T2 UCSD Storage Expectations for HL-LHC

Frank Wuerthwein

UCSD/SDSC

DOMA Access Meeting

November 29th 2020

Topic of this talk

- In backup is the talk I gave to DOMA Access on October 5th about disk types and usage expectation.
- Here I summarize, and then want to talk a bit about sizing of the T2 disk space for analysis facility, based on physics expectations.

Implications for disk @ UCSD

- Buffer space for processing workflows
  - JBOD only, we are not responsible for anything in here. If things get lost, not my problem.
  - Temporary space for AOD & RAW & output of processing
  - Want CMS to be organized => data stay here for ~ 2-4 days => O(10) speedup required from today
- Xcache space for analysis
  - JBOD only, we are not responsible for anything in here.

UCSD Implications for disk @ UCSD

- Origin space for Data Lake
  - – Erasure encoded CEPH with at least 3 disk security.
  - – Want CMS to automate recovery from disk losses.
- User data space for analysis
  - – Erasure encoded CEPH with at least 4 disk security.
  - – All columnar store as part of AF
    - Some HDD for volume, some NVMe for fast random access
    - Users decide what gets elevated into NVMe
    - Both HDD and NVMe spaces are quota'ed so that people know what to expect to have available to them.
      - – Group quotas because people work in groups on an analysis
  - – Focus on sizing this in next slides.

Estimating Size of Columnar Store

- Workflow assumption
  - – Have signal MC in MINI & NANO
    - 1% precision => 10,000 events
    - 10% efficiency => 100,000 events
    - x10 headroom => 1 Million events of MINI needed
    - 250kB * 1M events = 250GB sample size for MINI
  - – Develop NANO skim on MINI/NANO for signal & apply to ~ “all” NANO
    - Entire annual NANO ~ 2.4PB * O(1%) => 24TB NANO skim
  - – Develop extra needed from MINI

Develop Extra from MINI

- MINI event size ~ x125 NANO event size
  - Want as little MINI as possible but as much as needed
- 1% MINI = 300TB => selection assumption as before
- 10% per event from 1% MINI = 30TB
  - I believe this is very generous because average measured access fraction today for MINI analysis is <10%
- **Total data per analysis ~ 60TB/year of data taking**
  - 0.25TB + 24TB + 30TB ~ 60TB
- There have rarely if ever been more than 10 active analyses at UCSD T2 by local users.
- **Total columnar space ~ 600TB/year of data taking**

Why so small ???

- This is tiny in comparison with past experience. Why ?
  - We currently provision 1PB for user space
  - This is historic, going back to before the MINI when we used about as much because our ntuples were very large.
  - Today:
    - 116TB shared data for groups that are well organized
    - 308TB individual user spaces
      - Individual user spaces are cluttered, as expected.
    - One would predict x30 for HL-LHC => ~ 13PB
- **My Guess: We store O(10) more today because it is so damn hard to go back and get something you missed.**

If we had fast mechanism to execute the described workflow then the columnar data space of the UCSD T2 could be as small as PB/year of data taking

In particular, no reuse assumption for columnar space.
It is all user space for individual analyses.

Backup

This is October 5th talk as it was presented then.

No changes!

Situation Today

- We run HDFS2 with replica=2 and maximum 90% full.
  - RAW/usable space = 2.2
- We have power outages ~ 1-2 times per year => loose more than 2 disks ~ every time.
  - Data losses are very painful and much too frequent.
    - Manual recovery of losses for both user data and experiment data.
    - Providing NFS space for user data as backup.

CMS Data Format Reminder

- Annual nominal data volumes:

RAW 364PB	AOD	=>	Mini	=>	NANO
Simulation	240PB		30PB		2.4PB

- Aspirations:
  - – RAW & AOD accessible only via top-down workflows.
  - – MINI & NANO accessible to anybody in the collaboration via Analysis Facility and/or CRAB

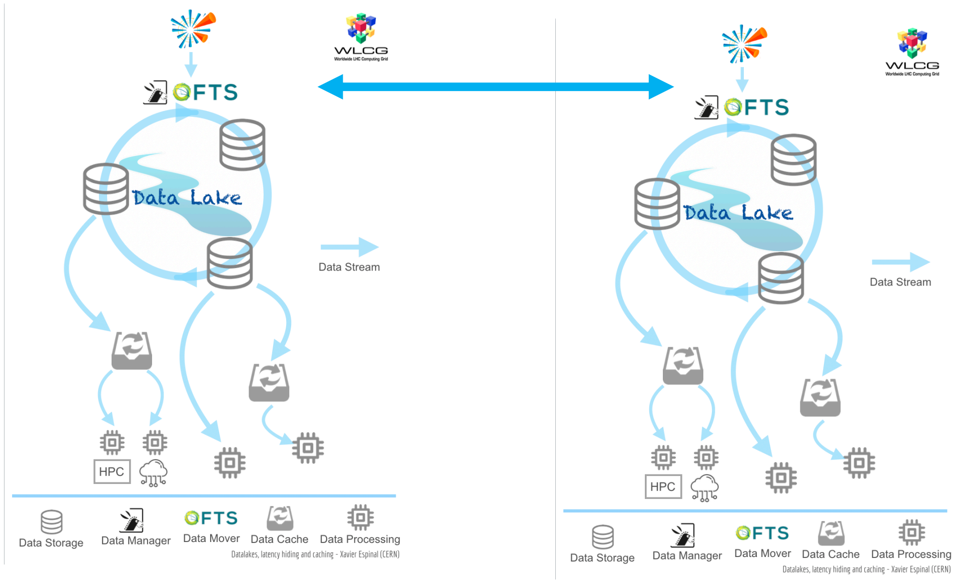
Aspirations vs Reality

- There will be a commissioning period during which detector, AOD, MINI, and NANO get commissioned.
- This will be done on a small fraction of the HLT output rate, less than 5%
- As MINI stabilizes, AOD will no longer be on disk, and the full HLT output rate will be available for analysis.
- This process has been suggested like this to the Collaboration by ECoM2x task force.

LHC Data Lakes Model

- More than one lake globally
  - E.g. USA as one lake per experiment seems plausible.
  - “Federation of lakes”
- Centrally managed replication between lakes.
- Intra lake data access via mix of:
  - Top-down placement, e.g. as part of workflows
  - Bottoms-up placement for cache misses
  - Streaming for remote file open

Cross lake transfers



Next: Enumerate implications for UCSD disk space.

Implications for disk @ UCSD

- Buffer space for processing workflows
  - JBOD only, we are not responsible for anything in here. If things get lost, not my problem.
  - Temporary space for AOD & RAW & output of processing
  - Expect that CMS is organized and data stays here for no more than 2-4 days.
- Xcache space for analysis
  - JBOD only, we are not responsible for anything in here.

Implications for disk @ UCSD

- Origin space for Data Lake
  - Erasure encoded CEPH with at least 3 disk security.
  - Am expecting CMS to automate recovery from disk losses.
- User data space for analysis
  - Erasure encoded CEPH with at least 4 disk security.
  - User level NANO derivatives only.
- Longer term Analysis Facility
  - Maybe NVME for fast random access in context of programmable CEPH storage supporting columnar data formats.
  - HDD user space still provides security against data loss.

Cost savings

- On average, more than x2 in RAW disk space.
- Ease of operations as the bulk of disk space is JBOD, and losses are handled automatically upstream.
- Ease of use for physicists that have user space assigned at UCSD because data loss is much much less frequent.
- Overall, spend larger fraction of total funding on CPU/GPU than today.

Comments & Questions