Object Storage for CMS in the HL-LHC era

Contact PI: Bo Jayatilaka (Fermilab)
Co-PI: David Mason (Fermilab)

1 Introduction

The HL-LHC will present unprecedented scientific opportunities as well as technical challenges in detector design and in computing infrastructure. One pressing challenge in computing infrastructure is how to address the data-storage needs that CMS will have in the HL-LHC era. A combination of higher pile-up and finer detector granularity will result in each collision event of data or simulation requiring considerably larger byte-storage [1]. The overall event rate to storage, currently at approximately 1 kHz, is also anticipated to increase to at least 7.5 kHz. In the current CMS computing model, the RAW output of the detector is committed to archival storage (usually tape) while active storage (usually magnetic disk) is dominated by intermediate derived/reduced formats such as AOD and miniAOD for both collider data and simulation, from which end-user ntuple formats are derived. A proliferation of end-user analysis formats has necessitated frequent access of AOD/miniAOD tiers up to the present day. More universal end-user ntuple formats such as nanoAOD (effectively a flat collection of columns) have helped reduce the need to access intermediate tiers. CMS computing plans for the HL-LHC assume that a majority of analyses will utilize centrally produced nanoAOD and that the bulk of intermediate formats currently kept on disk (and often with multiple copies) will not be kept on active storage. Even with such assumptions, the disk storage needs of CMS in the first year of Run 4 will exceed one Exabyte across all sites, representing a factor of four increase over the needs in Run 3 (Fig. 1) [2].

Formats such as nanoAOD always involve a trade-off in that only a selected set of information is kept for each event. Innovative physics analyses can, and often do, require quantities not saved in nanoAOD as they were not considered of wide importance at the time the centrally produced version was defined. Any analysis that requires columns not stored in centrally-produced nanoAOD will require running over AOD/miniAOD tiers to extract the necessary columns and add them to nanoAOD. This process is cumbersome and time-consuming and could possibly be computationally prohibitive in the HL-LHC era as well as prove to be a stumbling block for new analysis ideas.

One inherent limitation in the current CMS computing model is organization of data around files (and collections of files) stemming from a dependence on file-based storage systems early in the LHC era. File-based organization makes it highly inefficient to access event-level or sub-event-level information without reading through a significant portion of a file first. Object stores, which operate on a key-value principle, allow for efficient access to granular information via metadata lookup. The evaluation of sub-file or object-based

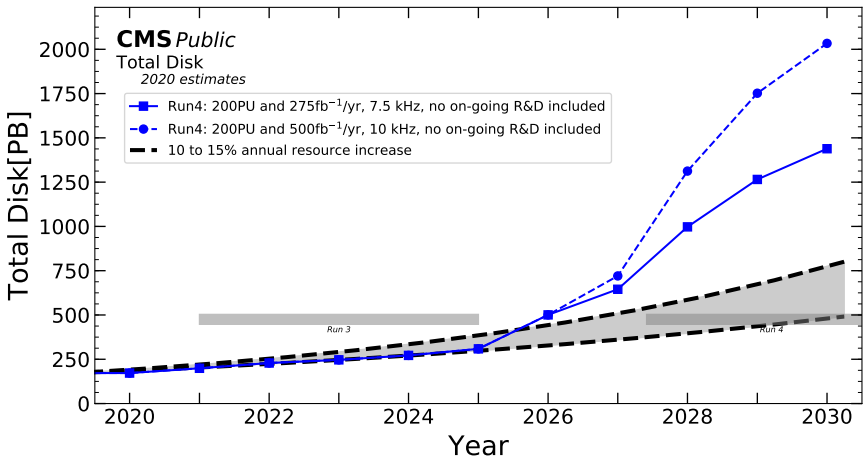

Figure 1: Annual disk space requirement estimated for CMS processing and analysis needs.

granularity is a key R&D goal for Data Organization Access and Management (DOMA) in the HEP Software Foundation Community White Paper [3]. A distributed object store of CMS data can potentially eliminate the need for analysis ntuple tiers, allowing physicists to access only relevant columns from storage. Object stores are widespread in industry and are near-universal in cloud storage with the Amazon S3 API emerging as a *de facto* access protocol across them. Thus, a storage architecture for CMS that exploits object storage would easily extend to cloud storage resources as well. An object store distributed across several sites, or a data lake [4], can also provide the backbone of a content delivery network (CDN) for CMS physicists and allow for analysis without need for an analysis ntuple format. An object store that allows for selected column access via mechanisms such as ServiceX [5] also provides a naturally compatible storage layer for columnar analysis tools such as COFFEA [6].

In this proposal we present the initial steps towards widespread use of an object paradigm for storing HL-LHC CMS data. Such adoption could drastically reduce the required storage footprint for CMS during the HL-LHC, provide considerable gains in speed and efficiency in accessing data at a sub-event level, and potentially allow for more innovation in physics analysis due to less overhead and constraints pertaining to analysis file format. We believe that embracing the object storage paradigm would allow for future CMS computing models to embrace the concept of "virtual data": where derived data formats are never stored and instead derivations and corrections are performed on subsets

of data as needed.

2 Proposed Research

This proposal seeks 50% remuneration for a postdoc to study the use of the Ceph [7] object store for CMS data. Ceph is a widely-used object storage system both in industry and in academia and has a host of features which make it suitable for a large distributed storage architecture such as the HL-LHC will require. A first stage will involve storing CMS data from centrally produced sources as objects and accessing those objects for physics analysis entirely within one site (Fermilab). Should that stage be successful, a second stage will involve distributed object storage of CMS data across multiple sites, effectively prototyping an object store data lake. The second stage may span beyond the initial year this proposal covers depending on collaboration with US CMS Tier-2 sites. While some CMS sites currently use Ceph as a basis of a filesystem (CephFS), there have been no R&D activities to date exploring the use of Ceph as an object store for CMS data.

A test instance of Ceph storage nodes is currently being established at Fermilab in order to assess the technology for storage of liquid argon time project chamber (LArTPC) event data such as from the DUNE experiment. The initial configuration of this set of servers will contain approximately one Petabyte of raw storage. It is envisioned this setup could be augmented or replicated with hardware expected to be retired out of the US CMS Tier-1 or Fermilab LPC storage. The postdoc will utilize this hardware to carry out the initial steps of the research plan. The Fermilab computing professionals who are setting up and administering these servers will help the postdoc gain expertise on the configuration and architectural side of Ceph.

The postdoc will first adopt a schema for storing centrally produced CMS data as a series of objects with associated metadata. The suitability of community tools such as SkyhookDM [8] will be explored for this step. A natural first format to attempt this with is miniAOD since many analysis ntuples are derived from it. One particular challenge will be efficiently serializing the objects from data. Ideally the modular data structures produced by CMSSW [9] (the EDM format) will facilitate both the necessary indexing into objects as well as creation of affiliated metadata. The postdoc will work with CMSSW developers and other experts at Fermilab (such as Chris Jones and Matti Kortelainen) to evaluate if there are any necessary changes to CMSSW as part of this program.

Once the ability to store and access CMS data as objects is established, the postdoc will choose a representative CMS physics analysis (ideally one that they are already involved in) and upload full data samples and a representative subset of simulation samples necessary for the analysis into Ceph. From these data, the postdoc should be able to run analysis code that accesses objects from Ceph directly rather than via reduced analysis ntuples. Benchmark physics quantities should be compared from the object-store derived analysis to an analysis using traditional reduced ntuples such as nanoAOD. Analysis access

should be made from traditional user analysis clusters such as the Fermilab LPC as well as from dedicated elastic analysis facilities currently being prototyped at Fermilab. Analysis performance (time-to-insight as well as resources consumed) using the new system should be compared to 1) "traditional" analysis frameworks using reduced ntuples, 2) columnar analysis frameworks using reduced ntuples, and 3) cases (1) and (2) requiring additional columns not in reduced ntuples. The postdoc will also work with Fermilab computing staff to develop automated workflows to perform bulk moves into the object store for initial population as well as be the foundation for data ingest in future production workflows.

Following the basic feasibility tests outlined above, the postdoc will move on to tackling scale-out issues regarding object storage. The primary challenges involve: multi-user/tenant environments and distribution across multiple sites. The postdoc will work with the elastic analysis facility team at Fermilab to devise multi-user testing of object store access. A stretch goal for this project will be working with US Tier-2 sites to test multi-site object storage, or data lakes. This work would ideally follow the direction established in the WLCG DOMA working group and data lakes and would likely be a focus of a second year plan, should funding be approved.

3 Milestones

Approximately quarterly milestones for the postdoc's work plan are listed below. These milestones are assumed to begin from the postdoc's start date and (if applicable) following an initial familiarization with CMS and CMS software.

- Month 1-3: Familiarization with Ceph and development of object/metadata scheme for miniAOD. Demonstrate ability to store and retrieve objects.
- Month 4-6: Upload of collision and simulation data to Ceph as objects/metadata. Development of analysis code to retrieve objects from Ceph.
- Month 7-9: Formulate an automatic workflow to move data in and out of this system. Benchmark performance of analysis code using object storage and compare to using analysis ntuples.
- Month 10-12: Scale testing with multiple users. Present results at international HEP Computing meetings/workshops. Stretch goal: Work with US Tier-2 sites to establish object store data lake prototypes.

4 Mentoring Plan

The Fermilab CMS Department has a well-established postdoc mentoring program [10]. Each postdoc in the group is given wide latitude in choosing physics analysis topics during

their time at Fermilab. Multiple Fermilab scientists are involved in mentoring and supervision of the postdoc, with one serving as an analysis guide (supervisor) for physics work, one as a technical guide, and one as a mentor. The mentor is usually a senior member of the group who regularly discusses career progression with the postdoc and typically is not directly involved in the postdoc's work allowing for a higher-level perspective. A mentoring committee (which PI Jayatilaka also serves on) meets with the postdoc and their mentoring/supervision team at least once a year and offers further feedback. Should this proposal be funded, PI Jayatilaka would serve as the postdoc's technical guide/supervisor. We would also involve Fermilab computing professionals involved in data storage and computing as well as other members of the Fermilab CMS group (both scientists and postdocs) working in computing operations in the postdoc's work and encourage regular meetings with them. Participation in international meetings and workshops in computing will be encouraged and supported as will taking a visible leadership role in these communities. It is expected that the postdoc will present this work at one CMS Software and Computing week and one international HEP computing meeting, at minimum. Postdoc career development and progression are of utmost importance to the Fermilab CMS group, a fact which contributes to over 60% of the group's postdocs moving on to tenure-track positions at universities or laboratories as their next position.

5 Summary

The scale of data storage needs remains one of the most significant computing challenges for the HL-LHC. Data storage can be a cost limitation on the physics potential of CMS in the HL-LHC with much of its needs driven by derivation of analysis formats. Widespread adoption of object storage paradigms can potentially reduce those derivation needs and drive down the cost of storage as well as increase efficiency in access. Our proposal seeks support for a postdoctoral researcher to lay the groundwork for object storage of CMS data.

References

[1] CMS Offline Software and Computing. Evolution of the CMS Computing Model towards Phase-2. Technical Report CMS-NOTE-2021-001, CERN, 2021.

[2] D. Piparo et al. CMS Offline and Computing Public Results. `https://twiki.cern.ch/twiki/bin/view/CMSPublic/CMSOfflineComputingResults`, 2020.

[3] J. Albrecht et al. A Roadmap for HEP Software and Computing R&D for the 2020s. *Computing and Software for Big Science*, 3(1):7, 2019.

[4] Bird, Ian, Campana, Simone, Girone, Maria, Espinal, Xavier, McCance, Gavin, and

Schovancová, Jaroslava. Architecture and prototype of a WLCG data lake for HL-LHC. *EPJ Web Conf.*, 214:04024, 2019.

[5] Galewsky, B., Gardner, R., Gray, L., Neubauer, M., Pivarski, J., Proffitt, M., Vukotic, I., Watts, G., and Weinberg, M. ServiceX A Distributed, Caching, Columnar Data Delivery Service. *EPJ Web Conf.*, 245:04043, 2020.

[6] N. Smith et al. Coffea Columnar Object Framework For Effective Analysis. *EPJ Web Conf.*, 245:06012, 2020.

[7] Sage Weil, Scott A. Brandt, Ethan L. Miller, Darrell D. E. Long, and Carlos Maltzahn. Ceph: A Scalable, High-Performance Distributed File System. In *Proceedings of the 7th Conference on Operating Systems Design and Implementation (OSDI '06)*, November 2006.

[8] Xiaowei Aaron Chu, Jeff LeFevre, Aldrin Montana, Dana Robinson, Quincey Koziol, Peter Alvaro, and Carlos Maltzahn. Mapping Datasets to Object Storage System. *EPJ Web Conf.*, 245:04037, 2020.

[9] C. D. Jones, M. Paterno, J. Kowalkowski, L. Sexton-Kennedy, and W. Tanenbaum. The New CMS Event Data Model and Framework. In *Proceedings of International Conference on Computing in High Energy and Nuclear Physics (CHEP06)*, 2006.

[10] L. Bauerdick et al. Guidelines for Fermilab CMS RA Supervisors, Mentors, and Guides. `https://cd-docdb.fnal.gov/cgi-bin/sso/ShowDocument?docid=5695`, 2017. FNAL-CS-doc-5695-v1.