Globus: Research Data Management as Service and Platform

Extended Abstract

Kyle Chard
The University of Chicago
5735 S Ellis Ave
Chicago, Illinois, USA 60637
chard@uchicago.edu

Ian Foster
The University of Chicago and
Argonne National Lab
5735 S Ellis Ave
Chicago, Illinois, USA 60637
foster@uchicago.edu

Steven Tuecke
The University of Chicago
5735 S Ellis Ave
Chicago, Illinois, USA
tuecke@globus.org

ABSTRACT

Globus offers a suite of data and user management capabilities to the research community, encompassing data transfer and sharing, user identity and authorization, and data publication. Globus capabilities are accessible via both a web browser and REST APIs. Web access allows Globus to address the needs of research labs through a software-as-a-service model; the newer REST APIs address the needs of developers of research services, who can now use Globus as a platform, outsourcing complex user and data management tasks to Globus cloud-hosted services. Here we review Globus capabilities and outline how it is being applied as a platform for scientific services.

CCS CONCEPTS

•Software and its engineering → Distributed systems organizing principles; •Security and privacy → Security services; •Computer systems organization → Distributed architectures;

KEYWORDS

Globus, research data management, science as a service

ACM Reference format:

Kyle Chard, Ian Foster, and Steven Tuecke. 2017. Globus: Research Data Management as Service and Platform. In *Proceedings of PEARC17, New Orleans, LA, USA, July 09-13, 2017*, 5 pages. DOI: 10.1145/3093338.3093367

1 INTRODUCTION

A significant percentage of research time is spent not on evaluating data, but rather on the acquisition, organization, and wrangling of data. As data sizes increase, and researchers rely on a distributed network of instruments, data storage, and compute resources, these challenges are likely to become prohibitive. In an effort to simplify the management of research data, we developed Globus (originally called Globus Online when released in 2010) [14] as a new model for delivering research data management capabilities as a service.

Globus adopted the software-as-a-service (SaaS) model that has proven to be transformative in industry [13]. The basic model is centered around a professionally hosted service that is always

available and can be easily accessed through a web browser. Thus, users need not download, install, and configure custom software on their PCs; developers of the software need not concern themselves with releasing and supporting different software versions for every possible combination of environment configurations. Data suggest that the SaaS approach can both reduce the cost of software delivery and increase the quality of experience for software consumers. We argue that similar benefits can be achieved in science [15].

Globus initially focused on data transfer [4]—a task performed by almost every researcher. It has since grown to offer a suite of services that include data sharing [10], data publication [9], user management [8], and authentication and authorization [19]. The value of these services is demonstrated by the more than 60,000 users who have collectively transferred more than 250 PB of data.

The success of Globus and other scientific SaaS projects has led other scientific software developers to adopt service-based models for delivering scientific software. However, there is at present little commonality between services. It is therefore difficult for users to move between services and for services to interoperate with one another. In each service developers replicate various components of other service architectures such as data management, user management, and authentication. Inspired by the success of platform-as-a-service models adopted in industry (e.g., services that build upon Amazon Web Services) we have enhanced Globus to provide platform capabilities through powerful APIs and easy to use software libraries that enable other service providers to securely leverage Globus' reliable and available platform services [5].

In this extended abstract we outline current Globus capabilities, focusing specifically on authentication and authorization, data transfer and sharing, and data publication. We highlight how Globus can be used as a platform and conclude with a summary of sustainability and user adoption.

2 AUTHENTICATION AND AUTHORIZATION

One of the most challenging problems in scientific computing is the need to establish a user's identity and manage authorization decisions. Globus, and other scientific services, rely on such fundamental capabilities in every operation they support. Service developers have traditionally implemented their own siloed solutions. This approach is increasingly problematic for several reasons: 1) it is clearly inefficient as every service must develop similar software; 2) users require a different account for every service and there is no support for single sign on; and 3) it is a potential security risk as best practices approaches are not always followed.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

PEARC17, New Orleans, LA, USA

© 2017 Copyright held by the owner/author(s). 978-1-4503-5272-7/17/07...\$15.00
DOI: 10.1145/3093338.3093367

To address these concerns, we developed Globus Auth, an identity and access management platform designed specifically to meet the needs of research services and users. Globus Auth is primarily designed to broker authentication and authorization between identity providers, resource services, and clients. By adopting Globus Auth, service providers can allow their users to authenticate with their service, using one of 407 supported identity providers, and to subsequently provide delegated authorization to access other services. They can integrate these capabilities without needing to develop software to directly manage users, provide peripheral workflow support (e.g., password reset), and keep up with security updates.

To reduce the complexity associated with the proliferation of identities, Globus Auth does not act as an identity provider. Instead, it allows users to authenticate using identities from existing identity providers, such as institutions, Google, and ORCID. To further simplify user experience, Globus Auth allows identities to be linked together such that authentication with one identity implies ownership of the set of linked identities. External services can use this information to grant users the permissions associated with the set of identities rather than the particular identity used to authenticate for a specific session.

Globus Auth implements two standard protocols for integration: OAuth 2 [16] and OpenID Connect [17]. These protocols enable services to request short-term access tokens on behalf of users. Access tokens are granted with respect to specific services with well-defined scopes that permit specific operations. Services are given access tokens following a web-based authentication process. Access tokens can then be used to perform operations on behalf of a user. For example, retrieving profile information about the user or transferring data between two of the user's endpoints. To support more complex workflows, Globus Auth allows delegated tokens to be issued. These tokens can be passed by a service to another service (if approved by the user). The second service can exchange the delegated token for an access token to perform operations.

3 DATA TRANSFER AND SHARING

Globus provides high-performance, secure, file transfer and synchronization between storage systems. As shown in Figure 1, users can coordinate secure transfers via the Globus service. Globus significantly simplifies large scale data transfer by handling all the difficult aspects of data transfer. For example, it coordinates authentication at source and destination, automatically tunes parameters to maximize bandwidth usage, provides automatic fault recovery, and notifies users of completion and problems.

Globus relies on two core components: 1) the Globus cloud-hosted service, which coordinates data transfer; and 2) the Globus Connect software that is deployed on storage systems to enable secure, high performance data transfer.

Globus Connect is distributed in two versions: 1) Globus Connect Personal, a lightweight single-user agent that can be easily installed on Windows, Mac, and Linux PCs; and 2) Globus Connect Server, a multi-user server available as native linux packages. Both Globus Connect Server and Globus Connect Personal include a GridFTP server that enables data transfer between systems using the high performance Grid FTP protocol [3]. Globus Connect Server

Figure 1: Globus reliable file transfer.

provides a number of advanced options that support parallel data transfer node (DTN) deployments and an extensible data storage interface (DSI) which can be used to efficiently access different storage systems. A major focus of our recent work has been the development of DSIs for systems such as Amazon S3, Ceph, HPSS, HDFS, and Google Drive.

Users are able to register their Globus Connect software as a Globus “endpoint.” The endpoint is uniquely identified (with a UUID) and maps the data access interface with a supported authentication service (most often a co-located MyProxy [6] or MyProxy OAuth server that supports PAM authentication). Furthermore, the endpoint includes metadata such as owner, name, and description that can be used to organize and discover endpoints.

3.1 Data sharing

As data sizes increase so too does the complexity involved with sharing data. Researchers desire the “Dropbox” experience, where data can be shared simply and immediately by applying permissions to data. However, once data sizes exceed the limits of Dropbox or data becomes too large to replicate in the cloud, there are few solutions available to researchers. While researchers typically have access to considerable storage resources, it is often difficult to provision accounts for collaborators due to restrictions specified by administrative policies. Globus data sharing is designed to specifically fill this gap, providing the ability for users to dynamically share data with others without needing to copy the data to the cloud and without needing to provision accounts on the machine. The general sharing flow is illustrated in Figure 2.

Globus data sharing builds upon the Globus data access fabric, by adding support for a virtual “shared endpoint.” Shared endpoints are defined upon an existing Globus endpoint and rooted at a particular path within that endpoint. The Globus Connect software provides advanced features that ensure that sharing is enabled by the endpoint owner, that shared endpoints are tied to a specific endpoint, and that users who access shared endpoints are restricted to the permissions granted by the owner.

Figure 2: Globus file sharing.

Shared endpoint owners are able to dynamically apply access control lists (ACLs) with any path within a shared endpoint. ACLs specify a principal (user, group, or public) and read/write permissions on a path. When accessing a shared endpoint, it appears to users exactly like any other Globus endpoint. Importantly, accessing users are unaware of the underlying host endpoint and they are unable to see any details of the root path. User permissions are evaluated on access, allowing changes to be reflected immediately.

3.2 HTTPS

The most recent addition to the Globus suite of data management capabilities is support for secure HTTPS access to endpoints. This support is particularly valuable as it allows users to access data directly through a web browser. It therefore does not require installation of Globus Connect to transfer data. Globus HTTPS supports data sharing, enforcing the same access control permissions as one would encounter through the GridFTP interface.

HTTPS access is supported via extensions to the Globus Connect Server software to include an HTTP server. This packaging of GridFTP and HTTP servers allows Globus Connect to share core components, such as the DSI interface and authentication model. Furthermore, the HTTPS server leverages the enhanced authentication and authorization models described above to seamlessly enforce authorization decisions based on user identity established through the OAuth workflow.

Access to HTTPS endpoints is possible through the Globus web interface via direct mapping of an endpoint to a unique DNS name that includes the endpoint's UUID. Endpoint names are exposed under a common domain, which allows association of an endpoint-specific SSL certificate and also DNS CNAME mapping for users' domains.

4 DATA PUBLICATION

Researchers, funding agencies, and publishers are increasingly concerned with matters of scientific reproducibility. As such, there has been a noticeable increase in the number of public data repositories designed for the sole purpose of enabling researchers to publish their data [1, 2, 11]. Unfortunately, most solutions focus on small scale data—that which can be easily uploaded and downloaded via

HTTP. Of course, this is suitable for some types of data and not for others. To address the gap in supporting large data, we developed a self-service publication service that can support arbitrarily large datasets. This is possible as we separate the software that manages the publication workflow from the software that manages the access and transfer of data.

Globus data publication capabilities build upon the DSpace institutional repository system [18]. The general model is as follows: administrators manage data collections in which other users can then publish data. Collections are defined with a number of policies regarding where data is to be stored (a Globus endpoint), which users can publish data to that collection, what metadata schema is to be used to generate metadata input forms, what curation workflow should be applied, what persistent identifier will be associated with published data (e.g., DOI), and which users are able to access published data.

The publication workflow leverages Globus data management capabilities extensively. Each collection maps to a Globus endpoint with sharing enabled. Globus controls this endpoint to facilitate data publication (data and metadata storage as well as access control and immutability). The general model is shown in Figure 3. When a user starts a publication, Globus will allocate space (a directory) on the endpoint and sets the write ACL for that user. The user may then transfer, potentially large, amounts of data to the directory. When the user completes their publication, Globus transfers a copy of the metadata to the endpoint and changes the ACLs to ensure data is immutable. Depending on the publication workflow, ACLs are applied to allow curators to access the data. After curators approve the publication, an identifier is assigned to the dataset and the ACLs are updated to match the collection's permissions to restrict who may access the published data.

Globus data publication has been used as the basis for creating the Materials Data Facility [7].

5 GLOBUS AS A PLATFORM

The capabilities described above were primarily designed to address the needs of individual researchers directly; however, they are equally valuable to those that wish to develop their own scientific services, many of whom have little desire to build and operate their own authentication, data access, and data transfer infrastructure.

The Globus platform is comprised of REST APIs and a Python Software Development Kit (SDK) that simplifies the process of building upon and integrating Globus into third party services. Over the past several years, projects such as the National Center for Atmospheric Research's Research Data Archive, the Department of Energy's Systems Biology Knowledge Base (Kbase), the FaceBase craniofacial research hub, and the Wellcome Trust Sanger Institute's Imputation service have all built upon Globus APIs to deliver advanced capabilities to their users. While these domains and services are designed to address significantly different use cases, each has common requirements to manage and authenticate users, provide access to data, and transfer data between locations.

One example of how Globus is used as a platform is as a basis for a research data portal. Research data portals aim to support the distribution and sharing of scientific data. While such portals are in use around the world, the architectures employed are primitive.

Figure 3: Globus data publication.

They are typically built on a simple web server that serves not only the web interface for discovering data but also the data itself. This is particularly inefficient as it is not well-suited to supporting access to large amounts of data. Building upon Globus we have observed a new development pattern for what we term a “modern research data portal.” In this case, portal developers separate the management and discovery service from the data access infrastructure. This separation enables investments in high performance data transfer infrastructure (the Science DMZ [12] and DTNs) to be efficiently utilized to serve data. The entire portal then leverages Globus Auth for user authentication and managing access to the data. The central data portal server can then be developed simply, leveraging Globus APIs for much of its core operations, and focusing on the specific interfaces required to address user needs. Given it is not required to serve data, it may be hosted on relatively modest infrastructure.

6 SUSTAINABILITY

Perhaps the greatest challenge faced in the scientific software community is that of sustainability: ensuring that software can be supported beyond the lifetime of grant funding. Over the last several decades we have explored several approaches for sustaining large-scale software projects: most notably the open source models used to support the Globus Toolkit and our recent subscription-based models for supporting the Globus service.

Operating high quality services for the community brings with it new challenges. No longer is sustainability concerned only with code in a repository. Instead, we must support the people and infrastructure who deliver the service. To support the Globus service, we have adopted a freemium model in which core Globus capabilities (e.g., data transfer) are completely free to the non-profit science community. Advanced capabilities (e.g., endpoint management and data publication) as well as enhanced support are available to those with subscriptions. Globus subscriptions are available to groups such as universities, national laboratories, research projects, and

companies. Each subscription is managed by a group of administrators who may then select which users (e.g., members of an institution) receive access to advanced Globus capabilities.

Our initial experiences with this subscription model are encouraging. Over the past two years, more than 60 research institutions, national laboratories, research projects, national cyberinfrastructure providers, and companies have subscribed to the Globus cloud service. As we describe below, we continue to see increasing usage, both in the free tier and for those with subscriptions.

7 GLOBUS BY THE NUMBERS

Globus is now relied upon by a large segment of the research community for a variety of purposes. We briefly review adoption and share high level usage statistics as of March, 2017.

Globus Auth manages 64,000 unique identities of which 33,352 have been registered in the last year. These identities are provided from 300 identity providers. Globus Auth supports 300 applications and services, both internal and external, and has issued almost 4.4 million access tokens.

Globus Transfer has been used to move more than 250 PB involving more than 41 billion files in the time since deployment in November, 2010. 65,861 Globus endpoints have been defined and used to transfer data, of which 59,718 are currently active (not deleted). These active endpoints comprise 47,195 Globus Connect Personal endpoints and 12,522 Globus Connect Server endpoints. In the past year 16,948 endpoints have participated in a transfer.

Globus users have created 24,578 shared endpoints, of which 4,463 are currently active and 5,510 have participated in a transfer in the past year. These endpoints have been shared by 1,167 unique users and have participated in 44,9534 transfers by 5,185 users. Cumulatively data transfers involving shared endpoints exceed 12 PB of data. While Globus sharing is primarily used for secure data sharing, it can also be used to share data publicly. More than 4.5 million files and 30 TB have been shared in the latter model. More than 750 unique users have transferred data to/from publicly shared endpoints.

Globus data publication has been accessed by 999 unique users across its 27 user-created collections. It hosts 184 unique publications that cumulatively exceed 6 TB of data. In total it manages 8,388 individual metadata items.

8 SUMMARY

Globus provides a set of core service and platform capabilities to the research community. These capabilities address important needs of both researchers and developers of scientific services, such as authentication and authorization, data transfer and sharing, and data publication. Over the past six years, we have seen rapid adoption of Globus both from end users and developers. We continue to work to expand the suite of data management capabilities provided, for example by adding support for flexible data search and by introducing features to enable data management workflows to be more easily expressed. We are also focused on developing an ecosystem of libraries, tools, and documentation that will allow others to more easily leverage these capabilities.

ACKNOWLEDGMENTS

We thank Globus subscribers for supporting the operation and development of Globus, and users of Globus services for their continued support. This research was supported in part by NSF grant ACI-1148484 (SciDaaS) and US Department of Energy contract DE-AC02-06CH11357.

REFERENCES

- [1] 2017. figshare. (2017). <http://figshare.com>.
- [2] 2017. Zenodo. (2017). <https://www.zenodo.org>.
- [3] William Allcock, John Bresnahan, Rajkumar Kettimuthu, Michael Link, Catalin Dumitrescu, Ioan Raicu, and Ian Foster. 2005. The Globus Striped GridFTP Framework and Server. In *ACM/IEEE Conference on Supercomputing (SC '05)*. IEEE Computer Society, Washington, DC, USA, 54–. DOI: <http://dx.doi.org/10.1109/SC.2005.72>
- [4] Bryce Allen, John Bresnahan, Lisa Childers, Ian Foster, Gopi Kandaswamy, Raj Kettimuthu, Jack Kordas, Mike Link, Stuart Martin, Karl Pickett, and Steven Tuecke. 2012. Software As a Service for Data Scientists. *Commun. ACM* 55, 2 (Feb. 2012), 81–88. DOI: <http://dx.doi.org/10.1145/2076450.2076468>
- [5] Rachana Ananthakrishnan, Kyle Chard, Ian Foster, and Steven Tuecke. 2014. Globus Platform-as-a-Service for Collaborative Science Applications. *Concurrency - Practice and Experience* 27 (2014), 290–305. Issue 2. DOI: <http://dx.doi.org/10.1002/cpe.3262>
- [6] Jim Basney, Marty Humphrey, and Von Welch. 2005. The MyProxy Online Credential Repository. *Software: Practice and Experience* 35, 9 (2005), 801–816. DOI: <http://dx.doi.org/10.1002/spe.688>
- [7] B. Blaiszik, K. Chard, J. Pruyne, R. Ananthakrishnan, S. Tuecke, and I. Foster. 2016. The Materials Data Facility: Data Services to Advance Materials Science Research. *Journal of the Minerals, Metals & Materials Society (JOM)* 68, 8 (2016), 2045–2052. DOI: <http://dx.doi.org/10.1007/s11837-016-2001-3>
- [8] Kyle Chard, Mattias Lidman, Brendan McCollam, Josh Bryan, Rachana Ananthakrishnan, Steven Tuecke, and Ian Foster. 2016. Globus Nexus: A Platform-as-a-Service provider of research identity, profile, and group management. *Future Generation Computer Systems* 56 (2016), 571–583. DOI: <http://dx.doi.org/10.1016/j.future.2015.09.006>
- [9] K. Chard, J. Pruyne, B. Blaiszik, R. Ananthakrishnan, S. Tuecke, and I. Foster. 2015. Globus Data Publication as a Service: Lowering Barriers to Reproducible Science. In *11th IEEE International Conference on e-Science*. 401–410. DOI: <http://dx.doi.org/10.1109/eScience.2015.68>
- [10] K. Chard, S. Tuecke, and I. Foster. 2014. Efficient and Secure Transfer, Synchronization, and Sharing of Big Data. *IEEE Cloud Computing* 1, 3 (Sept 2014), 46–55. DOI: <http://dx.doi.org/10.1109/MCC.2014.52>
- [11] Merce Crosas. 2011. The Dataverse Network: An Open-Source Application for Sharing, Discovering and Preserving Data. *D-Lib Magazine* 17, 1/2 (2011).
- [12] Eli Dart, Lauren Rotman, Brian Tierney, Mary Hester, and Jason Zurawski. 2013. The Science DMZ: A Network Design Pattern for Data-intensive Science. In *International Conference on High Performance Computing, Networking, Storage and Analysis (SC '13)*. ACM, New York, NY, USA, Article 85, 10 pages. DOI: <http://dx.doi.org/10.1145/2503210.2503245>
- [13] Abhijit Dubey and Dilip Wagle. 2007. Delivering software as a service. *The McKinsey Quarterly* 6 (2007).
- [14] I. Foster. 2011. Globus Online: Accelerating and Democratizing Science through Cloud-Based Services. *Internet Computing, IEEE* 15, 3 (May 2011), 70–73. DOI: <http://dx.doi.org/10.1109/MIC.2011.64>
- [15] Ian Foster, Vas Vasiliadis, and Steven Tuecke. 2013. *Software as a service as a path to software sustainability*. Technical Report. <http://dx.doi.org/10.6084/m9.figshare.791604>.
- [16] D. Hardt. 2012. The OAuth 2.0 Authorization Framework. <http://www.rfc-editor.org/info/rfc6749>. (October 2012).
- [17] N. Sakimura, J. Bradley, M. Jones, B. de Medeiros, and C. Mortimore. 2014. OpenID Connect Core 1.0. <http://openid.net/specs/openid-connect-core-1.0.html>. (Nov. 2014).
- [18] MacKenzie Smith, Mary Barton, Mick Bass, Margret Branschovsky, Greg McClellan, Dave Stuve, Robert Tansley, and Julie Harford Walker. 2003. DSpace: An open source dynamic digital repository. *D-Lib Magazine* 9, 1 (2003).
- [19] S. Tuecke, R. Ananthakrishnan, K. Chard, M. Lidman, B. McCollam, S. Rosen, and I. Foster. 2016. Globus Auth: A research identity and access management platform. In *12th IEEE International Conference on e-Science (e-Science)*. 203–212. DOI: <http://dx.doi.org/10.1109/eScience.2016.7870901>