EPJ Web of Conferences **226**, 01007 (2020) https://doi.org/10.1051/epjconf/202022601007
Mathematical Modeling and Computational Physics 2019

Enabling Data Intensive Science on Supercomputers for High Energy Physics R&D Projects in HL-LHC Era

Alexei Klimentov[1], *Douglas* Benjamin[2], *Alessandro* Di Girolamo[3], *Kaushik* De[4], *Johannes* Elmsheuser[1], *Andrej* Filipcic[5], *Andrey* Kiryanov[6,10], *Danila* Oleynik[7], *Jack C.* Wells[8], *Andrey* Zarochentsev[9,10], and *Xin* Zhao[1] on behalf of ATLAS Collaboration

[1] *Brookhaven National Laboratory, NY, USA*
[2] *Argonne National Laboratory, IL, USA,*
[3] *European Particle Physics Laboratory (CERN), Geneva, Switzerland*
[4] *University of Texas in Arlington, TX, USA*
[5] *Josef Stefan Institute, Ljubljana, Slovenia,*
[6] *Petersburg Nuclear Physics Institute NRC "Kurchatov Institute", Gatchina, Russia*
[7] *Joint Institute of Nuclear Research, Dubna, Russia*
[8] *Oak Ridge National Laboratory, TN, USA*
[9] *Saint-Petersburg State University, St. Petersburg, Russia*
[10] *Plekhanov Russian University of Economics, Moscow, Russia*

Abstract. The ATLAS experiment at CERN's Large Hadron Collider uses the Worldwide LHC Computing Grid, the WLCG, for its distributed computing infrastructure. Through the workload management system PanDA and the distributed data management system Rucio, ATLAS provides seamless access to hundreds of WLCG grid and cloud based resources that are distributed worldwide, to thousands of physicists. PanDA annually processes more than an exabyte of data using an average of 350,000 distributed batch slots, to enable hundreds of new scientific results from ATLAS. However, the resources available to the experiment have been insufficient to meet ATLAS simulation needs over the past few years as the volume of data from the LHC has grown. The problem will be even more severe for the next LHC phases. High Luminosity LHC will be a multi-exabyte challenge where the envisaged Storage and Compute needs are a factor 10 to 100 above the expected technology evolution. The High Energy Physics (HEP) community needs to evolve current computing and data organization models in order to introduce changes in the way it uses and manages the infrastructure, focused on optimizations to bring performance and efficiency not forgetting simplification of operations. In this paper we highlight recent R&D projects in HEP related to data lake prototype, federated data storage and data carousel.

1 Introduction: Scale of computing needs for particle physics

The largest scientific instrument in the world – the Large Hadron Collider (LHC) [1] operates at the CERN Laboratory in Geneva, Switzerland. The experiments at the LHC explore the fundamental nature of matter and the basic forces that shape our universe. To address an unprecedented multi-petabyte data processing challenge, experiments are relying on the deployed computational infrastructure of the Worldwide LHC Computing Grid (WLCG) [2]. More than 9000 scientists from 200

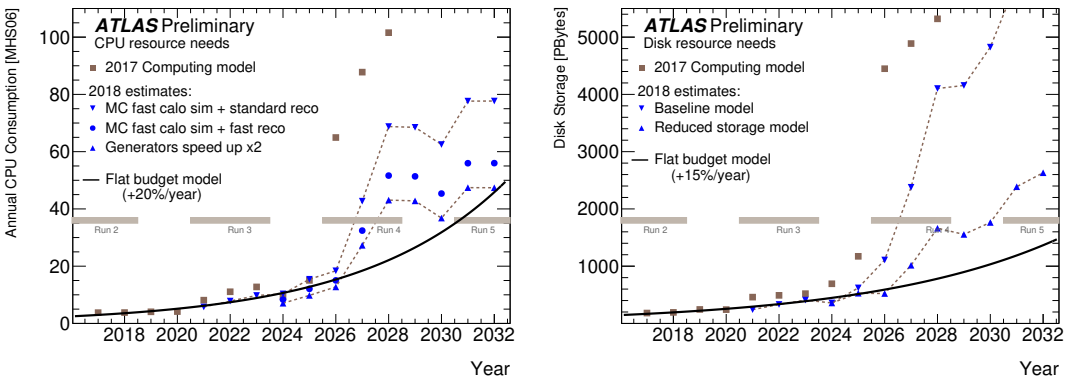

Figure 1. Projection of the ATLAS computing needs in the coming years [6]

universities and laboratories in 45 countries analyze the LHC data in search of new discoveries. ATLAS experiment [3] leads WLCG resource usage in the number of jobs completed, processed data volume, and in the core-hours used for High Energy Physics (HEP) experiments. Since the start of LHC data taking, ATLAS operates under conditions in which contention for computing resources among high-priority physics activities happen routinely.

Scientific priorities in High Energy and Relativistic Nuclear Physics (HENP) present Big Data challenges requiring state-of-the art computational approaches, and therefore, serve as drivers of an integrated computer and data infrastructure. For HEP, these priorities include investigating properties of Higgs boson candidates in an attempt to better understand the origin of mass and search for new laws of physics [4]. To avoid potential shortfalls in projected LHC Grid resources, ATLAS is actively using supercomputing-scale resources as an important supplement to keep up with the rapid pace of data collection and to produce simulated events for LHC experiments which are too complex and require enormous computing resources to produce them on the WLCG.

The LHC Run1 (2009–2013) and Run2 (2015–2018) have convinced physicists that their codes need fundamental re-engineering to address the realities of future commodity, highly parallel processors, and that, if they can achieve this re-engineering, they could emerge with codes ready to exploit High Performance and Supercomputing facilities quite well. Explicitly targeting supercomputing in the re-engineering efforts is a tactic that will trickle down to benefit the Ethernet cluster approach that is used for most HENP computing in the past two decades. The U.S. Department of Energy, Office of Science, High Energy Physics and Advanced Scientific Computing (DOE ASCR and HEP) funded the BigPanDA project [5] which has provided the first important demonstration of the capabilities that a workload management system (WMS) can have on improving the uptake and utilization of supercomputers from both application and systems points of view.

The high luminosity LHC (HL-LHC) run will begin operations in 2026 with expected data volumes to increase by at least an order of magnitude as compared with present systems (ATLAS computing and storage needs estimation [6] is presented in Fig. 1). HL-LHC needs for ATLAS and CMS are above the expected hardware technology evolution (15 % to 20 %/yr). Extrapolating from existing trends in disk and compute pricing, and assuming fixed infrastructure budgets, the implications for end-user analysis of the data are significant. The risk is that the current system architecture and tools will not be able to provide the abstractions and data management capabilities to scale to meet the expected growth.This challenge cannot be solved by simply extending the current LHC computing model. New state-of-art technologies need to be applied and potentially developed, leveraging the investments and research already being conducted in the commercial sector. In this paper we will describe two R&D projects to address technical challenges that need to be explored for the HL-LHC.

2 ATLAS High Performance Computing data processing and simulation

ATLAS has successfully integrated High Performance Computing (HPC) facilities into its distributed computing system and used HPCs (including Leadership Class Facilities such as Titan [7]) for more than 5 years. HPC facilities are integrated via different technologies because of the unique nature of HPCs, such as access to external network and WLCG storage endpoints. Several approaches have been developed and commissioned to address differences in the access, authorization and service requirements of the various HPC centers. The flexibility of the ATLAS processing, workload management system (PanDA [8]) and data management system (Rucio [9]) has enabled ATLAS to run most of the workflows on HPCs. ATLAS uses the HPCs in the following ways:

- Grid-like execution when an HPC provides the external connectivity on the nodes and enables access to CVMFS [10] software area either by directly mounting CVMFS on the nodes or using Parrot [11] with access to a Squid service [12]. Any ATLAS workflow can execute on such HPCs. Typically ATLAS agents (ARC-CE or Harvester) are used as a site service for job submission and control. Both payload push and pull can be used, with the push mode being preferred on HPCs without close Storage Element, where downloads and uploads are managed by a dedicated service on data transfer nodes.
- Limited connectivity execution when the nodes do not have outbound network access. The ATLAS software can either be installed locally on a shared file system or provided through fat containers on HPCs that support Singularity or Shifter. Such HPCs typically run ATLAS Event Generation or Geant4 Simulation where conditions data can be stored in a local SQLite file. A dedicated ATLAS software infrastructure is used for job submission and control. Those services can be installed locally on an HPC site or they can be used remotely through ssh connection to the batch system and SSHFS to manage input and output files.

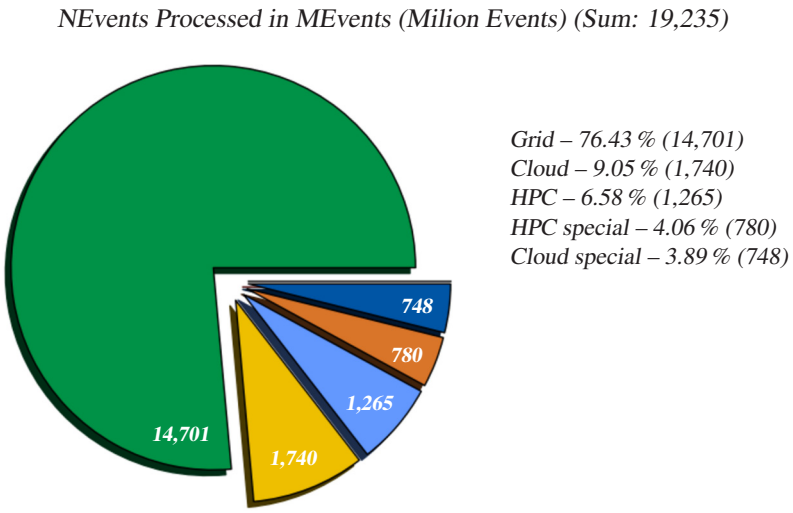

Figure 2. ATLAS Monte-Carlo on HPC and Leadership Class Facilities (LCF) in 2018

Some HPCs provide a subset of the services that are required on grid sites, e.g., Squid, CVMFS, Computing Elements (CE), disk cache. If ATLAS is able to transparently use any of these, the HPC services are then described in detail in information system (AGIS). Table 1 shows current workflows

Table 1. ATLAS Workflows on HPC

Workflow	**CPU/event HS06s**	**cores Architecture**	**CPU**
Event Generation	980	single, some multi-	x86_64, Power9
Geant4 Simulation	3250	multi-core	x86_64 Power9 ARM
MC Reconstruction	667	multi-core	x86_64
Data Reconstruction	230	multi-core	x86_64
Derivations	0.4	multi-core	x86_64
Analysis	0.4	single, multi-core	x86_64
FastChain MC	new WF	multi-core	x86_64
Machine Learning algorithms for tracking and calorimetry	new WF	multi-core	x86_64 Power9, ARM
Machine Learning in end-user analysis	new WF	multi-core	x86_64, Power9, ARM

running by the ATLAS experiment and workflows we are planning to port to new architectures before HL-LHC run. Figure 2 shows the number of events (in millions) processed by ATLAS on HPC and LCF (special-HPC) in 2018.

3 Data storage and data handling R&D projects

The HL-LHC will be a multi-exabyte challenge where the envisaged Storage and Compute needs are a factor 10 to 100 above the expected technology evolution and flat funding. The WLCG community needs to evolve the current computing and data organization models in order to introduce changes in the way it uses and manages the infrastructure, focused on optimizations to bring performance and efficiency not forgetting simplification of operations. These are the ingredients that will allow to drive down costs and be able to satisfy the HL-LHC requirements. Technologies that will address the HL-LHC computing challenges may be applicable for other scientific communities (SKA, DUNE, LSST, BELLE-II, JUNO, etc.) to manage large-scale data volumes. The evolution of the computing facilities and the way the storage will be organized and consolidated will play a key role how this possible shortage of resources will be addressed by the LHC experiments. The need for an effective distributed data storage has been identified as fundamental from the beginning of LHC, and this topic has became particularly vital in the light of the preparation for the HL-LHC run. WLCG has started several R&Ds within the WLCG Data Organization and Management (DOMA) project. Two of them are Data Lake and Data Carousel.

- Within the Data Lake project, DOMA, we are considering EOS and dCache storage systems as a backbone software for data federation and xCache for data caching. Synthetic tests and experiments specific tests have been developed by ATLAS and ALICE for a federated storage prototype in Russia. Data Lake for Science project has been launched in Russian Federation in 2019 to set up a National Data Lake prototype for HENP and to consolidate geographically distributed data storage systems connected by fast network with low latency.
- The aim of the ATLAS Distributed Computing Data Carousel project is to increase the usage of less expensive storage, i.e. tape or even commercial storage. The Data Carousel orchestrates data

processing between the workload management, data management, and storage services with the bulk data resident on offline storage.

3.1 Data Lake project

Data Lake is a set of sites, associated by proximity, providing together storage services, possibly accompanied by compute nodes to an identified set of user communities, capable to carry out independently well defined tasks. Proximity could be defined by geography, connectivity, funding or a shared user community. This requires that their combined storage capacity and network bandwidth can meet the demands of the designated task and that usage of the different sites is transparent to the users, which, in turn, implies some form of trust relationship between the sites and a way to locate data, ranging from a simple file catalogue to a full fledged namespace. While access for users is transparent, the population and management of the storages within the Data Lake is a planned and managed activity. This includes the transitions between Quality of Service (QoS) levels. These operations are done on the granularity of the Data Lake. Data is moved to or from the Data Lake as a whole, not to or from a specific site. Resource management within the Data Lake is the responsibility of the Data Lake. Taking the aforementioned into account we can come up with some basic but crucial requirements for the future WLCG data storage infrastructure:

- Common namespace and interoperability;
- Coexistence of different QoS;
- Geo-awareness;
- Fault tolerance through redundancy of key components;
- Scalability, with the ability to change the topology without stopping the entire system;
- Security with mutual authentication and authorization for data and metadata access;
- Optimal data transfer routing, providing the user direct access to the closest data location;
- Universality, which implies validity for a wide range of research projects of various sizes, including, but not limited to the LHC experiments.

In 2015 in the framework of the Laboratory "BigData Technologies for mega-science class projects" at NRC "Kurchatov Institute" a work has begun on the creation of a united disk resource federation for geographically distributed data centres, located in Moscow, St.Petersburg, Dubna, Gatchina (all above centres are part of the Russian Data Intensive Grid – RDIG of WLCG) and Geneva, its integration with existing computing resources and provision of access to this resources for applications running on both supercomputers and high throughput distributed computing systems (Grid) [13] .The objective of these studies was to create a federated storage system with a single access endpoint and an integrated internal management system. With such an architecture, the system looks like a single entity for the end user, while in fact being put together from geographically distributed resources. This work was continued as a part of an award granted by the Russian Science Foundation to the Laboratory of Cloud Computing of Plekhanov University. The concept of Russian Data Lake for Science is described in [14]. The resources used for the RF data lake prototype are located at PNPI (Gatchina), JINR (Dubna), SPbSU (St. Petresburg) and MEPhI (Moscow). Two primary topics to be addressed in the context of the data lake prototype (in Russia) in the next years:

- develop tests methodology, conduct and automate functional tests including:
  - synthetic tests
  - experiment's specific payloads including intensive I/O (derivation production) and CPU intensive (Monte-Carlo simulation) payloads;
- configure Russian LHC sites using EOS and xCache software

3.2 Data Carousel project

The ATLAS Experiment is storing detector and simulation data in raw and derived data formats across more than 150 Grid sites world-wide: currently, in total about 200 PB of disk storage and 250 PB of tape storage is used. Data have different access characteristics due to various computational workflows. Raw data is only processed about once per year, whereas derived data are accessed continuously by physics community (more than a thousand physicists). Data can be accessed from a variety of mediums, such as data streamed from remote locations, data cached on local storage using hard disk drives or SSDs, while larger data centers provide the majority of offline storage capability via tape systems. Disk is comparatively more expensive than tape, and even for disks there are different types of drive technologies that vary considerably in price and performance. Slow data access can dramatically increase the costs for computation.

The HL-LHC era data storage estimated requirements are several orders larger than the present forecast of available resources, based on a flat budget assumption. On the computing side, the ATLAS Distributed Computing (ADC) was very successful in the last years with HPC and HTC integration and using opportunistic computing resources for the Monte-Carlo simulation. On the other hand, equivalent opportunistic storage does not exist for HEP experiments. ADC started the "Data Carousel" project to increase the usage of less expensive storage, i.e. tape or even commercial storage, so it is not limited to tape technologies exclusively. Data Carousel orchestrates data processing between workload management, data management, and storage services with the bulk data resident on offline storage. The processing is executed by staging and promptly processing a sliding window of inputs onto faster buffer storage, such that only a small percentage of input date are available at any one time. With this project we aim to demonstrate that this is the natural way to dramatically reduce our storage costs. The first phase of the project was started in the fall of 2018 and was related to I/O tests of the sites archiving systems. Now we are at Phase II, which requires a tight integration of the workflow (Production System – ProdSys2), workload (PanDA) and data management (Rucio) systems. We plan to run the project at large scale in 2020, so results could be used during LHC Run3 (after 2021).

The basic considerations to address the scaling requirements for data storage at the HL-LHC for ATLAS are:

- "Opportunistic storage" basically does not exist for the LHC experiments;
- Format size reduction and data compression are both long-term goals and these will require significant efforts from the software and distributed computing teams. We also assume that it will be beneficial to work together with industrial partners on these. A common R&D program would definitely be useful and it is one of the topics we have identified as a potential joint research with Google Cloud Platform team;
- The increased usage of cold, less expensive storage (currently tape) relative to disk is a natural way to dramatically reduce the storage costs. A "data carousel" model should be evaluated as an R&D project within ATLAS distributed computing.

Our final goal is to evaluate the dependence of the impact on the execution time of various workflows on the input data access on disk or tape (or in more general terms on hot, cold and archive storage (possible scenarios are shown in Fig. 3). To reach this, we have identified several areas we plan to work on:

- Work with the WLCG Archival Storage working group to define the metrics to describe the data access on tape and their limits: define realistic expectations with the present sites configuration and evaluate the possible evolutions.
- In the ATLAS distributed data management system Rucio, study the feasibility to improve the tape usage. Examples are: monitoring the queued tape requests and estimating the time to retrieve the

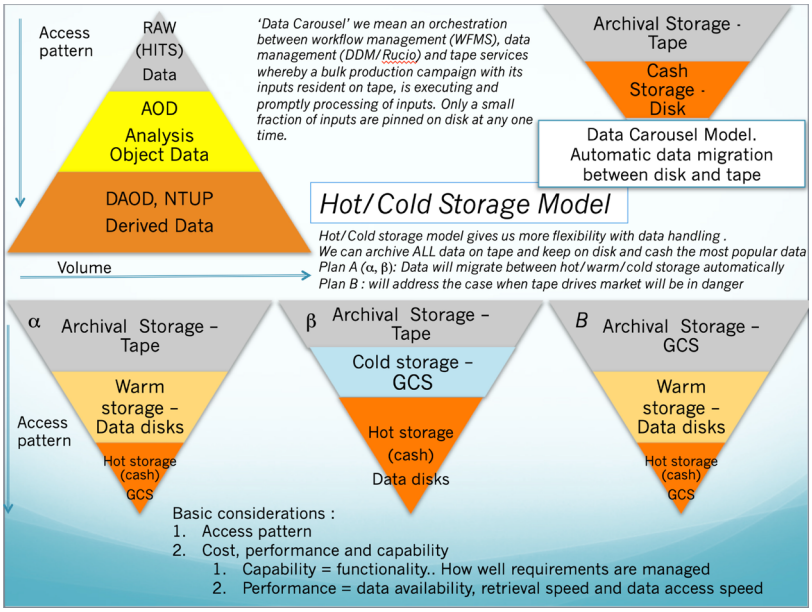

Figure 3. Possible ways of organization of different Quality of Service of hot/warm/cold storage

requested files and making bulk requests of tape staging, bulk size request tailored to the sites parameters.

- In the file transfer system (FTS), study the feasibility to optimize scheduling of transfers between tape and other storage endpoints.
- In the storage elements technologies used at the sites like, e.g., dCache, Storm, and other disk storage systems, identify any bottlenecks in their respective framework, study feasibility to improve their interface with the respective tape system.
- Evaluate and improve the way in how data is stored on tape and optimize their recall speed:
  - Well defined tape families for files which are known to be re-read from tape.
  - Use large file sizes and define the optimal file size.
- Evaluate and eventually propose the development of new functionalities in tape systems, e.g., support high priority tape recall requests with low latency.
- Explore new data transfer mechanisms using tape, for example, can we let tape talk to local cache directly?
- Study and optimize prompt processing of data as it appears off of tape – process immediately data or when X % of a data sample is staged in.

By "data carousel" we mean an orchestration between workload management, data management and tape services whereby a bulk production campaign with its inputs resident on tape, is executed by staging and promptly processing a sliding window of X % of inputs onto buffer disk, such that only X % of inputs are pinned on disk at any one time. There are several phases of this R&D:

- In the first phase we have conducted tests to understand tape systems performance at the various sites (ATLAS Tier1 and Tier0 centers) – for more than a decade the ATLAS policy was to use tapes at Tier1s only for data archiving, organized data reprocessing and simulation (and possible limitations related to, e.g., geographical limitations).

- In the second phase (which is currently ongoing) we are addressing the issue of data retrieval/exchange between tape/disk with our data management and workflow management systems. This second phase of R&D required a deeper integration between the two main distributed computing components, i.e. data and workflow/workload management.
- Finally during the third phase we will exercise the workflows using tape systems. Tape intensive usage will also complicate the workflow orchestration, because of protocol to be defined between three systems (workflow management, workload management and data management) how and where to place and process archived data from tape.

In terms of a workflow, we decided to start this R&D with ntuples production from tape. The reasoning behind is that derivations are highly demanding in terms of disk I/O and overall disk space usage. The understanding of this workflow can then easily be mapped to other (simpler from the infrastructure point of view) workflows. The input to derivation production are AODs which make up a third of the total ATLAS disk space thus a very relevant part which we want to shrink.

It is worth mentioning that Russian Data Lake and ATLAS Data Carousel are two out of the several storage-related and data management R&D projects conducted in parallel. Other R&D projects aimed to address proper handling of storages with different QoS include: data ocean, hot/cold storage, EU data lake prototype, Google-HEP, ...

Acknowledgments

This work was funded in part by the U. S. Department of Energy, Office of Science, High Energy Physics and Advanced Scientific Computing under Contracts No. DE-AC02-98CH10886 and DE-AC02-06CH11357, the Russian Data Lake R&D project is funded by the Russian Science Foundation under contract No.19-71-30008 (research is conducted in Plekhanov Russian University of Economics).

References

[1] *LHC – The Large Hadron Collider*, `http://lhc.web.cern.ch/lhc/`
[2] J. Shiers, Computer Physics Communications **177**, 219–223, (2007)
[3] ATLAS Collaboration, G. Aad, et al, J. Instrum. **3**, S08003 (2008)
[4] G. Aad, T. Abajyan, B. Abbott, et al, Physics Letters B **716**, 1, 1–29, (2012)
[5] A. Klimentov et al, *Next generation workload management system for big data*, In 16th International Workshop on Advanced Computing and Analysis Techniques in Physics Research, plenary talk, September 2014, Prague, Czech Republic
[6] ATLAS Experiment – Public Results, https://twiki.cern.ch/twiki/bin/view/AtlasPublic/ComputingandSoftwarePublicResults
[7] Oak Ridge National Laboratory Leadership Computing Facility. https://www.olcf.ornl.gov
[8] K. De et al, Physics of Particles and Nuclei Letters **13**, 5, 647–653 (2016)
[9] M. Barisitis et al, Computing and Software for Big Science **3**, 11 (2019)
[10] J. Blomer, T. Fuhrmann, *A Fully Decentralized File System Cache for the CernVM-FS*, in Proceedings of 19th International Conference on Computer Communications and Networks, August 2–5, 2010
[11] J. Blomer et al, Computing in Science and Engineering **17**, 6, 61–71 (2015)
[12] `squid-cache.org`
[13] A. Kiryanov, A. Klimentov, A. Zarochentsev, et al, J. Phys.: Conf. Ser. **898**, 062016 (2016)
[14] A. Klimentov, A. Kiryanov, A. Zarochentsev, Open Science Platforms, 4, 32–34 (2018)