The Quest to solve the HL-LHC data access puzzle

Xavier Espinal[1], *Stephane* Jezequel[2], *Markus* Schulz[1], *Andrea* Sciaba[1], *Ilija* Vukotic[3], and *Frank* Wuerthwein[4]

[1]European Organisation for Nuclear Research (CERN), Geneva, Switzerland
[2]Laboratoire d'Annecy de Physique des Particules, Annecy, France
[3]University of Chicago, Chicago, Illinois, US
[4]University of California, San Diego, La Jolla, CA, USA

Abstract. HL-LHC will confront the WLCG community with enormous data storage, management and access challenges. These are as much technical as economical. In the WLCG-DOMA Access working group, members of the experiments and site managers have explored different models for data access and storage strategies to reduce cost and complexity, taking into account the boundary conditions given by our community. Several of these scenarios have been studied quantitatively, such as the datalake model and incremental improvements of the current computing model with respect to resource needs, costs and operational complexity. To better understand these models in depth, analysis of traces of current data accesses and simulations of the impact of new concepts have been carried out. In parallel, evaluations of the required technologies took place. These were done in testbed and production environments at small and large scale. We will give an overview of the activities and results of the working group, describe the models and summarise the results of the technology evaluation focusing on the impact of storage consolidation in the form of datalakes, where the use of read-ahead caches (XCache) has emerged as a successful approach to reduce the impact of latency and bandwidth limitation. We will describe the experience and evaluation of these approaches in different environments and usage scenarios. In addition we will present the results of the analysis and modelling efforts based on data access traces of experiments.

1 Introduction

The WLCG strategy paper [1] set out the path towards computing for the HL-LHC era, building up from the input provided by the HSF [2] Community White Paper [3]. The estimates for the data volumes and computing show a major step up from the current needs and a program of work was established from the WLCG point of view to address this future challenge. One of the charges is addressed by the DOMA access working group to evaluate future data access scenarios.

The working group collected information from experiments future plans about analysis data formats, is following-up the work pioneered by the cost model working group to understand file placement and file usage statistics and is investigating data caches infrastructures and promoting the deployment of caching models leveraging data access from a consolidated storage infrastructure labeled as datalake (Fig. 1).

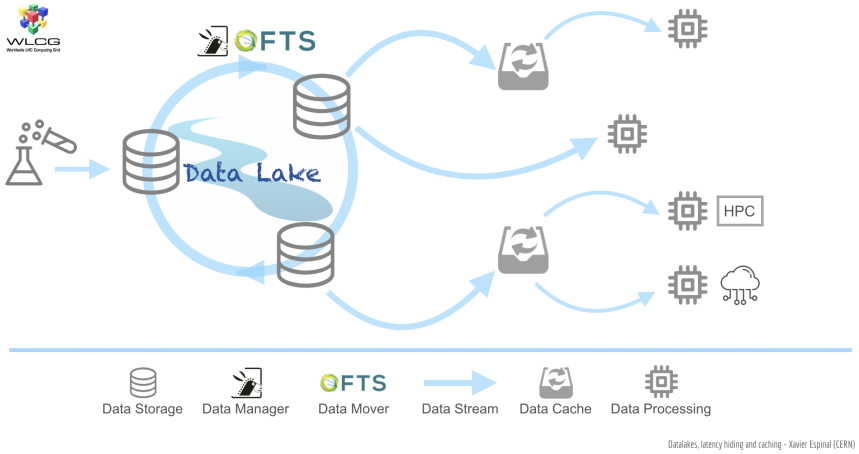

Figure 1. Conceptual sketch of the datalake idea

2 Compact analysis objects

CMS and ATLAS experiments are transitioning towards more compact datasets for analysis with event sizes in the order of the kB/event. This implies that a full analysis datasets will be close to 1PB per year taking CMS numbers on their compact analysis objects (nanoAOD), to be compared with the 50PB per year for older analysis objects (miniAOD), these sizes has been estimated taking as a reference LHC delivery of 80 billion events/year (data) and the production of 160 billion events/year (MC) together with expected sizes for the different data types of 7.4MB(RAW), 2.0MB(AOD), 200kB(miniAOD) and 4kB(nanoAOD). These compact objects open the window to evaluate new ways of doing computing and offer different options for the sites currently providing computing and storage resources. In particular storage has been identified as the main challenge for HL-LHC due to the increasing use of disk volumes, and also the costs from the site perspective to operate and maintain complex storage systems.

One of the goals of the working group is to study the feasibility to exploit these new analysis objects and its reduction in size to promote less demanding storage at the sites (e.g. stateless storage) while fostering the efforts towards computing resources (CPU, GPU, ML, etc.) with the possibility to access the full analysis objects from a centralized storage through caches to minimize latency and increase file re-usability at site or even at federation/regional level. At this point in time there is the clear need of the engagement from the physics community to converge using these new compact objects and need to be emphasized that the overall goal is to be able to maximise the physics potential of the machine and hence taking as much data as possible which means to maximise the use and efficiency of the global storage resources. Use of storage is a delicate concept, a full storage does not mean it is used in an efficient way.

3 File usability and data access patterns

One of the key parameters to assess how effective we are in using storage is to measure the access frequency after data placement. There are two extremes regarding data thermodynamics: a) cold data, where files are WORN (Write Once Read Never) and b) hot data, where

files are expected to be accessed continuously and with high concurrency. But after studying data access patterns at several sites we observed that large fraction of our files are neither cold nor hot. The analysis objects files seems to lose popularity with time and the access rate decreases significantly after days/weeks, in (Fig. 2) it is shown the file access rates and file popularity on a Tier-1 and a Tier-2 as a function of time as a representative example.

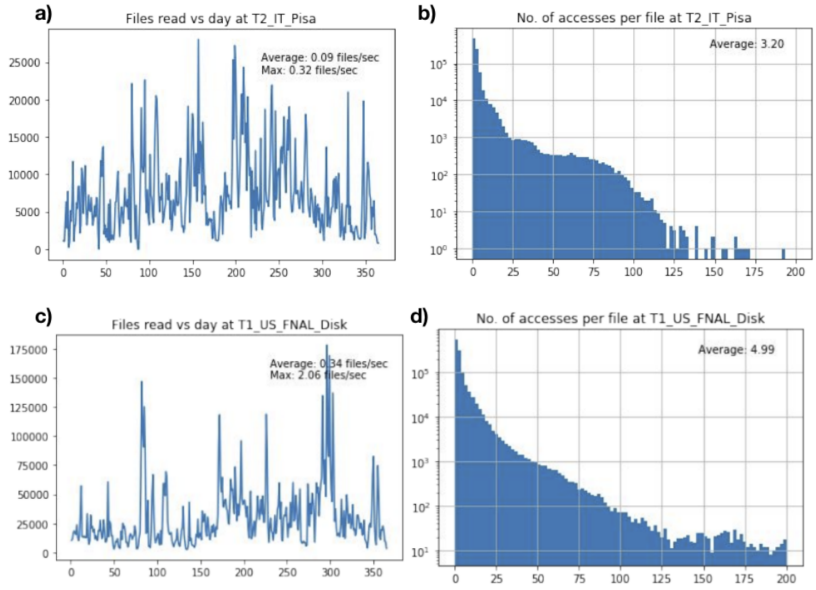

Figure 2. *(left)* File popularity on a Tier-1 and a Tier-2 as a function of time (300 days). The plots above indicate that data is not accessed very often, it is most likely to be re-read within days after placement then the access drops substantially, almost two orders of magnitude.

This provides an indication whether this type of data could be better handled with cache, so it is available when popular and gets superseded with newer files once they are less demanded. In this way the space on disk at the computing sites is optimised for data being actively used, can this be completely delegate to a stateless cache? In parallel less frequently used data might be re-fetch again from the datalake (disk or tape) where the experiments will handle with the required Quality of Service (QoS) label to make use of the best cost/usage ratio.

We also observed a fundamental difference between Analysis and Production data. Analysis has higher re-use while production files have very few re-reads. The net effect is that running combined workflows on a site has the effect that production file push analysis data out of the storage/cache. This made us think that in the case we do not change much of the current infrastructure we would nevertheless benefit by changing the current model and favoring running predictable and time-defined workflows at the sites with less storage and favor less-predictable user analysis on sites with larger storage services.

Need to keep in mind these are preliminary studies based on a period of half a year. This need to be extended in further studies and also need to combine with staging and data deletion information. Nevertheless the results provide hints towards a cache-oriented storage.

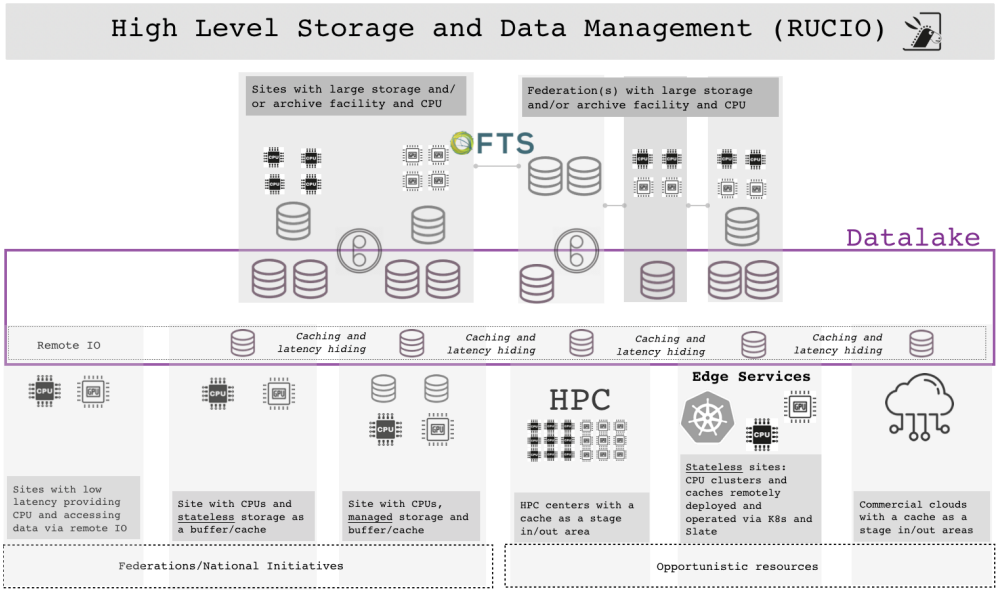

Figure 3. *(center)* Datalake sketch composed by sites and federations holding the bulk of the data regions, and the different types of computing-oriented sites, commercial clouds and HPCs accessing the datalake

4 Data caching: concept, infrastructure and initiatives

Simulations of caching layers based on reference WLCG workloads showed a very good ability for latency hiding even when data is read for the first time [1]. Within the root framework [6] there is the possibility to cache data (reading ahead) while the file start to be accessed, this has good ability after the first bytes arrive to the worker node and is very effective for latencies up to 10 ms when the job configuration parameters are adjusted to the job data accessing patterns (TTreeCache root configuration). For higher latencies the impact start to be noticeable and CPU inefficiencies grow with the increase on latency.

In WLCG we do have more than 160 sites all with different roles and scopes e.g. different experiments, local communities, analysis groups, etc. and with a big variety on network topologies and latencies. Sites might have different interests related to the future of their sites and in particular on their storage services. The new analysis models together with the datalake concept offers more flexibility for the sites to re-shape their services according to their needs and future perspectives. An example could be a modest Tier-2 currently providing storage and computing to WLCG experiments, needing to maintain a storage system which they have few interest on. If This site is close enough to a the datalake they can think about accessing data remotely, and if they are on a more distant (in network latency units) place they could interface with data by deploying a stateless storage as a caching layer for latency hiding and eventual file reusability. In Fig. 3 is shown a tentative sketch envisioning a datalake composed by sites and federations holding the bulk of the data regions, and the different types of computing-oriented sites, commercial clouds and HPCs accessing the datalake.

[1] The simulations have been conducted using using XCache technology (from the xrootd software framework [5])

The working group has promoted the deployment of several caching models to operate in a region and on a site level. We are investigating three different approaches: a) High performance caching servers in USA to feed a region Southern California (SoCal) and Chicago and b) caching federation to feed data to regional sites (France and Italy examples) and c) site caching mechanism as stateless Tier-2 storage (Munich and Birmingham). The results obtained confirm that caching is a promising mechanism to address the analysis challenge and help increasing an efficient usage of the storage and hence able to optimise the overall cost (meaning be able to store and deliver as much data as possible for the HL-LHC). The caching layer setup at SoCal demonstrate that three sites (Riverside, Pasadena and San Diego) can benefit of a common caching layer of 1PB (c.f. with the old model where the site had to deal with 5PB of statefull storage installation), this cache can serve 90% of the jobs/user request at 1/5th of the cost in hardware and alleviating the site to manage a complex storage service. The initiative in LMU Munich demonstrated that an old disk pool node, with simple hardware configuration (JBOD) and simple XCache deployment could serve up to 3k jobs of ATLAS workflows reading data from the neighbour site in Hamburg (DESY) and from a far site in China (IHEP in Beijing). The test concluded that reading from the neighbor site and from the far site is no longer a killing option for the CPU but reasonable when comparing distance/latencies (Fig. 4).

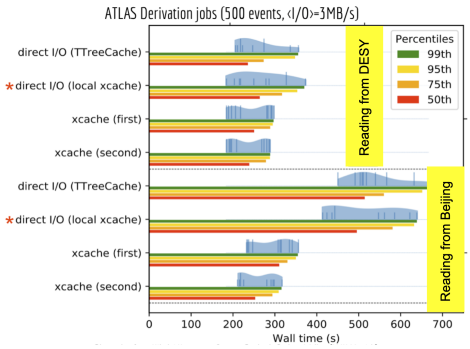

Figure 4. *(center)* XCache running on modest hardware at LMU. Successfully served 3.2k analysis and derivation jobs from ATLAS with and average I/O of 1MB/s and 3MB/s respectively (plot taken from N. Hartman (LMU). Effective latency hiding can be easily achieved for high latency data consumption.

5 Conclusions

The DOMA access working group is focusing the end of their mandate by December 2020. The goal is to wrap up on the investigations performed during the two years mandate (2019 and 2020) and provide input and recommendations about the possible future directions to address data access from the analysis data perspective. In this coming year we will learn more about the operations and performance of the different caching infrastructures running worldwide for ATLAS and CMS and we will also will have a clearer picture of experiments commitment for the envisaged compact data objects for analysis.

We do see that the implications on the datalake storage infrastructure as a data source and the data workflows are a dominant factor and hence we feel the working group would need to evolve and start addressing in detail the storage consolidation in the form of datalakes and looking at the data access taking into account the full picture: detectors, data distribution

and storage, experiment workflows and analysis workflows. We do think the experience and information gathered during this initial mandate will provide precious guidelines for this future work towards a new data storage infrastructure and new data processing modelling to start being evaluated during Run-III.

References

[1] Worldwide LHC Computing Grid, http://wlcg.web.cern.ch

[2] HEP Software Foundation, https://hepsoftwarefoundation.org/

[3] HEP Software Foundation, *A Roadmap for HEP Software and Computing R&D for the 2020s*, arXiv:1712.06982 (2018)

[4] I. Bird, S. Campana *https://cds.cern.ch/record/2621698*

[5] A. Hanushevsky *et al*, https://xrootd.slac.stanford.edu/

[6] Rene Brun and Fons Rademakers, ROOT - An Object Oriented Data Analysis Framework, Proceedings AIHENP'96 Workshop, Lausanne, Sep. 1996, Nucl. Inst. & Meth. in Phys. Res. A 389 (1997) 81-86. See also root.cern.ch/. D. Lange *et al*, *CMS Computing Resources: Meeting the demands of the high-luminosity LHC physics program*, these proceedings

[7] R. Vernet, J. Phys.: Conf. Ser. **664**, 052040 (2015)