Bit preservation
Update from CERN archival storage

Oliver Keeble

On behalf of the IT-ST-TAB (“Tape, Archive and Backup”) section at CERN

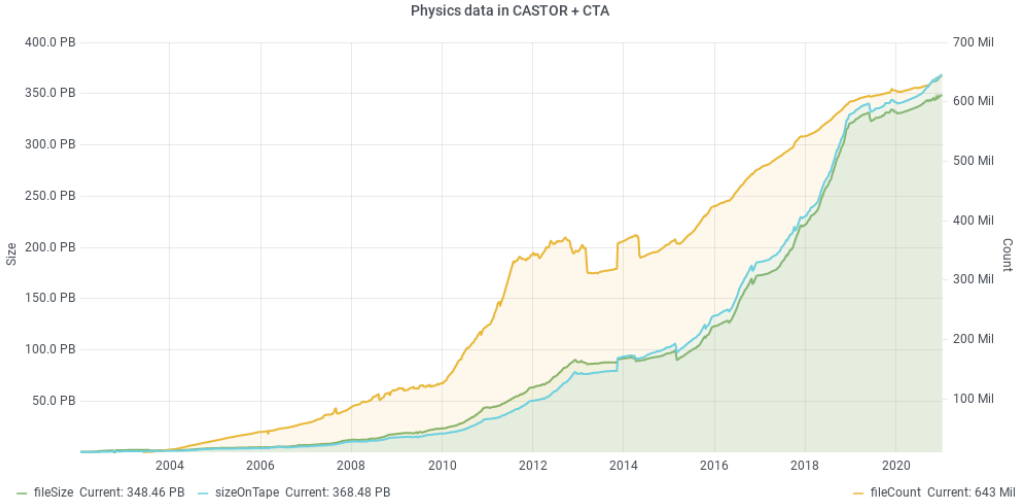
23/Jun/2021

Overview

- **Our archival system**
- **Aspects of bit preservation**
- **The future**

Since the last time

- **End of LHC Run 2**
  - Has left us with an archive of 350PB
- **Deployment of a new tape system ready for Run 3**
  - CASTOR → CERN Tape Archive (CTA)
  - CTA has reimagined a lot of CASTOR's disk management and scheduling
  - CTA has inherited all of CASTOR's bit preservation features
    - Data has not been moved between systems
      - Format on tape is the same



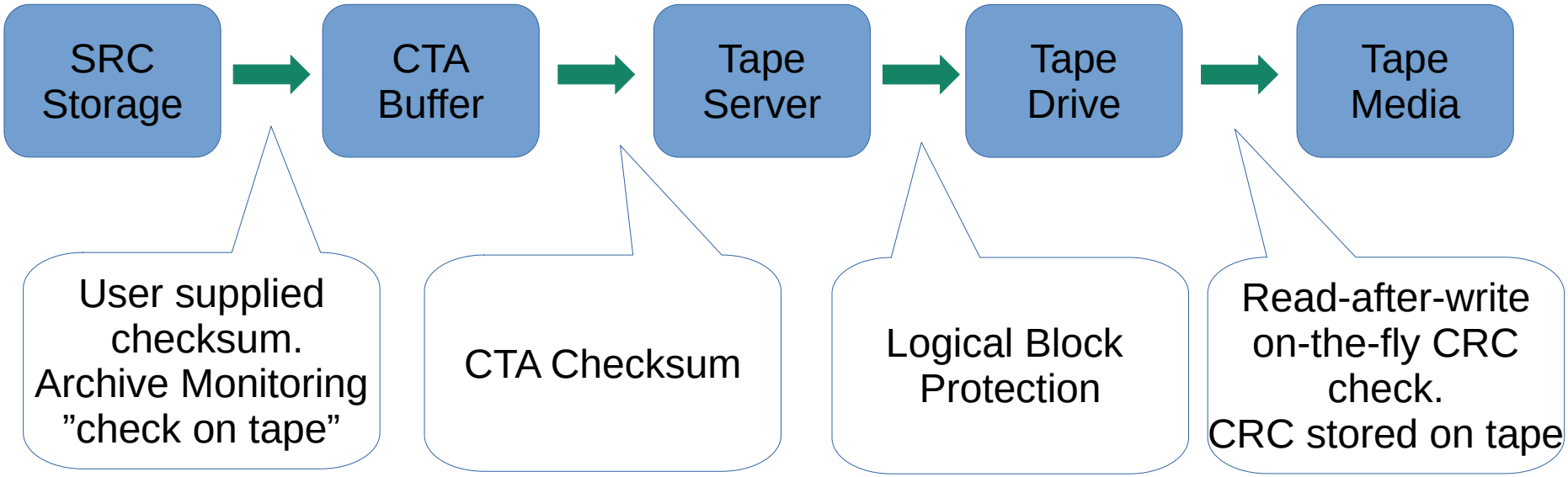
CERN Tape Archive

Our Archive

- **Libraries**
  - 2xIBM/Enterprise, 1xIBM/LTO, 1xSpectra/LTO
    - 1xSpectra/LTO awaiting delivery
    - Separate infrastructure for backup services (SpectrumProtect)
- **Drives**
  - ~100 Enterprise/LTO drives
- **Tapes**
  - ~30k tapes, 350PB
- **Dedicated tapeservers and SSD buffer**

- **Multi-media, multi-vendor policy**
- **2 buildings on the same site**

Transmission Integrity



FTS

Network layers have their own checksums - e.g. TCP and FC

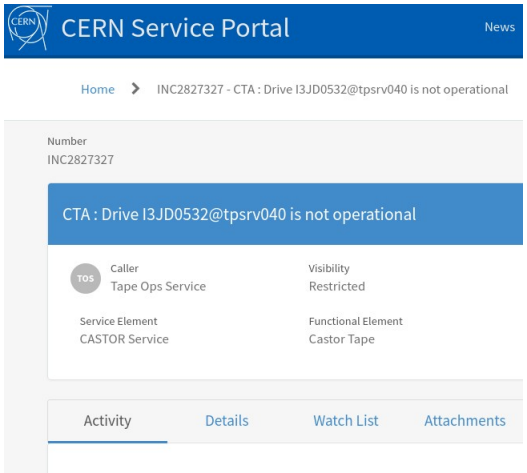
Promoting durability

- **Media properties**
  - Tape Bit Error Rate of 10^19
    - ...if handled according to environmental specifications and usage limits
    - Tape ECC (20% overhead)
  - Immediate read after write validation
- **Environment**
  - Climate controlled CCs, avoiding large environmental variations
    - Enterprise drives have humidity and temp sensors
    - SPECTRA air filters
    - Repack schedule means we don't reach nominal media lifetimes
- **Manufacturer support and warranties**
- **Redundancy**
  - Dual replica is possible
  - Geographical separation will be possible (different buildings, same site)
  - No "RAIT" plans
  - Raw LHC Physics data is single replica and has another copy at a separate centre (T1) with MoU

Monitoring and Warning signs

- **We track the operation of the infrastructure**
- **Certain conditions raise automatic tickets and trigger workflows**
  - Mount failure detection
    - Excessive mount failures for drives or tapes raise tickets
  - SCSI-3 Tape Alerts
    - CTA logs relevant tape alerts
    - Procedures in place to record and act on issues

```
[1624241876.462451000] Jun 21 04:17:56.462451 tpsrv040.cern.ch cta-taped: LVL="WARN" PID="14773" TID="14773" MSG="Tape alert detected" tapeAlert="Hard error" tapeAlertNumber="0" tapeAlertCount="2"
[1624241876.462954000] Jun 21 04:17:56.462954 tpsrv040.cern.ch cta-taed: LVL="WARN" PID="14773" TID="14773" MSG="Tape alert detected" tapeAlert="Read failure" tapeAlertNumber="1" tapeAlertCount="2"
```




```
Jun 21 04:17:56 tpsrv040.cern.ch kernel: [1017344.698965] st 11:0:0:0: [st0] Sense Key : Medium Error [deferred] Jun 21 04:17:56 tpsrv040.cern.ch kernel: [1017344.700613] st 11:0:0:0: [st0] Add. Sense: Medium format corrupted
```

Monitoring and Warning signs

- **SCSI info**
  - CTA logs mount stats, drive stats, volume stats
  - Contains info on recovered (and unrecovered) read and write errors
  - Semantics differ between manufacturers
  - Not clear how to use these predictively

```
2021/06/21 04:43:38.837803 tpsrv030.cern.ch info cta-taped:
LVL="INFO" PID="29686" TID="29779" MSG="Logging volume statistics"
thread="TapeRead" tapeDrive="I3551412" tapeVid="I51642"
mountId="103749" driveManufacturer="IBM" driveType="0359255F"
firmwareVersion="4B12" serialNumber="0000078D9FAD"
lifetimeBOTPasses="2151" lifetimeMOTPasses="1414"
lifetimeVolumeMounts="181"
lifetimeVolumeRecoveredReadErrors="2251"
lifetimeVolumeRecoveredWriteErrors="5"
lifetimeVolumeUnrecoveredReadErrors="3"
lifetimeVolumeUnrecoveredWriteErrors="0" validity="1"
volumeManufacturingDate="20141018"
```

Error Detection and Recovery

- **Validation operations**
  - Still ongoing in CASTOR, being ported to CTA
  - “Partial Tape Validation” workflow
    - Checks files and tapes with sparse sampling
- **Repacks**
  - Data is validated during media migrations
- **Read checks**
  - CTA checks the checksum of recalled files
- **Media repair workflow**
  - Integrated with the tape alerts
  - No outstanding issues with manufacturer or media
  - Easier after market consolidation, all roads lead to IBM

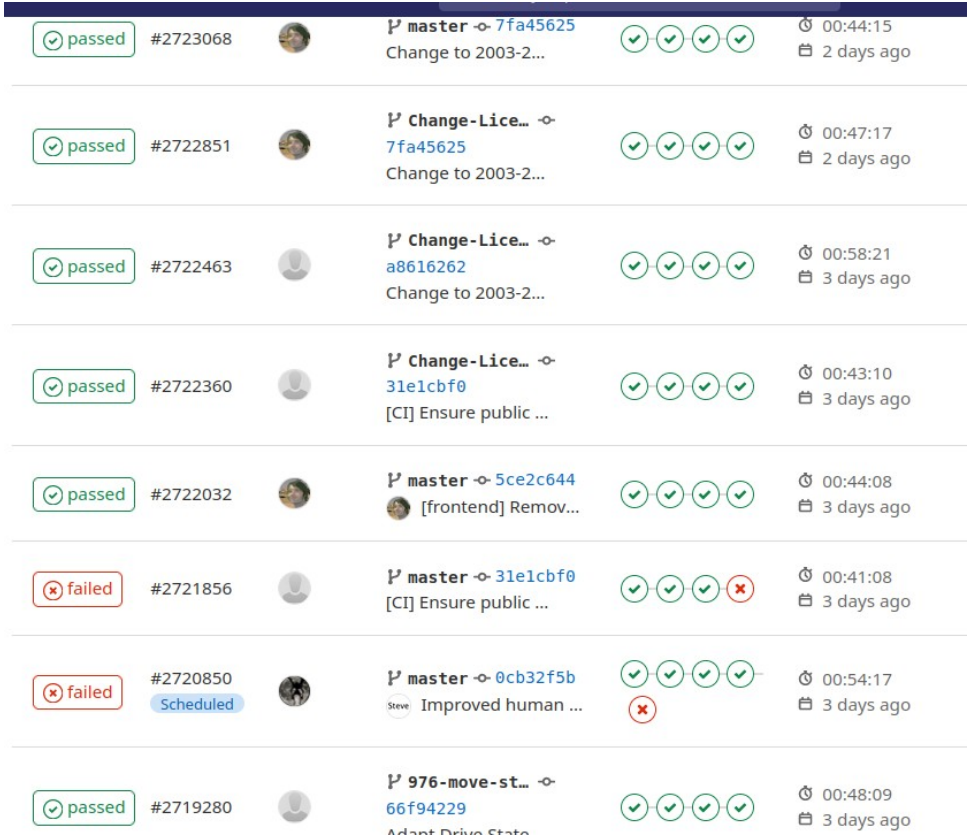
```
==================================
tape verify report for last 7 days
==================================

total tapes analyzed : 1190
total tapes with data: 1189 ; empty tapes: 1
tapes with [ERROR]   : 0
tapes with [WARN]    : 1
tapes with [OK]      : 1137
ongoing verifications: 45

total verified data (incl. ongoing jobs): 0 TB
average performance per drive and tape  :
90.683 MB/s
aggregate performance for all drives    : 0 GB/s
```

Even archives need a delete function

- **How to prevent accidental and malicious deletion?**
- **Tape**
  - Append-only nature of tape
  - Air gap possibility
- **Audit logs**
- **CTA**
  - ACLs
    - Separation of delete, archive and bring-online privileges (work ongoing)
  - Least privilege
    - User accounts with different privilege sets
  - Immutable files, modifications are not possible
  - CTA recycle log
    - Reference to data deleted from namespace remains in the catalogue, enabling easy recovery from tape

Service integrity and business continuity

- **Business continuity**
  - CTA catalogue
    - Oracle backup
  - Namespace
    - CTA backup instance
- **Software QA**
  - CI and tests
  - Pre-prod, staged rollout



The future

- **CERN has no plans to move out of tape**
- **Keep 50/50 LTO/Enterprise split**
- **We're not waiting for any particular technical developments**
- **Major procurements for LHC Run 3 before the end of the year**

- **Media capacity roadmap**
  - Durability per cartridge should stay flat (thanks to ECC and iterative decoding)
  - Capacity increases make dual copy easier to justify
- **Data Preservation use cases**
  - Digital Repositories
  - Open Data
  - Babar and LEP

Thank you
Questions?

home.cern