Karlsruher Institut für Technologie

dCache on top of GPFS

Samuel Ambroj Pérez (samuel.ambroj@kit.edu)

Steinbuch Center for Computing (SCC), Scientific Data Management (SDM)

dCache overview

- dCache 7.2.17.
- 4 independent production instances: ATLAS, CMS, LHCb and UrProd (Belle2, Auger and Compass).
- 2 single instances with all services on just one server (DOMA and DLT).
- Postgresql 14 (master – slave). One pair per instance.
- Zookeeper 3.8. One common cluster (3 members) for all instances.
- Java: java-11-openjdk
- Everything is puppetized.
- Billing logs in OpenDistro. Migrating now to OpenSearch.
- Pools on physical machines.
- Servers running non pool domains hosted in VMs.

Details about dCache pools

- One GPFS file system per instance, one directory per pool.
  - ATLAS (13,35 PB), CMS (7,87 PB), LHCb (6,63 PB)

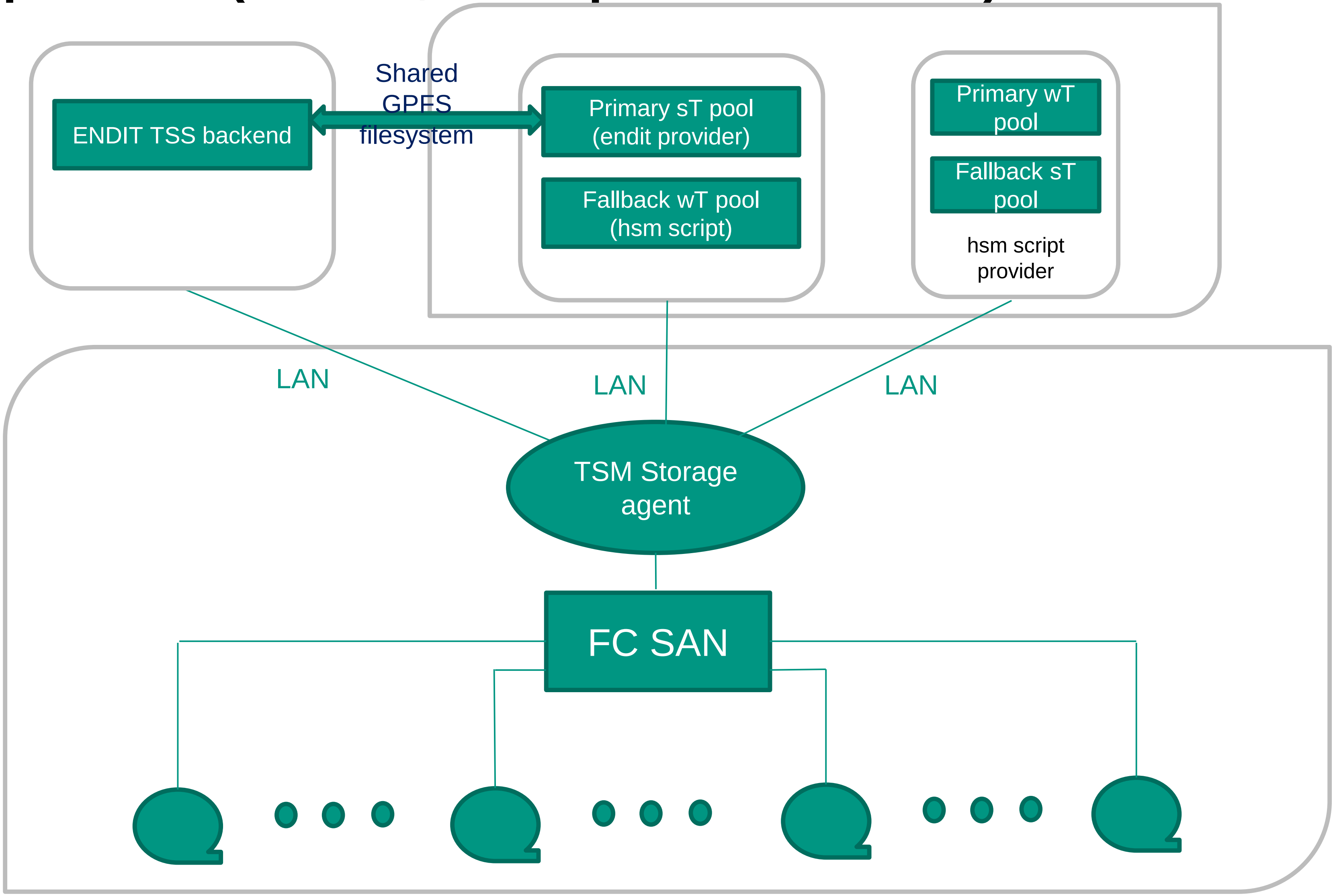
	ATLAS	CMS	LHCb	Belle II
# disk-only pools	11 (1,15 PB/pool)	6 (1,17 PB/pool)	6 (996 TB/pool)	2 (395 TB/pool)
Avg # files / used PB	2,69 million	718629	1,165 million	8,177 million
stage pools	2 (300/100 TB)	2 (275 TB/pool)	2 (175 TB/pool)	1 (50 TB)
write-tape pools	2 (200/100 TB)	2 (150 TB/pool)	2 (150 TB/pool)	1 (50 TB)
Avg file size on tape (GB)	1,46	3,37	3,26	2,29

Tape systems connected to dCache

- CMS, LHCb and Belle II are running HPSS in production.
- ATLAS is being migrated from IBM Spectrum Protect to HPSS.
- Staging from tape
  - dCache-endit provider plugin [1] on the stage pools. GPFS is a key part here. See Haykuhi’s presentation.
  - One ENDIT backend per instance (ATLAS, CMS, LHCb and Belle II).
- Writing to tape
  - HPSS: dc2hpss.py
  - IBM Spectrum Protect: dc2tss.pl
  - Haykuhi will provide more details about ENDIT-HPSS.

[1] https://github.com/neicnordic/dcache-endit-provider/

Tape conn. (ATLAS, IBM Spectrum Protect)



Problem while writing

- Uwe also commented on it.
- We have observed sometimes problems with written files.
    - In the DB: checksum and file size ok (consistent with what the experiment expects).
    - On the pool: file size mismatch (corrupted file). When the file is accessed again, dCache pool gets disabled.
        - The files must be set as broken and the VO must be informed.

- Why does it happen?
    - What does GPFS perform during a write?
    - Why does dCache think that the file was successfully written if the file size on the pool is not correct?

Could we have highly available pools?

- One GPFS file system per instance.
- Ideally: all disk-only/tape pools having access to all disk-only/tape files.
- One common ilocation in Chimera

```
chimera=> select count(1),ilocation from t_locationinfo where itype=1 GROUP BY ilocation;
  count  |      ilocation
---------+----------------------
  739792 | f01-120-126-e_D_cms  --->  disk-only
    1202 | f01-120-126-e_ops
  739842 | f01-120-127-e_D_cms  --->  disk-only
     302 | f01-120-127-e_S_cms
  737352 | f01-120-128-e_D_cms  --->  disk-only
   26378 | f01-120-130-e_sT_cms  --->  stage-tape
   47038 | f01-120-130-e_wT_cms  --->  write-tape
  720349 | f01-152-139-e_D_cms  --->  disk-only
  732036 | f01-152-188-e_D_cms  --->  disk-only
  733707 | f01-152-189-e_D_cms  --->  disk-only
   74974 | f01-152-190-e_sT_cms  --->  stage-tape
   43192 | f01-152-190-e_wT_cms  --->  write-tape
```

How to deal with the metadata?...

Thank you for your attention!
Questions?

Backup

- Monitoring