Operations

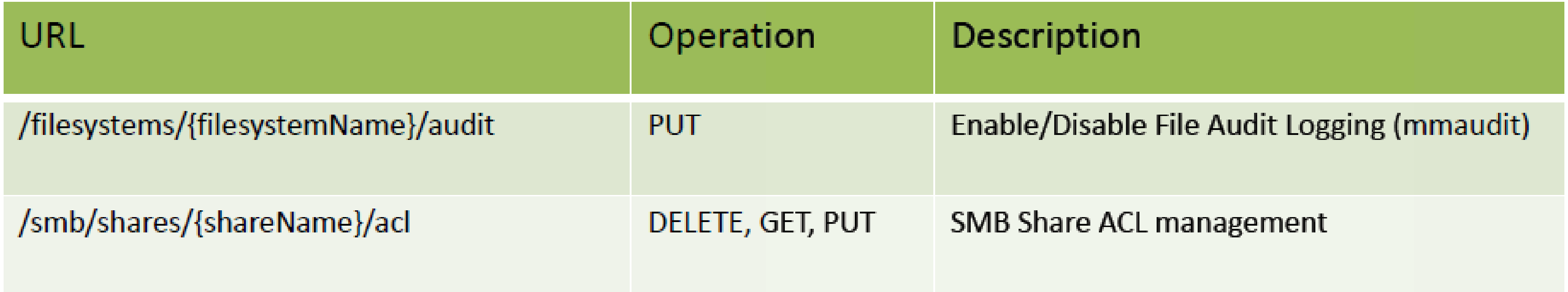
Spectrum Scale REST API v2 : APIs for managing a Spectrum Scale cluster — Show/Hide | List Operations | Expand Operations

GET	/scalemgmt/v2/access	Get details about your access token. If you are a security administrator a list of all access tokens is returned
POST	/scalemgmt/v2/access	Request access to this API
GET	/scalemgmt/v2/access/status	Get status about your access token.
GET	/scalemgmt/v2/ces/addresses	Get listing of CES Addresses
GET	/scalemgmt/v2/ces/addresses/{cesAddress}	Get detailed information about a CES Address
GET	/scalemgmt/v2/ces/services	Get listing of CES Services
GET	/scalemgmt/v2/ces/services/{service}	Get detailed information about a CES Service
GET	/scalemgmt/v2/cluster	Get current configuration information
GET	/scalemgmt/v2/config	Get cluster config
GET	/scalemgmt/v2/filesystems	List of filesystems in the cluster
GET	/scalemgmt/v2/filesystems/{filesystemName}	Get detailed information about a filesystem
GET	/scalemgmt/v2/filesystems/{filesystemName}/acl/{path}	Get access control list of file/directory
PUT	/scalemgmt/v2/filesystems/{filesystemName}/acl/{path}	Write access control list of file/directory
GET	/scalemgmt/v2/filesystems/{filesystemName}/afm/state	List afm state in a file system
GET	/scalemgmt/v2/filesystems/{filesystemName}/disks	Get listing of disks
GET	/scalemgmt/v2/filesystems/{filesystemName}/disks/{diskName}	Get detailed information about a disk
GET	/scalemgmt/v2/filesystems/{filesystemName}/filesets	Get listing of filesets
POST	/scalemgmt/v2/filesystems/{filesystemName}/filesets	Create a new fileset
DELETE	/scalemgmt/v2/filesystems/{filesystemName}/filesets/{filesetName}	Delete a fileset

REST API - Extra endpoints in 5.0.[1,2]



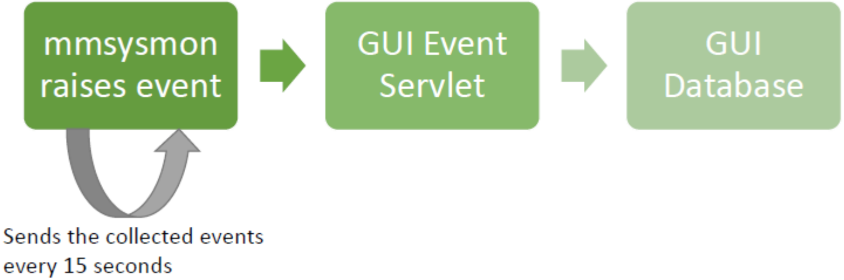
URL	Operation
/cliauditlog	GET
/config	PUT
/filesystems/{filesystemName}/filesets/{filesetName}/afmctl	POST
/filesystems/{filesystemName}/policies	GET, PUT
/nodes/{name}/services	GET, PUT
/perfmon/sensors	GET, PUT



URL	Operation	Description
/filesystems/{filesystemName}/audit	PUT	Enable/Disable File Audit Logging (mmaudit)
/smb/shares/{shareName}/acl	DELETE, GET, PUT	SMB Share ACL management

GUI optimizations

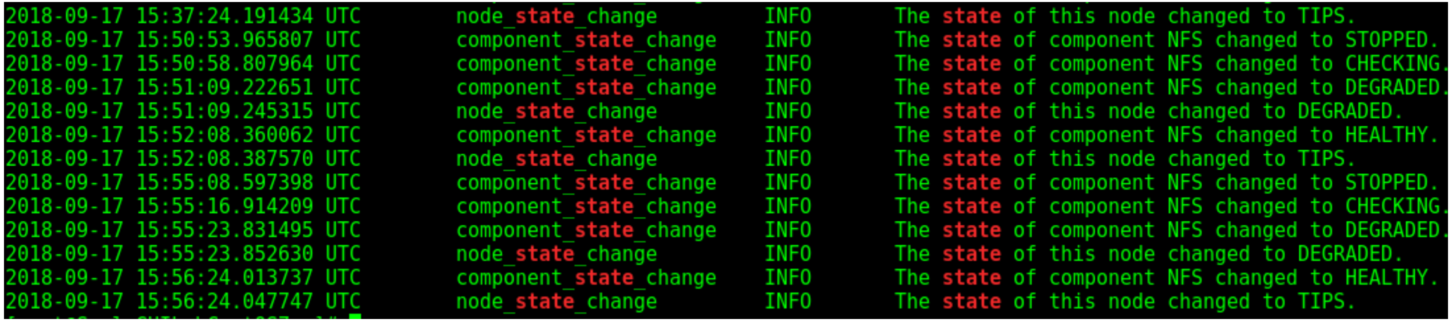
- Reduce call to mmhealth
- Reduce to 2 CPU cores for JAVA and postgres
- Reduce local I/O on GUI node
- Reduce memory on GUI node
- Should help with ESS EMS



System health

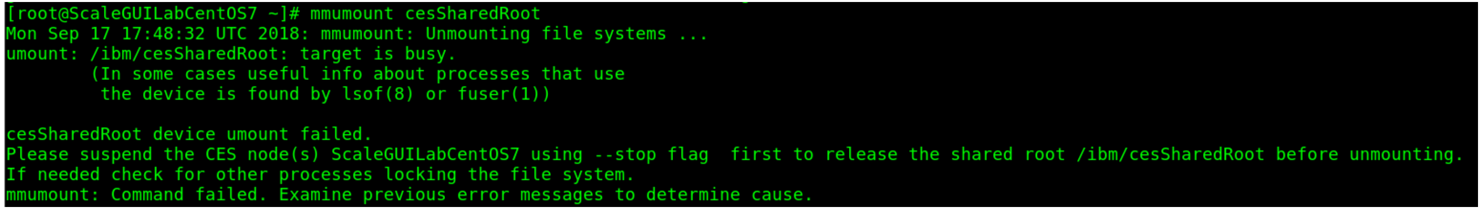
mmces address list can see who is preferred (--extended-list) and who cannot host (-- full-list)

mmhealth --show-state-changes can display state change



```
2018-09-17 15:37:24.191434 UTC        node_state_change        INFO       The state of this node changed to TIPS.
2018-09-17 15:50:53.965807 UTC        component_state_change   INFO       The state of component NFS changed to STOPPED.
2018-09-17 15:50:58.807964 UTC        component_state_change   INFO       The state of component NFS changed to CHECKING.
2018-09-17 15:51:09.222651 UTC        component_state_change   INFO       The state of component NFS changed to DEGRADED.
2018-09-17 15:51:09.245315 UTC        node_state_change        INFO       The state of this node changed to DEGRADED.
2018-09-17 15:52:08.360062 UTC        component_state_change   INFO       The state of component NFS changed to HEALTHY.
2018-09-17 15:52:08.387570 UTC        node_state_change        INFO       The state of this node changed to TIPS.
2018-09-17 15:55:08.597398 UTC        component_state_change   INFO       The state of component NFS changed to STOPPED.
2018-09-17 15:55:16.914209 UTC        component_state_change   INFO       The state of component NFS changed to CHECKING.
2018-09-17 15:55:23.831495 UTC        component_state_change   INFO       The state of component NFS changed to DEGRADED.
2018-09-17 15:55:23.852630 UTC        node_state_change        INFO       The state of this node changed to DEGRADED.
2018-09-17 15:56:24.013737 UTC        component_state_change   INFO       The state of component NFS changed to HEALTHY.
2018-09-17 15:56:24.047747 UTC        node_state_change        INFO       The state of this node changed to TIPS.
```

When unmounting CES FS, error if CES services are running



```
[root@ScaleGUILabCentOS7 ~]# mmumount cesSharedRoot
Mon Sep 17 17:48:32 UTC 2018: mmumount: Unmounting file systems ...
umount: /ibm/cesSharedRoot: target is busy.
        (In some cases useful info about processes that use
         the device is found by lsof(8) or fuser(1))

cesSharedRoot device umount failed.
Please suspend the CES node(s) ScaleGUILabCentOS7 using --stop flag  first to release the shared root /ibm/cesSharedRoot before unmounting.
If needed check for other processes locking the file system.
mmumount: Command failed. Examine previous error messages to determine cause.
```

Install Toolkit 5.0.2 New Features

Recall install toolkit introduced in 4.1.1.0

Mark nodes offline during upgrade
Do an offline upgrade for entire cluster
Exclude nodes from upgrade
(upgrade subset of nodes)
Resume a previously failed upgrade

Enhanced node listing
(NSD, client, protocol, audit, callhome …)

Enhanced CES shared root creation and detection (populate)

Ability to specify broker nodes for File Audit Logging

Removal of gpfs.ext on upgrade
(consolidated into gpfs.base)
(works with rpm/yum update too)

Upgraded chef for orchestration

Support Ubuntu 18.04 and 18.04.1
s390x installation support

Watch Folder installation
(via key enablement)

Windows 10 support! Pro and Enterprise

Both heterogeneous and homogeneous clusters
Currently, Secure Boot must be disabled on Windows 10 nodes

FAQ update: **A14.7: Windows 10 related advisories and recommendations:**

1. *User Access Control (UAC) must not be disabled on latest Windows versions such as Windows 10.* GPFS now runs with UAC enabled (default OS setting).

2. Latest versions of Windows such as Windows 10 now come with a built-in antivirus component known as **Windows Defender**. While performing real-time scanning of files, Windows Defender may memory-map these files even when they are not in use by any user application. This memory-mapping of files on GPFS filesystems by Windows Defender in the “background”, can result in performance degradation. *Therefore, it is recommended that GPFS drives/volumes be “Excluded”from Windows Defender scans all together.*

3. Windows 10 version 1803, now incorporates a native secure shell ‘**OpenSSH** for Windows’. GPFS requires 'OpenSSH for Cygwin', especially if the Windows node(s) join a GPFS cluster having Linux/AIX nodes. Therefore, before operating a Windows 10 node in a mixed GPFS cluster, please ensure that the Windows native 'OpenSSH SSH Server' is not enabled/running and that the 'Cygwin sshd' service is working reliably. Additionally, it is recommended that the Windows Subsystem for Linux (WSL) feature not be installed to avoid potential conflicts with Cygwin.

IBM Spectrum Scale 5.0.2

Other Protocols

“mmuserauth” enhancement for password

Example for FILE authentication

mmuserauth service create --type ad --data-access-method file --netbios-name test --user-name administrator --idmap-role master --servers myADServer **--pwd-file fileauth**

Contents of fileauth saved at /var/mmfs/ssl/keyServ/tmp/ are:

```
%fileauth:
password=Passw0rd
```

Example for OBJECT authentication

mmuserauth service create --type ad --data-access-method object --base-dn "dc=example,DC=com" --servers myADserver --user-id-attrib cn --user-name-attrib sAMAccountName --user-objectclass organizationalPerson --user-dn "cn=Users,dc=example,dc=com" --pwd-file objectauth

Contents of fileauth saved at /var/mmfs/ssl/keyServ/tmp/ are:%objectauth:

```
password=Passw0rd
ksAdminPwd=Passw0rd1
ksSwiftPwd=Passw0rd2
```

For FILE authentication now validates DNS records to AD severs as well

Samba update

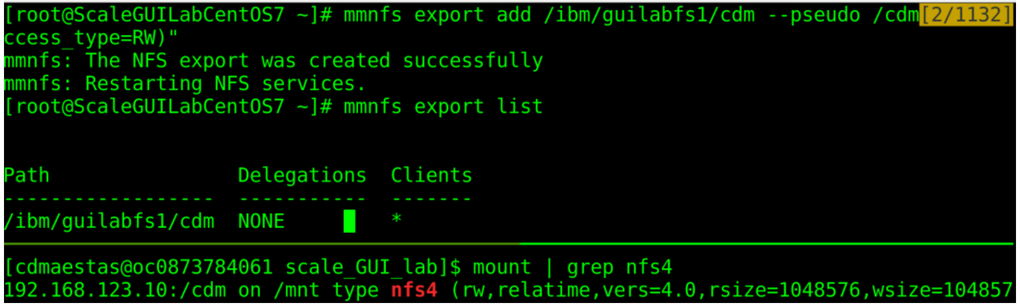
- Allow user to change min and max SMB protocols
- Reduce load on cache generation if a lot of idmap lookups occur
- Graceful behavior of ctdb during OOM
    - Log memory, change to unhealthy if swap > 95% used

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

Ganesha NFS update

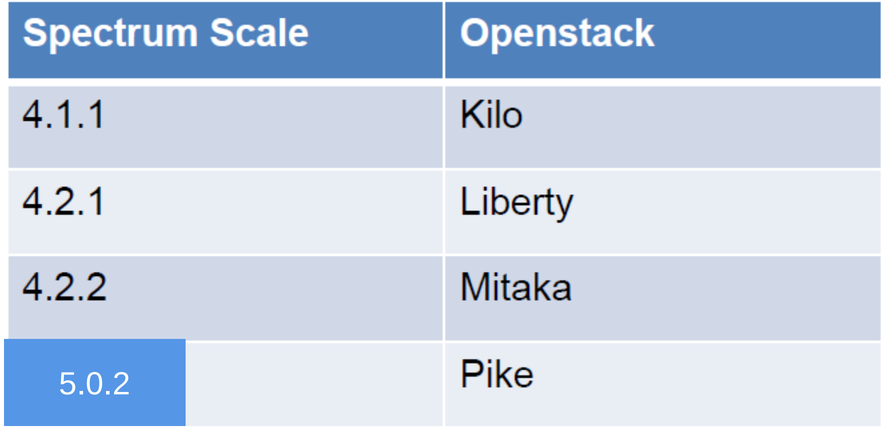
- Restructure code to “maybe” support more exports per filesystem
- Pseudo path for export at creation time*
- Performance counters (ganesha_stats)*

Integration with mm and GUI coming soon

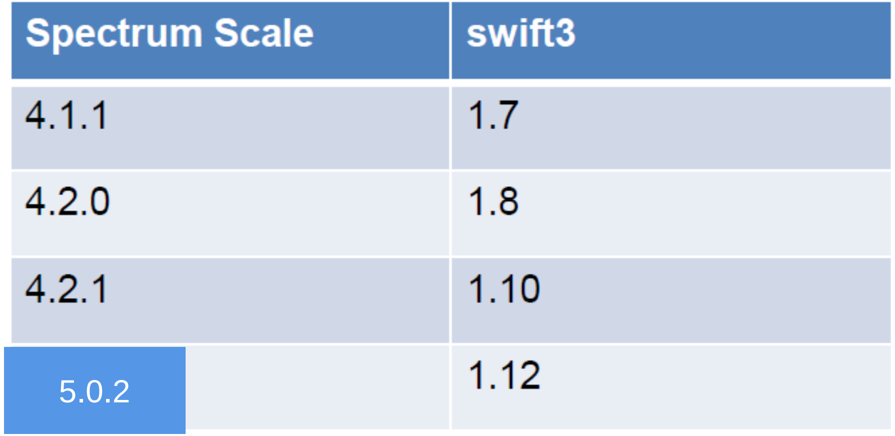


```
[root@ScaleGUILabCentOS7 ~]# mmnfs export add /ibm/guilabfs1/cdm --pseudo /cdm[2/1132]
ccess_type=RW)"
mmnfs: The NFS export was created successfully
mmnfs: Restarting NFS services.
[root@ScaleGUILabCentOS7 ~]# mmnfs export list


Path                Delegations  Clients
------------------  -----------  -------
/ibm/guilabfs1/cdm  NONE      █  *

[cdmaestas@oc0873784061 scale_GUI_lab]$ mount | grep nfs4
192.168.123.10:/cdm on /mnt type nfs4 (rw,relatime,vers=4.0,rsize=1048576,wsize=104857
```

Object Release Overview



Spectrum Scale	Openstack
4.1.1	Kilo
4.2.1	Liberty
4.2.2	Mitaka
5.0.2	Pike



Spectrum Scale	swift3
4.1.1	1.7
4.2.0	1.8
4.2.1	1.10
5.0.2	1.12

Amazon! - **http://ibm.biz/ScaleAWS**

AWS Quick Starts

IBM Spectrum Scale on AWS

High-performance storage solution for managing data at scale

Two models of deployment with **a good deployment guide!**

virtual private cloud (VPC) that spans ***two*** Availability Zones in your AWS account.

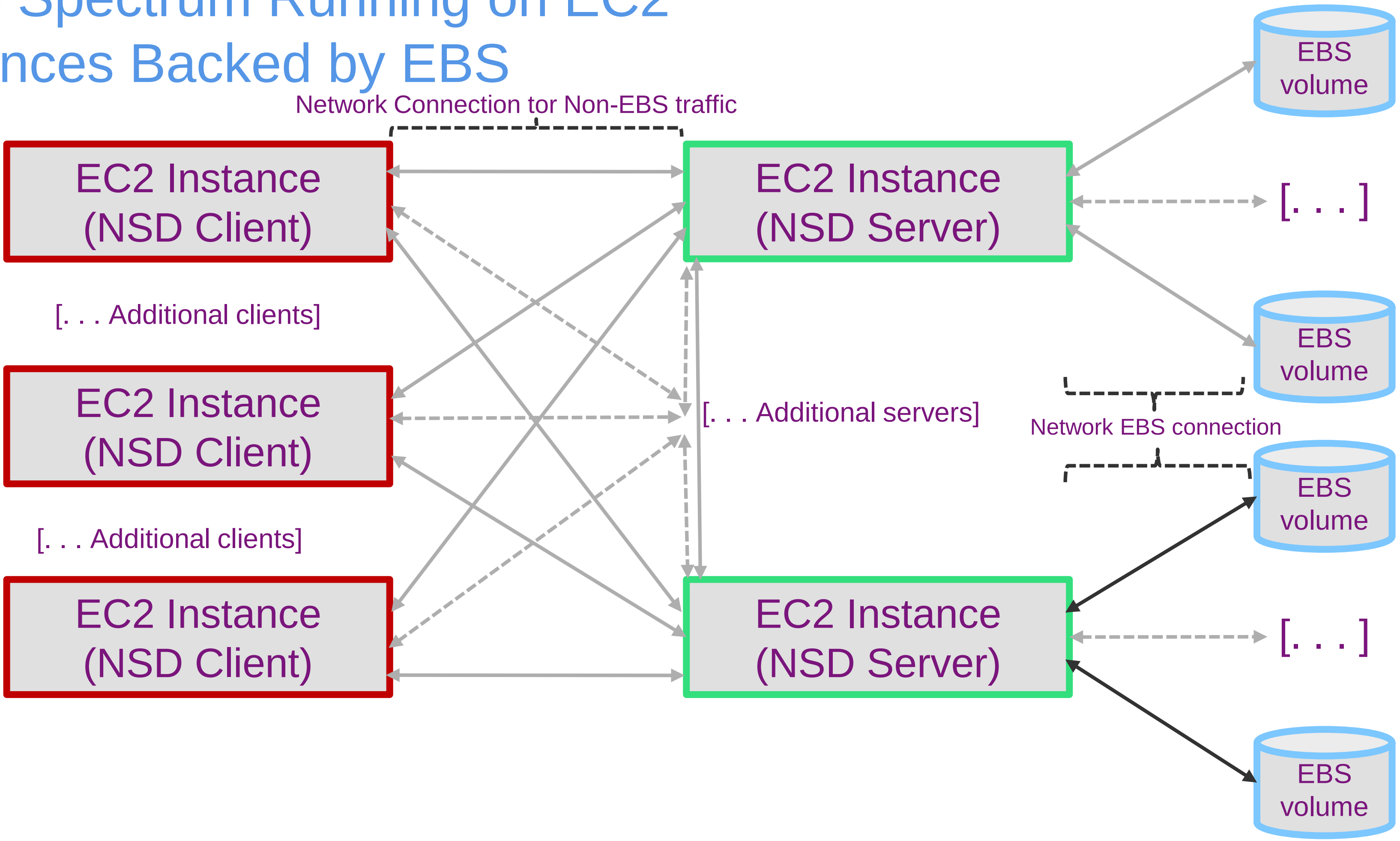
- Can build a new VPC for IBM Spectrum Scale, or
- Deploy the software into your existing VPC

Deployment and configuration tasks are automated by AWS CloudFormation templates

- Customizable prior to launch

AWS Spectrum Running on EC2 Instances Backed by EBS

IBM Storage & SDI


Network Connection tor Non-EBS traffic
EC2 Instance (NSD Client)
[. . . Additional clients]
EC2 Instance (NSD Client)
[. . . Additional clients]
EC2 Instance (NSD Client)
EC2 Instance (NSD Server)
[. . . Additional servers]
EC2 Instance (NSD Server)
EBS volume
[. . .]
EBS volume
Network EBS connection
EBS volume
[. . .]
EBS volume

mmaws - Managing IBM Spectrum Scale workflows on AWS

```
Usage:

    mmaws add_nodes      Add compute/server nodes

    mmaws remove_nodes        Removing compute/server nodes

    mmaws list_instances      Listing instances in the vpc

    mmaws start_nodes         Starting nodes

    mmaws stop_nodes     Stopping compute/all nodes

    mmaws create_lambda_functions   Create Lambda functions

    mmaws collect_debug_data    Collect AWS debug data


optional arguments:

-?, -h, --help, help          show this help message and exit
```

New in IBM Spectrum Scale 5.0/5.0.1

Security and Compliance!

DEFAULTNISTSP800131AFAST Encryption enhancement

DEFAULTNISTSP800131AFAST uses 128-bit key length and 128-bit keys are secure according to NIST publication SP 800.131A.

DEFAULTNISTSP800131AFAST can provide 5-20% speed up for certain I/O workloads (e.g. large block random reads, direct I/O) compared to DEFAULTNISTSP800131A

Encryption ALGO value – DEFAULTNISTSP800131AFAST
Maps to 'AES:128:XTS:FEK:HMACSHA512'

Sample Encryption policy

```
RULE 'EncPolicyGeneratorRule2' ENCRYPTION 'EncPolicyGenerator2' IS
ALGO 'DEFAULTNISTSP800131AFAST‘
KEYS('KEY-ABC..XYZ:sklmnRKM')
RULE 'EncPolicyGeneratorFileRule2' SET ENCRYPTION 'EncPolicyGenerator2‘
FOR FILESET('encryptedFSet_FAST_NIST')
```

For I/O > 2 MiB Write (> 15%) and Read (> 3%) performance is faster versus **DEFAULTNISTSP800131A**

Alert for Certificate Expiration in keystore

Problem: Spectrum Scale does not alert when client or key-server certificate in keystore is going to expire

Solution: Periodically check validation of all certificates in keystore. (Including client and key server certificates); generate alert and dump it into GPFS log when detect coming expiration, for example, in next 6 month.

Watch Folders 101 - /usr/lpp/mmfs/samples/util/tswf.C

Take actions based on filesystem events

- Run against folders, filesets (independent too)
- Modeled after Linux inotify, but works with clustered filesystems, and supports recursive watches for filesets (independent too)

2 primary components

- GPFS API (included within <gpfs_watch.h>)
- **mmwatch**– provides information of all watches running within cluster

A watch folder application uses API as a C program on cluster

- Utilizes message queue to receive events from multiple nodes and consume from the node running the program
- Events come in from all eligible nodes within cluster and from accessing clusters

Limitations and Requirements #include <gpfs_watch.h>

- **Requires key enablement in 5.0.2**
  - Development/Sales will provide approved use cases with a hidden configuration variable
- All Clusters and file system format
  - code level >= 5.0.2
- Message queue must be enabled on owning cluster of filesystem
  - Minimum 3 Linux quorum nodes and 3 nodes for brokers
  - Data Management Edition (DME)
    (yes advanced too)
- 25 watches per file system
  - 3 GB per watch of local disk space per watch
- 100 watches per cluster

Watch Folder Troubleshooting

mmwatch –
verify information about all currently running watches

/var/adm/ras/mmwf.log –
primary log file for watch API and mmwatch command

/var/adm/ras/mmfs.log
(major problems with policy, watches, etc.)

/var/adm/ras/mmmsgqueue.log
(problems with the message queue)

Watch Folder Performance

Streaming I/O is fine

Lots of reads (70/30) is fine

Lots of metadata performance, it depends

New in IBM Spectrum Scale 5.0.2

Data Movement (Compression, AFM and TCT)

Advanced File Management (AFM) enhancements

AFM Performance improvements:

- User defined gatway mapping with afmHashVersion=5

  Assign at fileset create or modify after

  afmGateway=NODENAME

AFM prefetch enhancements:

- Get statistics of transfer during pre-fetch
- --enabled-failed-file-list
- --retry-failed-file-list
- **--directory # build a list!**
- **--policy # policy syntax**

Transparent Cloud Tiering enhancements

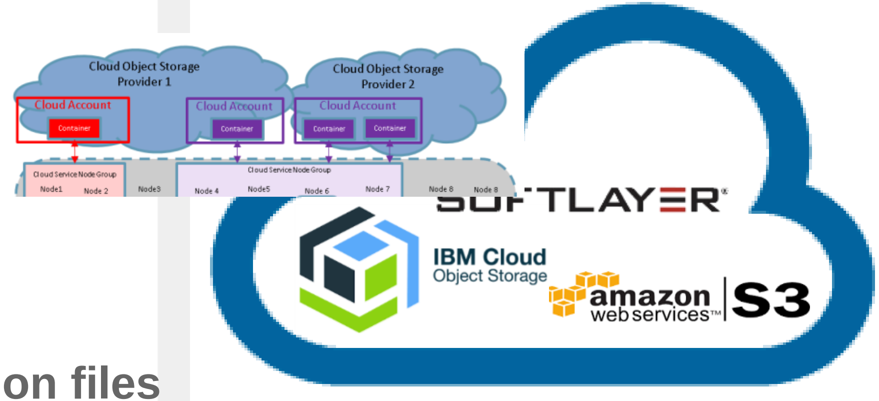
Support all IBM Storage

Remote mounted filesystem support

Clients can access tiered files on a remotely mounted filesystem

Ability to tier different filesets to different cloud containers

Yes, can now be **fileset** focused!



Enhanced support for multiple cloud accour containers

Pull and push to different cloud providers

Container spillover in same fileset > 100 Million files

Big Data and Analytics Enhancements

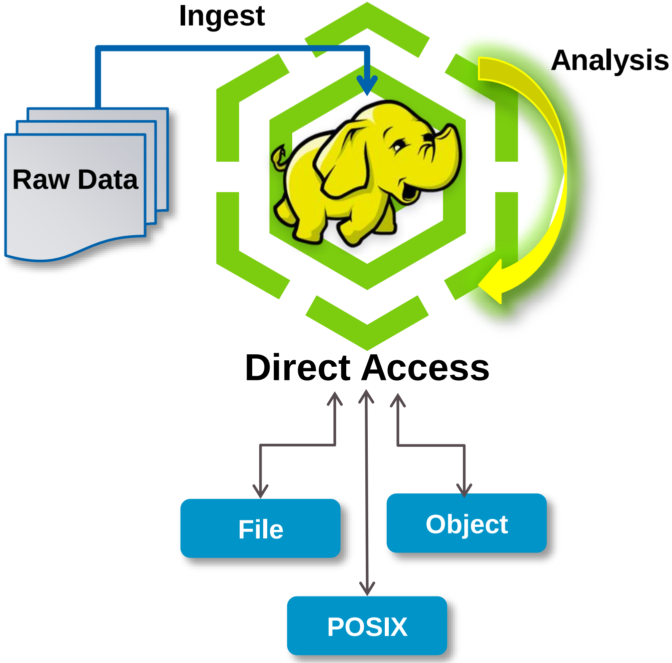
FPO v5.0.2

Try and resume suspended disks if requested

Check for replica mismatch mmrestripefile -c ***--read-only***

HDFS Transparency v3.0.0-0 GA

- Supports HDP 3.0 and Mpack 2.7.0
- Supports Apache Hadoop 3.0.x
- Support native HDFS encryption
- Spectrum Scale Configuration now in:
  - /var/mmfs/hadoop/etc/
  - /var/log/transparency



Thank You.

IBM Storage & SDI