FTS improvements for LHC Run-3 and beyond

Edward Karavakis[1,*], *Andrea* Manzi[1] , *Maria* Arsuaga Rios[1] , *Oliver* Keeble[1] , *Carles* Garcia Cabot[1] , *Michal* Simon[1] , *Mihai* Patrascoiu[1] , and *Aris* Angelogiannopoulos[1]

[1]CERN, Esplanade des Particules 1, Geneva, Switzerland

Abstract. The File Transfer Service (FTS) developed at CERN and in production since 2014, has become a fundamental component for the LHC experiments and is tightly integrated with experiment frameworks. Starting from the beginning of 2018 with the participation to the European Commission funded project eXtreme Data Cloud (XDC) and the activities carried out in the context of the WLCG DOMA TPC and QoS working groups, a series of new developments and improvements have been planned and performed taking also into account the requirements from the experiments in preparation for the LHC Run-3. This paper provides a detailed overview of these developments; more specifically, the integration with OpenID Connect (OIDC), the QoS integration, the service scalability enhancements, the support for XRootD and HTTP Third Party Copy (TPC) transfers along with the integration with the new CERN Tape Archive (CTA) system.

1 Introduction

The File Transfer Service (FTS) [1, 2] is distributing the majority of the Large Hadron Collider (LHC) [3] data across the Worldwide LHC Computing Grid (WLCG) [4] infrastructure and is integrated with experiment frameworks such as Rucio [5], PhEDEx [6] and DIRAC [7]. It is used by more than 30 experiments at CERN and in other data-intensive sciences outside of the LHC and even outside the High Energy Physics (HEP) domain. FTS is part of the prototype of the European Commission funded ESCAPE [8] project offering a shared solution to computing challenges, targeting Astronomy and Particle Physics facilities and research infrastructures and focusing on developing solutions for handling Exabyte scale datasets.

FTS is a low-level data management service, responsible for scheduling reliable bulk transfer of files from one site to another while allowing participating sites to control the network resources usage. It can be accessed through CLI or REST API. FTS provides simplicity by allowing easy user interaction for submitting transfers, a WebFTS [9] portal, which is a web-based file transfer and management solution that allows users to invoke reliable, managed data transfers on distributed infrastructures from within their browser, a real-time monitoring that is rich in content and a Web Admin interface to be able to modify the internal settings of the service such as to configure access rights and limits on storages and links.

* Corresponding author: edward.karavakis@cern.ch

© The Authors, published by EDP Sciences. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (http://creativecommons.org/licenses/by/4.0/).

It also provides reliability as it ensures the data integrity since checksums are compared and failed transfers are individually retried. Moreover, features like multiprotocol support (WebDAV/HTTPS [10], GridFTP [11], XRootD [12], SRM [13]), diversity on the ways that clients can access the service (REST APIs, python bindings, CLI), transfers from and to different storages (EOS [14], DPM [15], Object Storages, STORM [16], dCache [17], CASTOR [18] and CTA [19]) and its support for tapes with the bringonline component make it flexible and scalable. Finally, one of the biggest advantages of FTS is its ability to be run without manually adding link and channel configuration with parallel transfer scheduling and optimisation to get the most from the network without saturating the storages, with support for intelligent priorities, activity shares and VO shares for classification of transfers.

In 2019, the centrally monitored FTS instances transferred more than 800 million files and a total of 0.95 Exabyte of data. The FTS team has been very active in performing several significant performance improvements to its core to prepare for the LHC Run-3 data challenges, supporting the new CERN Tape Archive (CTA) system, supporting a more user-friendly authentication and delegation method using tokens [20] and supporting the Third Party Copy (TPC) [21] and storage Quality of Service (QoS) [22] activities within the WLCG Data Organisation, Management and Access (DOMA) [23] project.

2 Performance enhancements in preparation for the LHC Run-3 data challenges

The performance of the scheduler and the optimiser of FTS was seriously affected when more than 2000 links were active (source/destination pairs) in combination with more than 2 million queued transfers. The database queries did not return the result in time and, as a consequence, a reduced number of transfers was scheduled when the instance was under heavy load. Performance improvements were needed in order for the experiments to be able to handle the increased load during the LHC Run-3 and also to be able to use fewer FTS instances to reduce operational costs.

The first set of improvements, that were released early on the 3.9 series, included the addition of missing indices, by examining all slow queries and passing them through a profiler, and various optimisations to get rid of expensive joins. Although gains were in the order of 5%-10%, the FTS instance was performing noticeably better under heavy load. The second set of improvements that were released at a later stage introduced table partitioning in MySQL, bringing significant performance gains in the order of 20%-30% as each node behind an FTS instance was accessing and writing to its very own partition within the same table, avoiding unnecessary locks on the entire table. Various optimisations were also performed by the CERN IT DBA team on the MySQL instances of FTS at CERN such as increasing the size of the InnoDB buffer pool and increasing the size of the InnoDB log file, performing less checkpoint flush activity and saving some disk I/O.

As a result of the aforementioned performance enhancements and optimisations, FTS regularly reached more than 6000 active links without observing any performance issue. These optimisations were documented and shared with the other FTS instances.

3 Integration with the CERN Tape Archive (CTA) system

CTA is the new tape based solution implemented at CERN and integrated with the EOS system. CTA will receive new data from the LHC experiments during Run-3 and all the existing data from CASTOR will be imported in order for CASTOR to be phased out

before Run-3. It exposes an XRootD interface and supports Third Party Copy (TPC) transfers.

3.1 FTS support for staging with XRootD

The FTS support for CTA is production ready and it has been successfully stress-tested during the multiple ATLAS Data Carousel [24] exercises. Most importantly, the interface to FTS remained the same for the experiments as everything is handled transparently by FTS.

FTS manages stage-in and transfers between EOS and CTA. The staging activity was implemented via the XRootD protocol with support for "staging and multihop" transfers in order to first stage from the tape, then to copy from the tape disk buffer to EOS and from there to handle the data export to the Tier-1s. In addition, disk copy eviction was implemented such that once the staged file is copied successfully to the destination, it is evicted from the source disk cache in order to better handle the reduced buffer size of CTA.

3.2 Monitoring of migration to tape

FTS has been designed in order to optimise WAN transfers between Storage Endpoints via different protocols. In the case of transfers to a tape-backed storage, the entire process of tape migration was not taken into account by FTS, which considered the transfer to be successfully completed at the storage system disk buffer level only.

Clients submitting transfers to a tape-backed system via FTS had to explicitly check on the destination storage if the file has been correctly migrated to tape before issuing, for example, any clean up on their side. An "archive to tape" monitoring feature was therefore needed as part of the migration of CMS from PhEDEx to Rucio and by CTA as a way to throttle active transfers. This new feature enables the reporting of a transfer to a tape-backed storage as completed only when the file has been migrated to tape successfully.

FTS was extended to monitor the migration of a file to tape and to fail the transfer if this does not complete within a given timeout. The updated FTS transfer state machine can be seen in Figure 1.

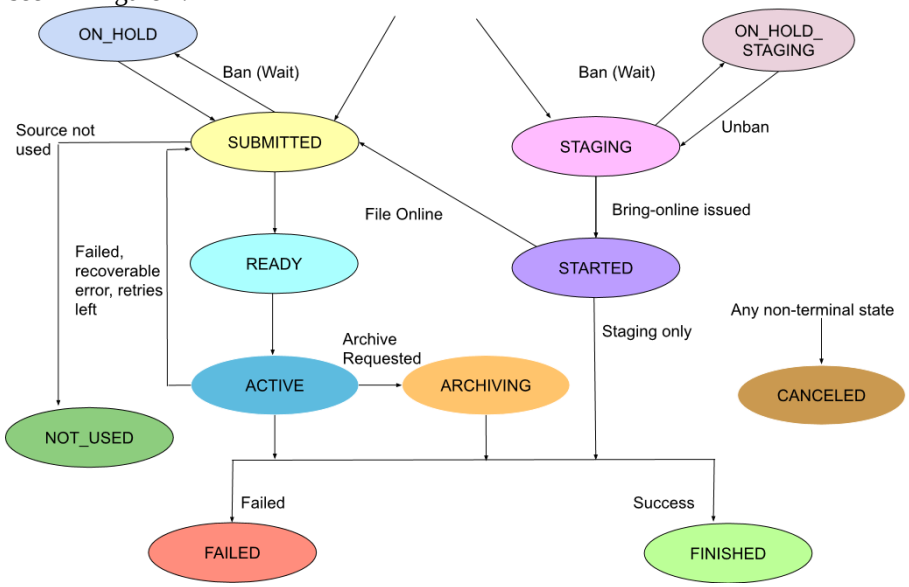

Fig. 1. The new FTS transfer state machine to support the migration to tape.

- *SUBMITTED*. Initial state of a file as soon as it is inserted into the database
- *READY*. File is ready to become active
- *ACTIVE*. Transfer is running
- *FAILED*. Transfer has failed, and the number of retries have been exhausted
- *FINISHED*. Transfer has finished successfully
- *STAGING*. When staging of a file is requested
- *STARTED*. Bring online request sent to the server
- *ARCHIVING*. File is migrating to tape
- *NOT_USED*. For multiple replica jobs, those replicas that have not been used for the transfer
- *ON_HOLD*. When either the storage or the destination is banned temporarily, transfers won't go through
- *ON_HOLD_STAGING*. Same as before, but for files that need to be staged first.

A prototype was designed and implemented that included a new FTS transfer state machine to support the new "ARCHIVING" state, with changes to the DB schema, to the Messaging component which publishes to external systems such as the CERN IT MONIT project [25] and to the clients (REST and CLI). Users need to enable this feature when they submit a transfer as FTS is not aware if a storage has a tape backend or not. The new "ARCHIVING" state, is a non-final state and it is reported back to clients polling for transfer status and as well as a transfer state change via Messaging. This first prototype was tested with both XRootD (for CTA) and SRM capable endpoints. At a later stage, a back-pressure mechanism could be implemented along with limits on the maximum number of "ARCHIVING" requests to a tape-enabled storage in order to have buffer-aware scheduling.

4 Work to support the WLCG DOMA TPC activity

The purpose of the Third Party Copy (TPC) working group of the WLCG DOMA project is twofold: to find a viable replacement to the GridFTP protocol for bulk transfers and to replace the X.509 certificate-based authorisation with token-based authorisation.

4.1 HTTP and XRootD TPC

Alternative protocols to GridFTP enable the community to diversify; explore new approaches such as alternate authorisation mechanisms; and reduce the risk due to the retirement of the Globus Toolkit [26], which provides a commonly used GridFTP protocol implementation. Two alternatives were selected by the TPC group: HTTP/WebDAV and XRootD. Each approach has multiple implementations, allowing to demonstrate interoperability between distinct storage systems. Each major storage system utilised by WLCG sites has at least one functional non-GridFTP protocol for performing third-party-copy. FTS fully supports XRootD TPC with X509 delegation and HTTP TPC with X509 delegation and various bearer token technologies (see 4.2).

While waiting for the completion of the storage upgrade campaigns at the WLCG sites, a testbed was set up for functional and stress testing of HTTP and XRootD TPC transfers with daily reports for each of the sites and storage solutions. An example DOMA TPC dashboard can be seen in Figure 2, showing the efficiency of transfers between sites to evaluate the current status of HTTP TPC.

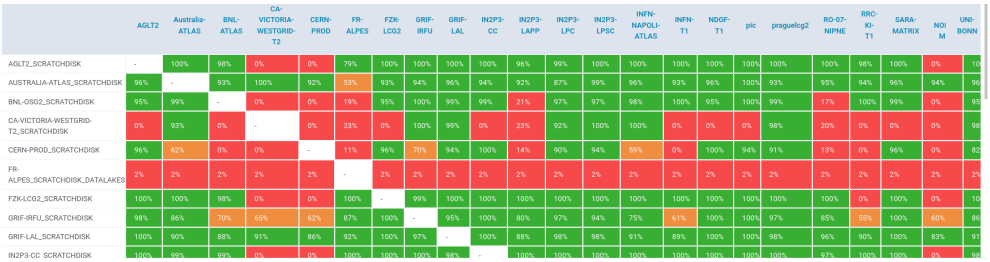

Fig. 2. An example DOMA TPC dashboard showing the HTTP TPC functional test results.

4.2 HTTP Token-based authentication and authorisation

There is an increasing interest on the replacement of X.509 certificates with tokens issued by a centralised identity provider. Developments in both FTS and GFAL2 were performed in order to enable the token retrieval of Macaroons [27] and Scitokens [28] (development contributed by the Scitokens project) and their usage in HTTP TPC transfers. As part of the European Commission funded Horizon2020 eXtreme Data Cloud (XDC) [29] project, integration with OpenID Connect (OIDC) has been implemented. This allows the users to authenticate to FTS REST via an OIDC access token and FTS contacts the storages with that token to perform the transfers. FTS can be configured to support multiple identity providers such as the XDC Identity and Access Management (IAM) and the WLCG IAM.

In the WLCG scenario, as shown in Figure 3, the client's identity is delegated. Taking Rucio as an example, Rucio gets a token from IAM with the minimum privileges needed to interact with FTS. Rucio submits a transfer job to FTS (1), along with the token obtained from IAM. FTS then checks if the token is valid, either offline (using the cached keys from IAM) or online (using token introspection via IAM). If valid, the transfer is accepted. FTS now needs a token impersonating Rucio that will be used for authentication and authorisation at the storage elements. The token it already has cannot be used for the transfer as it does not provide the necessary rights to read and store files at the storage elements. FTS then exchanges the obtained token for an access token and a refresh token that will be used to manage the transfer (2, 3). FTS will use the refresh token to get a fresh access token when the transfer is about to start. FTS then submits the third-party transfer against Storage Element (SE) 2 by including the token in the request (4). The same token will be used for authentication/authorisation at SE 1 and SE 2. SE 2 will then use the obtained token for authentication/authorisation against SE 1 (5).

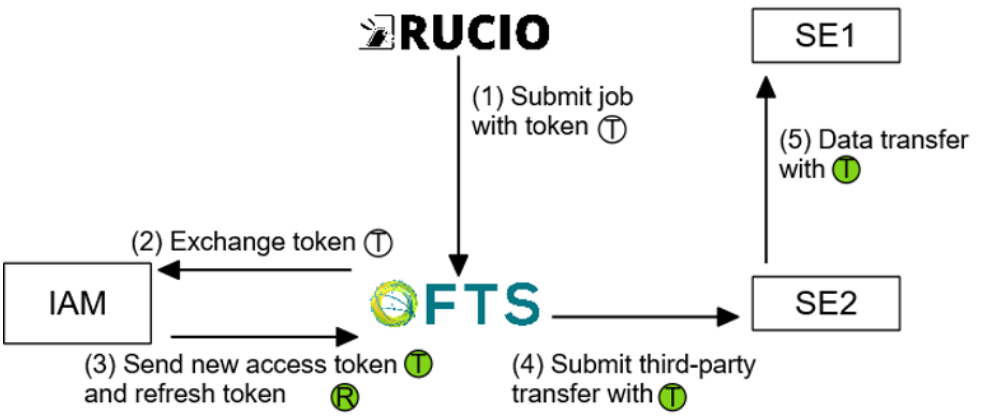

Fig. 3. The token workflow in the WLCG scenario, taking Rucio as an example client.

5 Work to support the WLCG DOMA QoS activity

Given the expected flat budget for High-Lumi / LHC Run-4, the mandate of the QoS working group is to create a mechanism to allow a diversity where sites can offer specific QoS options through innovative solutions that save cost. It aims to drive down the total cost of the storage, while allowing experiments to optimise their storage usage. A Storage QoS represents a common agreement between storage providers and the scientists using that storage on how that storage system should behave. A QoS class is typically understood in terms of access-latency, bandwidth, and likelihood of data loss. Some storage systems may provide a single QoS, while others may provide several QoS. Data may require different QoS classes at different times. Moving away from simple descriptions (DISK and TAPE) to more general concepts may allow sites to better manage the provided storage in order to drive down costs. It may also allow trade-offs, such as providing increased storage capacity but with an increased risk of data loss.

Work on the FTS project has started since 2018 as part of the XDC project in order to support QoS transitions by integrating the Cloud Data Management Interface (CDMI) [30] protocol. A first version of the FTS QoS daemon was released for XDC. GFAL2 was also extended to implement CDMI operations. The QoS daemon is a bringonline daemon managing two QoS values; “disk” and “tape”, making use of the GFAL2 CDMI API to query, perform and monitor QoS operations to storages. At the moment, this is supported by dCache and EOS.

A proof-of-concept (PoC) was implemented demonstrating support for a basic QoS functionality; to request and be able to monitor a QoS transition. The FTS submission interface and state machine have been updated to support QoS transitions. FTS will receive a transfer with some QoS metadata (target QoS). If the destination directory does not exist, it will create it with the appropriate QoS. If the destination directory exists, it will check if the directory has the requested QoS. If it doesn’t have it, it will check if the required QoS transition is permitted for a file in that directory and, if not, it will fail. It will then put the file and change the QoS in necessary.

6 Conclusion

Various performance improvements and new features were put in place in preparation for the LHC Run-3. FTS is the workhorse for asynchronous point-to-point data transfers behind the WLCG DOMA activity, and more specifically behind the TPC, the token authentication and authorisation and QoS working groups, and the European Commission funded ESCAPE project. The CERN Tape Archive (CTA) system will receive new data from the LHC experiments during Run-3 and the existing data from CASTOR will be imported in order for CASTOR to be phased out before Run-3. FTS has implemented the integration with CERN’s new tape archival system and it has been successfully stress tested on multiple occasions during the ATLAS Data Carousel exercises.

FTS continues to evolve with the infrastructure as WLCG’s principal data movement service and, at the same time, it expands its community and adoption by upcoming data-intensive projects.

References

1. FTS website, http://fts.web.cern.ch
2. A. A. Ayllon et al, *FTS3: New Data Movement Service For WLCG*, J. Phys.: Conf. Ser. **513** 032081 (2014)

3. L. Evans and P. Bryant, *LHC Machine*, JINST **3** S08001 (2008)
4. I. Bird, *Computing for the Large Hadron Collider*, Annual Review of Nuclear and Particle Science **61** :99-118 (2011)
5. M. Barisits, T. Beermann, F. Berghaus et al, *Rucio: Scientific Data* Management, Comput Softw Big Sci **3**: 11 (2019)
6. M. Giffels, Y. Guo, V. Kuznetsov et al, *The CMS Data Management System*, J. Phys.:Conf. Ser. **513** 042052 (2014)
7. S K Paterson and A Tsaregorodtsev, *DIRAC optimized workload management*, J. Phys.: Conf. Ser. **119** 062040 (2008)
8. S. Campana et al, *ESCAPE prototypes a Data Infrastructure for Open Science*, EPJ Web of Conferences (to be published)
9. A. Kiryanov, A. A. Ayllon and O. Keeble, *FTS3 / WebFTS – A Powerful File Transfer Service for Scientific Communities*, Procedia Computer Science **66** 670-678 (2015)
10. G. Bernabeu et al, *Experiences with http/WebDAV protocols for data access in high throughput computing*, J. Phys.: Conf. Ser. **331** 072003 (2011)
11. W. Allcock et al, *The Globus Striped GridFTP Framework and Server*, SC '05: Proceedings of the 2005 ACM/IEEE Conference on Supercomputing, Seattle, WA, USA, 2005, pp. 54-54, doi: 10.1109/SC.2005.72 (2005)
12. A. Dorigo et al, *XROOTD - a highly scalable architecture for data access*, WSEAS Transactions on Computers, 4(4):348--353 (2005)
13. L. Abadie et al, *Storage Resource Managers: Recent International Experience on Requirements and Multiple Co-Operating Implementations*, 24th IEEE Conference on Mass Storage Systems and Technologies (MSST 2007), San Diego, CA, 2007, pp. 47-59, doi: 10.1109/MSST.2007.4367963 (2007)
14. A. J. Peters and L. Janyst, *Exabyte Scale Storage at CERN*, J. Phys.: Conf. Ser. **331** 052015 (2011)
15. A. Alvarez et al, *DPM: Future Proof Storage*, J. Phys.: Conf. Ser. **396** 032015 (2012)
16. A. Carbone et al, *Performance Studies of the StoRM Storage Resource Manager*, Third IEEE International Conference on e-Science and Grid Computing (e-Science 2007), Bangalore, pp. 423-430 (2007)
17. P. Fuhrmann and V. Gülzow, *dCache, Storage System for the Future*, Euro-Par 2006 Parallel Processing. Euro-Par 2006. Lecture Notes in Computer Science, vol **4128** (2006)
18. G Lo Presti et al, *CASTOR: A Distributed Storage Resource Facility for High Performance Data Processing at CERN*, 24th IEEE Conference on Mass Storage Systems and Technologies (MSST 2007), San Diego, CA, 2007, pp. 275-280 (2007)
19. E. Cano et al, *CERN Tape Archive: production status, migration from CASTOR and new features*, EPJ Web of Conferences (to be published)
20. A. Ceccanti et al, *WLCG Authorisation; from X.509 to Tokens*, EPJ Web of Conferences (to be published)
21. A. Forti et al, *Modernizing Third-Party-Copy Transfers in WLCG*, EPJ Web of Conferences (to be published)
22. M. Lassnig et al, *Quality of Service (QoS) for cost-effective storage and improved performance*, EPJ Web of Conferences (to be published)
23. D. Berzano et al, *HEP Software Foundation Community White Paper Working Group -- Data Organization, Management and Access (DOMA)*, arXiv:1812.00761 (2018)

24. X. Zhao et al, *The ATLAS Data Carousel Project*, EPJ Web of Conferences (to be published)
25. E. Karavakis et al, *Unified Monitoring Architecture for IT and Grid Services*, *J. Phys.: Conf. Ser*. **898** 092033 (2017)
26. I. Foster and C. Kesselman, *Globus: A Toolkit-Based Grid Architecture*, The Grid: Blueprint for a New Computing Infrastructure, pp. 259-278, (1999)
27. A. Birgisson et al, *Macaroons: Cookies with Contextual Caveats for Decentralized Authorization in the Cloud*, Network and Distributed System Security Symposium, Internet Society (2014)
28. A. Withers, B. Bockelman et al, *SciTokens: Capability-Based Secure Access to Remote Scientific Data*, PEARC '18: Proceedings of the Practice and Experience on Advanced Research Computing **24** pp. 1-8 (2018)
29. P. Fuhrmann et al, *The eXtreme-DataCloud project - solutions for data management services in distributed e-infrastructures*, EPJ Web of Conferences (to be published)
30. Cloud Data Management Interface (CDMI) website, https://www.snia.org/cdmi