Large Scale Data Handling experience at INFN-CNAF Data Center

Alessandro Cavalli, Antonio Falabella, Enrico Fattibene, Federico Fornari, Lucia Morganti,[*] Andrea Prosperini, Andrea Rendina and Vladimir Sapunenko

INFN-CNAF,
viale Berti Pichat 6/2, Bologna, Italy

E-mail: lucia.morganti@cnaf.infn.it

INFN CNAF is the National Center of INFN (National Institute for Nuclear Physics) for research and development in the field of information technologies applied to high energies physics experiments. CNAF hosts the largest INFN data center which also includes a Worldwide LHC Computing Grid (WLCG) Tier1 site (one of 13 around the world), providing resources, support and services needed for computing and data handling in the WLCG framework. The data center also represents a key data facility for many astro-particle and nuclear physics experiments.

The Data management team manages and makes available all the Tier1 storage resources to the scientific community. Currently, such resources consist of more than 50 PB of disk storage and more than 110 PB of tape storage.

We describe the adopted technologies for Data Management and Data Transfer and how our services are evolving to cope with the requirements imposed by the High Luminosity-LHC era, in the context of a worldwide transition to new protocols and authorization approaches for bulk data transfers between WLCG sites.

Also, we report on our work to provide POSIX filesystems with different technologies: along with the bulk of data center storage, which is based on GPFS deployments, we provide CephFS as well as object and block storage service for data access requirements beyond WLCG use cases.

41st International Conference on High Energy physics - ICHEP2022
6-13 July, 2022
Bologna, Italy

[*] Speaker

1. Introduction

CNAF, located in Bologna, is the INFN National Center dedicated to Research and Development on Information and Communication Technologies. CNAF hosts the main INFN data center, providing services and resources to more than 40 scientific collaborations and representing the INFN Tier-1 in the WLCG e-infrastructure.

In July 2022, the installed resources amount to ~ 42k cores, ~ 50 PB of disk space and ~ 116 PB of tape space (see Figure 1). However, a significant increase of resources is foreseen in the coming years: by 2025, ~ 130k cores, ~ 110 PB of disk space and ~ 250 PB of tape space, and even more starting in 2027, in support of the High-Luminosity LHC project.

The present data centre is located within the premises of the Physics Department of the University of Bologna. With the current IT technology, we have estimated that we would be able to host IT resources to cover the requirements up to the end of LHC Run 3 (2024). This fact, coupled with the opportunity offered by the building of a new district promoted by the Emilia-Romagna region and devoted to research, innovation and technological development, represents the main driver for the transfer of CNAF resources to such new Tecnopolo area, where our data center will be able to meet the growing requirements in terms of space and data rate of the data taking of the HL-LHC experiments up to 2035 and beyond, providing as well services and resources to the many other experiments, projects, and activities in which INFN is involved.

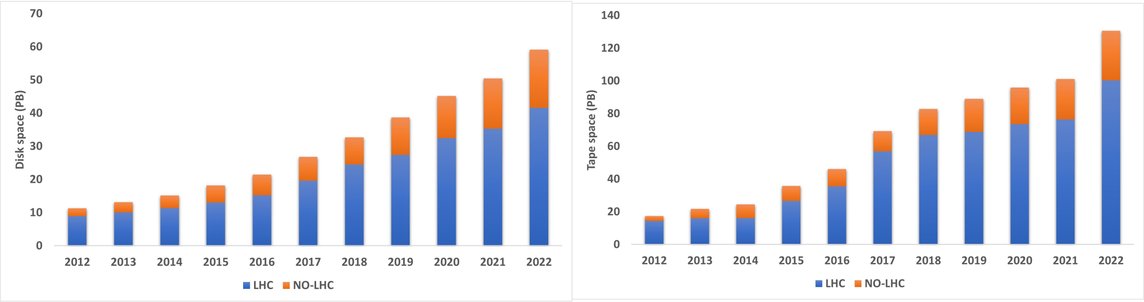

Figure 1: Increase in storage resources installed at INFN CNAF for both LHC (blue) and no-LHC experiments (orange) over the last 10 years. Disk storage on the left, tape storage on the right.

2. Architectural choices

Our storage infrastructure, based on industry standards, relies on a solution which has been well consolidated over the years, allowing the implementation of a completely redundant data access system from a hardware point of view and capable of very high performances.

The disk storage systems in production are DDN SFA12K and SFA7900, Huawei OS18800v5, 6800v5 and 5800v5, and DELL MD3860F, supporting an average data transfer rate of ~ 700 TB in, ~ 1.7 PB out per day (see e.g. Figure 2), with ~ 300k transferred files per day.

At the hardware level, Medium Range or enterprise level storage systems are interconnected via Storage Area Network (16 Gbps Fiber Channel or FDR/EDR InfiniBand) to the disk-servers.

At the file-system level, the parallel file system IBM Spectrum Scale (formerly GPFS - General Parallel File System) is adopted as POSIX interface and backend for all data management and data transfer services. The GPFS file-systems, one for each main experiment, are directly mounted on the compute nodes (> 1000 clients for each file-system), so that jobs from users have direct POSIX access to all data.

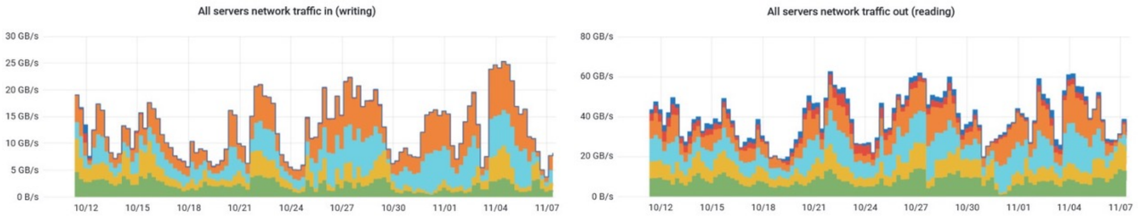

Figure 2: Network traffic in (left) and out (right) our disk storage systems over the last 30 days.

For what concerns tape storage, data are hosted in two tape libraries, with 10000 slots fully filled in the elder ORACLE SL8500, equipped with 16 T10KD tape drives (250 MB/s bandwidth per drive), and about 6200 slots filled in the newer IBM TS4500, equipped with 19 TS1160 tape drives (400 MB/s bandwidth per drive). On average, $10k$ files ($\sim$ 20 TB) are recalled and $20k$ files ($\sim$ 70 TB) are migrated every day.

As Hierarchical Storage Manager (HSM), we use GEMSS (Grid Enabled Mass Storage System, [1]), a thin software layer developed in house, together with IBM Spectrum Protect (formerly TSM - Tivoli Storage Manager) to manage the tape resources and provide all the data movements between disk and tape.

3. Data access

Our daily challenge is hosting data and resources for tens of experiments and projects, hence dealing with many and varied computing models, which means different storage usage, different requirements and different solutions. The majority of these computing and data models are experiments-driven, and thus somehow managed, even though users-driven (unmanaged) computing models are also possible. In general, software and middleware are mainly in the hands of experiments or of small groups of developers, and therefore very specific, far from being industry-standard.

A very common requirement for most users and experiments is POSIX access (mainly read POSIX access) to the data from the worker nodes and the user interfaces. On top of that, an ecosystem of heterogeneous protocols is adopted for data transfer, ranging from gridftp (with and without srm) to xrootd and https (with an without srm), coupled with different authentication and authorization methods, e.g. X509 proxies with and without VOMS attributes, tokens.

For what concerns data management, CNAF is a StoRM[2] site. We run a dedicated StoRM endpoint for each major experiment, plus two endpoints which are shared among the many others. Each StoRM endpoint has a dedicated pool of StoRM WebDAV transfer nodes (14 in total) and GridFTP transfer nodes (14 in total, see below). While data management servers (StoRM front-end and back-end) are virtualized, data movers run on dedicated high performing hardware.

In addition to the srm and https protocols, the experiments and users of INFN-CNAF data center are also using XrootD to distribute and access data [3].

In particular, among the WLCG experiments, ALICE has adopted XrootD since the very beginning of its activity, and we deploy an XrootD installation specifically optimized for ALICE to work on top of GPFS. Also, a specific plugin was developed at CNAF to manage tape recalls.

CMS uses an XrootD federation, with INFN-T1 hosting national and local redirectors, plus several servers [4].

ATLAS and LHCb also have dedicated XrootD instances, as they make use of XrootD sparingly for streaming data access; several other experiments have dedicated XrootD instances, e.g. AMS, DAMPE, JUNO, PADME.

VIRGO uses a dedicated Stashcache instance to access remote storage with a POSIX-based interface thanks to an extension to the CernVM-Filesystem (CVMFS, [5]) allowing to publish datasets instead of software [6].

These, together with various XrootD proxies and caching proxies in support of the HPC datacenter integration for jobs running in a dedicated partition of CINECA, Marconi, without external networking connectivity [7], all add up to 35 XrootD instances.

3.1 Supporting GridFTP protocol replacement

In 2017, Globus announced they would stop supporting the Globus Toolkit, whose end-of-life was then targeted for 2022. Over the last decades, WLCG has been using two major features from the Globus toolkit, namely GridFTP as the primary third-party-copy transfer protocol for the WLCG infrastructure, and the GSI authentication.

The latter is being transitioned to OAuth2.0 [8] token-based authorization in all relevant WLCG workflows, following the WLCG Authorisation Working Group initiative (see e.g. [9], [10]), although a realistic timeline for the full transition from X.509 certificates to JSON Web Tokens [11] seems to be slowly advancing towards 2024. In our storage services, we provide support for token-based authentication/authorization with StoRM WebDAV, and some experiments, most notably Belle II, are already accessing part of their data with X.509 VOMS proxies and JWT tokens used at the same time within the same directory structure.

Instead, for what concerns the GridFTP protocol replacement, the third-party-copy sub-group of the DOMA working group was created in order to investigate alternatives for bulk transfers across WLCG sites, and its efforts led to the ultimate decision that all storage elements support HTTP/WebDAV- or XrootD-based 3rd party data transfers. In particular, at INFN-T1 we provide support for HTTP-TPC with StoRM WebDAV.

The HTTP-TPC transition is most advanced, and the 2021 Network Data Challenge was carried out using HTTP-TPC for at-scale production transfers among the disk storage endpoints as the final step of the commissioning process. In that occasion, INFN Tier 1 performed very well.

In spring 2022, a WLCG Tape Data Challenge was also carried on using srm+http to write data to tape storage endpoints across WLCG sistes. INFN-T1 achieved target rates, however we reported two issues, namely a known bug affecting FTS, causing all gfal copy operations to always target the same host machine behind a given DNS alias, and the saturation of available StoRM WebDAV threads for one peculiar LHCb workflow. These issues prompted us to start investigating a load-balancing strategy for StoRM WebDAV endpoints, and to push LHC experiments for a definitive replacement of the gsiftp protocol (still used for data traffic from some Tier 2s and from the worker nodes, in a few workflows), given a significant increase in the efficiency of the transfers is measured when GridFTP and StoRM WebDAV do not execute at the same time on the same gateway server.

No-LHC experiments, on their hand, still rely heavily on gsiftp as data transfer protocol, and a joint initiative between our Data management team and the User support team at CNAF will be needed soon to enforce a transition away from gsiftp.

4. Future challenges

At INFN Tier 1, storage operations are proceeding smoothly.

In line with the WLCG DOMA efforts, support is provided to the ongoing transitions in the WLCG framework, for what concerns both data transfer protocols and token-based authentication/authorization. Currently, this translates into deploying (too) many services, so that we hope such ecosystem gets simplified.

Several R&D activities are also ongoing, most notably Ceph is being considered as a viable alternative for disk-only solution, i.e. without tape backend. A dedicated Ceph file system (CephFS) was deployed and is currently used by ALICE experiment. CephFS is optimized for relatively big file data transfers (of the order of few GB) mainly with XrootD. We foresee a possible usage also with StoRM WebDAV software. The usage of Ceph is not limited to CephFS though. We are also looking at Ceph object storage as an additional service for users with specific use cases.

Most importantly, our storage team is actively planning and working for the move of our data center to the Tecnopolo area, an essential step in order to cope with the increase of resources (by a factor of 2 in 2025, and by a factor of 10 starting in 2027 with the High-Luminosity upgrade of the LHC) demanded by the next challenges of science.

References

[1] Ricci, PP, Bonacorsi, D, Cavalli, A, dell'Agnello, L, Gregori, D, Prosperini, A, Rinaldi, L, Sapunenko, V and Vagnoni, V, The Grid Enabled Mass Storage System (GEMSS): the Storage and Data management system used at the INFN Tier1 at CNAF. In Journal of Physics: Conference Series, Vol. 396. IOP Publishing, 042051.

[2] Magnoni, L, Zappi, R, and Ghiselli, A (2008). StoRM: A flexible solution for Storage Resource Manager in grid. In IEEE Nuclear Science Symposium Conference Record, 2008, pp. 1971-1978, doi: 10.1109/NSSMIC.2008.4774772.

[3] Gregori, D, Boccali, T, Noferini, F, Prosperini, A, Ricci, PP, Sapunenko, V, and Vagnoni, V, Xrootd data access for LHC experiments at the INFN-CNAF Tier-1. In Journal of Physics: Conference Series, Vol. 513, Issue 4. IOP Publishing, 042023.

[4] Boccali, T, Donvito, G, Diacono, D, Marzulli, G, Pompili, A, Ricca, G, Mazzoni, E, Argiro, S, Gregori, D, Grandi, C, Bonacorsi, D, Lista, L, Fabozzi, F, Barone, L, Santocchia, A, Riahi, H, Tricomi, A, Sgaravatto, M, and Maron, G (2014). An Xrootd Italian Federation. In Journal of Physics: Conference Series, Vol. 513. IOP Publishing, 042013.

[5] Predrag, B, C Aguado, S, Jakob, B, Leandro, F, Artem, H, Pere, M, and Yushu, Y. 2010. CernVM-a virtual so ware appliance for LHC applications. In Journal of Physics: Conference Series, Vol. 219. IOP Publishing, 042003.

[6] Weitzel, D, Bockelman, B, Brown, D, Couvares, P, Würthwein, F, and Hernandez, E (2017). Data Access for LIGO on the OSG.

[7] Boccali, T, Ciangottini, D, Noferini, F, Bozzi, C, Perazzini, S, Valassi, A, Stagni, F, Doria, A, dell'Agnello, L, Maron, G, De Salvo, A, Zani, S, Morganti, L, Cesini, D, Sapunenko, V, Spiga, D and Dal Pra, S (2021). Enabling HPC systems for HEP: the INFN-CINECA Experience. In PoS: Vol. ISGC2021, pages 003.

[8] Hardt, D, The OAuth 2.0 Authorization Framework, RFC 6749 (2012), https://rfc-editor.org/rfc/rfc6749.txt

[9] Ceccanti, A, Vianello, E, and Giacomini, F (2020). Beyond X.509: Token-based authentication and authorization in practice. In EPJ Web Conf., Vol. 245, pages 03021, doi 10.1051/epjconf/202024503021

[10] Bockelman, B, Ceccanti, A, Dack, T, Dykstra, D, Litmaath, M, Sallé, M, and Short, H (2021). WLCG Token Usage and Discovery. In EPJ Web Conf., Vol. 251, pages 02028, doi 10.1051/epjconf/202125102028

[11] Jones, M B, Bradley, J, and Sakimura, N, RFC 7519, IETF Tools (2015), https://tools.ietf.org/rfc/rfc7519.txt