Evolution to a HL-LHC Computing Data Model

Elizabeth Sexton-Kennedy, Fermilab

Introduction

Unlike when the last computing model document was written [1], CMS is no longer in a startup phase. In fact as of this writing, the luminosity doubling time is a year long and will only grow going forward. We have reached a phase of stable operations. Largely because of this CMS no longer needs to widely distribute all of the reconstructed data. This is fortunate because with the high pileup conditions in Run 2, that format, many MB/event, has become prohibitively expensive. In fact this example illustrates a question. Should the computing model drive the cost model?, or should costs and our willingness to fund software and computing, S&C ,for HL-LHC drive the computing model? I believe it has to be the latter. Given a funding profile for S&C it is our duty to maximize the physics output of the HL-LHC. A strategy for achieving this has to include field wide common software, middle-ware, services, and infrastructure projects.

The Data Challenge

Today the physics reach of the experiment is coupled to and limited by the available resources for triggering, event processing and *data access*. Data collection volumes are dominated by the live time of the accelerator (or number of seconds / year spent in stable beams), the event size, and the experiment's agreed upon average trigger rate per store. CMS will continue to use dynamic pre-scaling strategies to maximize the physics opportunities of the detector, until it is no longer needed due to the LHC succeeding in devising a luminosity leveling scheme that does not decrease the integral amount of data collected (if that is ever possible). Event size scales with instantaneous luminosity but this effect is weaker then data collection live time. Recently the LHC has managed to double it's uptime; this is a remarkable accomplishment. It is the best ever achieved at a hadron collider, and there is no reason to suspect that this will change over the course of the next 20 years of LHC operation. Given all of the above it is easy to predict that the HL-LHC will record exabytes of RAW data, which needs to be processed into additional derived data, and matched by simulated data. The challenge can be states as, how can we improve our services and infrastructures over the next 10 years such that we can collect, process and make accessible to physicist doing data analysis, this huge amount of data?

Traditional Data Handling Methods

The metric that different data handling methods should be judged against is minimizing time to delivery between data simulation/collection and creation of publication quality physics results. Most make publication plots from data set sizes that can fit onto a laptop or some other local resource that is small enough to run on interactively multiple times per day. The path to getting to that size from the 10s of PB of data collected and simulated has traditionally been done with some mixture of skimming, slimming and thinning.

Skimming

Skimming is the term applied to event selection. Data flows from front end crates, to event switch/builder/trigger, to storage manager, to archival storage in the data grid. While it is flowing through these systems it is aggregated into primary datasets based on features of the events built by the trigger. In this way primary dataset definitions are a function of the physics objects reconstructed in real time that will not change by definition. It is in the nature of hadron colliders that it is possible to write O(50) physics primary datasets with minimal overlap. In a way this is a collaboration wide first level skim of all of the data produced by the experiment. While secondary skims based on prompt reconstruction information have been possible in CMS, they get very little for physics analysis. The primary usage of secondary skims is for calibration and detector studies. The skimming done for analysis happens at a much later stage in the workflow after analysis level processing. These analysis level selections are an integral part of the creativity that goes into the science.

Slimming

Slimming is the removal of object collections from events. Experiments do this centrally when they create analysis formats that are smaller then all of the output of the central reconstruction. What is dropped are intermediate object collections such as calorimeter clusters, not generally needed by analysis users.

Thinning

Thinning is the filtering of the remaining collections according to some selection criteria. An example is “thinning the track collection by applying a pT cut which only keeps tracks above the threshold”.

Exactly when in the data processing flow these data reduction methods are employed by the central system and when they are employed by users varies from experiment to experiment but they can all be categorized in this way. The choices are partially based on history and the culture of the experiment, but mostly it is based on what can be afforded. As time goes on and we integrate more data the nature of what can be afforded will change. In CMS our primary analysis format, the AOD, is a thinned event, aggregated into a primary dataset for collected data, and the AODSIM, which is the same event content simulated for a specific physics process. By Run 2 standards this is a “fat” format of order 300-500KB/ event which we keep two copies of in order to assure availability to analysis users around the world. For CMS this is becoming unaffordable, and we are considering moving to a miniAOD format that has been further slimmed and thinned to reach 30-50KB /event, as the only disk resident format. Users needing the fatter AOD format will probably need to move to a model in which a coordinated re-staging from tape and data reduction processing pass is scheduled through the central systems.

Data Handling Methods for HL-LHC

Even with the miniAOD, data sets will become too large in the era of the HL-LHC. While it is possible to imagine that in 10 years there will be exascale computing facilities available for science, there is no one talking about exascale storage facilities. We will need even more compact data formats that currently exist today if we continue with this model. An alternative model being investigated by some is to use big data tools that can create virtual datasets that serve the needs of multiple analysis. Centrally produced miniAOD outputs of

production are loaded into these systems which in turn can skim, slim, and thin the events and event content on the fly in flexible ways that service the needs of multiple analyses. It is not clear yet if the tools provided by industry can scale to the performance needs of HEP at an affordable price, given the 100s of datasets and the large size of each needed for many analyses. It is certainly worth investigating, but it would be foolish to trust that such alternative solutions are the only options to consider.

Summary

There is a lot of R&D work needed over the next 10 years to meet the physics needs of HL-LHC. While it is possible to see a number of promising initiatives in the domain of compute technologies, the clear path forward for world wide federated storage technologies has not materialized yet. Even if some of the current R&D is very successful, traditional data handling methods will be needed to feed those systems. I do not think our current systems will scale to this need. It is imperative that the HEP community come together to find ways of solving these very difficult problems.

[1] CERN-LHCC-2005-023 CMS TDR 7, 20 June 2005