Understanding Disk needs at the HL-LHC for CMS

Frank Würthwein

SDSC/UCSD

DOMA Access

November 20th 2018

Disclaimer !!!

- The ideas presented herein are NOT necessarily representing current thinking in CMS.
- They are nothing more than rumblings of an individual to stimulate discussions.
- However, those rumblings are informed by many discussions with a long list of colleagues from within CMS. Too long to list them all here, as I would undoubtedly forget to mention some people.

Start with Data Formats and their expected use

Data Tier	Data
RAW [MB]	7.4
AOD [MB]	2.0
MiniAOD [kB]	200
NanoAOD [kB]	4

Courtesy David Lange
Present Model of CMS
HL-LHC resource planning

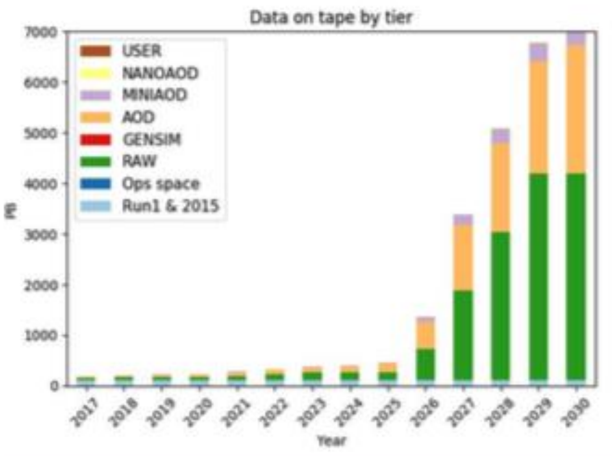
Primary Processing:
RAW -> AOD -> Mini -> Nano

Processing Assumptions:

- All events are prompt reconstructed
- 25% of events are re-reconstructed (eg for startup)
- There is a reprocessing each year of the current years data
- MC is always made starting from scratch (eg, GEN-SIM is redone)
- In shutdown years, all events in the last 3 years are reprocessed and corresponding MC remade
  - Take 2 years to do this reprocessing as it doesn't fit into 1 year without a resource bump (first shutdown is 2030..)

Data formats span x1000 in size per event.
Files in large data formats are touched at most twice a year.

May 2018 Tape Estimate



Use Tape estimates as guidance of size of the total available data.

Dominated by RAW and AOD

Another way of looking at it:

80+160 Billion events/year (Data+MC) = 240B events/year

⇒7.4MB x 8e10 ~ 6e11 MB ~ 0.5 Exabytes/year of RAW

⇒2.0MB x 2.4e11 ~ 5e11 MB ~ 0.5 Exabytes/year of AOD

⇒0.2MB x 2.4e11 ~ 0.5e11 MB ~ 50 Petabytes/year of Mini

⇒0.004MB x 2.4e11 ~ 0.01e11 MB ~ 1 Petabyte/year of Nano

The data that is accessed 1-2 times per year is x1000 larger than the data that dominates data analysis use !!!

Caching Objectives

- Want to define a working set of data that is accessible from all CPU anywhere.
  - We think this requires regional caches where region is defined by maximum RTT within the region to avoid latency that significantly deteriorates CPU/wall time for analysis.
    - This requires continued detailed measurement of latency dependence of user analysis on the production system.
  - This requires an understanding of the time evolution of the working set. And the resulting tape recall bandwidth.
- Want to reduce administrative effort of supporting storage infrastructure for analysis.

Straw Proposal for optimizing US T2 disk space usage

The geography makes for two obvious cache collaborations.

UNL maybe close enough to Midwest.
Florida and MIT unlikely to be close enough.

~ 4 caches total in US CMS

Equivalent Distances in EU

Good goal to set for IO stack to be sufficiently latency tolerant to lose less than 10% in CPU time for access distances of ~500 Miles (RTT UNL-Purdue).

Gains from regional caches

- T2s at Caltech, UCSD, UNL, … today use HDFS with replica = 2.
  - Disk failures are a major operational concern as it can lead to data corruption.
- Xrootd caches are run as JBODs
  - Disk failures in caches are of no concern.

There is an immediate x4 increase in useable disk space for cache deployment across 2 sites in SoCal (or elsewhere).

Caching Model

- One high quality disk copy of the “working set” for all analysis in the Americas.
  - As the working set changes over time, we recall data from tape.
    - Need to measure the global analysis working set !!!
    - Need to measure data lifetime and transients to estimate tape recall needs.
- Each T2 has zero redundancy disks inside caches.
- T2s are grouped into regional caches within distances that have less than 10% degradation of CPU/Wall due to access latencies.
  - Need to understand latency tolerance for analysis

Summary & Conclusions

- Need to distinguish 4 storage uses
  - Cheapest possible Archive
  - Golden disk copy (likely to be most expensive storage)
  - Caches with zero redundancy
  - Buffers for processing campaigns
    - Single vs distributed buffers requires understanding of latency tolerance of processing applications.
- Need to measure/estimate
  - Working set for analysis globally
  - Tape recall needs for processing buffers
  - Tape recall needs for golden disk copy
  - Latency tolerance of analysis applications