Archival Storage in CMS

Christoph Wissing (DESY)
for CMS Computing and Offline

DOMA General Meeting, November 2018

HELMHOLTZ RESEARCH FOR GRAND CHALLENGES

Content

Data Tiers and usage

Data & MC processing

SRM-less Archival Storage

Quality of Service (QoS)

Recent Performance Figures

Tape Families

Disclaimers:

This is a snapshot of the status of present discussions!
Present CMS transfer system is being replaced in LS2, forward extrapolations are very limited.

Data Formats, Volume and Usage

	Run2		Run4 - Expectations
	Volume Order of Mag.	Life time / archiving	Anticipated usage
RAW	Some 10PB/y	2 archival copies On disk ~60 days after recording Re-staged, when needed for re-RECO	2 archival copies Mostly on archival storage only
RECO	O(10PB/y) getting less	Only stored for selected data No archival copy Deleted after ~90 days (when created)	Like fully transient
GEN-SIM	10PB/y	One archival copy Staged in, when needed	Archiving of GEN, SIM transient
AOD(SIM)	10PB/y	One archival copy Kept on disk for 100 days (last access) Typically 1-2 disk replicas (dynamic)	Similar to Run2 Likely less disk replicas
miniAOD(SIM)	1PB/y	One archival copy Kept on disk for 300 days (last access) Typically several disk replicas (dynamic)	Similar to Run2
NanoAOD(SIM) in commissioning in 2018	2018: ~0.5PB (total)	One archival Presently fully kept on disk Typically several disk replicas (dynamic)	Run3 experiences will pave the NanoAOD way for Run4

Note: This is a simplified table. Complex compositions for data tiers for special purposes were neglected.

Processing of Data

PromptRECO

- 48h after recording
- Output: AOD, miniAOD in future likely also nanoAOD (presently produced in separate workflow)

Re-RECO of data

- Typically once or twice per year
- Input: RAW – Output: AOD, miniAOD and nanoAOD
- Usually requires staging of most RAW data (life time on disk is 60 days only, if kept on disk at all)

Re-miniAOD of data

- Typically two times in addition to re-RECO of data
- Input: AOD – Ouput: miniAOD, nanoAOD
- Due to reduced life time of 100 days of AOD significant fraction needs staging

Re-nanoAOD of data

- Strategy not yet fixed – Could be “frequent” or on demand

Processing of Monte Carlo

Traditional production

- Produce GEN-SIM from generator input
- GEN-SIM is archived and re-staged for following DIGI-RECO workflow
- DIGI-RECO writes AODSIM, miniAODSIM, nanoAODSIM
  - Some requests for GEN-SIM-[DIGI|RECO|RAW] output, operationally rather heavy

Planned approach

- Produce only miniAODSIM, nanoAODSIM from generator input
  - No saving of GEN-SIM nor AOD (with some justified exceptions probably)
  - For MC production CMS needs CPU for very roughly 50% for Geant4 and another 50% for DIGI-RECO
  - In recent years less than 50% of GEN-SIM input was re-used for another DIGI-RECO
- Trading archival storage and more importantly its related operations against CPU cycles

SRM-less Archival Storage – non-GridFTP WAN Transfers

For processing input data is always subscribed to disk

- CMS transfer system need to put data on disk
- Transfer system cleans disk after processing
- No particular dependence on any method or protocol

Staging

- CMS needs a file on buffer disk just for the purpose of transfer
- Can be handled in FTS
- CMS transfer system would talk to FTS

Space reporting

- SRM not involved here for CMS

Transition to Rucio in LS2

- Opportunity to start an infrastructure without SRM

WAN transfer protocol

- GridFTP to be replaced
- Fine for CMS – inter-operability between all SEs needs to be ensured

Some (very initial) Thoughts on QoS

We understand QoS as an intend by sites

- Are there plans to monitor and verify the promised QoS? Who?

Some possible QoS classes:

Archival	High I/O Disk	Resilient Disk	Non-redundant Disk
- Long term archiving - Minimal data losses - Understood recall rates	- Fast spinning disk - SSD - Capability to serve most demanding Workflows Pileup Mixing	- Medium I/O - RAID or duplication against disk failures - Site attempts recovery of files	- Medium I/O - Maximum capacity per cost - Experiment recovers (expected) file losses

Presently Tape

Presently Disk (not distinguishing any QoS)

Other relevant QoS metrics

- WAN connectivity: at least coarse classification (1Gb/s, 10Gb/s, 100Gb/s)
- Minimum effective read size
  - CMS application sends vectors of many smallish read requests
  - Too large minimum read sizes lead to good throughput, but still inefficient applications

Some Recent Performance Figures

Last month of pp data taking 2018

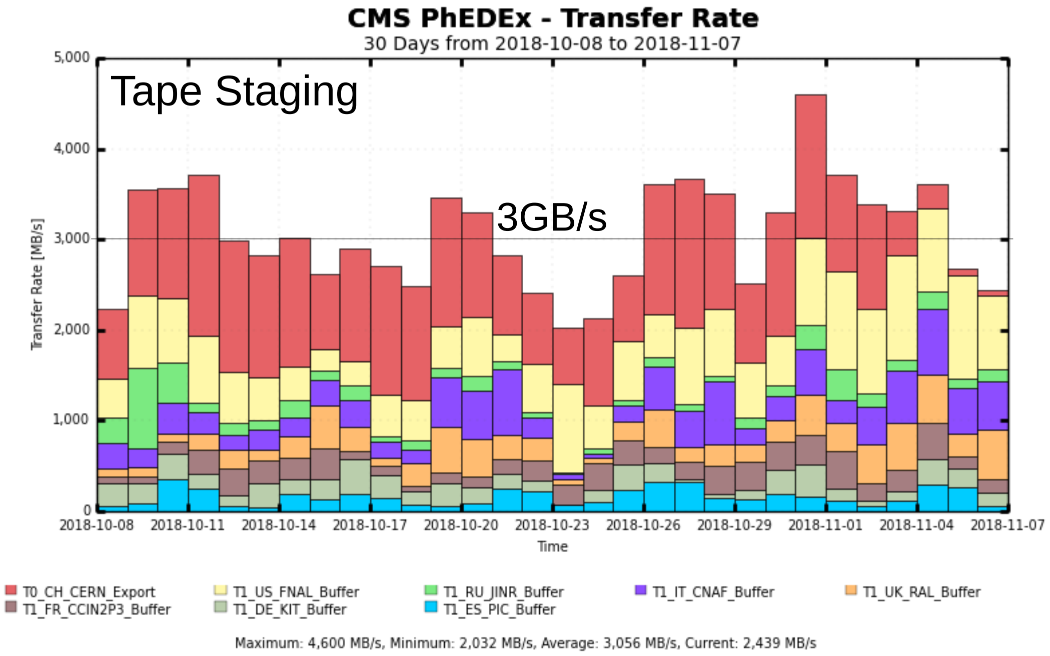
- No dedicated campaign
- Archiving of RAW, data Prompt- and Re-Reco and MC output
- Recall for MC production and data Re-Reco
- Earlier in Run2 CMS had a few exercises with sites
  - Increased queue depth for routing Phedex

Both rates aggregated on average between 2.5-3GB/s

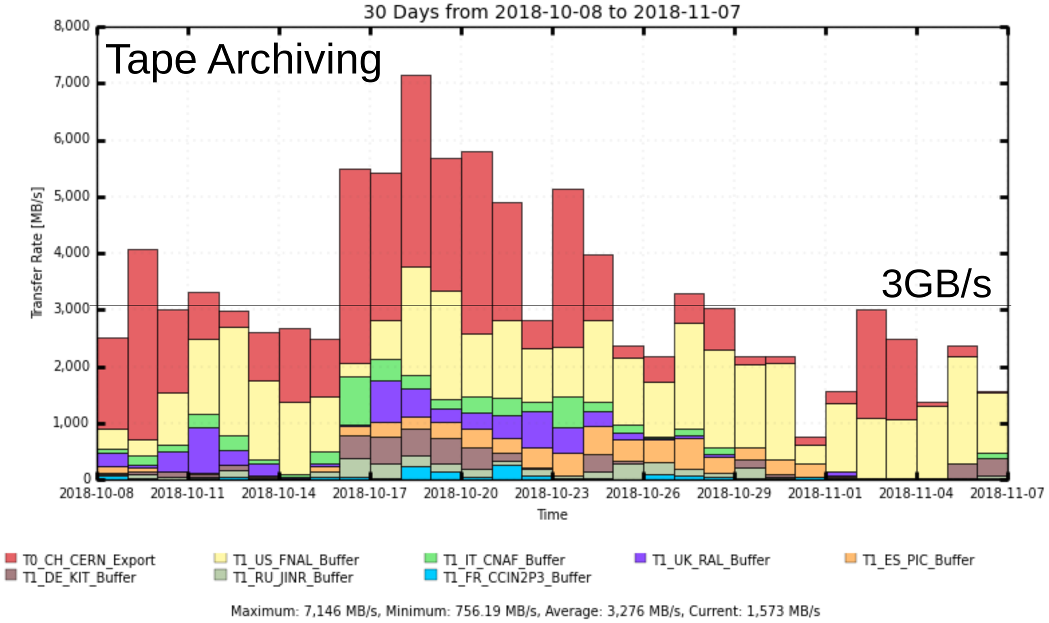
- Dedicated exercises with sites could push rates further

Operation effort goes into the tails

- Regular struggles with last few files to come from tape
- Sometimes help needed from local admins


CMS PhEDEx - Transfer Rate
30 Days from 2018-10-08 to 2018-11-07

Maximum: 4,600 MB/s, Minimum: 2,032 MB/s, Average: 3,056 MB/s, Current: 2,439 MB/s


30 Days from 2018-10-08 to 2018-11-07

Maximum: 7,146 MB/s, Minimum: 756.19 MB/s, Average: 3,276 MB/s, Current: 1,573 MB/s

Other Operational Items

Deletion campaigns

- CMS had typically two deletion campaigns per year
- Total volume of each campaign in the order of 10PB over all sites
  - Deletion of superseded re-reco versions
  - Deletion of MC samples produced with obsolete generator versions
- Some tapes become available immediately, others require repacking

Distributed Agent infrastructure

- Present CMS transfer system Phedex requires local Agents on (almost) each site
- Storage access can be highly customized locally
- Requires effort by site administrators
- CMS is going to move to Rucio for Run3
  - No local agents required
  - Centrally managed

LFN Structure

Physical File Name (PFN): Site specific prefix + Logical File Name (LFN)

Configurable via TFC

Primary dataset
or
MC / Generator

data or mc

“Processing String”

/site/specific/prefix/store/data/Run2018A/Tau/AOD/17Sep2018-v1/00000/12E46F3B-D183-7144-9F32-A8CC32123913.root

Acquisition Era
- Data: About 5 per year
- MC: Rough 10 campaigns per year

Data tier

Tape Families

- Assign group of tapes below certain directory
- Actively discussed with sites (before) Run1 times
  - Nothing changed in CMS LFN structure
- Perhaps worth revisiting during LS2

Summary

Some data tiers live mainly on archival storage

- GEN-SIM: only staged for MC reconstruction
- RAW: fraction of datasets stays on disk for 60 days, usually staging required for reconstruction
- AOD: life time on disk got reduced to 100 days, expect even less AOD on disk in the future

Protocols

- CMS can live in a world without SRM and GridFTP, but effort is required to get there

Quality of Service

- Interesting concept offering a number of options
- Discussions have just started and several aspects would require clarification

Operational aspects

- Archival storage requires effort centrally and at the sites
- CMS had same limited campaigns in Run2 to improve performance of archival storage, also because

Rucio replaces Phedex latest by Run3

- Opportunity to revisit certain site configurations
- Possible synergies with ATLAS