Globus: Recent Enhancements and Future Plans

Kyle Chard
Computation Institute
University of Chicago &
Argonne National Laboratory
chard@uchicago.edu

Steven Tuecke
Computation Institute
University of Chicago &
Argonne National Laboratory
tuecke@globus.org

Ian Foster
Computation Institute
University of Chicago &
Argonne National Laboratory
foster@anl.gov

ABSTRACT

Globus offers a broad suite of research data management capabilities to the research community as web-accessible services. The initial service, launched in 2010, focused on reliable, high-performance, secure data transfer; since that time, Globus capabilities have been progressively enhanced in response to user demand. In 2015, secure data sharing and publication services were introduced. Other recent enhancements include support for secure HTTP data access, new storage system types (e.g., Amazon S3, HDFS, Ceph), endpoint search, and administrator management. A powerful new authentication and authorization platform service, Globus Auth, addresses identity, credential, and delegation management needs encountered in research environments. New REST APIs allow external and third-party services to leverage Globus data management, authentication, and authorization capabilities as a platform, for example when building research data portals. We describe these and other recent enhancements to Globus, review adoption trends (to date, 40,000 registered users have operated on more than 150PB and 25B files), and present future plans.

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

1. INTRODUCTION

Researchers are increasingly overwhelmed by the volume, velocity, and variety of scientific data. It was this realization that first motivated the development of Globus [15] in 2010. Our goal was then, and continues to be today, to leverage the benefits of the increasingly pervasive Software-as-a-Service (SaaS) paradigm to produce better and cheaper solutions to research data management problems. SaaS has been shown in industry to reduce barriers to the use of advanced software, due to easy-to-use web interfaces and reduced costs for both consumer and provider. We believe (and assert that we have demonstrated) that the SaaS paradigm can deliver similar benefits in research data management.

Globus now provides a suite of research data management capabilities. It was first developed to provide high performance, secure, and reliable data transfer—a need for many, if not all, researchers. As of mid 2016, Globus has been used to move more than 150 PB of data in 25 billion files, is regularly used by more than 2,300 unique people per month, and supports more than 40,000 endpoints—storage systems accessible via Globus. Given the increasing reliance on Globus we have made a number of improvements to the transfer capabilities while also adding new capabilities to address other research data management challenges. Examples of new capabilities include the ability to securely share data in place as well as to publish and discover data using persistent identifiers and flexible metadata.

Here we provide an update on some of these new Globus capabilities, focusing in particular on those listed in Table 1. The remainder of the paper is structured as follows. In Sections 2 through 5 we describe enhancements to Globus transfer capabilities, secure in-place data sharing, data publication, and authentication and authorization, respectively. In Section 6 we explore user adoption of Globus. In Section 8 we describe forthcoming R&D plans. Finally, we present related work in Section 7 and summarize in Section 9.

2. RAPID, RELIABLE TRANSFER

Movement of data is at the core of many scientific activities, including analysis, collaboration, publication, and archival. However, given its importance and ubiquity, this task remains surprisingly challenging in practice: storage locations have different security configurations, achieving good transfer performance is non-trivial, and as data sizes increase the likelihood of errors increases. Globus transfer [6] simplifies the process of moving data between pairs of storage systems or *endpoints*. This cloud-hosted service handles the complexity involved in transfers, such as authenticating and authorizing user access to endpoints, creating a high-speed data connection between endpoints, and recovering from faults while a transfer proceeds. Importantly, Globus implements a third-party transfer model in which no data is transferred via the Globus service: instead, all data is transferred directly between the two participating endpoints.

XSEDE '16 Miami, Florida USA
© 2016 ACM. ISBN 978-1-4503-4755-6/16/07…$15.00
DOI: http://dx.doi.org/10.1145/2949550.2949554

Table 1: Recent Globus transfer service enhancements, new services, and platform capabilities.

Service	Enhancement	Description
Transfer	Management console	Support for monitoring and managing endpoints
	Pause rules	Ability for administrators to pause transfers
	Endpoint discovery	Extended endpoint metadata and rich endpoint search
	Amazon S3 endpoints	Support for Amazon Simple Storage Service (S3) endpoints
	Ceph S3 endpoints	Support for Ceph S3 endpoints
	HDFS endpoints	Support for Hadoop Distributed File System (HDFS) endpoints
	HPSS endpoints	Support for High Performance Storage Service (HPSS) endpoints
	BlackPearl endpoints	Support for Spectra Logic BlackPearl object storage archive endpoints
	Google Drive	Support for Google Drive endpoints (scheduled for later in 2016)
	HTTPS protocol	Authenticated and authorized HTTPS access to Globus endpoints
Sharing (New)	Shared endpoints	Support for creating shared endpoints in Globus Connect
	Secure access management	Management and enforcement of user and group access permissions
	Sharing workflows	Email and web-based sharing workflows
Publication (New)	Publication service	Support for publication submission, curation, and access workflows
	Distributed storage	Management of distributed storage via Globus sharing
	Metadata management	Schema definition and input form generation
	Persistent identifiers	Support for Handle and DOI identifiers
Auth (New)	External identity	Authentication via external identities (i.e., no need for a Globus identity)
	Broad IDP support	Support for various identity providers (e.g., CILogon, Google, OAuth2)
	Identity set	Federated accounts with linked identities
	OAuth, OpenID Connect	Standards-based secure identity and access management
Platform (New)	Secure REST APIs	REST APIs for programmatic invocation and integration in external services
	Delegated Auth	OAuth 2 APIs for delegated and consented use
	Python SDK	Software Development Kit (SDK) for integration in third-party applications

Globus, like many hosted services, comprises two core components: the hosted service and agent software. The agent software (called Globus Connect) implements mechanisms for authentication and data access. Globus Connect is available in two versions: Server and Personal. Globus Connect Server is a Linux package that is designed to be deployed on storage servers. Globus Connect Personal is a lightweight single user agent that can be deployed on Windows, Mac OS, and Linux computers.

As the number of Globus users and endpoints grows we have responded by developing new capabilities that support increasingly sophisticated data transfer scenarios. Two such improvements are the administrator management console for managing transfers involving specific endpoints and a new endpoint discovery model. We have continued to expand support for different storage systems such as Amazon S3, OpenStack Ceph, and HDSS. Finally, to provide broad access to Globus-accessible data we have developed a secure Globus HTTPS server via which data residing on Globus endpoints can be accessed using HTTPS.

2.1 Transfer Management Console

Globus endpoints are deployed on many of the world's largest compute and storage systems, such as those in XSEDE [24]. They are also deployed at many institution research computing centers. Some Globus endpoints regularly manage transfers totaling hundreds of terabytes per day. Large-scale deployments often include dozens of parallel data access nodes behind a single Globus endpoint. Given the mission-critical nature of these endpoints combined with the massive usage, administrators have long desired advanced methods for managing their endpoints. They have also requested such capabilities be offered via intuitive interfaces rather than require low-level server configurations. To meet these needs we have developed a web-based management console that allows administrators to oversee and mange their endpoints.

The Globus management console supports real-time monitoring across endpoints, providing the ability to identify and troubleshoot faults and performance degradation, and drill down into individual transfers to interrogate performance. To ease the task of finding individual tasks within many thousands of active tasks the management console provides a search-based interface. This interface supports querying across endpoints, users, and transfers. The management console also supports fine-grain management of individual transfer tasks. For example, it allows administrators to pause, resume, or cancel a single (or all) transfers that involve one of their managed endpoints. Thus, administrators can identify and then pause or cancel transfers that are experiencing difficulties or that are over-consuming resources. It also allows all transfers to an endpoint to be paused for the purposes of maintenance, troubleshooting, or upgrade.

Administration capabilities have been developed entirely in the managed Globus service. As such, they can be applied to any Globus endpoint and require no modifications to Globus Connect. The Globus service has been extended to enable pause rules to be defined, stored, and enforced. A new REST API and user interface have been developed to enable monitoring and management of endpoints.

2.2 Endpoint Discovery

The number of Globus-accessible endpoints has ballooned to over 40,000, rendering the initial namespace-based naming system and client-side search mechanisms inadequate. To address this need, we implemented new mechanisms that allow users to discover endpoints based on richer criteria than simply the endpoint name. As the basis for this model we extended the set of metadata that can be associated with an endpoint, allowing for attributes such as descriptions, owner, organization, and keywords. We developed a new free-text search index (using PostgreSQL search) to index endpoint metadata and provide rich search capabilities to users. Globus users can now quickly search across endpoints, filtering based on partial text matches. They can perform scope-based queries (e.g., "endpoints owned by me" or "endpoints shared with me") and access a list of recently used endpoints. We also added endpoint bookmarking support, so that users can save a particular endpoint, and path within that endpoint, to a list of bookmarks. These bookmarks can then be used to quickly access commonly used endpoints.

2.3 Storage System Support

Increasingly, high performance computing and research computing centers are investing in alternative storage architectures, such as object stores and archival storage systems (e.g., tape). To address the changing storage landscape we are integrating new storage systems.

Object stores, such as Amazon Simple Storage Service (S3), are one class of storage that is becoming increasingly prevalent. To support S3 in Globus we have developed a model that allows Globus endpoints to be created on S3 buckets. Our model relies on integration with Amazon's Identity and Access Management (IAM) service to facilitate delegated access using Globus-managed identities. We extended the Globus Connect software to support direct connections to S3 buckets, utilizing secure HTTPS upload and download. Over the past year we have implemented a number of performance enhancements to this model, such as multi-part upload for large files and the use of parallel TCP data streams. Evaluation of Globus S3 file transfer has shown throughput peaks of over 5Gbps.

Globus Connect Server, more specifically the GridFTP server, offers a modular Data Storage Interface (DSI) for interfacing with arbitrary storage systems. Implementation of a DSI requires implementation of a set of interface functions that are used to access the specific storage system. In collaboration with researchers at the National Center for SuperComputing Applications we have developed a DSI that supports the High Performance Storage System (HPSS). This DSI enables high performance access to large-scale tape archives. We have also developed DSIs to support the Spectra Logic BlackPearl Object storage appliance, the Hadoop Distributed File System (HDFS), and OpenStack Ceph via the Object Gateway S3 API. We are currently developing support for Google Drive and expect to release this capability in the Fall of 2016. We are also developing an S3 DSI to support advanced transfer features and direct transfers between S3 endpoints.

2.4 HTTPS access to Globus Connect data

The most frequently requested feature by Globus users has been the ability to download Globus-accessible data via a web browser. In response, we have extended the Globus Connect Server software to incorporate a secure HTTPS server. This new capability leverages the new authentication and authorization support described in Section 5 to provide standard HTTPS data access (e.g., as used in a browser) while enforcing the same Globus authorization model. Thus, a single storage system can store both smaller files designed for web browser access and bigger files for which GridFTP transfer is preferred; users and applications can choose to deliver data via either HTTPS or GridFTP, depending on size, performance requirements, and context.

The core component required to support HTTPS access is a new Globus Connect HTTPS server. This server is deployed alongside the Globus Connect GridFTP server and shares the same authorization and control flows. It also shares the same underlying data access interface, which supports arbitrary DSIs. Thus, when a request is submitted using HTTPS, Globus Connect first ensures that the user is authenticated and authorized to access data. (It uses the Globus Auth system to direct the user to authenticate via Globus before checking the user's permissions to access the data.) If the request is approved, Globus Connect will serve the file from the underlying filesystem, using any of the supported DSIs.

Access to an HTTPS endpoint requires a unique Domain Name System (DNS) name. To construct the DNS name we leverage the UUID that uniquely identifies each endpoint. This UUID is then exposed under a common domain, <ep-uuid>.glob.us. This construct has two important benefits: it allows each endpoint to have its own SSL certificate, and it allows users to associate their own DNS name with their endpoints using a DNS CNAME.

We are currently upgrading the Globus web application to support HTTPS endpoints. These enhancements will recognize when an endpoint supports HTTPS and customize the options presented to users. The potential implications of this work are immense. Users will soon be able to visualize files within the Globus web application using inline viewers, directly download files via the Globus web application or their own web application without the need for a Globus endpoint, and develop new third-party applications that interface directly with Globus-accessible data.

Globus HTTPS access is currently in a pilot deployment phase and will soon be released as part of the Globus Connect Server package.

3. SECURE DATA SHARING

Data sharing underpins collaboration, yet sharing data efficiently becomes increasingly challenging as data sizes increase. While a multitude of data sharing models exist, all have inherent limitations. For example, cloud-hosted storage providers (e.g., Dropbox) charge users beyond a free tier and require that data is replicated to the cloud in order to be shared. In many cases, researchers are unwilling to pay for cloud storage as they have significant storage allocations within their institution or at national cyberinfrastructure providers; however, it is often impossible (or at least difficult) to share data publicly from such storage infrastructure and administrative policies often restrict the creation of accounts for collaborators.

Globus data sharing [13] aims to simplify the process of sharing, even large amounts, of data without copying it to a cloud service (i.e., it facilitates in-place data sharing). The Globus data sharing model is based on virtual "shared endpoints" rooted at a specific path within an existing endpoint (called a "host endpoint"). A shared endpoint looks exactly like a host endpoint to those that can access it and recipients can transfer data to/from the endpoint using the same Globus transfer interfaces. Sharing is supported on both Globus Connect Server and Personal endpoints, although it must be activated in the respective configurations.

Shared endpoints provide their owners with a set of enhanced management options. For example, owners may select paths within an endpoint and associate access control lists (ACLs) with these paths. ACLs define read and/or write permissions for a specific path and with respect to a specific user or group, or public access. When sharing data, the owner can use an email-based sharing workflow in which the recipient is notified via email with a unique link. The unique link allows the recipient to claim access to the shared endpoint by authenticating using their Globus-supported identity. ACLs can be changed and modified at any time by the owner. Owners can also assign management roles on their shared endpoints. For example, the "Access Manager" role delegates the ability to assign and manage ACLs on the endpoint.

Given the potential security vulnerabilities of data sharing we have developed a secure multi-layer security implementation. Sharing is built upon extensions to GridFTP that control how shared endpoints are created and used. When accessing shared endpoints both the owner's local permissions and accessor's shared permissions are evaluated. The GridFTP server has been extended to securely pair a shared endpoint to its host endpoint thus negating attack vectors that aim to move a shared endpoint's underlying host endpoint. Finally, the Globus service has been extended to support the secure storage of shared endpoint definitions and their associated ACLs.

The Globus data sharing model also has a number of user-focused security configurations. For example, endpoint owners must enable sharing on their endpoints before data can be shared. Owners have complete control over what paths may (or may not) be shared and what permissions can be assigned to those paths (read or write). When shared endpoints are created they are rooted at the selected path, thus, path information is hidden from other users (e.g., "/hidden/" is seen by others as "/") Finally, ACLs are assessed in real-time when users access the shared endpoint this enables ACLs to be revoked with immediate effect.

4. DATA PUBLICATION

There are growing societal, institutional, and funding agency pressures to publish research data. While there are an increasing number of data publication systems (e.g., Dataverse [2], Zenodo [5], and figshare [3]) they are not without limitations. For example, they are often designed for a particular domain or type of data and most—if not all—are unable to deal with large datasets.

Globus data publication [12] provides a self-service model via which researchers can create and manage *collections* of data publications. Collections are defined by a number of policies that specify: data storage location, accessible via Globus, where all data published in the collection will be stored; metadata schemas and input forms that will be used to describe data; persistent identifier to be assigned to data published in the collection (e.g., DOI or Handle); workflows for submission and optional curation; and permissions for submitting and accessing data in the collection.

Globus data publication is built on the DSpace institutional repository [23] and the Globus data management fabric. DSpace provides the user interfaces and workflows for publication, curation, and administration. The Globus data fabric provides the mechanisms for associating remote data storage with a collection, assembling a submission from multiple locations (using Globus transfer), and managing access to both unpublished (in curation) and published data. Globus data publication uses Globus groups to implement access control (e.g., to control submission or access to a collection) and roles (e.g., administrator and curator). The Globus data publication service extends DSpace in several other ways, most notably to support operation as a multi-tenant hosted service. In addition, it supports user-oriented creation and management of collections, and per-collection configuration of metadata schemas, input forms, workflows, and persistent identifier providers.

We have identified the materials science community as a first target community in need of large-scale publication capabilities. To address this need we have deployed the Materials Data Facility (MDF) [10], which offers domain-specific schemas and persistent identifier providers.

We are in the process of releasing our extensions to DSpace under an open source license, so that other organizations that use DSpace can leverage Globus in their instances.

5. GLOBUS AS A PLATFORM

Over the course of the last five years it has become increasingly apparent that Globus can provide enormous value as a platform [7]. Already, Globus data management and identity and group management [11] capabilities are used by a number of services including the National Center for Atmospheric Research (NCAR) Research Data Archive (RDA), the University of Exeter, the Department of Energy Systems Biology Knowledge Base (KBase), and the National Institutes of Health FaceBase. Based on our experiences leveraging Globus as a platform we note that developers of new research services face two major infrastructure challenges: first, providing sophisticated identity and access management (IAM) functionality; and second, integrating with multiple other services, that have been developed by independent parties. To address these requirements we have developed Globus Auth: a flexible identity and access management service that can be leveraged by external applications.

5.1 Globus Auth

Globus Auth is a foundational identity and access management platform service designed to address unique needs of the science and engineering community. It serves to broker authentication and authorization interactions between end-users, identity providers, resource servers (services), and clients (including web, mobile, and desktop applications, and other services). Globus Auth thus makes it easy, for example, for a researcher to authenticate with one credential, connect to a specific remote storage resource with another identity, and share data with colleagues using their institution identity. By eliminating friction associated with the frequent need for multiple accounts, identities, credentials, and groups when using distributed cyberinfrastructure, Globus Auth streamlines the creation, integration, and use of advanced research services.

Globus Auth implements the OAuth 2 [17] and OpenID Connect specifications [22]. In its standard operating model Globus Auth issues short-term access tokens to a client (a third party service or application). Access tokens are granted after a user successfully authenticates with a supported identity provider and obtains an authorization (consent) for the client to access a resource server on behalf of the user. The client may subsequently make requests to the resource server by presenting the access token as part of a request. Globus

Auth can act in multiple roles. First, it may act as the authorization server to an extensible set of resource servers. It is this model that is used by Globus transfer and publication, with both services acting as resource servers. In some cases Globus Auth also acts as a resource server allowing clients to access Globus Auth-managed resources, such as identities and access tokens.

Globus Auth supports more than 70 different identity providers, from InCommon institutions to Google. It represents each identity issued by an identity provider by a unique case-insensitive username (e.g., user@example.com). A user demonstrates possession of an identity via an authentication process directly with the identity provider. Globus Auth neither defines its own identity usernames nor verifies identity authentication (e.g., via passwords). Rather, it acts as an intermediary between external identity providers and clients and services that want to leverage identities issued by those providers.

One unique contribution of the Globus Auth model is the ability to link different identities. Linked identities may be used to support login with one identity by using another linked identity. To support a "federated identity" model, Globus Auth creates a transparent *account*. A Globus account has a single primary identity and a set of linked identities. An identity can be the primary identity of at most one Globus account. However, one identity may be linked to any number of other primary identities, and thus Globus accounts. Identity linking allows for authentication via one identity to imply login to a Globus account with a different primary identity (i.e., federated identity login).

The Globus Auth service is implemented as a stand-alone Python web application. It is hosted on Amazon Elastic Compute Cloud (EC2) and leverages the Amazon Relational Database Service (RDS) for storing state. Globus Auth exposes interfaces for managing tokens, consents, identities, identity providers, and clients. It also implements OAuth 2 and OpenID Connect interfaces for authentication and obtaining and using delegated tokens.

5.2 Globus Platform APIs

All Globus services offer REST APIs via which their capabilities can be used programmatically. Collectively, the Globus service APIs offer a flexible platform that can be used by developers to outsource functionality such as identity management, data transfer and sharing, and group management. Outsourcing this functionality to a reliable and highly available platform has several benefits: developers need not reimplement functionality themselves, they can ensure that best practices (e.g., for security) are followed, technical debt and support burden are reduced, users can employ the same identity and other state (e.g., groups) across services, and users and developers alike benefit from high availability services. As in other domains (e.g., mobile phone applications), the Globus platform reduces costs, improves quality, and promotes usability, extensibility, and interoperability.

In an effort to support usage of Globus as a platform we have developed new service APIs, created detailed reference documentation, released a new Python SDK, and created interactive training materials (e.g., workshops and Jupyter notebooks). Globus APIs are categorized by their level of maturity and support. Currently the Globus Auth, transfer, and sharing APIs are publicly available with full documentation, SDK support, and training materials. Globus groups and publication APIs are currently private and are only used by internal Globus services and pilot external users. In the near future we will will release these APIs publicly.

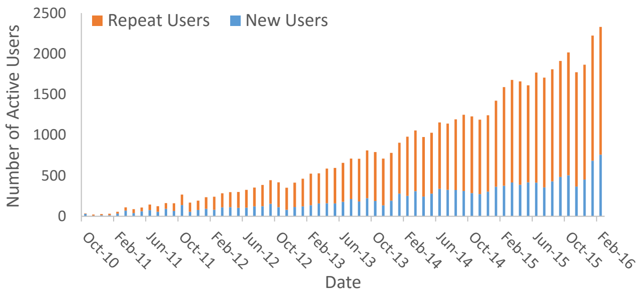

Figure 1: Number of registered Globus users who have performed at least one transfer in the given month. New users are those for which this is their first transfer.

6. ADOPTION

Globus now supports more than 40,000 registered users, with more than 2,300 active each month. Both the number of registered users and usage of Globus continues to grow, as we explore in this section.

6.1 Users

Figure 1 shows the number of "active users"—users who have performed at least one transfer in a given month. The graph also shows the number of new and repeat users, where new users are those that have not before transferred data with Globus. It is worth noting the increase for both categories of users. Of course, this graph shows only one area in which a user might be active; however, many users use their Globus account for other purposes, such as in third party applications that use Globus identities.

6.2 Transfer and Sharing

Globus has more than 40,000 registered endpoints, of which more than 20,000 have been used to conduct at least one transfer. Curiously, approximately 20,000 registered endpoints have not used to transfer data. We are exploring why this is the case. Figure 2 shows the number of endpoints that have been used to conduct at least one transfer in a given month, categorized by type: Globus Connect Personal, Globus Connect Server, or Shared. We see that the total number of such "active" endpoints is increasingly steadily, now exceeding 3,000 per month. As one would expect, the number of Globus Connect Personal endpoints far exceeds the number of Globus Connect Server endpoints. The number of shared endpoints is increasing less rapidly, perhaps because sharing is a relatively new feature and is only accessible to institutions with a Globus subscription.

Figure 3 shows another perspective of Globus endpoint adoption, that is deployment of Globus Connect Server endpoints across the globe. This map shows registered endpoints grouped by the number of endpoints in a single location. To create this map, we performed a DNS lookup of the server addresses of each Globus Connect Server installation. We then used a common geolocation database to resolve IP addresses to their locations (latitude and longitude).

Figure 4 shows the total amount of data and the number

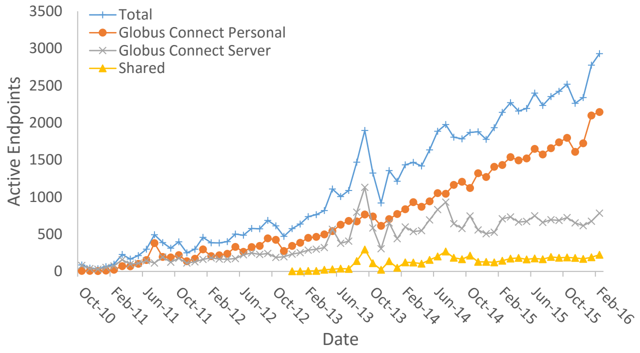

Figure 2: Number of active endpoints by type. Active endpoints are those that have participated in a file transfer within the month.

of unique users who have performed a transfer involving one of the more than 3,500 Globus shared endpoints. Given that sharing has only been publicly available to users for approximately two years, and the fact that sharing is enabled only for users with a Globus subscription, these results are encouraging. We see a linear increase in the number of shared endpoint users per month with more than 400 unique users in March, 2016. Similarly, the amount of data transferred to/from a shared endpoint is increasing steadily, highlighted by the more than 1.4 PB moved in February, 2016.

6.3 Publication

Globus data publication was first made available for pilot usage in mid-2015 and was released publicly in early 2016. Like data sharing, data publication is available only to users with a Globus subscription; thus, adoption is likely to be slower than that of other Globus services. To allow users to evaluate Globus data publication we operate two versions of the publication service: the first serves as an open trial environment in which any user can test and evaluate capabilities, while the second is a production service for those who are using the service to publish real data. Over the last nine months since initial pilot deployment 554 (411 production, 143 trial) users have registered, and 16 (4 production, 12 trial) communities and 40 (19 production, 21 trial) collections have been created. In total, 428 (33 production, 395 trial) datasets have been published. The Materials Data Facility has been used to publish more than 7 TB of data in the two months since its public release.

7. RELATED WORK

Globus was one of the first services developed specifically for the scientific community. Since this time, others have developed and deployed various scientific services, notable examples include Agave [14] and Apache Airavata [20]. Both Agave and Apache Airavata provide a range of platform capabilities that can be leveraged by other services for authentication and authorization, user and resource management, job submission, data management, and workflow development and execution.

There are a plethora of file transfer tools available. However, most do not support third party transfers (e.g., RSync and SCP). They also lack the high performance optimizations and security features offered by Globus. High performance data transfer tools such as bbftp [16] and Stork [18] support high performance transfer of large data, however they are not offered as services and lack the extensive network of endpoints accessible via Globus.

Cloud-based data storage and sharing providers, such as Dropbox, Box, and Google Drive, are increasingly used in science. However, as mentioned previously, they do not support large datasets and require costly transfers and replication. The EUDAT B2SHARE [8] service is focused on research data but, like commercial cloud providers, requires that data be uploaded to cloud infrastructure for sharing.

The quest for reproducible research has underpinned the deluge of data publication systems. There are several varieties of systems. For example, institution data publication repositories (e.g., the Purdue University Research Repository [4]), community-specific publication systems (e.g., the database of Genotypes and Phenotypes [19]), and commercial and non-commercial data publication services such as figshare [3] and Zenodo [5]. These systems offer support for specific data types or user communities. The Globus data publication service is differentiated by its support for arbitrarily large data, domain-independent publication, and self-service configuration and management.

Commercial identities, such as those provided by Google and Facebook, are often used by third-party commercial applications for authentication. The CILogon [9] authentication framework provides federated institutional identity management. It federates campus identity providers and provides an OAuth interface for third party applications. Auth0 [1] shares similar goals with Globus Auth, but with a focus on commercial applications and limited consent and delegation support.

8. FUTURE PLANS

Increasing usage and evolving requirements drive our continued investigations of new capabilities. We highlight here three areas of future work: advanced data search, a new policy-based data collection model, and active data management.

8.1 Search

Rapid growth in the number of Globus endpoints and in adoption of Globus data sharing means that individual users can access increasingly large volumes of data distributed across Globus endpoints. With the aim of providing a global data management solution we are investigating methods for enabling search of large quantities of Globus-accessible data.

Users are now accustomed to powerful search interfaces and thus expect high-quality results from deeply indexed files. To support such use cases we want to support search on both file system metadata and internal file structure and content. Given the large size and often rich structure of scientific files we are researching methods to index files intelligently, extracting and preserving structure where possible and supporting adaptive selection of the granularity of data to be indexed. A viable search implementation must also provide for fine-grain access control, without which we expect many users would be unlikely to use the service. In addition, we will need to provide methods to enable users to scope their searches, for example searching over user-owned endpoints, shared endpoints, or restrictive sets such as XSEDE endpoints.

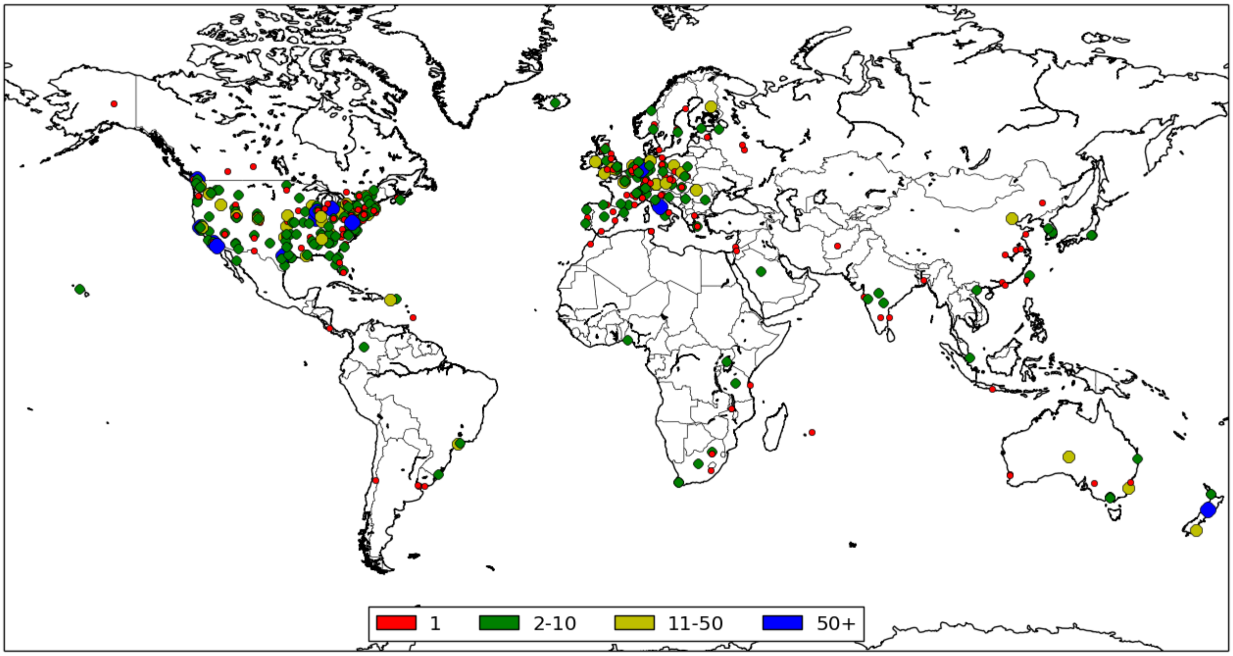

Figure 3: Globus Connect Server deployments, grouped by the number of deployments in a single location.

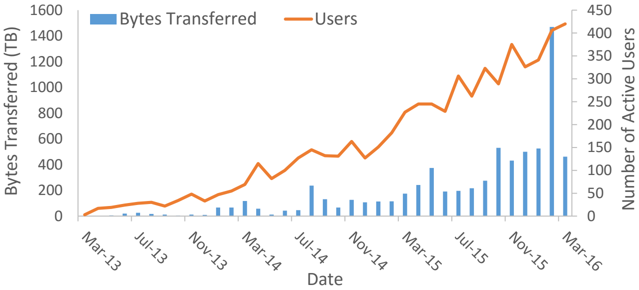

Figure 4: Amount of data and number of users who have transferred data involving a shared endpoint.

8.2 Collections

As shared endpoint usage grows, so too do requests for more advanced sharing policies. Globus sharing and publication provide similar capabilities (e.g., access control, workflows, identifiers) at different ends of the data sharing spectrum. However, the two systems are currently disjoint: a user cannot, for example, apply persistent identifiers to data within a shared endpoint. We envision integrating the two approaches by building on the flexible data sharing model to develop a policy-based approach to instantiating and managing user-managed collections of data.

The Globus collection model will enable policies to be applied to user-defined sets of data. Users will be able to define a collection by selecting all or a subset of the data within one or more Globus-accessible endpoints. They will then be able to optionally associate policies with a collection: for example, who can access and manage the collection, what metadata can or must be associated with data in the collection, what metadata should be extracted from files in the collection, and what persistent identifier will be assigned to the collection.

8.3 Active data management

Our third research activity aims to support a more active data management model. We see many people writing custom scripts that use Globus APIs to perform certain tasks repeatedly or in response to another event: for example, data archiving, transfer to geo-replicated storage, and movement to analysis services for automated quality control.

To address these use cases we will develop a modular active data management environment that will allow users to define rules that result in actions within the Globus ecosystem. These rules may be expressed in terms of periodic events (e.g., cron jobs) or based on events generated by Globus services. We will develop a subscription model via which users can subscribe to events generated from various services. We will develop a service that implements a rules engine to process events. The service will provide interfaces that allow users to define their actions in an intuitive manner. This approach has some similarities to iRODS [21], but differing in its ability to work with data on arbitrary endpoints and with respect to the type of rules we intend to support and the mechanisms by which we implement and enforce these rules.

9. SUMMARY

Globus has become, for many researchers and institutions, the primary means of managing both large and small amounts of research data. Over its first five years of operation Globus has grown to support tens of thousands of users and has been used to transfer more than 150 PB of data. During this time, we have continued to both enhance existing capabilities and add new research data management services. In particular, we have added support for advanced transfer management, interfaces to new storage systems, and native HTTP support. We have also released data sharing and publication services to fill important gaps in the research data management ecosystem.

Looking towards the future we are focused on three major activities: supporting our growing user community via continued operation and enhancement of existing capabilities; development of new services that address research data management challenges; and providing Globus capabilities as a platform. Globus Auth is one step in this direction: it enables a more seamless user experience for current users while also putting in place a model via which external applications can securely access Globus services.

10. ACKNOWLEDGMENTS

We thank Globus subscribers for supporting the operation and development of Globus. We also thank users of Globus services for their continued support. The development of Globus services has been supported in part by DOE DE-AC02-06CH11357; NSF ACI-1053575 and ACI-1148484; NIH U24 GM104203 and 1-U54EB020406-01; the University of Chicago; a grant from the Sloan Foundation; and a grant of computer time from Amazon Web Services.

11. ADDITIONAL AUTHORS

Additional authors: Bryce Allen (Computation Institute: CI), Rachana Ananthakrishnan (CI), Joe Bester (CI), Ben Blaiszik (CI), Vytas Cuplinskas (CI), Raj Kettimuthu (CI), Jack Kordas (CI), Lukasz Lacinski (CI), Mattias Lidman (CI), Mike Link (CI), Stu Martin (CI), Brendan McCollam (CI), Karl Pickett (CI), Dan Powers (CI), Jim Pruyne (CI), Brigitte Raumann (CI), Gigi Rohder (CI), Stephen Rosen (CI), Dave Shifflett (CI), Teresa Sutton (CI), Vas Vasiliadis (CI), and Jason Williams (CI).

12. REFERENCES

[1] Auth0. http://auth0.com. Web site. Accessed: April, 2016.

[2] Dataverse. http://thedata.org. Web site. Accessed: April, 2016.

[3] figshare. http://figshare.com. Web site. Accessed: April, 2016.

[4] PURR: Purdue University Research Repository. http://purr.purdue.edu. Web site. Accessed: April, 2016.

[5] Zenodo. https://www.zenodo.org. Web site. Accessed: April, 2016.

[6] B. Allen, J. Bresnahan, L. Childers, I. Foster, G. Kandaswamy, R. Kettimuthu, J. Kordas, M. Link, S. Martin, K. Pickett, and S. Tuecke. Software as a service for data scientists. *Communications of the ACM*, 55(2):81–88, Feb. 2012.

[7] R. Ananthakrishnan, K. Chard, I. Foster, and S. Tuecke. Globus platform-as-a-service for collaborative science applications. *Concurrency - Practice and Experience*, 27:290–305, 2014.

[8] S. B. Ardestani, C. J. Håkansson, E. Laure, I. Livenson, P. Stranák, E. Dima, D. Blommesteijn, and M. van de Sanden. B2SHARE: An open escience data sharing platform. In *Proceedings of the 11th IEEE International Conference on e-Science*, pages 448–453, Aug 2015.

[9] J. Basney, T. Fleury, and J. Gaynor. CILogon: A federated X.509 certification authority for cyberinfrastructure logon. *Concurrency and Computation: Practice and Experience*, 26(13):2225–2239, 2014.

[10] B. Blaiszik, K. Chard, R. Ananthakrishnan, S. Tuecke, and I. Foster. The Materials Data Facility: Data services to advance materials science research. *Journal of the Minerals, Metals & Materials Society*, to appear, 2016.

[11] K. Chard, M. Lidman, B. McCollam, J. Bryan, R. Ananthakrishnan, S. Tuecke, and I. Foster. Globus nexus: A platform-as-a-service provider of research identity, profile, and group management. *Future Generation Computer Systems*, 56:571–583, 2016.

[12] K. Chard, J. Pruyne, B. Blaiszik, R. Ananthakrishnan, S. Tuecke, and I. Foster. Globus data publication as a service: Lowering barriers to reproducible science. In *Proceedings of the 11th IEEE International Conference on e-Science*, pages 401–410, Aug 2015.

[13] K. Chard, S. Tuecke, and I. Foster. Efficient and secure transfer, synchronization, and sharing of big data. *IEEE Cloud Computing*, 1(3):46–55, Sept 2014.

[14] R. Dooley, M. Vaughn, D. Stanzione, and E. Skidmore. Software-as-a-service: The iPlant Foundation API. In *5th IEEE Workshop on Many-Task Computing on Grids and Supercomputers (MTAGS)*, 2012.

[15] I. Foster. Globus Online: Accelerating and democratizing science through cloud-based services. *Internet Computing, IEEE*, 15(3):70–73, May 2011.

[16] A. Hanushevsky, A. Trunov, and L. Cottrell. Peer-to-peer computing for secure high performance data copying. In *Proceedings of the International Conference on Computing in High Energy and Nuclear Physics*, 2001.

[17] D. Hardt. The OAuth 2.0 authorization framework. http://www.rfc-editor.org/info/rfc6749, October 2012.

[18] T. Kosar and M. Livny. A framework for reliable and efficient data placement in distributed computing systems. *Journal of Parallel and Distributed Computing*, 65(10):1146–1157, Oct. 2005.

[19] M. Mailman, M. Feolo, Y. Jin, M. Kimura, K. Tryka, R. Bagoutdinov, L. Hao, A. Kiang, J. Paschall, L. Phan, N. Popova, S. Pretel, L. Ziyabari, M. Lee, Y. Shao, Z. Wang, K. Sirotkin, M. Ward, M. Kholodov, K. Zbicz, J. Beck, M. Kimelman, S. Shevelev, D. Preuss, E. Yaschenko, A. Graeff, J. Ostell, and S. Sherry. The NCBI dbGaP database of genotypes and phenotypes. *Nature Genetics*, 39(10):1181–1186, 2007.

[20] M. Pierce, S. Marru, L. Gunathilake, T. Kanewala, R. Singh, S. Wijeratne, C. Wimalasena, C. Herath, E. Chinthaka, C. Mattmann, A. Slominski, and P. Tangchaisin. Apache Airavata: Design and directions of a science gateway framework. In *Proceedings of the 6th International Workshop on Science Gateways (IWSG)*, pages 48–54, June 2014.

[21] A. Rajasekar, R. Moore, C.-y. Hou, C. A. Lee, R. Marciano, A. de Torcy, M. Wan, W. Schroeder, S.-Y. Chen, L. Gilbert, P. Tooby, and B. Zhu. *iRODS Primer: Integrated Rule-Oriented Data System*. Morgan and Claypool Publishers, 2010.

[22] N. Sakimura, J. Bradley, M. Jones, B. de Medeiros, and C. Mortimore. OpenID Connect Core 1.0. http://openid.net/specs/openid-connect-core-1_0.html, Nov. 2014.

[23] M. Smith, M. Barton, M. Bass, M. Branschofsky, G. McClellan, D. Stuve, R. Tansley, and J. H. Walker. DSpace: An open source dynamic digital repository. *D-Lib Magazine*, 9(1), 2003.

[24] J. Towns, T. Cockerill, M. Dahan, I. Foster, K. Gaither, A. Grimshaw, V. Hazlewood, S. Lathrop, D. Lifka, G. D. Peterson, R. Roskies, J. R. Scott, and N. Wilkins-Diehr. XSEDE: Accelerating scientific discovery. *Computing in Science Engineering*, 16(5):62–74, Sept 2014.