
CERN

CERN Site Report

Giuseppe Lo Presti
On behalf of the CASTOR disk operations team

Outline

- Main activities since June 2016
  - Evolution / consolidation
  - Interesting incidents…
- Upcoming work and plans

Main Activities since June 2016

- RFIO write access retirement
  - Announced in November 2016, performed in January 2017
  - Very limited impact – complete RFIO retirement expected Q1 2018
- Consolidation, consolidation, …
  - July 2016: the Public instance now runs on 3 disk pools, following a large hardware retirements campaign
    - 100+ hosts retired (1 Gbps)
  - March 2017: the Stager databases have been regrouped to 3 clusters
  - Now: "only" 154 disk servers in CASTOR
    - ~constant disk-cache size in front of tapes, less larger bricks
    - Increasingly challenging to keep up with the increasing bandwidth requirements
    - *By the way: July 2016 was our record month with 11+ PB to tape*

Main Activities (cont'd)

- Retirement of remaining D1T0 pools
  - All instances but Public configured with 1 pool for DAQ and 1 small pool for stage-in/production activities (LHCb: shared; ALICE: on Ceph, *cf. later*)
  - Public pool for stage-in and user activities is actually `default`: today's CASTOR largest pool @ 3.5 PB, "EOS mode" = very large number of slots
- Tuning of most disk pools for fast streaming performance, but without sacrificing parallelism
  - Large read-ahead setting (80 MiB) to make the reading:seeking time ratio >= 1:1
    - Limit disk thrashing when 10+ streams hit them
  - RAID 6, no striping
    - Don't move too many spindles to access required performance

```
[root@p05798818j78094 ~]# /bin/mount | grep castor | cut -d ' ' -f 1 | xargs blockdev --getra
163840
163840
163840
163840
[root@p05798818j78094 ~]# df -h | grep castor
/dev/md124              33T   30T  3,4T  90% /srv/castor/01
/dev/md126              33T   30T  3,4T  90% /srv/castor/02
/dev/md127              33T   30T  3,5T  90% /srv/castor/03
/dev/md125              33T   30T  3,4T  90% /srv/castor/04
```

Main Activities (cont'd)

- *Puppexit* (© S. Traylen)
  - Moved configuration management to Puppet 4. Painful and time consuming.
- *DAQ2FTS* (© Xavi)
  - Smaller CERN experiments don't have an established framework for their Tier0 / DAQ workflow
    - Typically, they relied on legacy scripts, which used **rfcp** to copy data to CASTOR…
  - Leveraging on the RFIO write retirement, we have piloted DAQ2FTS with the NA62 experiment: run a mini-FTS instance + globus-gridftp-server on the pit and write data to CASTOR via SRM
  - The SRM overhead is not significant given their requirements, but nevertheless once the service is established, it is easy to reconfigure it to use e.g. xrootd without SRM

“Interesting” Incidents

- Tape recalls exercise from ATLAS, Dec 2016
  - Unveiled more (to us) or less (to them) known limits when accessing tapes in random order
    - Seek time largely dominating, 30mins – 1h
  - **RAO** (Recommended Access Ordering) will come to help
    - *cf. German’s presentation tomorrow*
- File corruption on recalls (with xrootd)
  - Turned out that the `all.export / nolock` entry in `/etc/xrd.cf.server` really meant “multiple parallel writes are allowed, no protection from xrootd”
  - Changing to `/ lock` fixed the problem
- Socket errors in `tapeserverd` for both recalls and migrations
  - Tapeserverd did not retry on connection losses because of an old xrootd bug (RFIO did…)
  - Network can get busy inside the data center – in particular between tape and disk servers
  - Being fixed now...

Ops Plans *(from face-to-face June 2016)*

- Support for SL6: vanishing…
  - Tape servers are already all CC 7
  - Remaining disk servers will be migrated to CC 7, after retirements are completed
  - Head nodes will be last, expected end of 2016
- Support for CentOS 7 is there
  - Full cert. tests now done in CentOS 7 along with SLC6
  - Ceph stress test is 100% CentOS 7
- As of 2017, SL6 will be unsupported server-side

Dev + Ops Plans 2017 (and not beyond...)

- Dev:
  - Tape: RAO will be implemented in CASTOR
  - General: release 2.1.17 = server-side packages for CentOS 7 only with Ceph support
- DevOps:
  - A `c2cgw` (CASTOR-to-Ceph gateway) docker container has been developed to make use of the Ceph diskservers
    - Can't directly deploy as CASTOR requires Kraken and Ceph servers run Jewel
    - Added value: limit CASTOR's xrootd memory consumption, shield from Ceph
  - Explore the possibility of a Ceph Luminous pool with Bluestore / 8+3 EC for Public/default
- Ops:
  - Migration to CentOS 7 still to be done + (minimal) extra capacity to be deployed
  - Grafana-based monitoring to be put in place
  - ~ready for the 2017 Run

(More) Questions?