EPJ Web of Conferences **245**, 04045 (2020) https://doi.org/10.1051/epjconf/202024504045
CHEP 2019

Implementation and performance of a DPM federated storage and integration within the ATLAS environment

Claire Adam Bourdarios[1], *Jean-Claude* Chevaleyre[2], *Frédérique* Chollet[1], *Sabine* Crépé-Renaudin[3], *Christine* Gondrand[3], *Stéphane* Jezequel[1,*], *Edith* Knoops[4], and *Philippe* Seraphin[1]

[1]LAPP, Université Grenoble Alpes, Université Savoie Mont Blanc, CNRS/IN2P3, Annecy; France.
[2]LPC, Université Clermont Auvergne, CNRS/IN2P3, Clermont-Ferrand; France.
[3]LPSC, Université Grenoble Alpes, CNRS/IN2P3, Grenoble INP, Grenoble; France.
[4]CPPM, Aix-Marseille Université, CNRS/IN2P3, Marseille; France.

Abstract. With the increase of storage needs at the High-Luminosity LHC horizon, data management and access will be very challenging. The evaluation of possible solutions within the WLCG Data Organization, Management and Access (DOMA) is a major activity to select the most optimal from the experiment and site point of views. Four teams hosting Tier-2s for ATLAS with storage based on DPM technology have put their expertise and computing infrastructures in common to build a testbed hosting a DPM federated storage called FR-ALPES. This note describes the infrastructure put in place, its integration within the ATLAS Grid infrastructure and presents the first results.

1 Introduction

Over the last two decades, the WLCG organisation [1] has developed and operated a world wide computing and storage infrastructure. Its success contributed to the quick announcement of the Higgs discovery shortly after the end of the data taking in July 2012 [2, 3]. Currently, the main challenge consists in maintaining the same level of excellence with a continuously increasing capacity and to integrate new heterogeneous computing facilities like High Performance Computers. The next major challenge will arise in the next decade: the High-Luminosity LHC (HL-LHC) program will increase the number of registered events by an order of magnitude while the computing budget will have to remain constant [4]. This triggered the R&D program called DOMA [5] with the objective to optimise the organisation and usage of the distributed storage. One of the main ideas is to aggregate computing and storage components in Data Lakes. By enabling the remote access to storages, the current data access model of DOMA gives the opportunity to reduce the number of storage endpoints.

This logic has triggered the interest of four teams (system administrators and ATLAS physicists) located within the same southeast region of France (CPPM-Marseille, LAPP-Annecy, LPC-Clermont-Ferrand et LPSC-Grenoble) to evaluate the reliability and performances of a shared federated storage based on the Disk Pool Manager (DPM) technology [6]

*e-mail: stephane.jezequel@lapp.in2p3.fr

with a single head-node but disk servers located in the different sites. Its performance was evaluated within the Grid infrastructure of the ATLAS experiment [7]. This activity is part of the french contribution to DOMA called DOMA-FR and also evaluated with the european ESCAPE [8] project.

The first part of the note overviews the existing distributed storage federations and presents the French one called FR-ALPES. The second part describes the integration of the infrastructure within the ATLAS Grid environment. The last section concludes by presenting preliminary results.

2 Existing federated storage

2.1 Existing Grid storage federations in ATLAS

The NorduGrid [9] federation was designed from the beginning as a federated Grid infrastructure. It has been running successfully as a Tier-1 over the last decade. It is based on the dCache technology [10] associated to the ARC cache [11] component to preplace input files on a shared file system close to the worker nodes (WN).

The possibility to build a federated storage with the DPM technology was also exploited. The Swiss federation has constructed an asymmetric storage federation [12]. The University of Bern hosts a standard DPM storage with its head-node as well as physical disk servers which are exposed to the ATLAS Grid system. The University of Geneva is just providing disk storage accessed only by local users. Another setup was built in Italy consisting in a DPM federated storage [13] associated with volatile pools to offer the possibility to easily add caching pools for user analysis.

2.2 FR-ALPES storage federation

The FR-ALPES federation was initiated by the interest of the local site administrators in France to contribute to the DOMA R&D effort through the DOMA-FR coordination. As the current main model issued from DOMA recommends to group sites in Data Lakes with fewer storage endpoints than computing facilities, the teams decided to build a storage federation aggregating their four IN2P3-CNRS sites located in the southeast part of France (CPPM, LAPP, LPC and LPSC). The already existing collaboration to operate WLCG Tier-2 computing facilities, directly or through the LCG-France technical coordination, and the existing expertise in Grid technology allowed us to setup this testbed in a few months. The DPM storage technology was selected since the site administrators were already operating DPM storages.

As the validation of the Data Lake model requires one to demonstrate that the effect of remote accesses to storages can be minimised, grouping disk-servers separated by roundtrip-time < 10 ms gave the opportunity to build a first element of a future European Data Lake with low latency.

The four sites are connected to 10 Gbps Wide Area Network links mostly dedicated to the WLCG production and analysis activities. Due to their heavy usage mainly by the LHC experiments, some links can be saturated. In all four sites, the disk servers are connected to the Local Area Network through 10 Gbps links while the worker nodes have 1 Gbps connection.

Figure 1 describes the different components of the FR-ALPES federated storage: the headnode located in LAPP and the disk servers deployed in the four sites.

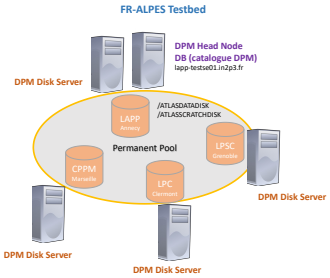

Figure 1. Components of the FR-ALPES storage.

3 Testing infrastructure

The first objective was to measure the impact of a geographically distributed storage on the processing performances as well as its impact on the regional network activity. To evaluate the performance of this federation and benefit from existing ATLAS tools, it was integrated in the ATLAS Grid infrastructure. This was straightforward as these sites were already ATLAS Tier-2s. In addition, the expertise in ATLAS tools for job submission and monitoring was available through local ATLAS experts in ATLAS Computing. As a consequence, the integration of FR-ALPES into the ATLAS Grid infrastructure was straightforward.

For the purpose of this study, the ATLAS testing infrastructure based on HammerCloud [14] was adapted for the evaluation of the federated storage. This tool submits regularly a set of identical ATLAS jobs and records the duration of each of the following steps:

- Software environment set up
- If requested within the job configuration, input file(s) copy to the local disk of the worker node
- Data processing
- Copying the output file to a Grid storage

The challenge has been to split the collected results from the testing infrastructure to report accounting according to the effective location of the input files. As the DPM catalog does not give easy access to a list of files according to their physical location, different set of files were aggregated in separated datasets registered into the Rucio [15] Data Management tool. Each dataset was transferred to one geographical location within the FR-ALPES storage. To ensure that each dataset was going to the selected location, the disk servers located at other places were set in read-only mode. Finally, a HammerCloud template was built for each pair (WN location)-(datasets associated to a site) so that performances could be isolated. Then, the HammerCloud submission was started and results were collected separately for each pair. For the first campaign of measurements presented in this note, input datasets were pre-positioned only in LAPP and LPSC.

4 First results

The first campaign, which occurred in autumn 2019, was targeted to quantify the impact of distributing files among the FR-ALPES federated storage on the ATLAS job CPU efficiency. At the same time, a comparison was done between two accessing modes :

Table 1. Duration of simulation job components for different WN locations, and access mode for input files located at LAPP and LPSC.

Input file location		LAPP	LPSC
	Action	Duration (seconds)	
WN location		CPPM-LAPP-LPC-LPSC	CPPM-LAPP-LPC-LPSC
Copy and processing	Input copy	5-10	5-10
	Processing	500-700	700-900
Direct access		500-800	700-1000

- Copy and processing: the input file is copied from the Grid storage to the WN internal disk and processed
- Direct access: the input file is directly accessed from the Grid storage

The first mode optimises the CPU efficiency by minimising latency effect during event processing. The second one maximises the overall CPU efficiency by accessing only the relevant part of the input file. For this campaign, no caching mechanism was implemented and the read-ahead mechanism relied on the one implemented within the ATLAS software.

The first set of measurements consisted in running simulation jobs: since, in this case, the input file is small (38 MB in this case) and the Input/Output rate is small, the impact due to latency between WN and the input file location is expected to be small. Table 1 summarises the results.

For the same input file location, the spread of measurements among random WNs at the same location was much larger than the difference of the central values between locations. This is the reason why global range for the same input file location is displayed. As expected, the difference in processing duration between the two types of accesses is negligible compared to the spread of the processing time for the local-access method. A significant difference in processing time is observed between LAPP and LPSC which is still under investigation.

A similar campaign was run with ATLAS reconstruction jobs whose Input/Output activity is much larger than simulation. Events within a hard-scattered sample of 3 GB are overlayed with two pile-up event files of 2 and 1 GB each and reconstructed to produced AOD files. Only two events are processed. The results are presented in Table 2.

Table 2. Duration of reconstruction job components for different WN locations, and access mode for input files located at LAPP and LPSC.

Input file location		LAPP	LPSC
	Action	Duration (seconds)	
WN location		CPPM-LAPP-LPC-LPSC	CPPM-LAPP-LPC-LPSC
Copy and processing	Input copy	100-1000	100-1000
	Processing	1500-2000	2000-3000
Direct access		1600-3000	2000-2500

As for simulation, the difference in processing time between accesses of remote and local storages is small compared to the spread of the processing time with a local copy. Variation by an order of magnitude is observed on the duration of the input file copy. It could be induced by the saturation of the WAN link due to concurrent accesses from the Grid production jobs running in parallel in the infrastructure. The lack of tools to correlate network activity with data access performances prevented to confirm this hypothesis. In early 2020, three out of the

four sites will have their WAN connectivity upgraded to 20 or 40 Gbps which should reduce the suspected occurrence of this problem.

5 Conclusion

The FR-ALPES infrastructure has demonstrated the possibility to build a federated storage based on the DPM technology and to integrate it within the ATLAS Grid infrastructure. On the site administrator side, its deployment and operation has been straightforward with the latest version of DPM software. It was the opportunity for the site administrators to share the maintenance and operation of a common Grid infrastructure. On the ATLAS side, the first results have shown minimal impact of the different latencies between the sites of the federation. A longer campaign of measurements is foreseen to get more precise measurements. To optimise the network occupancy and further minimise the impact of latency, tools like Xcache [16] will be included and new measurements will be done.

To extend the evaluation beyond the WLCG program, the FR-ALPES federation will also be integrated into the ESCAPE [8] European Data Lake.

References

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

[2] ATLAS Collaboration, *Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC* Phys. Lett. B, **716**, 2012 1-29, arXiv:1207.7214 (2012).

[3] CMS Collaboration, *Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC* Phys. Lett. B, **716**, 2012 30, arXiv:1207.7235 (2012).

[4] *A Roadmap for HEP Software and Computing RD for the 2020s*, The HEP Software Foundation, Albrecht, J. et al., Comput Softw Big Sci (2019) 3, 7 .

[5] DOMA, https://twiki.cern.ch/twiki/bin/view/LCG/DomaActivities

[6] M. Hellmich et al, *DPM efficient storage in diverse environments*, 2014 J. Phys.: Conf. Ser. 513 042025; http://inspirehep.net/record/1302100/files/jpconf14_513_042025.pdf

[7] ATLAS Collaboration, *The ATLAS Experiment at the CERN Large Hadron Collider*, JINST **3** S08003 (2008).

[8] ESCAPE, https://projectescape.eu//

[9] NorduGrid, http://www.nordugrid.org/

[10] P. Fuhrmann and V. Gulzow, *dCache, storage system for the future*, in European Conference on Parallel Processing. Springer, pp. 1106–1113, (2006)

[11] M. Ellert et al., *Advanced Resource Connector middleware for lightweight computational Grids* Future Generation Computer Systems **23**, Issue 2, 219-240 (2007)

[12] G. Sciacca *Multi-site DPM - The BERN case*, presentation accessible at https://indico.cern.ch/event/776832/contributions/3378586/attachments/1861907/3060285/MultiSiteDPM.pdf

[13] A. Doria et al. *Distributed caching system for multi-site DPM storage*, EPJ Web Conf. **214** 04056 (2019).

[14] HammerCloud, http://hammercloud.cern.ch/

[15] M. Barisits, T. Beermann, F. Berghaus et al., *Rucio: Scientific Data Management* Comput Softw Big Sci (2019) 3: 11

[16] A. Hanushevsky et al. *Xcache in the ATLAS Distributed Computing Environment*, EPJ Web Conf. **214** 04008 (2019).