EPJ Web of Conferences **245**, 04027 (2020) https://doi.org/10.1051/epjconf/202024504027
CHEP 2019

The Quest to solve the HL-LHC data access puzzle

X. Espinal[1], *S.* Jezequel[2], *M.* Schulz[1], *A.* Sciabà[1], *I.* Vukotic[3], and *F.* Wuerthwein[4]

[1]European Organisation for Nuclear Research (CERN), Geneva, Switzerland
[2]LAPP, Université Grenoble Alpes, Université Savoie Mont Blanc, CNRS/IN2P3, 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 evaluated quantitatively, such as the Data Lake 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 Data Lakes, where the use of streaming caches 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 the experiments.

1 Introduction

The WLCG strategy paper [1] set out the path towards computing for the High-Luminosity LHC (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.

During the first year, the DOMA Access Working Group collected information from the experiments about the evolution of their computing models and future plans for user data analysis.

The working group is investigating the potential benefits of data caching infrastructures and promoting their deployment within a consolidated storage infrastructure labeled as WLCG-Data Lake (Fig. 1), from now on referred to as the Data Lake.

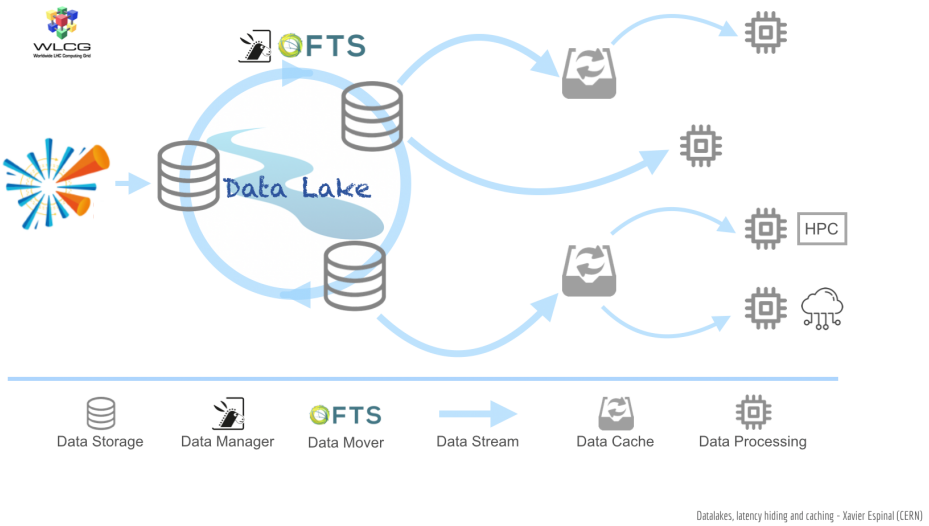

Figure 1. Conceptual sketch of the WLCG-Data Lake

2 Compact analysis objects

To speed up the analysis process cycle and to optimise the storage, the CMS and ATLAS collaborations are transitioning towards more compact datasets for analysis with event sizes in the order of the kB/event. Based on the numbers for compact analysis objects (nanoAOD [4]) provided by CMS, a full analysis dataset will be close to 1 PB per year. This is largely lower than the 50 PB per year for older analysis objects (miniAOD)[1]. Compact objects open the window to evaluate new ways to address the user analysis challenge and propose different scenarios for the grid computing sites currently providing both computing and storage resources. In particular storage has been identified as the main challenge for HL-LHC due to the increasing volume of disk storage used, 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 propose and evaluate new scenarios for data access leveraging the potential benefits of a Data Lake model:

- Benefit from these new analysis data format and their reduction in size which are expected to be heavily accessed
- Reduction of number of site with Grid storage (e.g. stateless storage).
- Provide efficient access to analysis datasets to diverse computing resources (CPU, GPU, Machine Learning, etc.) to access the full analysis datasets.

One of the current approached is to investigate content delivery through caching layers infrastructures to minimize latency impact and increase file re-usability, at the site level or at regional level. The engagement of the physics community will be crucial to converge on these new compact objects.

[1] these sizes have 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 7.4 MB(RAW), 2.0 MB(AOD), 200 kB(miniAOD) and 4 kB(nanoAOD)

3 File usability and data access patterns

One of the key parameters to assess our effective storage usage is to measure the access frequency after data placement. The two extremes regarding data thermodynamics are *cold data* where files are WORN (Write Once Read Never) and *hot data* where files are accessed quickly after data placement and with high concurrency. After studying data access patterns at several sites we observed that large fraction of our files are neither totally *cold* nor *hot*. The analysis files lose popularity with time and the access rate decreases significantly after days/weeks, in (Fig. 2) the file rates and file popularity on a Tier-1 and a Tier-2 are shown as a function of time as a representative example. This provides an indication whether this

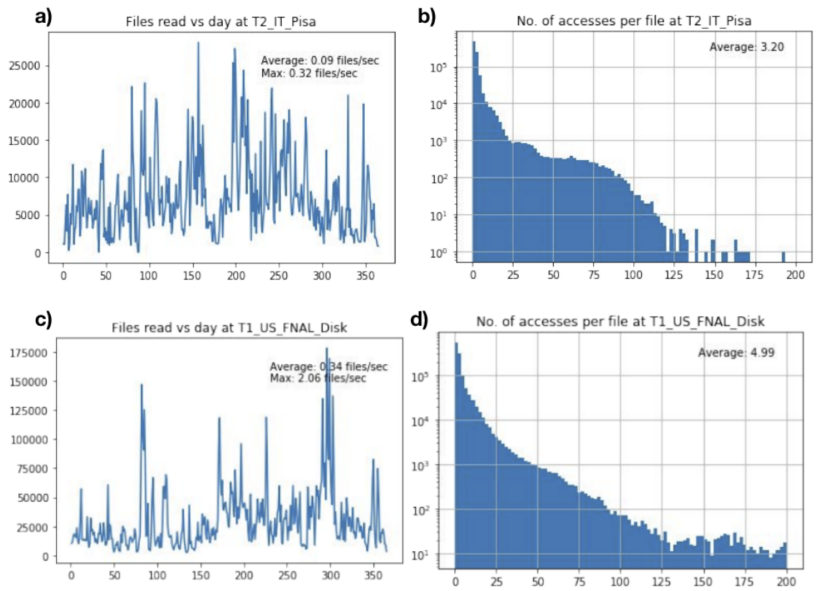

Figure 2. File popularity on a Tier-1 and a Tier-2 as a function of time (300 days). The plots above indicate that data is not accessed very often, it is most likely to be re-read within days after placement then the access drops substantially, almost two orders of magnitude.

type of data could be better accessed through a cache, so it is available when is popular and gets super-seeded 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 and this can potentially be completely delegated to an *stateless* cache. In parallel less frequently used data might be re-fetched again from the Data Lake (disk or tape) where the experiments will handle the popularity with the required Quality of Service (QoS) to make use of the best cost/usage ratio for the storage.

We also observed a fundamental difference between analysis and production data. Analysis has higher re-use while production files have very few re-reads. As a result running combined workflows on a site has the effect to push analysis data out of the cache.

It should be noted that these observations are based only on a period of six months but they provide hints towards a cache-oriented storage. Further studies should be done on longer period and also combined with staging and data deletion information.

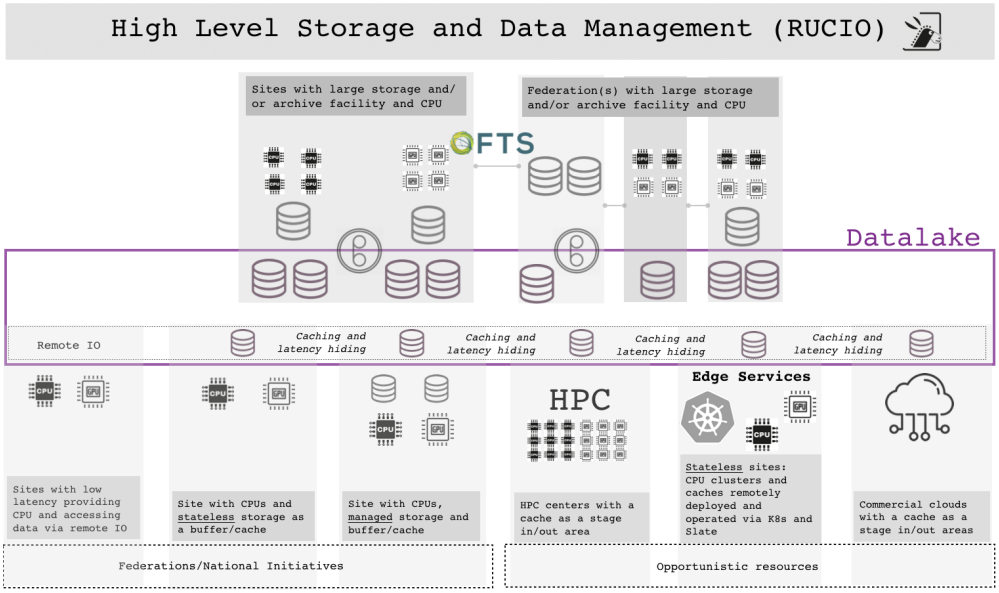

Figure 3. *(center)* Data Lake sketch composed by sites and federations holding the bulk of the data regions, and the different types of computing-oriented sites, commercial clouds and HPCs accessing the Data Lake

4 Data caching: concept, infrastructure and initiatives

Simulations of caching layers based on reference WLCG workloads showed the ability to hide latency even when data is read for the first time. The simulations have been conducted using using XCache technology (from the xrootd software framework [5]).
It should be noticed that Within the root framework [6] it is also possible to cache data from the client side while the file starts to be accessed, this is usually as as reading ahead. Read ahead ability is very effective for low latencies and enables the remote reading option for sites close to the main source of the data, the Data Lake and enables sites to opt for an storage-less approach. For moderate to high network latencies the impact of Round Trip Time (RTT) start to be noticeable and CPU inefficiencies grow with the the increasing latency.
In WLCG there are around 160 facilities spread worldwide contributing to the global storage and computing infrastructure. The sites has different roles and scopes, they may supporting different experiments and many of them have local scientific communities. Among these many sites there is big variety on network topologies and hence a wide range of network performance metrics. The working group is investigating how to enable storage-less or state-less storage approach for sites that are interested mainly in processing facilities without the burden to operate and maintain a full fledged storage system. These sites will use the Data Lake as the main source of data and will access these data via caches (state-less approach) or remote access (storage-less approach) form their data processing facilities.
In Fig. 3 is shown a tentative sketch envisioning a Data Lake 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 Data Lake.

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 perfor-

mance caching servers in US, feeding three sites in Southern California; b) caching federation to feed data to regional sites, this is teh case for INFN/Italy: CNAF, Bari and Legnaro ; and c) a site caching mechanism as state-less Tier-2 storage: Munich (LMU) and Birmingham (BHAM).

The caching layer setup at SoCal demonstrated the three sites UCSD, Caltech and Riverside can benefit from a common caching layer of around 1 PB (c.f. with the old model where the site had to deal with 5 PB of state-full 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 a simple hardware configuration (JBOD) and simple XCache deployment could serve up to 3k concurrent 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 the difference in CPU efficiency when reading from the neighboring site and from the far site is no longer a show-stopper taking into account the distance and the latency (Fig. 4).

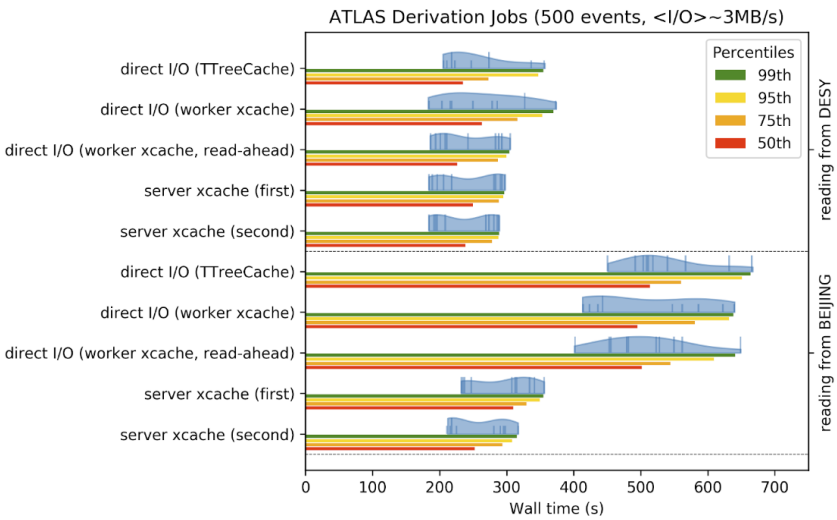

Figure 4. *(center)* XCache running on modest hardware at LMU. Successfully served 3.2k analysis and derivation jobs from ATLAS with and average I/O of 1 MB/s and 3 MB/s respectively. Effective latency hiding is achieved for high latency data consumption.

5 Conclusions

The DOMA Access Working Group is providing input and recommendations about the possible future directions for addressing the data access challenges in the HL-LHC scenario from the user analysis perspective. In the upcoming year we will increase our understanding about the operations and performance of the different caching infrastructures that are running worldwide for ATLAS and CMS. We are also engaged with the physics analysis community and liaising with the HSF Analysis working group to understand the future needs from the physics community for the HL-LHC. The goal is to provide and prototype the infrastructures

able to cater with the HL-LHC data processing demands.
The implications on the Data Lake storage infrastructure as a data source and its role in the experiment data workflows are the dominant factor. The DOMA ACCESS working group need to evolve and start addressing in detail the full picture from data storage, data distribution and data access to the combined impact of workloads and their requirements on the infrastructure.
The experience and information gathered during the initial mandate of the working group will provide precious guidelines for this future work towards a new data storage infrastructure and new data processing models that should start being evaluated during LHC Run-III.
The obtained results on content delivery with the different caching infrastructures confirm it as a promising mechanism to address the analysis challenge. This approach also promote an efficient use of the storage at the sites and hence help to optimize the overall storage cost while still meeting the HL-LHC data storage needs.

References

[1] I. Bird, S. Campana *https://cds.cern.ch/record/2621698*
[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] A. Rizzi, G. Petrucciani and M. Peruzzi for the CMS Collaboration *Further reduction in CMS event data for analysis: the NANOAOD format* EPJ Web of Conferences 214, 06021 (2019)
[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)
[8] Worldwide LHC Computing Grid, http://wlcg.web.cern.ch