Data Challenges

Frank Wuerthwein
UCSD/SDSC

AF workshop

October 26th 2020

3 Types of Data Transfers

- Moving datasets between archives and data lakes.
- Data transfers coupled with processing campaigns
- Data subsetting as part of AF use case

Will present one example for each to set scale.

All examples from CMS

Need to define DC's for each of these.

Numbers from US CMS input to ESnet requirements process.

Archives & Data Lakes

- Large volumes get moved irrespective of any processing activities.
- E.g.: RAW data from CERN to Archives@T1
  - 6.5MB * 7.5kHz ~ 400 Gbit/sec
  - To keep up, build in x2 peak/avg, i.e. plan for 50% duty cycle.

Goals:

800Gbit/sec out of CERN for CMS alone.

320 Gbit/sec for CERN – FNAL

200 Gbit/sec Tape write @ FNAL

HPC RAW Processing Campaign

- Coordinated use of archive, disk buffers, network, and compute resources.

Goals:
200 Gbit/sec read from archive
Up to Tbit/sec bursts out of FNAL
400Gbit/sec avg processing speed with large fluctuations
Roughly x3 ratio of input/output
Up to Tbit/sec bursts back in to FNAL
~70 Gbit/sec write to archive

- Processing fluctuations large for multiple reasons
  - Inherent CPU/event fluctuations
  - Batch scheduling fluctuations
- Plan for larger ratio of peak/avg to allow for fluctuations

Data Subsetting requires more explanation

Recall:

30PB per year MINIAOD

2.4PB per year NANOAOD

Take Inclusive Hadronic SUS searches as example

- Requires roughly 30% of all the data
  - Single lepton, di-lepton, and jet-met triggers
  - Top, Wjets, Zjets, SUS, … simulations
  - All of these samples are among the largest there are in CMS !!!
- If an analysis requires a single float from MINIAOD to proceed then this would require 30PB x 30% ~ 10PB of data transfer for just one analysis of one year's worth of data.
- However, typically, analysis requires <10% of events and <10% of objects per event, most often much less. => <<100TB of data actually needed.

Data subsetting can save O(100) to O(10,000) factor in data transfers.

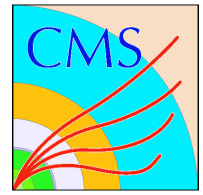
Can turn a week long data transfer into a lunchbreak.

Data Subsetting workflow

- Develop core analysis “skim” on NANOAOD
  - Might need small signal samples in MINIAOD for efficiency measurements of skim.
- Define “event list” and “objects per event” this way.
- Use “vector of byte read” retrieval over WAN as a large scale scheduled data transfer during “lunch break/overnight/weekend” depending on scale.
- Floats from MINI get added to NANO in AF as a custom NANO format for this analysis.
- Needs to get redone a couple times per year or so if physicists think carefully. Rest of the time, work locally at AF at full speed of NANO@AF.

Summary & Conclusion

- Identified 3 different types of data transfer use cases that are worth defining data challenges for.
- First two work on file level using Third Party Transfer via HTTPS
- Third works on XRootD partial file level.
- Next step:
  - Think carefully through the use cases and define an iterative set of DC's that cover all needs.
  - Build up complexity over time.



Comments & Questions