CMS Feedback to LHCC on Analysis Facilities

Nick Smith, on behalf of

Feb 3, 2025

Background

- The LHCC:
  - recommended that experiments engage in the process of developing and defining the structure of the future Analysis Facilities
  - requested they produce a document which defines the use cases in order to establish realistic benchmarks
    - This process should be coordinated with the HL-LHC Computing and SW review panel
    - The document is expected to be regularly updated in the process towards HL-LHC

Experiments are requested to answer the following questions in a focus session at the upcoming LHCC in February 2025:

- Description of the current Run-3 analysis model
- Future analysis model in Run-4 and Run-5
- Managing the evolution of the Analysis Infrastructure

1a. Current Run-3 analysis model

- Main analysis workflows and data reduction steps, including how closely chained they need to be.
  - Main analysis workflows are:
    - (Private) Signal MC production: generation (gridpack, LHE, Pythia, etc.), detector simulation, digitization, reconstruction, reduction (Mini/NanoAOD); much signal MC is handled centrally but individual analysts also produce some additional samples.
    - NTuple production: read central MC and data, produce private format or (custom) nano
    - Primary analysis: slimming/skimming, corrections, histograms
    - Interpretation: fitting (Combine) - usually on aggregated data
  - These workflows are not chained in that they need not be executed in a single workflow. However there are expectations on provenance to ensure that the correct datasets flow through each workflow.
    - Within the Common Analysis Tools (CAT) group, initiatives are underway to capture the analysis pipeline, towards automation and reproducibility/preservation.

1b. Current Run-3 analysis model

- Data formats used for analysis, including their size and level of adoption (current and Run-3 final goal)
  - The most common data format for analysis is NanoAOD, which contains high-level physics object information and is 1 (2) kB/event for data (simulation). It is estimated that over half of CMS analyses currently use NanoAOD. The format is designed to be produced in a flexible manner from upstream MiniAOD to allow customization of the available data products when needed.
  - MiniAOD, 40-50 kB/event, is also in active use primarily as an input to custom data reduction steps, where analysis codes have not migrated to the NanoAOD format.
  - AOD is rarely accessed (~10% as often as MiniAOD) but still made available (including automated tape recall) with CRAB. However, as AOD is 400-500kB/event, the total volume actively accessed by analysis is similar to that of MiniAOD.

1c. Current Run-3 analysis model

- How much compute, storage and network resources are used for Run-3 analysis. Which fraction are pledged and which fraction are used in interactive mode (as opposed to batch).
  - Analysis currently uses on average 20% (~85k cores) of the CMS global pool, varying with production job pressure. The global pool contains pledged or beyond-pledge resources made available to CMS. Analysis compute outside the CMS global pool is challenging to quantify. Interactive resources are typically modest in comparison to the above.
  - Input data for analysis by CRAB is locked in Rucio, and varies from 15-20 PB.
    FTS-orchestrated data transfers for the purpose of user analysis amount to 1-2 PB per week.
    Streaming data transfers for analysis are challenging to quantify, as xrootd monitoring is not robust.

1d. Current Run-3 analysis model

- Comment on what is working well and what is not, both from the point of view of users as well as providers (experiment S&C teams and sites).
  - Positives: analysis is getting done, there are no major restrictions on what people are able to implement.
  - Negatives: robustness of data access: over the past year, approximately 15% of all analysis jobs submitted to the CMS global pool failed due to file open or read errors.

2a. Future analysis model in Run-4 and Run-5

- Comment on which aspects of the current Run-3 analysis model will not scale for Run-4
  - We expect that private signal Monte Carlo production will not scale
  - Linear scaling of primary analysis step turnaround (3-5x) will be challenging for human productivity
  - Scale of ML training growing larger and may require a different infrastructure
  - Esoteric (i.e. low tier) data access via copy-forward nTuple reproduction

2b. Future analysis model in Run-4 and Run-5

- Describe the relevant changes in the model and their impact in resources: policies for number of versions and replicas, fraction of data which is managed vs. unmanaged (e.g. caches), remote vs. local data access, batch vs. interactive cpu/gpu access, need of access to external DBs, or any other.
  - For larger data tiers (AOD) we could not afford to have them on disk
    - Present modeling: 10% AOD on disk 1 miniAOD copy
  - More caches
    - Fractions are subject of R&D but not known yet
  - Pushing more for smaller data tiers
  - We would like to keep batch and also interactive (both cpu and gpu) and provide automated tools for managing batch infrastructure as a part interactive workflows
  - We will also need to provide network infrastructure commensurate to interactive timescales on expected datasets

2c. Future analysis model in Run-4 and Run-5

- Annual volume expected for the different data formats, both data and MC.
  - In Run 4, approximately 6e10 data and 14e10 simulation events per year are expected.
  - The Run 4 size estimates for AOD/Mini/Nano are 1400/180/4 kB per event, corresponding to a total 280/36/1 PB per year.

3a. Evolution of the Analysis Infrastructure

- Describe the user requirements for analysis in HL-LHC and the processes that will be used to track their evolution in the next few years.
  - Feedback cycle between facility designers and users https://arxiv.org/abs/2404.02100
  - Surveys

3b. Evolution of the Analysis Infrastructure

- Comment on which new technologies or emerging paradigms you expect to be needed or have a relevant impact on the future Analysis Infrastructure and which mechanisms can be set up to manage this evolution as new technology will appear (e.g. ML, GPUs/FPGAs, etc).
    - The interface between analysis facilities and ML training needs further and broader consideration. Should this be built in as a first class operation, or only a specialized case offered at specific (but not all) facilities? Hyperparameter scanning in particular is highly compute-intensive and benefits from central coordination at dedicated facilities.
    - As ML becomes more commonplace, there needs to be facility(s) with the right hardware to do large-scale training, which is something not readily available from the grid.
    - More actively to investigate object stores for fine-grained data access
    - SBI / differentiable analysis, inference at HL-LHC scale: jet taggers, basic classifiers, etc.

3c. Evolution of the Analysis Infrastructure

- Describe the plans to develop specific use cases that can be used to benchmark different building blocks of the Analysis Infrastructure so that a comparison can be made between different implementations.
  - IRIS-HEP is proposing to host a workshop in March 2025 together with ATLAS, CMS to identify a representative set of physics analyses, described in terms of workflow and computational needs. The workshop will be based on a survey in February to gather input for physics analysis examples and should result in a document summarizing these analyses, alongside an extrapolation of how we expect them to evolve at the HL-LHC.

3d. Evolution of the Analysis Infrastructure

- Comment if you think that support for analysis workflows in Run-4 will need specialized infrastructure different from the Grid. If so, please describe what features that Analysis Infrastructure will need to provide to expand the one in the Grid.
  - The original columnar analysis facility concept was to keep recently and/or frequently accessed columns in fast-access memory rather than having to cold start the analysis from disk every time, in order to allow rapid iteration and promote column sharing among analyzers/groups
  - Some interactive or machine learning workflows in order to match the HL-LHC magnitude will be not compatible with current grid specs
    - https://doi.org/10.1016/j.cpc.2023.108965 some systematic tests on T2 current specs

3e. Evolution of the Analysis Infrastructure

- Describe the current status and the R&D work that is underway towards implementing relevant Analysis Infrastructure functionality.
  - Exploring models where a central hub provides a seed of resources and scales out over heterogeneous resources based on user needs (INFN, Coffea-casa)
  - Hardware benchmarking activities (INFN, Purdue AF)
  - Investigation and adoption of WLCG bearer tokens in AF (Coffea-casa early adopter)
  - Scaling computing capabilities on AFs to HL-LHC rates: introducing 200 Gbps and 400 Gbps challenges (Coffea-casa)
  - Working on the improvements of “Analyst”-”Facility”-”Framework” feedback loop, one of examples was to introduce the benchmark analysis such as AGC