Tips for Production Workflows

Dirk Hufnagel

Muon g-2/SCD Computing Workshop

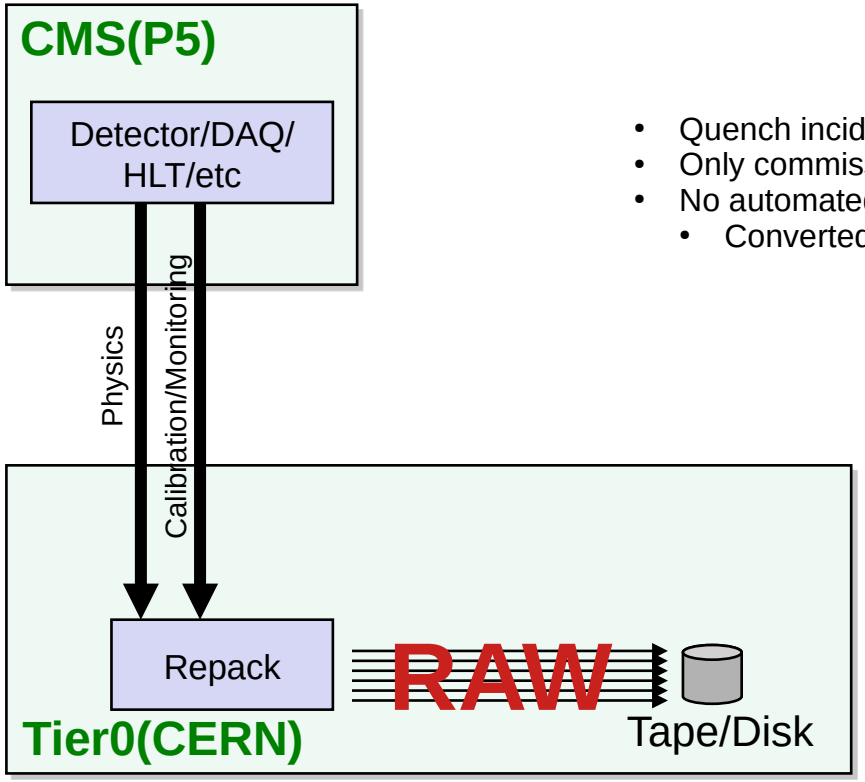
2nd March 2022

Overview

- Don’t expect targeted tips for g-2, but I plan to give a short overview how CMS has handled data taking/processing over the last 10+ years and how the system has evolved over time. Some of what CMS did could be useful/applicable. Had to dig through many old documents and talks, I apologize if I got some details wrong.

- Two main things I’ll talk about:
  - Express/Prompt Reconstruction and Prompt Calibration
  - Disk/Tape separation

2008 : First very simple system

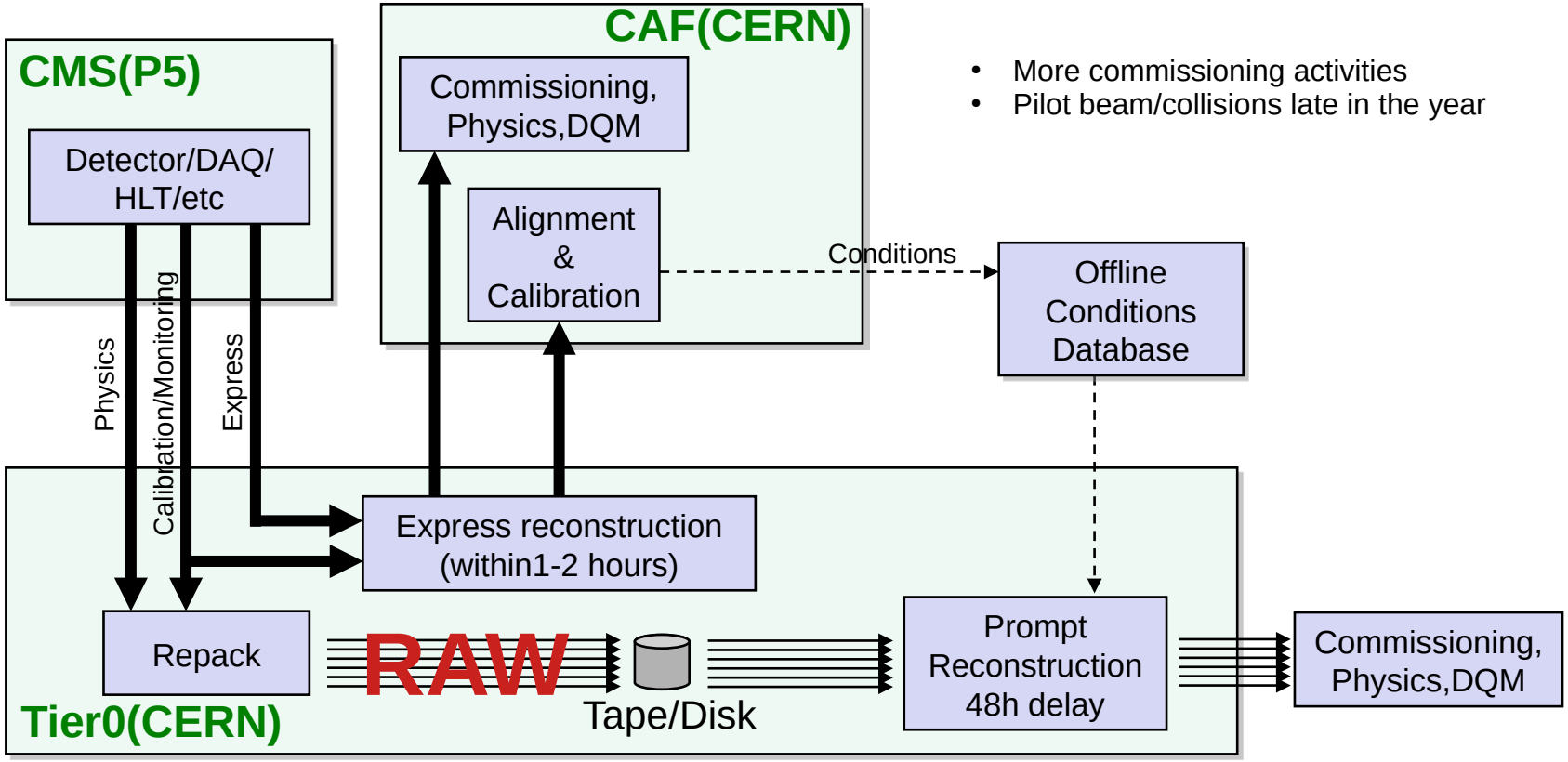


- Quench incident in September 2008
- Only commissioning activities
- No automated data processing
  - Converted data and wrote it to tape/disk

Work in 2008/2009

- Lots of work implementing the workflow management systems to technically run a prompt reconstruction
- Lots of work implementing an express processing mode that could process a subset of the data within O(1h)
- In parallel lots of work from the detector/physics groups to define critical calibration that had to be done before the prompt reconstruction
  - Output of the express processing used for calibration
  - Calibration itself not fully automated/integrated with rest of system
  - No real handshake, rely mostly on prompt reco delay

2009 : Adding Express and Prompt Reconstruction



- More commissioning activities
- Pilot beam/collisions late in the year

Fermilab

What we use(d) PromptReco and Express for

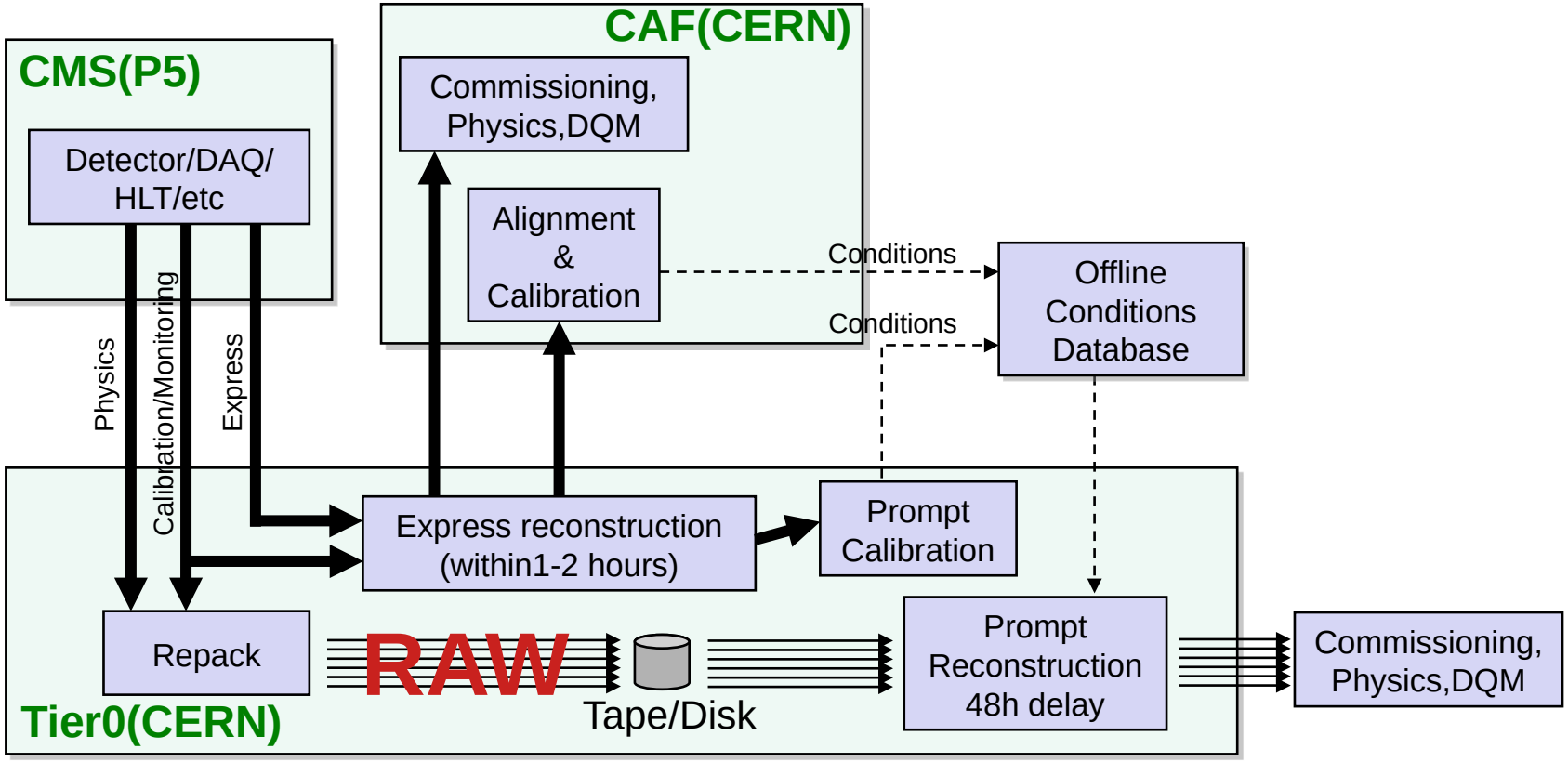
- Express reconstruction output is kept on disk for 30 days (much longer in the past since we were less disk space constrained) and can be used for many purposes (in addition to calibration).
- Prompt reconstruction output is used for data certification, i.e. we look at the quality of every run we take and certify it for physics use (or not).
- PromptReco output was also used extensively by the detector groups to get a better understanding of the detector (and improving the reco algorithms). During the first years of data taking, we kept the full prompt reconstruction output on disk for a relatively long time. Nowadays we don’t even write the largest output anymore (to save tape space) and really only keep the very slimmed down output suitable for analysis on disk.

Work in 2010

- Added a Prompt Calibration that can run multiple algorithm in parallel as part of the Express processing and also automated conditions upload and syncing of conditions uploads with Prompt Reconstruction release.

- Initial system in use since 2011, but (as far as I remember), the full featured Prompt Calibration wasn’t deployed until 2012 or 2013 (also because there was a complete rewrite of the whole Tier0 framework during that same time and the fully featured version only made it into the rewrite).

Since 2011 : 'Final' system



Fermilab

PromptCalibration since 2011

- We started with a single algorithm, the beamspot calculation (since getting that right makes a huge difference for the quality of the tracking in the prompt reconstruction).

- Since then have added many more (current Tier0 runs 6 algorithms).

- The goal was/is to make Prompt reconstruction usable for physics analysis. Not sure to what extent this happened, but even if it’s not used for the final publication, it can still be very useful for initial steps. Also, as data rates have gone up, the frequency of re-reconstructions of CMS data has gone down (now only at the end of year and then at the end of an LHC Run). You want to use this years data, you need Prompt.

How we use(d) tape

- CMS storage at CERN and our T1 were using disk/tape HSM system.
- Early LHC was a resource rich environment compared to the amount of data we had. A disk/tape HSM system can work well in such an environment and is convenient to use (illusion of infinite space).
- As our data volume increased, it worked less well (shattering that illusion of infinite space since in reality the disk buffer is limited, as is the tape recall capacity). We only ran production workflows at CERN and T1 (no analysis), but still had problems keeping workflow input disk-resident.
- We switched to managing disk/tape separately (CERN in 2012, the T1 in 2013).

Working in a separate disk/tape world

- Need to keep track of separate disk and tape endpoints in your data transfer system (easy for CMS since we already had lots of different storage endpoints, adding a few more wasn’t an issue).
- Need to integrate data management into workflow management since organized production activities now always have a data management component, potentially dealing with input data (pre-staging) and always dealing with output data (transfer/cleanup).
  - For CMS this was (and is) hard since we have O(1000-10000) workflows/datasets actively used in production at any given time.
  - For an experiment with (much) less granularity of processing/data it could be easier.