PROCEEDINGS OF SCIENCE

PoS(ISGC&HEPiX2023)005

Extension of local dCache instance capacity using national e-infrastructure

J. Chudoba,[a,*] **M. Svatoš,**[a] **A. Mikula,**[a] **P. Vokáč**[a] **and M. Chudoba**[b]

[a] *FZU - Institute of Physics of the Czech Academy of Sciences*
Na Slovance 1999/2, 182 00 Prague 8, Czech Republic

[b] *Faculty of Mathematics and Physics, Charles University,*
Ke Karlovu 2027/3, 121 16 Prague 2, Czech Republic

E-mail: chudoba@fzu.cz, svatosm@fzu.cz, mikula@fzu.cz, vokac@fzu.cz, michal.chudoba.praha@gmail.com

*Speaker

The Czech WLCG Tier-2 center for LHC experiments ATLAS and ALICE provides computing and storage services for several other Virtual Organizations from high energy and astroparticle physics. The center deployed Disk Pool Manager (DPM) for almost all (only ALICE VO uses xrootd servers) supported VOs as a solution for storage until recently. The local capacity was extended by a separate instance of dCache server which was operated by CESNET Data Storage unit in a remote location. The exact location has changed during the project, the distance was between 100 to 300 km. This storage extension was based on HSM and was mapped as a separate ATLAS space token where higher latencies were expected. The intended usage was for a non-automatic backup of the ATLASLOCALGROUPDISK spacetoken used by ATLAS users from the Czech Republic. Since the usage was relatively low and the system had only one group of users from the ATLAS VO, the effort required for maintenance and frequent updates was not effective.

The DPM project announced the end of support, and we migrated the main Storage Element in CZ Tier-2 to dCache. This brought the possibility of a unified solution for an SE. The dCache system at CESNET was stopped and we started to test a new solution with only one endpoint for all users. CESNET Data Unit also changed the underlying technology for data storage - they moved from HSM to CEPH. We mounted one file system as a RADOS block device (RBD) on a test dCache server and measured properties of the system to compare with storage based on local disk servers. This solution differs from a solution used in the Nordugrid Tier-1 center, where distributed dCache servers use caching on local ARC Computing Elements. Tests included long term stability of network throughput, duration of transfers of files with sizes from 10 MB to 100 GB and changes in duration of transfers when several simultaneous transfers are executed. The network tests were first executed on an older diskless server and later on a new dedicated test server with surprisingly different results. We used the same tools also to measure differences in transfer performance between local disk servers which are of different age and connected by different speeds. Since the results of tests were satisfactory, we will use the external storage first as a dedicated space token for ATLAS and later as a part of a space token located also on local disk servers. We may also use the solution for other Virtual Organizations if the external available space is increased by a sufficient volume.

International Symposium on Grids & Clouds (ISGC) 2023 in conjunction with HEPiX Spring 2023 Workshop, ISGC&HEPiX2023
19 - 31 March 2023
Academia Sinica Taipei, Taiwan

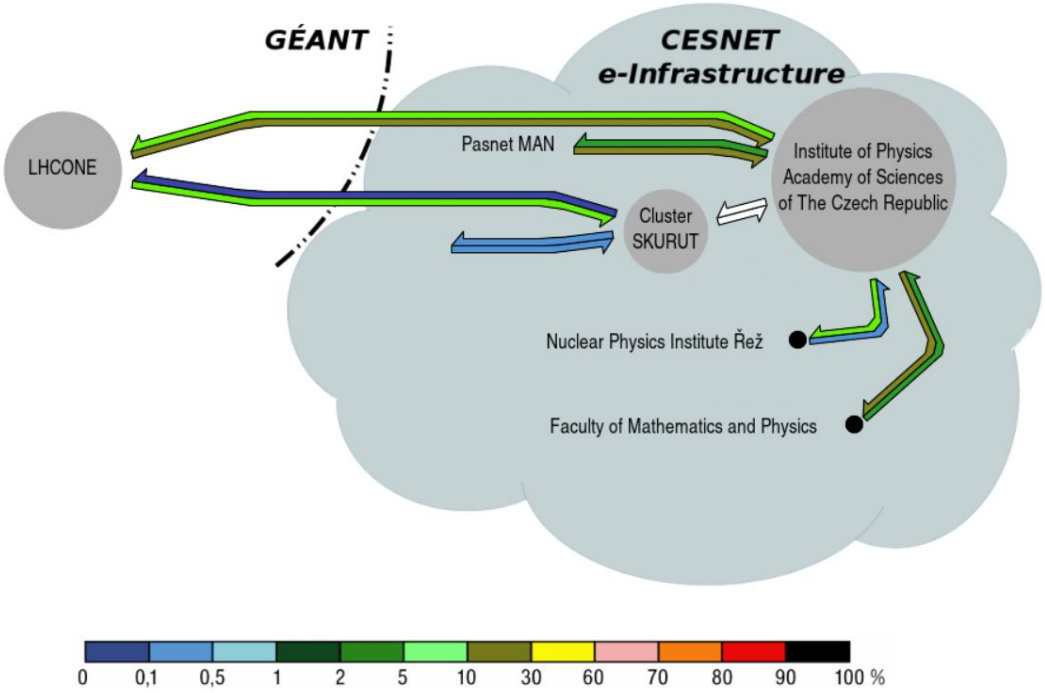

Figure 1: Network connection between the main site of CZ-Prague-T2 and other locations.

1. Introduction

The Czech WLCG Tier-2 center [1], [2] is distributed in several places. All main controlling and monitoring components and the main part of the computing and storage capacities are located at the Institute of Physics of the Czech Academy of Sciences (FZU) in the northern part of the Czech capital Prague. Storage capacity for the ALICE experiment consists of several xrootd servers placed in the Institute of Nuclear Physics of the Czech Academy of Sciences (NPI) in a small town Řež, in distance about 10 km from the FZU server room. More than 10 % of the computing capacity is available via servers running in the server room of the Faculty of Mathematics and Physics (MatFyz) in Troja Campus, just 2 km from the FZU building. These locations are connected via a dedicated network provided by CESNET, the Czech NREN. The main part is connected by a dedicated 100 Gbps link to LHCONE VPN and by 40 Gbps link to generic internet. Site at MatFyz is connected by a 10 Gbps link and NPI part by 30 Gbps link. The very good network connection with possible upgrades of capacities in future is a key element in consideration of extending current capacities by remote resources. The constant need of LHC collaborations for more and more resources forces us to search for additional external resources. The usage of computing servers of the Czech national supercomputing center IT4I in Ostrava was discussed in several papers [3], [4] and [5]. Here we will concentrate on storage extension.

2. Remote RBD

CESNET Storage department (CESNET DU) [7] develops and operates the national infrastructure data storage for science and research. Standard services include backup, archive and

environment for data sharing. Storage facilities are distributed in several towns in the Czech Republic, currently in Jihlava, Brno and Ostrava. The department operated dCache Grid Storage Element [8], which was used by the LHC ATLAS experiment and tested by astro-particle physics project Pierre Auger Observatory. It was based on disk cache with a tape backend. This service was mostly used as a backup for ATLASLOCALGROUPDISK space token provided by FZU Tier-2 center. Since the service was not used by other groups and the operation required relatively big effort (but still less than 1 FTE), we decided to stop the service when new protocols and settings for Third Party Transfers were required.

The FZU Tier-2 center used DPM [9] as the Grid Storage Element implementation. Since this project was discontinued, we migrated to another widespread solution, dCache. dCache unlike DPM supports also tape backends and this was the main reason why CESNET DU used it in the past (although in the meantime the CESNET facilities completely migrated to disk based solutions). We considered several options on how to use remote Ceph based capacity for a transparent use in the WLCG Tier-2 center. The NDGF uses distributed dCache [6] in such a way that it has just one central entry point and several dCache disk servers on each remote site. This solution would require a dedicated server on the CESNET DU site. Easier for maintenance and operation can be a model, where the remote site creates a Rados Block Device (RDB) on Ceph and the central site mounts it. This solution was tested with a project of CESNET Development Fund FR 712/2022.

2.1 Setup, configuration and first tests

The RDB of the size of 100 TB was created on the cl3 facility in Ostrava. At the FZU Tier-2 site we created a new dCache test instance as a virtual server while we were waiting on delivery of a new dedicated hardware. We did recommended customisation of network parameters, which are important for reading from RBD:

```
echo 12000000 > /sys/block/rbd0/queue/read_ahead_kb .
```

First tests were to copy 50000 of 1 GB files to the remote storage using dd command:

```
dd if=/dev/zero of=/mnt/ceph0/chudoba/50TB/test.1G.$i bs=10M count=100 .
```

Results showed big variation among individual transfers. When we used rolling average over 50 consecutive results, we saw that the transfer speed significantly varied although conditions on the test dCache server were not changing (see Figure 2). After some investigation we concluded that the variations may be due to the local network heavy utilization. We verified this hypothesis when shortly after these initial tests we got and installed the new physical hardware. The new server was connected directly to the main router using a 40 Gbps link. Results of new transfer tests were much more stable. The transfer time was mostly under 2 seconds and the speed exceeded 600 MB/s, although we still could observe some variations in time (Figure 3). Many more tests were done to test the read speed, to check variations in the speeds for several concurrent transfers and the frequency of deletion of files. Details of these results will be available in the final technical report of the project.

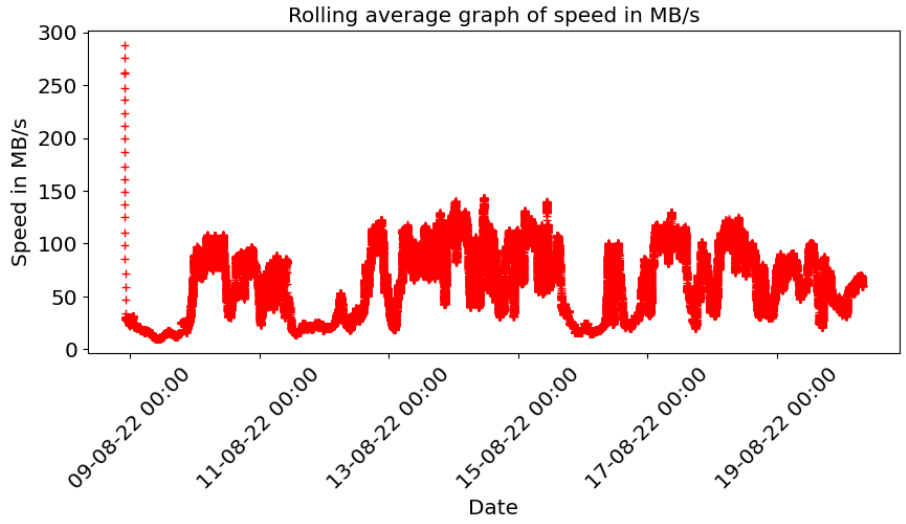

Figure 2: Rolling average of the transfer file of 1 GB file to the remote storage from a virtual server with 10 Gbps link connected to the router via several switches.

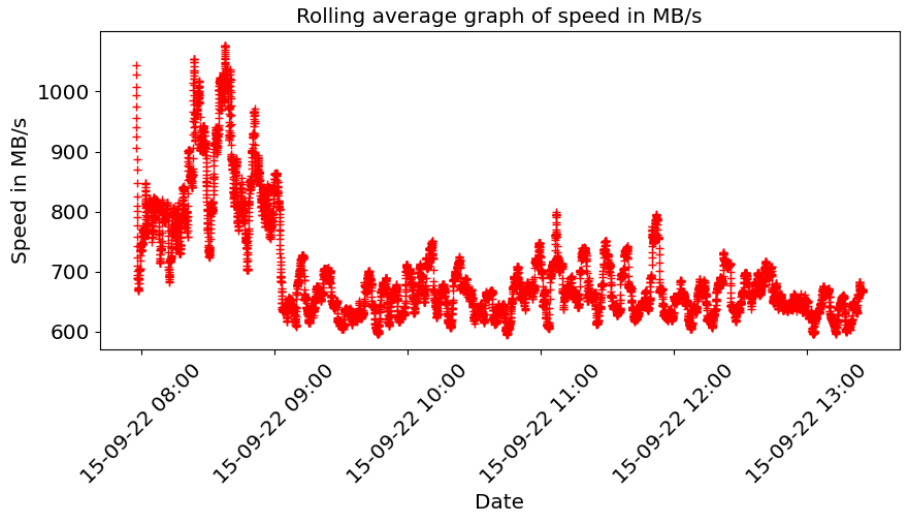

Figure 3: Rolling average of the transfer file of 1 GB file to the remote storage from a physical server with 40 Gbps connection directly to the site router.

2.2 Tests with a real analysis

We tested the system with a real ATLAS analysis performed on 2.2 million of ATLAS events stored in 594 files with a total volume 91.6 GB. To avoid several effects not related to the storage performance we used software installed on local scratch of a dedicated execution server. We made 500 repetitions for data located on a local disk of the test dCache server. Most jobs finished under 700 s, the shortest job took 558 s. The longest one took 1381 seconds, but it was a single outlier. Later investigation revealed that the reason was in a switch maintenance carried out during this test. Same tests were performed for the identical input data read from RBD and for comparison also for data read from our standard dCache instance, which uses 10 servers in a local network.

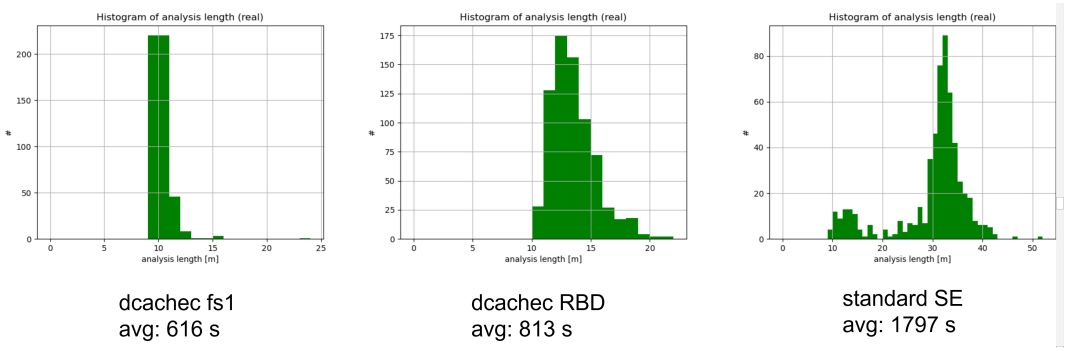

Figure 4: Duration of tested ATLAS analysis on a dedicated execution server when reading data from test dCache local disk, RBD and standard production dCache. Speed was limited by 1 Gbps connection of the execution server.

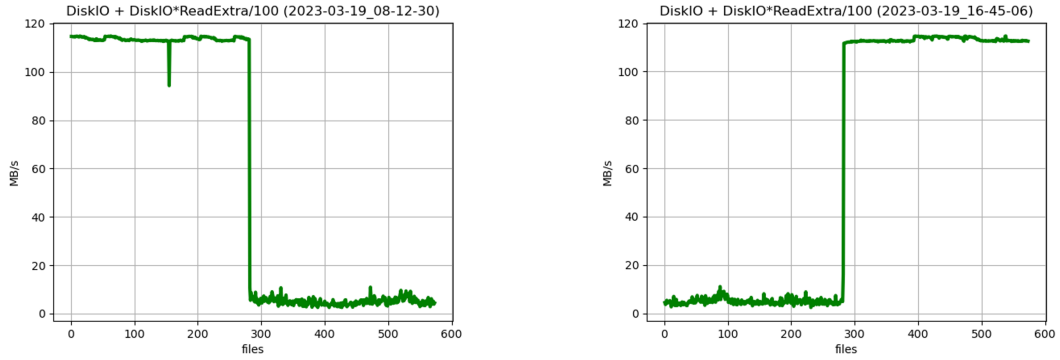

Figure 5: Effect of the memory cache on the test dCache server.

Reading from remote RBD was on average 32 % slower than from the local test dCache server disk. Production dCache servers provided the worst performance (292 % slower than local disk on the test server in average). We explain this result by heavy load in the production environment, whereas the test dCache server was dedicated only to these tests (Figure 4).

2.3 Memory cache effect

The results of the analysis above were significantly impacted by the memory cache. The test disk server has 128 GB RAM and so all analyzed dataset (91.6 GB) fit into memory cache. To check for this effect we cleaned the memory cache after reading about 300 files:

```
sysctl -w vm.drop_caches=3.
```

The speed for data reading dropped from 110 MBps (limit given by 1 Gbps connection of the dedicated execution server) to 5 MBps. Repetition of the test proved that the speed increased again to the previous value 110 MBps when the test started to read already cached files (Figure 5).

3. Summary

We have done several tests of performance of a remote storage capacity provided by RBD on Ceph instance. We achieved good transfer speeds in case of a single or a few concurrent transfers

which are sufficient for originally intended usage as an ATLASLOCALGROUPTAPE spacetoken. More real operational experience will be acquired when we use the remote RBD in the production environment.

Acknowledgments

This work was supported by the Ministry of Education, Youth and Sports of the Czech Republic project CERN-CZ LM2023040 and by project CESNET FR 712/2022.

References

[1] M. Adam, D. Adamová, J. Chudoba, A. Mikula, M. Svatoš, J. Uhlířová and P. Vokáč, EPJ Web Conf. **245** (2020), 03034 doi:10.1051/epjconf/202024503034

[2] J. Chudoba, D. Adamová, A. Mikula, L. Míča, M. Svatoš, P. Šesták, J. Uhlířová and P. Vokáč, PoS **ICHEP2022** (2022), 1146 doi:10.22323/1.414.1146

[3] J. Chudoba and M. Svatos, PoS **ISGC2018 & FCDD** (2018), 025 doi:10.22323/1.327.0025

[4] M. Svatoš *et al.* [ATLAS], EPJ Web Conf. **245** (2020), 09010 doi:10.1051/epjconf/202024509010

[5] M. Svatoš *et al.* [ATLAS], EPJ Web Conf. **251** (2021), 02008 doi:10.1051/epjconf/202125102008

[6] G. Behrmann, P. Fuhrmann, M. Gronager and J. Kleist, J. Phys. Conf. Ser. **119** (2008), 062014 doi:10.1088/1742-6596/119/6/062014

[7] CESNET Storage Department; https://du.cesnet.cz/en/start, accessed April 29, 2023

[8] dCache; https://www.dcache.org/about/, accessed April 29, 2023

[9] Disk Pool Manager DPM; https://lcgdm.web.cern.ch/dpm, accessed April 29, 2023