Data transfer for STAR grid jobs

Irakli Chakaberia[1], **Jerome Lauret**[2], **Michael Poat**[2], **Jefferson Porter**[1]

[1] Lawrence Berkeley National Laboratory, Berkeley, California, USA
[2] Brookhaven National Laboratory, Upton, New York, USA

E-mail: `iraklic@lbl.gov`

Abstract. The Solenoidal Tracker at RHIC (STAR) is a multipurpose experiment at the Relativistic Heavy Ion Collider (RHIC) with the primary goal to study the formation and properties of the quark-gluon plasma. STAR is an international collaboration of member institutions and laboratories from around the world. Yearly data-taking period produces PBytes of raw data collected by the experiment. STAR primarily uses its dedicated facility at BNL to process this data, but has routinely leveraged distributed systems, both high throughput (HTC) and high performance (HPC) computing clusters, to significantly augment the processing capacity available to the experiment.

The ability to automate the efficient transfer of large data sets on reliable, scalable, and secure infrastructure is critical for any large-scale distributed processing campaign. For more than a decade, STAR computing has relied upon GridFTP with its x509-based authentication to build such data transfer systems and integrate them into its larger production workflow. The end of support by the community for both GridFTP and the x509 standard requires STAR to investigate other approaches to meet its distributed processing needs.

In this study we investigate two multi-purpose data distribution systems, Globus.org and XRootD, as alternatives to GridFTP. We compare both their performance and the ease by which each service is integrated into the type of secure and automated data transfer systems STAR has previously built using GridFTP. The presented approach and study may be applicable to other distributed data processing use cases beyond STAR.

1. Introduction

STAR [1] is a large multipurpose detector that collects tens of PBytes of data each year from high energy proton or heavy ion collisions provided by RHIC. To reconstruct the raw physics data into analysis level objects requires a large amount of computing power. The Scientific Data and Computing Center (SDCC) at Brookhaven National Laboratory (BNL) provides the majority of computing and storage resources for STAR's needs. However, STAR has often supplemented it's computing needs by using external computing resources such as the PDSF cluster at NERSC. PDSF was heavily used for raw data reconstruction and embedding simulation jobs. PDSF has now been retired and we have migrated the STAR data production workflow to the High-Performance Computing (HPC) system known as Cori at NERSC. This transition has required a revision of the workflow for provisioning both software and data from STAR to NERSC. Our workflow was developed by adopting containerization, CERN's software distribution service, and optimization of job management at the HPCs [2, 3, 4].

2. Data Management

Central data processing campaigns at STAR consist of raw data reconstruction and simulated data embedding - two distinctly different workloads as far as data management is concerned. Embedding jobs require relatively small input-data that is reused during the multiple embedding productions. Conversely, raw data reconstruction entails processing PBytes of data over an extended production campaign. While the former could be achieved by staging the necessary data once and transferring the modest-sized produced datasets back as needed, the latter necessitates a robust, dynamic data management system between the two sites.

In the past, our data management system relied upon the gridFTP protocol with its x509 standard certificates used for authentication that could be embedded into the workflow. Specifically, data transfers between the SDCC and NERSC sites were continuously managed by experiment-specific service daemons run at each site, as shown in the workflow chart on Fig. 1. The daemons are designed to manage their local resources such that the daemon on the NERSC side pulls raw data from SDCC for processing at NERSC as long as staging disk and processing capacities are available, while the SDCC daemon pulls the reduced processed data from NERSC and stores the data for use by STAR physicists. These operations are done file-by-file and include handshakes between the remote daemons used to delete each file that is successfully pulled.

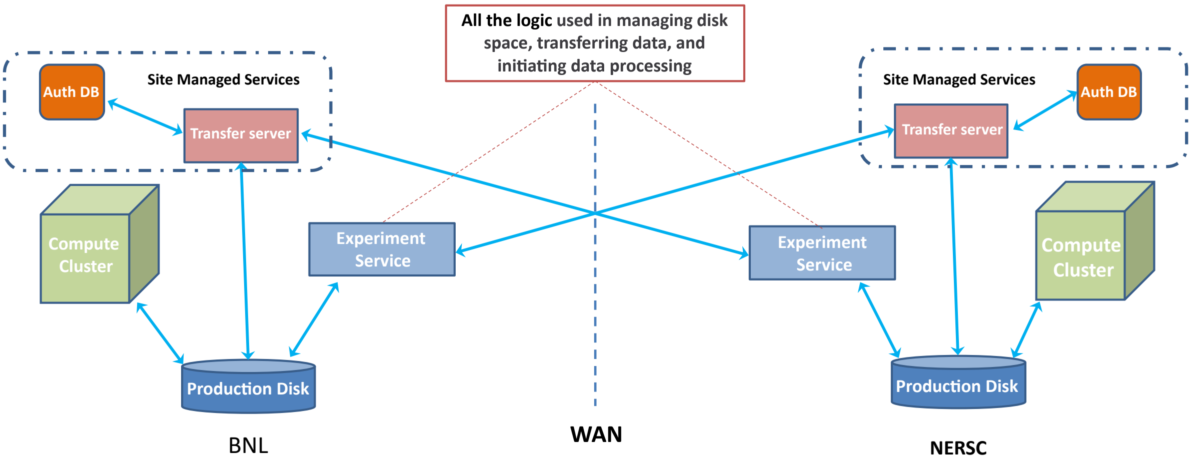

Figure 1. Diagram for data workflow between SDCC and NERSC.

In January 2022, both gridFTP and x509 reached the end of support by the community, therefore STAR is looking to replace the data transfer tool to resume the raw data reconstruction campaigns on Cori HPC at NERSC. A key requirement for such tool is that it can support data transfer speeds sufficient to satisfy STAR reconstruction campaign needs. We estimate the required transfer-rate for the STAR dataset production at NERSC based on our experience for maintaining continuous running of 12k concurrent jobs on NERSC Cori. In that case, we managed 60 jobs per node with each node consuming a 5 GB input DAQ file per 2-hour processing time. This yields a target data transfer speed of about 150 MB/s for our dynamic data management system, which allows for burst production of as high as $\sim$25k concurrent jobs. Another key feature for the data management workflow is the access to the authentication database that can authenticate the transfer between the two sites. The previous system, based on x509 certificates, will need to be rethought after the deprecation of that protocol. The data transfer services require site level privileges and any authentication system must be supported by the sites.

To investigate replacement options for gridFTP, we chose to evaluate XRootD [5] and Globus.org [6]. XRootD is a software framework that works with various kinds of data repositories and has long been used in STAR for its internal data management system, but not as a data transfer tool. Globus.org is a subscription-based service that provides secure data transfers of research data.

3. Performance Measurement

To obtain a baseline for the data transfer rates and verify the feasibility of using each tool, we have measured data transfer rates between SDCC and NERSC using XRootD, Globus.org, GridFTP, and SCP. The number of threads launched by the transfer and the number of parallel transfers were varied when possible. Globus.org endpoints are maintained and configured by NERSC, therefore, we could not vary the parameters (number of threads used by the service, number of data transfer nodes (DTN) used for a transfer, etc).

Figure 2 shows the transfer speed measurements as a function of the number of threads for XRootD and gridFTP. Both transfers show saturation over 4 threads.

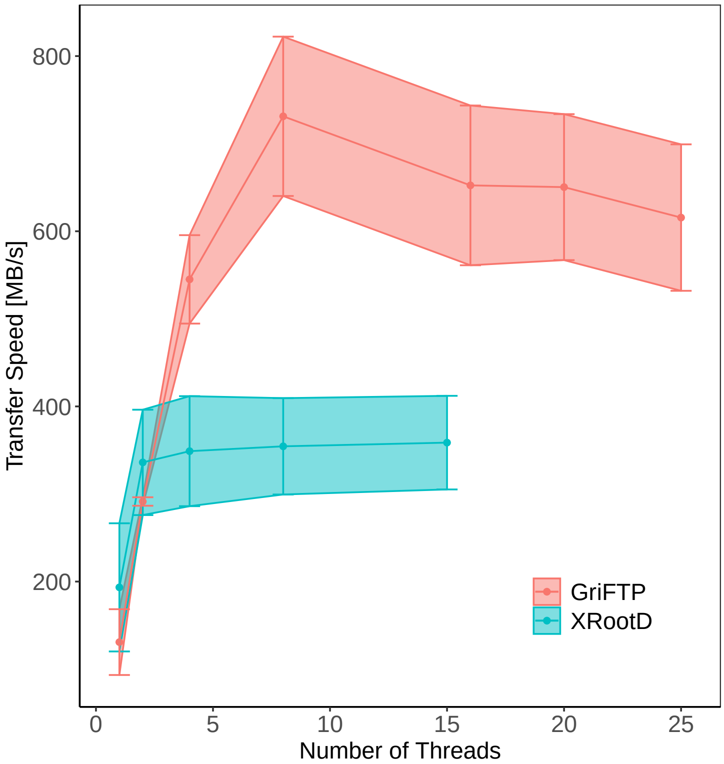

Figure 2. Data transfer rate as a function of number of threads used by the transfer service.

To have a proper comparison with the Globus.org, we measured the transfer rates with an endpoint set up on a single DTN. This DTN was used for all the tests. The thread configuration for this endpoint was set to four threads. The summary comparison for all the tested transfer services for both single and four thread cases is shown in Fig. 3.

In the case of XRootD, the launching of parallel transfer processes enabled us to push the transfer rates and maximize the available connection bandwidth between the two sites. Figure 4 shows the single-process transfer speed and the aggregate total as a function of the number of parallel processes launched. The aggregate total shows a saturation at the level of available

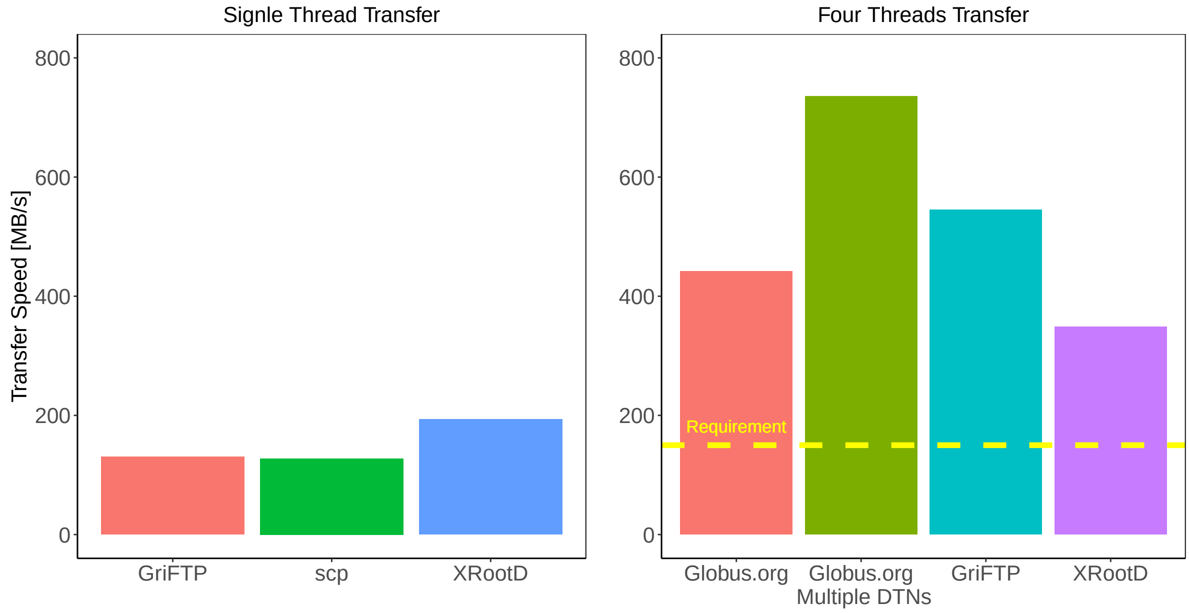

Figure 3. Data transfer rate measurements between SDCC and NERSC.

bandwidth. Similar tests with Globus.org, were multiple parallel transfers were requested from the service, did not yield such conclusive results.

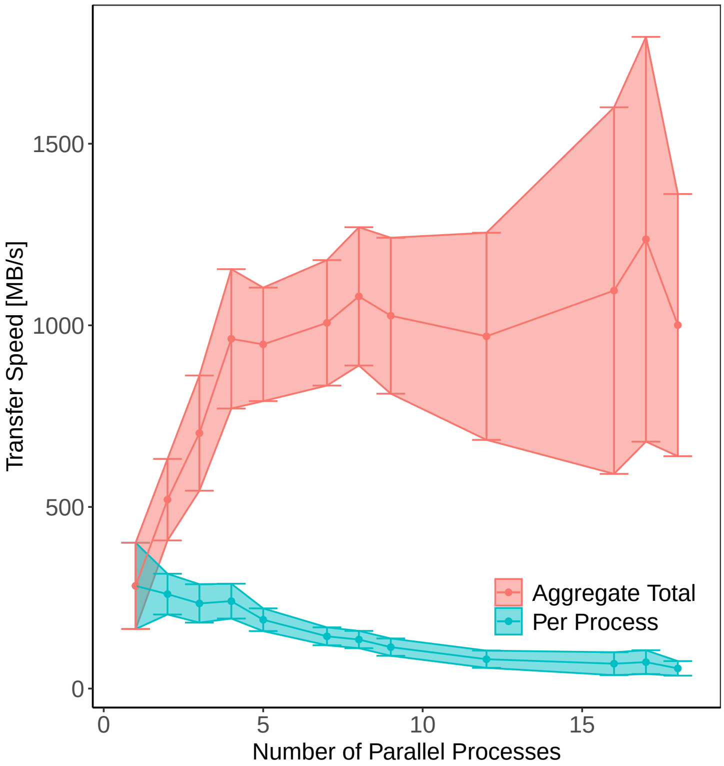

Figure 4. Data transfer speed for parallel XRootD transfers. Each parallel process is run with 15 threads.

The above plots show the comparison of transfer speeds and demonstrate the scalability for XRootD and GridFTP based transfers with the number of threads. We also show that both XRootD and Globus.org meet STAR processing requirements when running multiple threads (4 and over).
Globus.org, while being supported on major sites like NERSC and SDCC, operates under a commercial license and might not be readily available on other laboratory or institutional clusters. On the other hand, XRootD is widely used as a data management tool in the HEP/NP community. We summarize the rate measurements and basic features of the XRootD and Globus.org in Table 1.

Tool	License	CLI	WEB GUI	Single Process <speed> [Mb/s]	Authentication options
XRootD	Free	✓		350	Multiple
Globus.org	Paid	✓	✓	450	Globus Auth

Table 1. Basic features of the XRootD and Globus.org along with the speed measured for this study.

One major aspect of experimental data transfer between computing sites is the authentication component. The community is moving from x509 based token authentication to the WLCG/OSG type token authentication, which XRootD does support. Conversely, Globus.org uses its own token profile that differs from the WLCG/OSG ones.

4. Summary

While STAR continues to successfully use Cori for its embedding simulation needs, we are working on resuming the data reconstruction workflow at NERSC. Such operation requires dynamic and automatic data management. The use of Globus.org and XRootD tools has been considered and their performance studied. Both tools, while having pros and cons in terms of features and ease of availability on large computing sites, have met the criteria for the transfer speed. NERSC and SDCC support and maintain the Globus.org endpoints, however the open-source license of XRootD offers a more versatile and easier to deploy transfer services, allowing us to access a wider pool of computing sites than the National Lab Facilities.

5. Acknowledgements

We thank the RCF at BNL, the NERSC center at LBNL, and the Open Science Grid consortium for providing resources and support. This work was supported in part by the Office of Nuclear Physics within the U.S. DOE Office of Science.

References

[1] *STAR Experiment webpage.* `https://www.star.bnl.gov/`. Accessed: 2022-02-15.

[2] Mustafa Mustafa et al. “STAR Data Reconstruction at NERSC/Cori, an adaptable Docker container approach for HPC”. In: *Journal of Physics: Conference Series* 898 (Oct. 2017), p. 082023. DOI: `10.1088/1742-6596/898/8/082023`. URL: `https://doi.org/10.1088/1742-6596/898/8/082023`.

[3] Poat, M D et al. “Physics Data Production on HPC: Experience to be efficiently running at scale”. In: *EPJ Web Conf.* 245 (2020), p. 09003. DOI: `10.1051/epjconf/202024509003`. URL: `https://doi.org/10.1051/epjconf/202024509003`.

[4] M Poat et al. “STAR Data Production Workflow on HPC: Lessons Learned & Best Practices”. In: *Journal of Physics: Conference Series* 1525 (Apr. 2020), p. 012068. DOI: `10.1088/1742-6596/1525/1/012068`.

[5] *XRootD webpage*. `https://xrootd.slac.stanford.edu/`. Accessed: 2022-02-15.

[6] *globus webpage*. `https://www.globus.org/`. Accessed: 2022-02-15.