EPJ Web of Conferences **295**, 01040 (2024) https://doi.org/10.1051/epjconf/202429501040
CHEP 2023

New XrootD Monitoring Implementation

Borja Garrido[1,*], *Julia* Andreeva[2,**], *Derek* Weitzel[3,***], *Alessandra* Forti[4,****], and *Shawn* McKee[5,†]

[1]CERN, Meyrin, Switzerland
[2]University of Nebraska-Lincoln, Lincoln, US
[3]University of Manchester, Manchester, UK
[4]University of Michigan Physics, 450 Church St, Ann Arbor MI 48109, USA

Abstract. Complete and reliable monitoring of the WLCG data transfers is an important condition for effective computing operations of the LHC experiments. WLCG data challenges organized in 2021 and 2022 highlighted the need for improvements in WLCG data traffic monitoring. In particular, it concerns the implementation of remote data access monitoring via the root protocol. It includes data access to native XRootD storage, as well as to other storage solutions. We refer to it as XRootD monitoring. This contribution describes the new implementation of the XRootD monitoring flow, the overall architecture, the deployment scenario, and the integration with the WLCG global monitoring system.

1 Introduction

From 2021 WLCG[1] is organizing data challenges in order to scale out the infrastructure for the HL-LHC targets[2] for which a large increase in data volume is forecast. Following the outcome of the data challenges in 2021 and 2022, the WLCG Monitoring Task Force has been set up. The goal is to improve the completeness and accuracy of the WLCG data traffic monitoring, which implies data transfer and remote data access on the WLCG infrastructure. For data transfers, LHC experiments mainly rely on the File Transfer Service (FTS[3]), which has been well instrumented for monitoring purposes and as a result the monitoring of data traffic handled by FTS is well established. Another important contribution to network traffic on the WLCG infrastructure is remote data access using the XRootD protocol[4]. It includes data access to native XrootD[5] storage, as well as other storage solutions. We refer to it as XRootD monitoring. Data challenges and analysis of XRootD monitoring data detected serious issues in the monitoring of XRootD data flow and therefore improvements in this area were considered to be an important area of work of the WLCG Monitoring Task Force[6]. In the following sections, changes in the overall XRootD monitoring architecture will be described along with the current status of its implementation and future plans.

*e-mail: borja.garrido.bear@cern.ch
**e-mail: Julia.Andreeva@cern.ch
***e-mail: dweitzel@unl.edu
****e-mail: alessandra.forti@cern.ch
†e-mail: smckee@umich.edu

2 Overview of the monitoring landscape

In order to tackle the known issues of the current transfer monitoring solutions, an extensive evaluation has been made to identify the components that would require modifications. In the schema presented in figure 1 various monitoring data producing and processing systems are presented.

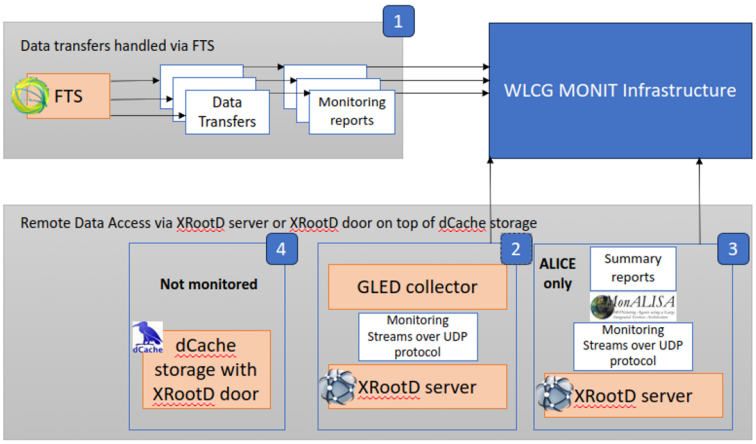

Figure 1. Current transfers monitoring infrastructure overview

Monitoring of WLCG data traffic is mostly based on the monitoring reports produced either by transfer systems (FTS) or by servers that manage storage and data access, for example, XRootD or dCache. These monitoring reports are being collected, processed, and exposed by the MONIT[7] infrastructure at CERN. As has already been mentioned, the data flow managed by FTS is well monitored. In the gray block in the lower part of Figure 1, various use cases are presented that are not managed by FTS but contribute to data traffic on the WLCG infrastructure. These use cases are being addressed by the WLCG Monitoring Task Force. In this section, each of these use cases will be described in detail.

2.1 Monitoring flow producer by XRootD server

The processing of the monitoring flow produced by the XRootD server is shown in blocks 2 and 3 in Figure 1. XRootD server reports monitoring information via the UDP protocol. ALICE experiments are fully dependent on the MonALISA[8] monitoring system for all computing activities, including data traffic. There is a MonALISA instance deployed at every ALICE site that performs processing of the monitoring data produced by the site. The processing and aggregation of the data in the ALICE global scope is done centrally by a central MonALISA instance. The advantage of this approach is that primary processing and aggregation occur on site using a local area network; therefore, the probability of data loss is low. The validation of MonALISA monitoring data confirmed its reliability.

For other experiments, monitoring data reported by the XRootD server with UDP packets are collected and processed centrally by the so-called GLED collectors. This implementation has several issues which result in poor quality of monitoring information. Studies[9] that have been performed to estimate the quality of the data showed that the volume of data transferred calculated by the GLED collector is approximately 60% lower compared to the real volume.

UDP fragmentation[10] is mentioned as the main reason for the low reliability. A single transfer operation is described in several UDP packets produced by the XRootD server. Lack of acknowledgment for UDP packets as well as the big size of the XRootD streams cause UDP packet loss, which creates a critical condition, since loss of a single packet makes the reconstruction of a given transfer operation impossible. Moreover, some limits were discovered with parallel processing of the collector, causing the collector to start dropping streams when the throughput reaches 100 streams per second. A new implementation has been proposed to solve the mentioned technical limitations. It is described in Section 3.

XCache[11] is an XrootD disk-based file proxy cache. From the implementation point of view, XCache is a'regular' XRootD server, therefore, monitoring of data traffic based on XRootD and XCache monitoring streams is similar.

2.2 dCache with XRootD door

To allow file transfers in and out of dCache[12] storage using XRootD, the so-called XRootD door which acts as the entry point to all XRootD requests must be set up. XRootD door is not equivalent to the XRootD server implementation and is not enabled with monitoring. Up to now this use case, which is shown in block 4 in figure 1, is completely missing in the global WLCG data traffic monitoring. The technical solution to implement monitoring of traffic entering and exiting dCache with XRootD door is described in Section 4.

3 New implementation of the monitoring flow produced by XRootD servers

3.1 Overall architecture

In order to address the issues summarized in Section 2.1, OSG[13] came with a proposal for a new architecture based on two new components: shoveler and collector.

The shoveler is thought to be a very lightweight daemon that resides next to the XRootD server and enables reporting of the monitoring data via message queue. Shoveler deployment on the same local network as XRootD server dramatically decreases the probability of UDP packet loss.

The new implementation of the collector is similar to the concept of the GLED collector, except that the new collector consumes data from the message queue, ensuring fault tolerance and reconstruction capabilities in case of long outage. A new collector performs data aggregation on the level of a single transfer operation. It reads messages correlating by their content to the individual XRootD UDP packet and reconstructs from them the monitoring data required for a single transfer operation. This is required since a single transfer document is composed of several "streams" or UDP packets that need to be combined together in order to reconstruct the transfer information. Data aggregated on a single transfer operation level are then reported to a dedicated queue of the message bus. From there it can be consumed by other clients. For OSG sites, the aggregated data is consumed by the GRACC OSG accounting system GRACC[14].

In order to benefit from OSG development and to avoid duplication of effort, WLCG considered the new architecture proposed by OSG with the only exception of message queue implementation. OSG uses a commercially operated RabbitMQ[15] message bus residing in the United States, while the European part of the WLCG infrastructure relies on the ActiveMQ[16] supported by CERN. The data aggregated by a US collector is then imported into a dedicated queue of CERN ActiveMQ message bus to reconstruct the complete picture. The new architecture is depicted in figure 2.

Please note that the figure 2 shows a shoveler running as part of the OSG block, this is a central shoveler depeloyed to ease the integration of sites that can't run the component on their own, this is anyhow discouraged as it brings back on of the issues tried to be solved by the new infrastructure regarding UDP fragmentation.

In order to converge ALICE XRootD monitoring with the common implementation, the ability of XRootD servers to forward monitoring streams to two different endpoints will be used. In this scenario, ALICE XRootD monitoring streams will be delivered to both MonALISA and the site shoveler to enable ALICE data processing in a common way. In the long term, this will allow us to reduce maintenance costs.

Finally, another flow that will be covered by this new architecture is the XCache one. XCache generates streams similar to classical XRootD servers. OSG has already confirmed that the XCache flow is compatible with the shoveler/collector approach.

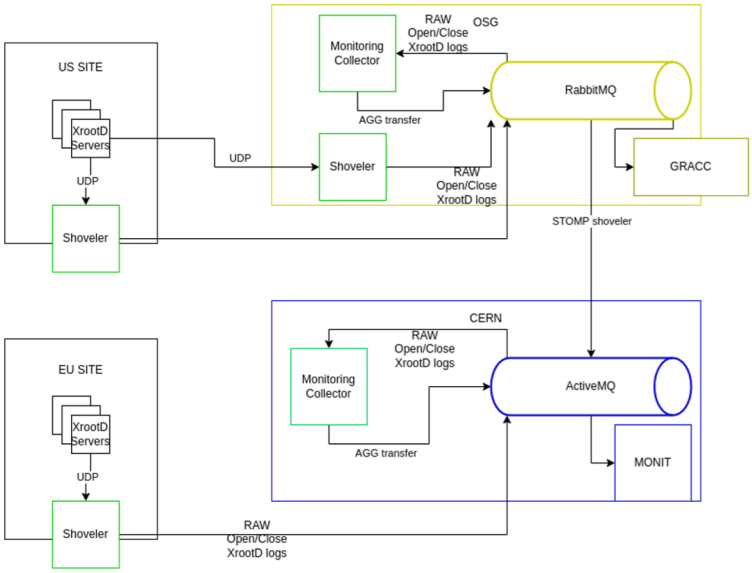

Figure 2. Proposed architecture for new XRootD monitoring flow implementation

3.2 Required modifications

Apart from using a different mechanism for transferring monitoring data, a couple of other modifications had to be implemented on the collector level in order to overcome limitations of the previous implementation. New XRootD collector ingests monitoring data from the XRootD server, aggregates it into one monitoring record per transfer operation, and sends a resulting JSON formatted record into a message bus. For backward compatibility, the collector is capable of consuming UDP packets as well as data from the message bus, either RabbitMQ or ActiveMQ.

The initial version of the shoveler developed for OSG has been modified to communicate with the message queue using the STOMPprotocol [17], as a requirement for the CERN messaging service. This includes a secure connection over TLS using x509 certificates. The evaluation of token support has been discussed with the CERN messaging team and will be studied as a potential medium-term improvement.

3.3 Current status

The new collector overcomes the parallelism limitation, allowing it to run a single collector instance per US or European infrastructure. The implementation implies spawning of several subprocesses in various threads, improving overall throughput.

According to Section 3.2, the key modification for the new shoveler/collector components to work with the CERN infrastructure is to adapt them to talk to the ActiveMQ message bus using the STOMP protocol. This has been completed, and newer versions of these components are already being shipped with the required changes.

On top of this, in order to start CERN EOS servers integration a setup has been configured at CERN running in Kubernetes with a battery of shovelers for high availability. The central collector for non-US sites is also running at CERN and is already producing data.

4 Implementation of monitoring for dCache with XRootD door use case

In this section, we will cover the need to integrate the dCache monitoring solution with the XRootD door use case. The part of traffic managed by FTS is well monitored, while data access to dCache storage services via the XRootD protocol is completely missing in the WLCG monitoring system. The goal is to provide an accurate monitoring picture for DC24 as much as possible. In particular it is important to enable monitoring for FNAL dCache storage, as so-called "pileup" data samples are hosted by FNAL and therefore many CMS jobs will access these data remotely.

The monitoring for dCache is completely different from that provided by the XRootD servers. dCache produces monitoring messages and forwards them to Apache Kafka real-time data streaming technology. The content and format of the monitoring data is completely different from that agreed for XRootD monitoring. The sites which are running dCache and want to enable monitoring are supposed to deploy a local Kafka instance. The proposed technical solution foresees a new component which will consume dCache monitoring data from a local Kafka instance, translate them into a format agreed for XRootD monitoring flow and report the reformatted message to the message bus at CERN. This technical solution came up as a result of collaboration of the WLCG Monitoring Task Force, the dCache development team, and the FNAL dCache experts. It is described in figure 3.

One of the issues that has to be resolved in this scenario is the fact that the current set of dCache monitoring data does not contain information about the source and destination of the transfer operation. This information is mandatory for WLCG transfer monitoring. It has been agreed with dCache developers that at least IP addresses of the data transfers source and destination will be enabled in the dCache monitoring set. This feature has been enabled in dCache 9.12. However, in reference to source/destination, WLCG transfer monitoring operates with WLCG sites rather than with IP addresses. Therefore, it is necessary to translate the source / destination of the IP addresses into WLCG site names. It has been agreed that this functionality will be provided by the WLCG CRIC topology system[18]. An API which returns the name of the WLCG site taking the IP address as an input parameter has been enabled in CRIC. The first prototype of the new data flow is currently being tested.

5 Future work

Current effort is focused on testing and debugging the new workflows. By the end of autumn 2023, new components should be deployed in production at least at CERN and FNAL and

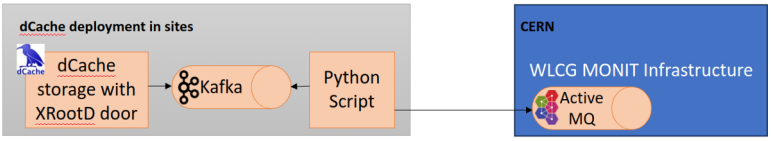

Figure 3. Proposed architecture for dCache with XRootD door monitoring flow implementation

several pioneer sites. The massive deployment campaign is planned for the beginning of next year. The progress of deployment at the WLCG sites should be followed by a thorough validation of the monitoring data.

There are several issues that must be addressed in the longer term. One of those issues is the eventual lack of information about the virtual organization that initiated the transfer. It was discovered that it is not always properly defined in the XRootD data flow. WLCG Monitoring Task Force needs to work with XRootD developers to resolve this issue, as it is crucial for WLCG monitoring. Enabling token authentication for communication with the message bus is another area that has to be addressed with the CERN WLCG message team. This doesn't apply to dCache monitoring flow as they are able to set the VO properly when sending this data.

Another important goal is to enable visibility of traffic for XRootD data flows. This is followed as part of the "scientific tags"[19] initiative driven by the Research Networking Technical Working Group[20].

6 Conclusions

Following the outcome of the 2021-2022 data challenges, the WLCG Monitoring Task Force is contributing to the task of improving the WLCG data transfer monitoring. An important effort has been made to monitor remote data access through the XRootD protocol, which is currently not well covered by the WLCG monitoring system. An analysis of the problems of the current system has been performed and a new architecture has been proposed and prototyped. The validation of the new implementation is ongoing. The start of deployment of the new implementation in production is planned for autumn 2023. It will ensure our ability to reliably monitor all data traffic on the WLCG infrastructure.

7 Acknowledgements

We want to explicitly acknowledge the support of the National Science Foundation, which supported some of this work through OSG: NSF MPS-1148698 and IRIS-HEP: NSF OAC-1836650 grants. Furthermore, we acknowledge our collaborations with the CERN IT, WLCG DOMA, and LHCONE/LHCOPN communities, who also participated in this effort.

References

[1] *WLCG: Worlwide LHC Computing Grid*, https://wlcg.web.cern.ch/, [Online; accessed 20-September-2023]

[2] *HL-LHC: High Luminosity Large Hadron Collider*, https://hilumilhc.web.cern.ch/, [Online; accessed 20-September-2023]

[3] *FTS: File Transfer Service, Open source software for reliable and large-scale data transfers*, https://fts.web.cern.ch/fts/, [Online; accessed 20-September-2023]

[4] A. Hanushevsky, *The XRootD Protocol version 4.0.0*, https://xrootd.slac.stanford.edu/doc/dev49/XRdv400.htm (2018), [Online; accessed 20-September-2023]
[5] *XRootD: Fully generic suite for fast, low latency and scalable data access*, https://xrootd.slac.stanford.edu/, [Online; accessed 20-September-2023]
[6] *WLCG Monitoring Task Force*, https://twiki.cern.ch/twiki/bin/view/LCG/MonitoringTaskForce, [Online; accessed 20-September-2023]
[7] A. Aimar, A. Aguado, P. Andrade, J. Delgado, B. Garrido, E. Karavakis, D. Marek, L. Magnoni, *MONIT: Monitoring the CERN Data Centres and the WLCG Infrastructure*, in *EPJ Web of Conferences* (CERN, 2019), p. 8
[8] *MonALISA: ALICE Grid monitoring with MonALISA*, https://alien.web.cern.ch/content/alice-grid-monitoring-monalisa, [Online; accessed 20-September-2023]
[9] D. Weitzel, D. Davila, *XRootD Monitoring Validation*, in *Zenodo* (2020)
[10] *Broken packets: IP fragmentation is flawed*, https://blog.cloudflare.com/ip-fragmentation-is-broken/, [Online; accessed 20-September-2023]
[11] *XCache: Caching of data accessed using xrootd protocol*, https://slateci.io/XCache/, [Online; accessed 20-September-2023]
[12] *dCache: Distributed storage for scientific data*, https://www.dcache.org/, [Online; accessed 20-September-2023]
[13] *OSG: Open Science Grid*, https://osg-htc.org/, [Online; accessed 20-September-2023]
[14] K. Retzke, D. Weitzel, S. Bhat, T. Levshina, B. Bockelman, B. Jayatilaka, C. Sehgal, R. Quick, F. Wuerthewin, *GRACC: New generation of the OSG accounting*, in *IOP Conf. Series: Journal of Physics* (Fermilab, University of Nebraska, Indiana University and University of California, 2017), p. 9
[15] *RabbitMQ: Easy to use, flexible messaging and streaming service*, https://www.rabbitmq.com/, [Online; accessed 20-September-2023]
[16] *ActiveMQ: Flexible and powerful open source multi-protocol messaging*, https://activemq.apache.org/, [Online; accessed 20-September-2023]
[17] *STOMP: Simple Text Oriented Messaging Protocol*, https://stomp.github.io/, [Online; accessed 20-September-2023]
[18] A. Anisenkov, J. Andreeva, A.D. Girolamo, P. Paparrigopoulos, V. A, *CRIC: A unified information system for WLCG and beyond*, in *EPJ Web of Conferences* (Budker Institue of Nuclear Physics, Novosibirsk State University, CERN and PIC, 2018), p. 8
[19] G. Attebury, M. Babik, D. Carder, T. Chown, A. Hanushevsky, B. Hoeft, A. Lake, M. Lambert, J. Letts, S. McKee et al., *Identifying and Understanding Scientific Network Flows*, in *EPJ Web of Conferences* (EDP Sciences, 2023)
[20] *Research Network Technical Working Group*, https://twiki.cern.ch/twiki/bin/view/LCG/ResearchNetworkingWG, [Online; accessed 20-September-2023]