HUK-COBURG

Backup of Hadoop Clusters with Spectrum Scale, Spectrum Archive and Tape at HUK-Coburg (Implementation Status)

Coburg, 04.03.2020

HUK-COBURG

Agenda:

- **Introduction HUK-Coburg**
- **Use-Case Definition**
- **Chosen HW & SW Solution**
- **Time-line/Solution Experiences**
- **What‘s Next**

The parent company HUK-COBURG a.G.

Mutual
insurance company

Operating in the
insurance sector for
85
years

only for
public servants

Principle of mutuality

The company is owned by the insured persons

The object of the company is solely geared to the interests of the insured persons.

Germany's largest insurer
for public servants
3.6
million members

The entire group works
according to the
Principle
of mutuality

Largest german motor insurer

with more than 11,6 million insured vehicles

traditionally offering favourable prices

12 million customers

a large insurer for private households

Second place

in legal expenses insurance for private households

Health insurance & Life insurance

low costs, low lapse rates, high benefits

First place

in personal liability and home contents insurance

Being close to the customer is of utmost importance to us ...

Flexible sales channels
free choice for our customers

About 2,500 permanent employees in 38 branch offices
Expert advice, underwriting and

About 680 customer service offices: self-employed full-time field service

Around 3,000 local part-time “trusted counsellors”
regional near our customers

More than 100 consultants
specialising in life, health and accident insurance

Customer service centres
with highly qualified experts for questions regarding different classes of insurance

Comprehensive range of insurance products available on the internet our online-only subsidiary HUK24 offers particularly

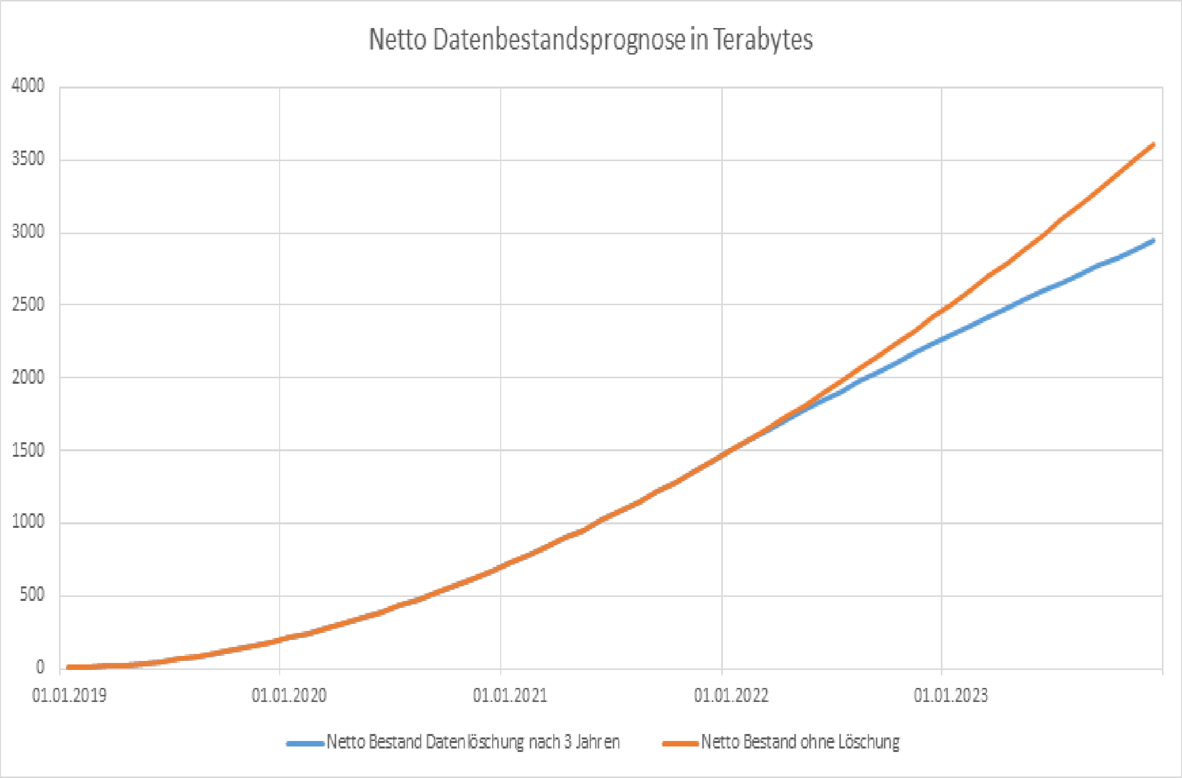
insurers in the field of churches
about 380 full-time customer contact persons as well as 1,400 trusted

Use Case Definitions and Requirements

- Creation of a Backup-Interface to all hadoop-clusters
- special request for our telematik use case
- Tiering of hadoop data
- Implementation of cost effective storage media

Requirements hadoop-Backup

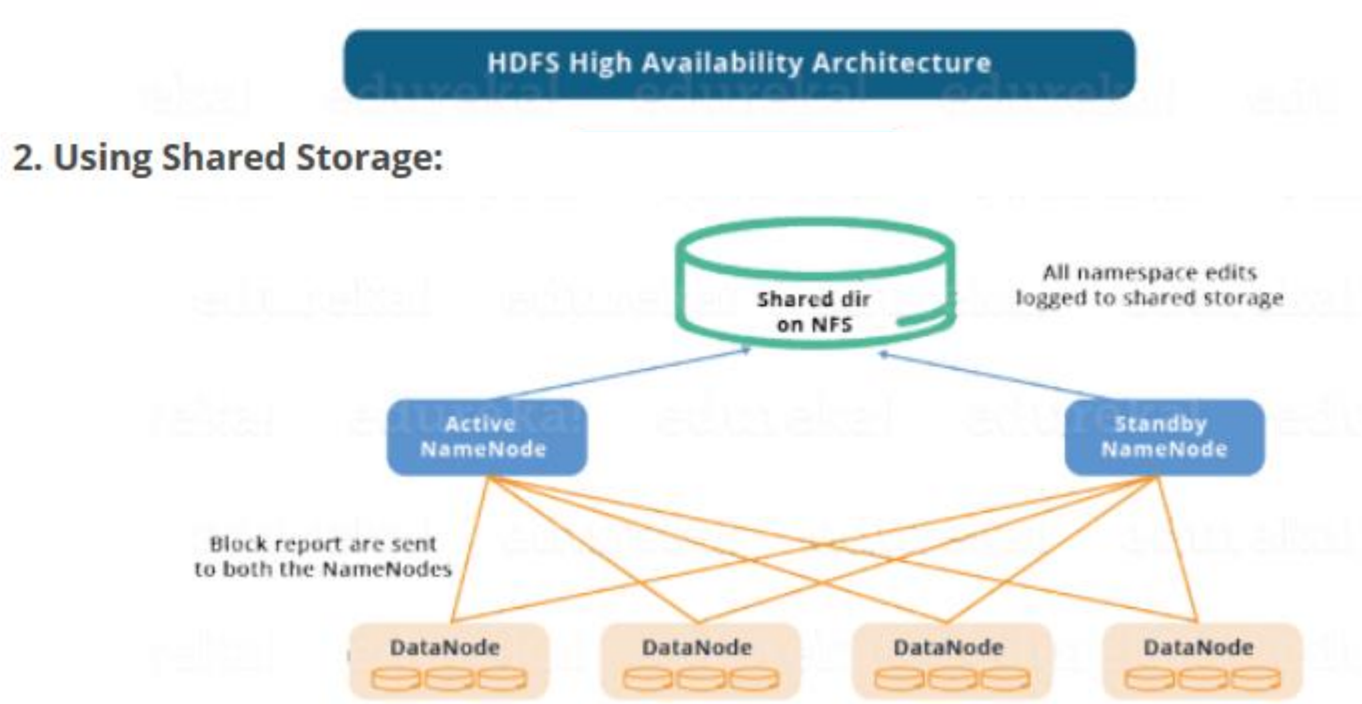
- Backup-Requirments
  - daily fullbackup of Hbase scoring
  - daily differential backup of TripData with media break



Hadoop Features in use

- Name-Node HA
- Namenode Federation with usage of discp from source-clusters
- Ambari only for Handling Kerberos and Zookeeper
- Transparency Connector in Version gpfs.hdfs-protocol-3.1.0-4.x86_64

Hadoop HA-Features Overview traditional vs. Scale transparency

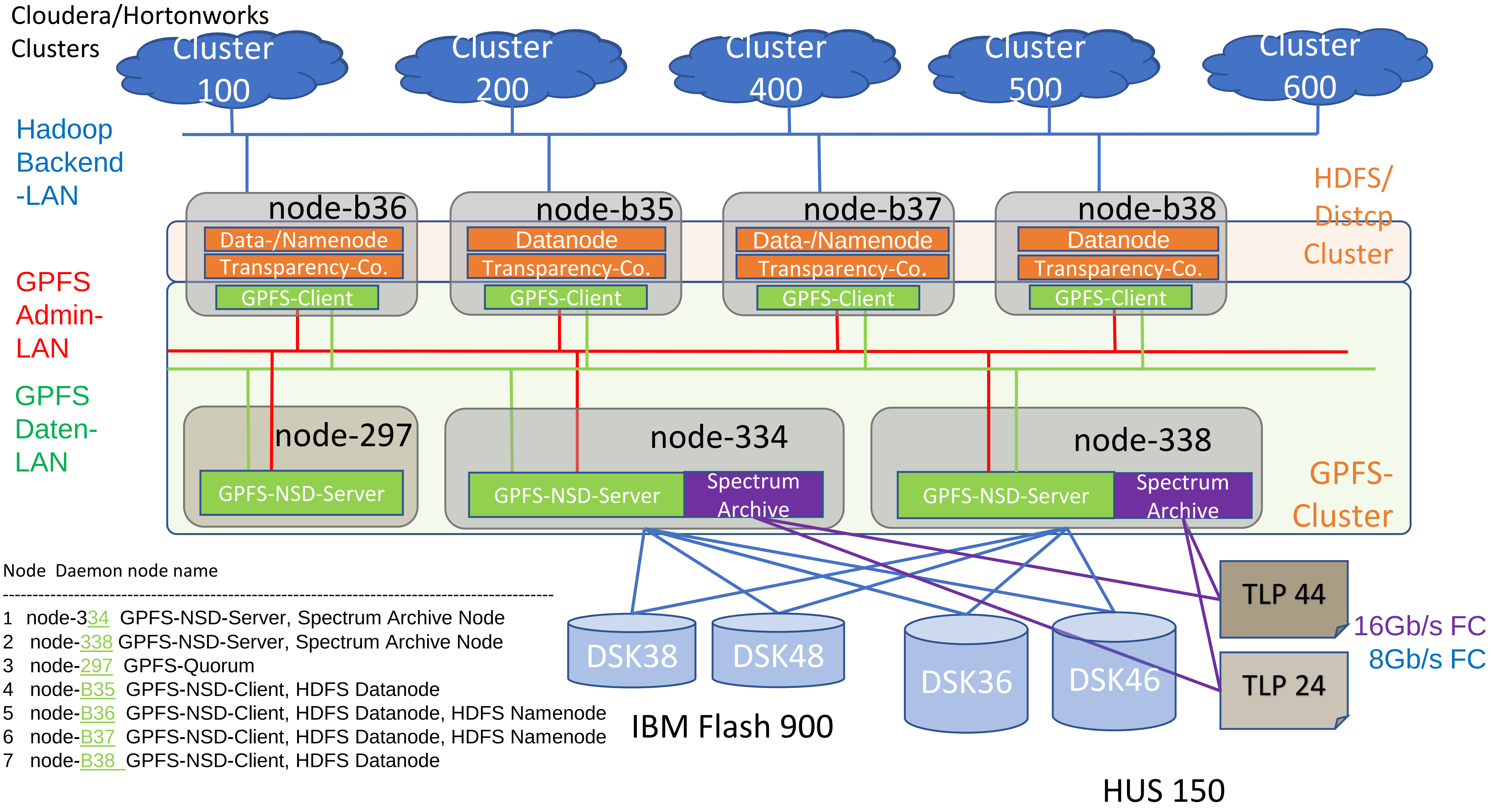


HDFS High Availability Architecture

2. Using Shared Storage:

Chosen HW & SW Solution

Architecture Overview- Overview



Node Daemon node name

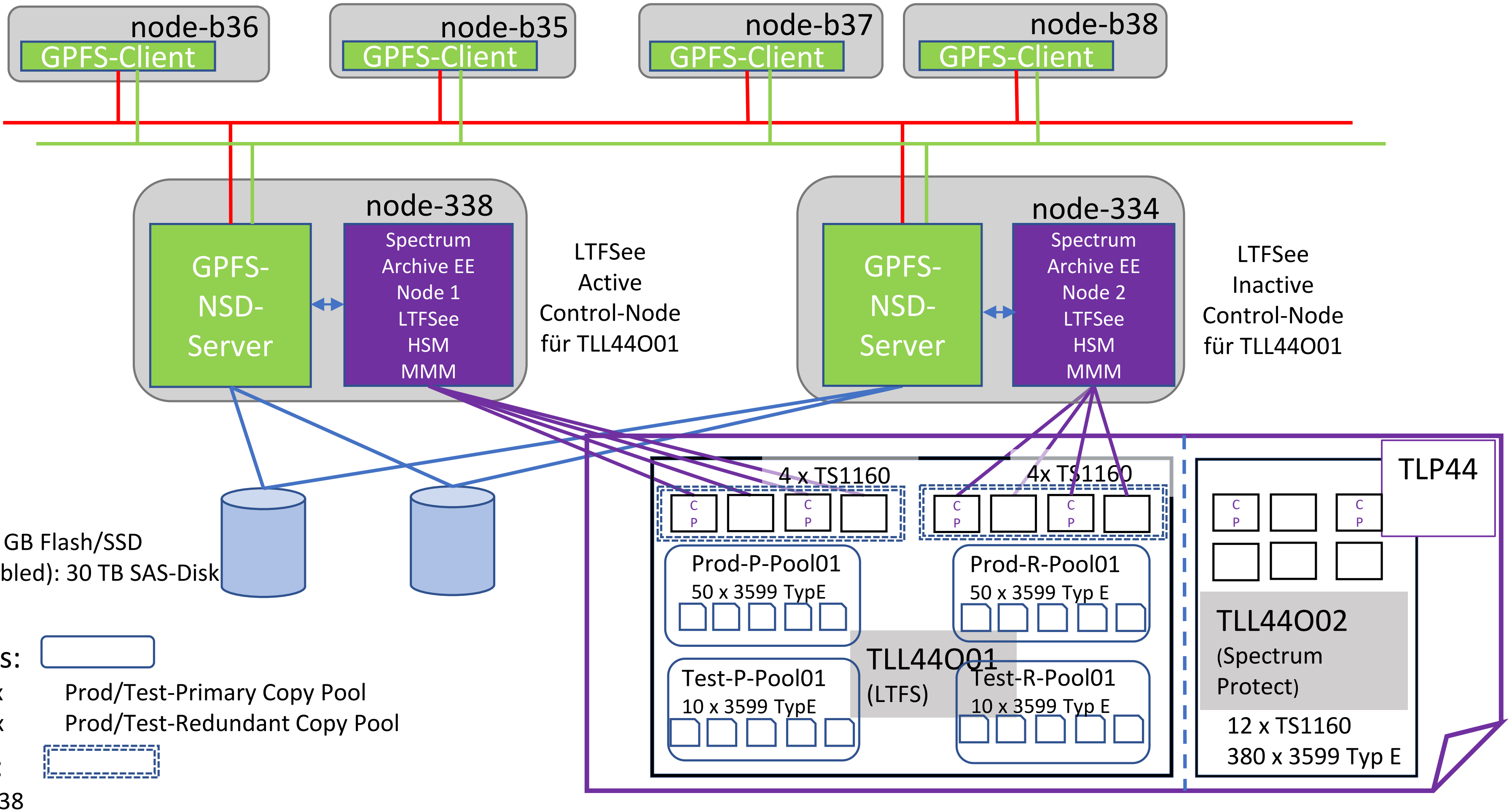
---

1 node-334 GPFS-NSD-Server, Spectrum Archive Node
2 node-338 GPFS-NSD-Server, Spectrum Archive Node
3 node-297 GPFS-Quorum
4 node-B35 GPFS-NSD-Client, HDFS Datanode
5 node-B36 GPFS-NSD-Client, HDFS Datanode, HDFS Namenode
6 node-B37 GPFS-NSD-Client, HDFS Datanode, HDFS Namenode
7 node-B38 GPFS-NSD-Client, HDFS Datanode

Chosen HW & SW Solution

Architecture Overview- Details

Hadoop-
Backup Cluster



Filesystems:

ltfsmeta01: 2 x 500 GB Flash/SSD

hdfs01 (DMAPI enabled): 30 TB SAS-Disk

LTFS-Tape-Pools: ▭

Prod/Test-P-Pool0x	Prod/Test-Primary Copy Pool
Prod/Test-R-Pool0x	Prod/Test-Redundant Copy Pool

Tapedrive Sets: ⬚

4 x TS1160 SAP00338

4 x TS1160 SAP00334

TS4500 [3 Frames, HA, 1100 Slots]

Chosen HW & SW Solution

Evaluation period

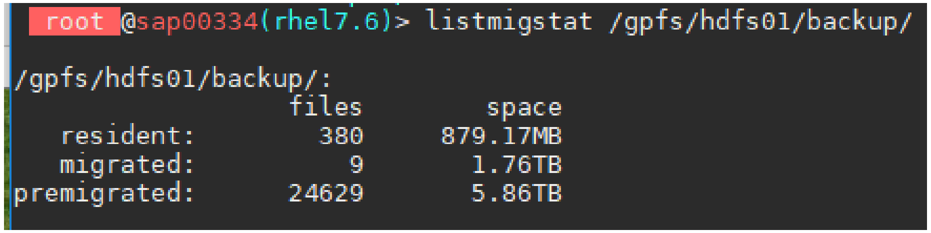
- POC since Feb. / April 19
  - Test categories
    - Basic Functions
    - Monitoring
    - Availability
    - Performance
    - Maintenance and Operations
  - Tests are documented in a spreadsheet with column headers test-categorie, test-number, test-name, test-description, test-expectation, test-result and test-status
- Production since Aug. 2019

Test Results in our Evaluation period

- Basic Functions
  - 9 of 12 defined tests performed
  - migrate, recall via eeadm or via GPFS-Policy and something like that
  - main features premigration via threshold or scheduled premigration were successful like all other performed tests here
- Monitoring
  - hit `lowspacewarningthreshold` to provoke SNMP-trap
- Maintenance and operational tests have also passed with some issues
- Some performance tests were defined, but serveral infrastructure tweakings are still going on
- Availability
  - performed 3 of 4 defined tests
  - only „broken Tape“ test passed

Test Results so far

- Availability tests not passed
  - Test „Drive Failure“ - task was aborted, issue LIP (Loop Initialization Protocol) for each HBA was required to recognize the drive
  - Test “Library Failure” – task were aborted, library rescan failed, issue LIP for each HBA was required as well as a EE-cluster restart
- created a tool based on GPFS-List-Policy to see how many files are migrated, premigrated or still resident and how much space they occupy – `listmigstat <ltfs-path>`



```
 root @sap00334(rhel7.6)> listmigstat /gpfs/hdfs01/backup/

/gpfs/hdfs01/backup/:
                    files         space
   resident:          380      879.17MB
   migrated:            9        1.76TB
premigrated:        24629        5.86TB
```

Test-Plan-Details I

Test plan Spectrum Archive / LTFSEE 2019

Category	Test-Nr.	Test	Description	Expectation	Execution	Result	Status
Function	101	Migrate with eeadm	By filelist files are migirert with the eeadm CLI on tape	Files are moved to tape. Space is released in the filesystem / pool.	21.06.2019 12:34 at Node node-334 in /gpfs/hdfs01/backup/test	Status of the files changes to migrated. Space in the file system becomes free	passed
	102	Recall with eeadm	Filelist files are recalled using tape's eeadm CLI	Files are copied back from tape and Space is allocated in the filesystem / pool	21.06.2019 12:15 at Node node-334 in /gpfs/hdfs01/backup/test	Status of migrated changes to premigrated. Files were written back from tape	passed
	103	GPFS Premigration with Policy/Scheduled	The resident files should be premigrated via the GPFS rule with mmapplypolicy The resident files should be premigrated via the GPFS rule with mmapplypolicy	Files are copied to tape. Filesystemspace remains the same. Tape occupancy grows. File attributes go to "premigrated"	25.06.2019 16:16 at Node node-334 in /gpfs/hdfs01/backup/test	Age weighting has been added. All files were premigrated as expected.	passed
	104	GPFS Migration with Policy/Scheduled	The GPFS-Rule is supposed to premigrate the resident files with mmapplypolicy and release premigrated files up to Threshold (70%)	Resident files will be copied to tape and older premigrete files will be released up to the threshold (70%)	25.06.2019 14:32 at Node node-334 in /gpfs/hdfs01/backup/test	the oldest premigrated files were released, but also the newest resident file was migrated, not premigrated! After adjusting the Premig-Rule in order	passed
	105	GPFS Active-Migration with Threshold (premigrated)	By GPFS policy and callback is to be released when the pool threshold (90%) premigrated files	Premigrated Files go to "migrated" status, so that the space in the file system is released immediately until the lower limit (70%) of the pool is reached	24.06.2019 16:30 at Node node-334 in /gpfs/hdfs01/backup/test	When filling the test pool, the callback was triggered and the oldest premigriet files were released	passed
	106	GPFS Active-Migration with Threshold (not premigrated)	GPFS policy and callback are used to move resident files to tape when the pool threshold (90%) is reached	If no premigrieten files exist and 90% pool usage is exceeded, the oldest resident files must be moved to tape	25.06.2019 13:16 at Node node-334 in /gpfs/hdfs01/backup/test	all resident files were up to the pool usage at 70% mirgiert	passed
	107	Recall with File-Access	A migrietes file is opened with an application (cat / head / vi etc.) for reading	The file is automatically moved back from tape. Filestatus of "migrated" changes to "premigrated"	01.07.2019 at Node node-334	File was read back transparently at the Open of Tape. File state changes to premigrated	passed
	108	Delete migrated Files	The stub files will be deleted in GPFS	logical space in the FS becomes free and Tape-Space remains allocated	27.06.2019 11:15 at Node node-338	FS-Space was freed, Tape Space was not	passed
	109	Reconcile after Delete	Reconcilition of the tapes of a pool per eeadm / eeadm-Scripting	Tape Space becomes free after reconciliation. Reconciliation notation rate is updated	27.06.2019 11:50 at Node node-334	Tape Space becomes free after reconciliation. Reconciliation notation rate is updated	passed
	110	Directory Backup	A list rule should be used to save the file system structure via "eeadm save"	List-Rule generates all necessary directories and files that are backed up by eeadm save			
	111	Directory Tape-Restore	If the directory structure no longer exists in the file system, it should be restored from tape	Directory structure is present again as before			
	112	Stubfile Restore	If stub files are deleted in the file system, they can be restored from the tape	Deleted stub files or premigrated files has been restored			
Monitoring	201	Pool lowspacewarningthreshold	If less space is available in the tape pool than specified in the "lowspacewarningthreshold" paron Nodeeter, an SNMP trap should be triggered.	An SNMP trap is generated and the low tapespace is reported.	01.07.2019 14:00 at Node node-334	When the tape pool usage passed the "lewspacewarning" -threshold, an entry of net-snmp was logged in "/var/log/ltfsee.log"	passed

Test-Plan-Details II

Test plan Spectrum Archive / LTFSEE 2019

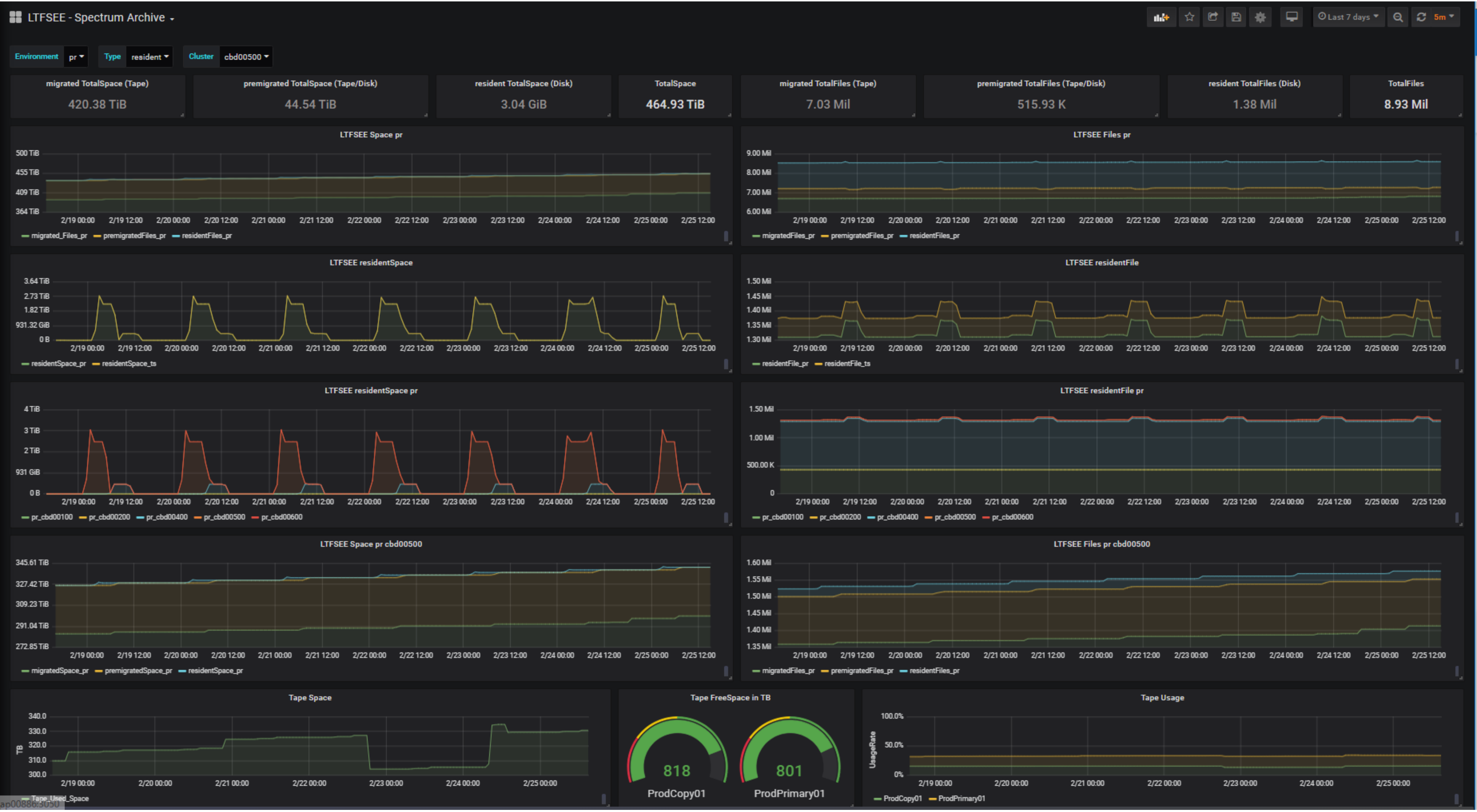
Availability	301	Drive failure	The failure of a drive is supposed to simulate a tape drive failure. At the son Nodee time, an eeadm task is started on this drive.	Task is held and resumed to another path	11.07.2019 at nodes node-334 and SAP00338 by IB11. At the library TLP44 by IBM service technician	Drive failure is not detected, task aborts with tape error. After installing the drive, the FC paths must be read in again. (ILP)	error
	302	Cardridge Rebuild	If a cardridge is not available, the copy can be used to create a backup again	New back cardridge is created. The files are updated in the GPFS with the new Cardridge			
	303	Tape error / not available	Simulation of a defective tape by removing it from the library	The library detects that the tape is broken or missing. Accesses to the data on the tape are intercepted via the copy	11.07.2019 at nodes node-334 and SAP00338 by IB11. At the library TLP44 by IBM service technician	Library recognizes when it accesses the tape that it does not exist and marks it with "warning". Access to the data is via the copy. After re-inserting the tap, this must be validated.	passed
	304	Library failure	By de-emphasizing all tape drives, a library failure is to be simulated. At the son Nodee time an eeadm task is started on this drive.	Spectrum Archive recognizes that the library is gone and reports it on	11.07.2019 at nodes node-334 and SAP00338 by IB11.	Tasks break with error. Library rescan fails with incorrect error message. After paths are back online, an ILP must be performed and the EEADM cluster must be restarted	error
Performance	401	daily backup from IB13	IB13 creates a daily backup and pushes it to the backup cluster via Distcp	Limitation by Ethernet channel vs. IO throughput in the GPFS			passed
	402	daily backup restore from Disk	IB13 brings back a premigrated daily backup	Limitation by Ethernet channel vs. IO throughput in the GPFS			passed
	403	daily backup restore from Tape (transparent)	IB13 retrieves a migrated daily backup	Limitation by Ethernet channel vs. IO throughput of the tape drives / GPFS	29.07.2019 IB13 by distcp	Recall starts, 6000 large files restored in 65 minutes, remaining 11000 small files on Node the following day not finished, task aborted by server reboot - so impractical	not passed, canceld
	403	daily backup restore from Tape (Bulk-Restore)	Migrated data will be retrieved via bulk restore	Limitation by Ethernet channel vs. IO throughput of the tape drives / GPFS	30.07.2019 15:30 at Node node-334 in /gpfs/hdfs01/backup/pr/cbd00400/20190730/	Migrated data (156GB) after 20 minutes completely on disk again - 13 minutes later at the IB13	passed

Production Status:

- since 1.8.2019 in production
- SW-Levels current: Scale at 5.0.4.2,
  Archive at ltfs-mig-1.3.0.6-51A19.x86_64.
- daily Ingest: 1,9 TB mostly, but also to many small files.
- If all HW-Component are working properly, there are no problems with this implementation.
- If you have some HW errors, then manual tasks are necessary.
- The policy-engine works without any problems.
- There is a threshold-migration rule active that is triggerd by a callback.
- Every 6 h a pre-migration is started per cron (future maybe per mmjob).

- Thanks to Takeshi Ishimoto and his Team for all the requested changes (syslog-ng, Admin-interface Support, LTFS-Mountpoint permission, rpcbind-port,..).
  Thanks for Nils Haustein for his help in the evaluation phase.

Capacity View:



LTFSEE - Spectrum Archive

Environment: pr
Type: resident
Cluster: cbd00500

Last 7 days
5m

migrated TotalSpace (Tape)	premigrated TotalSpace (Tape/Disk)	resident TotalSpace (Disk)	TotalSpace	migrated TotalFiles (Tape)	premigrated TotalFiles (Tape/Disk)	resident TotalFiles (Disk)	TotalFiles
420.38 TiB	44.54 TiB	3.04 GiB	464.93 TiB	7.03 Mil	515.93 K	1.38 Mil	8.93 Mil

Whats Next?

Dual-Library Implementation with a second tape-location

- *Installation of a new nodegroup with two additionally ee-Nodes and migration of the the copy-pool data to the second library.*

To be re-evaluated with next version 1.3.0.x in 2020

- *Component Interfaces (Scale,HSM,LTFS):*
  When a user makes a transparent recall storm, there is no easy way to stop this task completely. Mostly we have to wait until the tasks are finished or we restart the whole ltfsee-cluster
  Missing: interface for dmapi request to cleanly clear all queues, or a command to cancel tasks in the queues for all migration, pre-migration and recall-tasks. In 1.3.0.6 there are first enhancements available, but not complete yet.
- *Maintenance requirements for a rolling update procedure*
  Missing: multi version support in LTFSEE, and rolling upgrade procedure.
- *If a tape cartridge has problems, LTFSEE will still try to use this tape for future tasks.*
  Missing: Logic to exclude defective cartridges from successor tasks. Only manually task are possible here
- *if recalls are happen*, some status drive useage indicators are not updated in time.
  Missing: Each task type should reflect the status of its tape drives usage.
- *Handling smal files* for disaster recovery purposes
  Missing: eeadm save cmd for handling small files located on tier0 or on the inode
- *Performance:* Some known restrictions should be enhanced in Scale / HSM / LTFS
  dmapiWorkerThreads 64 <- Maximum ??.
  Enhancements on Sort Performance during bulk recalls!!
  Are the LTFSEE queues are big enough for growth what we see??
  Tape-Mount-optimisation like can we also read from copy-pool?

Questions ?

Backup: Policy-Files

Treshold-Policy:

```
mmlspolicy hdfs01 -L
/* Migration Rules */
/********************/

/* define exclude rule*/
RULE 'exclude' EXCLUDE WHERE
(
 PATH_NAME LIKE '%/.SpaceMan/%' OR
 PATH_NAME LIKE '%/.ltfsee/%' OR
 PATH_NAME LIKE '%/.mmSharedTmpDir/%' OR
 PATH_NAME LIKE '%/.snapshots/%' OR
 PATH_NAME LIKE '%/current/%' OR
 NAME LIKE '.mmbackupShadow%' OR
 NAME LIKE 'mmbackup%'
)

/* macro to define access age */
define( access_age,(DAYS(CURRENT_TIMESTAMP) - DAYS(ACCESS_TIME)) )

/* define external pool PROD */
RULE 'extpool_prod' EXTERNAL POOL 'ltfs_prod' EXEC '/opt/ibm/ltfsee/bin/eeadm'
OPTS '-p ProdPrimary01@TLP44 ProdCopy01@TLP44' SIZE 10485760

/* define external pool TEST */
RULE 'extpool_test' EXTERNAL POOL 'ltfs_test' EXEC '/opt/ibm/ltfsee/bin/eeadm'
OPTS '-p TestPrimary01@TLP44 TestCopy01@TLP44' SIZE 10485760

/* Migration rule PROD */
RULE 'threshMig_Prod' MIGRATE FROM POOL 'data01' THRESHOLD(90,70) TO POOL 'ltfs_prod' WEIGHT (access_age)
WHERE (KB_ALLOCATED > 0)

/* Migration rule TEST */
RULE 'threshMig_Test' MIGRATE FROM POOL 'test01' THRESHOLD(90,70) TO POOL 'ltfs_test' WEIGHT (access_age)
WHERE (KB_ALLOCATED > 0)

/* default placement policy */
RULE 'test-placement' SET POOL 'test01' FOR FILESET ('test')
RULE 'default' SET POOL 'data01'
```

Backup: Policy-Files

Sheduled Policy:

```
/* define exclude rule*/
RULE 'exclude' EXCLUDE WHERE
(
 PATH_NAME LIKE '%/.SpaceMan/%' OR
 PATH_NAME LIKE '%/.ltfsee/%' OR
 PATH_NAME LIKE '%/.mmSharedTmpDir/%' OR
 PATH_NAME LIKE '%/.snapshots/%' OR
 PATH_NAME LIKE '%/current/%' OR
 NAME LIKE '.mmbackupShadow%' OR
 NAME LIKE 'mmbackup%'
)

/* define is_premigrated */
define(
  is_premigrated,
     (MISC_ATTRIBUTES LIKE '%M%' AND MISC_ATTRIBUTES NOT LIKE '%V%')
)
/* define is_migrated */
define(
  is_migrated,
     (MISC_ATTRIBUTES LIKE '%V%')
)
/* define is_resident */
define(
  is_resident,
     (NOT MISC_ATTRIBUTES LIKE '%M%')
)
/* macro to define access age */
define(
  access_age,
     (DAYS(CURRENT_TIMESTAMP) - DAYS(ACCESS_TIME))
)
/* define external pool */
RULE 'extpool' EXTERNAL POOL 'ltfs'
EXEC '/opt/ibm/ltfsee/bin/eeadm'
OPTS '-p ProdPrimary01@TLP44 ProdCopy01@TLP44' SIZE 10485760

/* Migration rule */
RULE 'preMig' MIGRATE FROM POOL 'data01' THRESHOLD(0,70,0) TO POOL 'ltfs'
WEIGHT (access_age)
WHERE ( (KB_ALLOCATED > 0) AND (NOT (is_migrated)) )
```