Globus Toolkit & Globus Cloud Service

Report

Course: Clusters, Grids and Clouds

Prepared by Anastasiia Grishina

Professor: Andrey Shevel

ITMO, Saint Petersburg, PERCCOM

2018

Table of Contents

Globus System: Introduction

Globus system is an open source software toolkit used for building grids which was developed by the Globus Alliance. The system was primarily aimed at secure data transfer capability which then developed and extended to Globus Cloud Services for transferring, sharing, publishing data as well as developing applications with the help of Globus pre-built authentication and data management application template with REST APIs.

In its history, Globus has two stages: Globus Toolkit development and Globus Cloud Services which was being developed in parallel with the toolkit after around 12 years of its existence and finally attracted all the resources of developers in early 2018. The Globus Toolkit has grown through an open-source strategy similar to the Linux operating system's, and distinct from proprietary attempts at resource-sharing software. This encourages broader, more rapid adoption and leads to greater technical innovation, as the open-source community provides continual enhancements to the product.

Similar to the majority of grid services, Globus is oriented at providing secure collaboration tools for researchers involved in fundamental domains. They comprise astronomy, particle physics, fluid dynamics and other spheres which would not suffer from the lack of confidentiality while sharing the data with scientific world before acquiring the patent for their inventions, like chemistry.

The Globus Alliance has also contributed to the research in grid technologies analyzing issues and prospects of grid implementation and wide use which are reflected in the papers. They are listed on the official website of the Globus Toolkit [1] and are representatives of the first scientific works on the grid research area.

Genesis of Globus Toolkit development

In late 1994 Rick Stevens, director of the mathematics and computer science division at Argonne National Laboratory, and Tom DeFanti, director of the Electronic Visualization Laboratory at the University of Illinois at Chicago, proposed establishing temporary links among 11 high-speed research networks to create a national grid (the "I-WAY") for two weeks before and during the Supercomputing '95 conference.

A small team led by Ian Foster at Argonne created new protocols that allowed I-WAY users to run applications on computers across the country. This successful experiment led to funding from the Defense Advanced Research Projects Agency (DARPA, and 1997 saw the first version of the Globus Toolkit, which was soon deployed across 80 sites worldwide.

The U.S. Department of Energy (DOE) pioneered the application of grids to science research, the National Science Foundation (NSF) funded creation of the National Technology Grid to connect university scientists with high-end computers, and NASA started similar work on its Information Power Grid.

The following progress of the project is stamped by the sequential releases of Globus Toolkit (GT) updates from 1998, the year of the first official GT 1.0.0 release to 2016, the last GT 6.0 release, after which GT community was proposed to move to the Globus Cloud Service. The developing team has announced that, starting in January 2018, the Globus team at the University of Chicago no longer supports the open source Globus Toolkit, except for its use with the Globus cloud service by Globus subscribers.

By the end of 2018, all endpoints connected to the Globus cloud service using the open source Globus Toolkit GridFTP server must migrate to Globus Connect. At the end of 2018, the Globus team will discontinue all maintenance (including security patches) and distribution of the open source Globus Toolkit. Endpoints using Globus Connect Server or Globus Connect Personal will be unaffected, as long as they continue to perform routine software updates.

The decision is based on two main reasons: lack of funding and Globus Cloud Service being available and fully maintained. The team claims that the open source Globus Toolkit, like any software, requires constant effort to answer support requests, apply security patches, and perform other maintenance whereas grants from U.S. National Science Foundation end in the autumn of 2018. Moreover, Globus Connect is quickly diverging from the Globus Toolkit.

Globus Toolkit Features

The toolkit includes software for security, information infrastructure, resource management, data management, communication, fault detection, and portability. It is packaged as a set of components that can be used either independently or together to develop applications. The Toolkit is aimed at solving data transfer and multiple downloading issues like hang up or rebooting problems by providing aforementioned software for the computing and data grids.

The advantages of the Toolkit comprise:

- Effective content sharing
- Efficient downloading and transfer
- Maximum utilization of resources
- Applicability for scientific and satellites applications

The disadvantages can be seen in:

- Limited applicability for small applications
- Chance of data misuse, even though double encryption is performed

Globus Cloud Service Features

Globus is considered as one of leading providers of secure, reliable research data management services. Namely, with Globus, users can move, share, publish & discover data via a single interface. It supports data management stored on a supercomputer, lab cluster, tape archive, public cloud or your laptop. Access is available from any existing identities of users on their PC and via just a web browser. Another important Globus feature is provision of a platform for application integration and gateway development leveraging for advanced identity management, single sign-on, search and authorization capabilities.

The system is available in Globus Connect Personal and Server distributions for different purposes. Personal distribution is aimed at individual researchers and other end users which need to manage data transfer and management of large amount, i.e. single users with personal machines even without administrator privileges on it. This distribution can be installed on Linux, Windows and MacOS. Server version is dedicated for network and system administrators, who get possibility to manage multi-user computing and storage resources. The system is only available for Linux users.

Main four features of Globus Cloud Service are described below. All of them require subscription for user authentication and only data transfer is free of charge feature, whereas other functionalities need investment. The prices can be acquired upon submitting an order via Globus official website and they are claimed to depend on the user status, being a student, researcher or other type of end user.

Data Transfer

Globus provides a secure, unified interface to research data of many scientific collaborations and individual researchers. Globus lets users choose a web browser or command line interface to submit transfer and synchronization requests, optionally choosing encryption.

Users' data is transferred directly between the source and destination systems while Globus tunes performance parameters, maintains security, monitors progress, and validates correctness. It is possible to check the transfer status at any time via the Globus activity page. Users receive notification in the form of email when the transfer completes.

If a network or system involved in the transfer goes down, Globus automatically resumes the transfer when the component comes back online. If an issue requires action from a user, such as an expired credential or exceeded disk quota, Globus resumes the transfer after a user remedies the problem. If a transfer has not made progress after a period of time (usually 3 days), the transfer will expire and a user will be notified.

The transfer process is schematically represented in the Figure 1. A user selects endpoints and submits transfer request. Globus then transfers the data and notifies a user about the task completion.

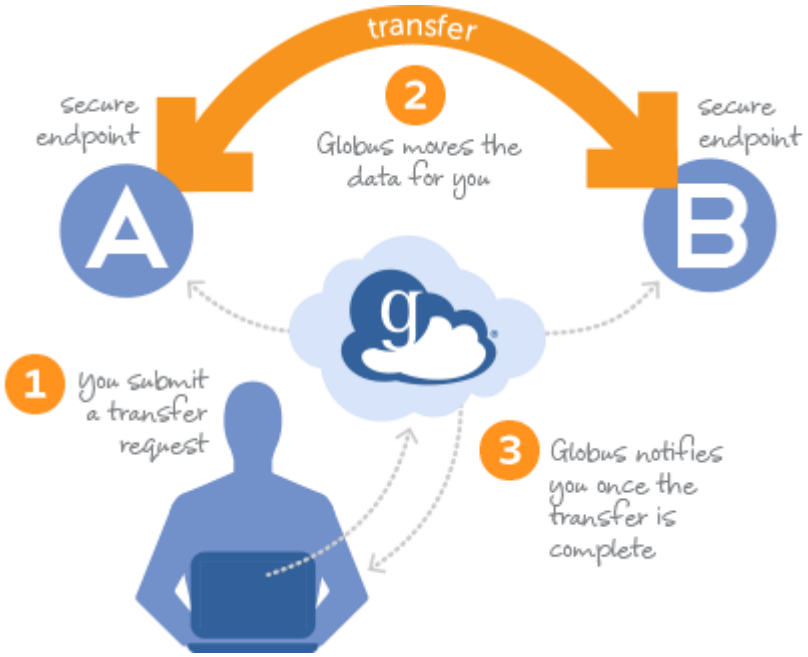

Figure 1. Globus Data Transfer process

Data Sharing

Globus provides users easy and secure data sharing capabilities with collaborators without requiring them to create temporary accounts or transfer the data to be shared to an external storage system.

The problem which is successfully solved in Globus is the fact that most file sharing services require users to copy your data to an external storage system that they manage and which is usually hosted in the cloud. This sharing model is reasonable for a few megabytes of data, but can be slow and expensive for big research data.

Globus, by contrast, does not require movement of data in order to share it. Project developers explain that Any storage system that has a Globus connection, including public clouds like Google Drive and Amazon S3, can be configured to allow secure data sharing directly by users of the system.

Globus uses widely-adopted industry standards such as OAuth2 and OpenID Connect for authentication/authorization, and uses trusted protocols such GridFTP

and HTTPS. Globus recognizes that administrators need to configure their systems for sharing in a secure manner, and provides the tools to do so.

The sharing functionality allows users to select directory paths which will be securely shared with external remote collaborators. Globus allows to manage sharing groups and granting them different types of access. The process of data sharing is shown in Figure 2. This process works in the way that a user shares the files directly from their storage place. Collaborators receive a link to the shared files and have a chance to transfer files between themselves.

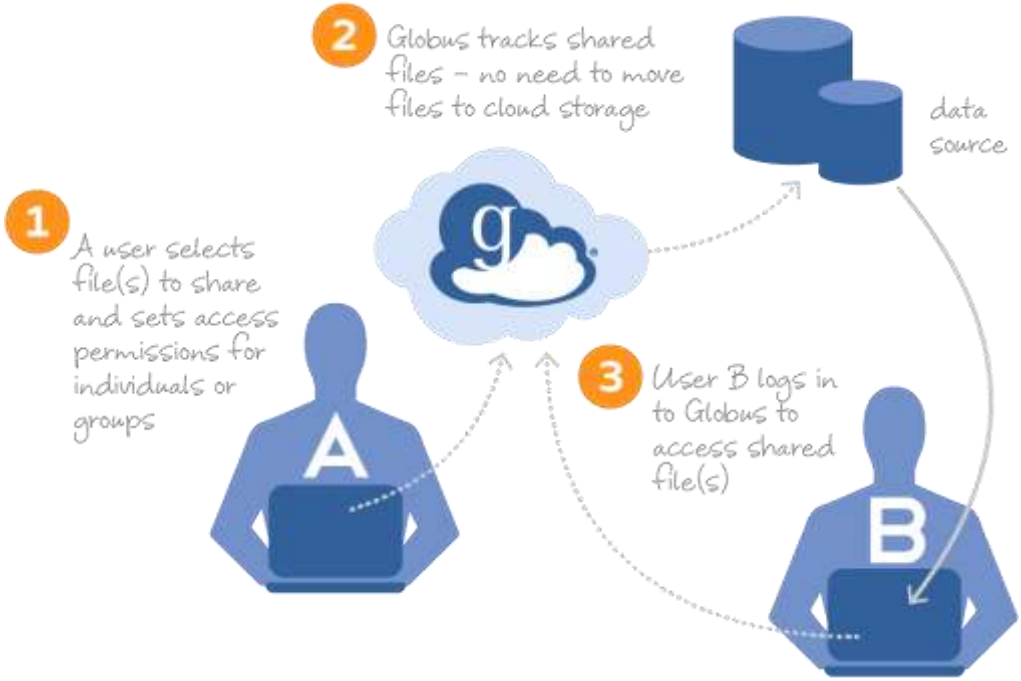

Figure 2. Data Sharing Process performed by Globus

Data Publishing

The problem which Globus tends to solve with regard to data publishing is that there are limited tools currently available to digital media managers and others in campus organizations tasked with managing data publication. Typical approaches involve developing, installing, and configuring various software components, and integrating these with existing campus identity and storage systems. This is a costly and time-consuming activity.

Globus publication capabilities are delivered through a hosted service (Figure 3). Published data is stored on campus and institutional resources that can be managed by different administrators. To associate storage resources with a data collection Globus shared endpoints can be used and they should be further associated with the data repository to publish. Globus users can create and manage their own communities and collections through the data publication service and place rules for different user groups.

Datasets undergo curation based on a workflow defined by the community that will publish the data. Workflows may be customized by each community to capture their

specific metadata and to reflect the community's review process. After the dataset is published, it is discoverable using a faceted search that allows the researcher to progressively filter results and rapidly focus in on the data of interest.

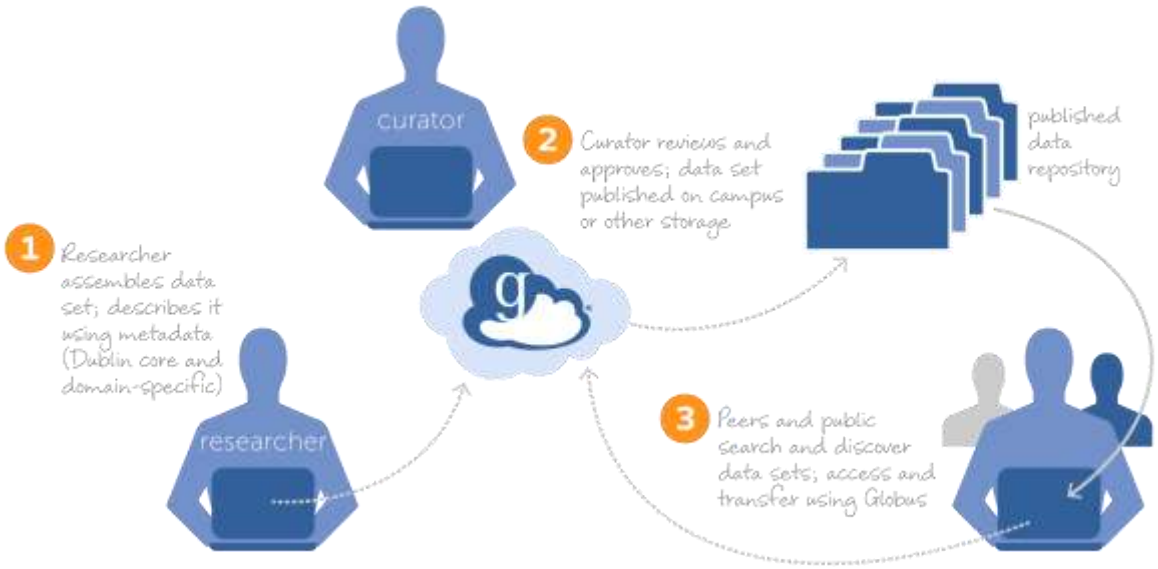

Figure 3. Data Publishing Process performed by Globus

Develop Applications and Gateways

The principle idea of this Globus functionality is to free researchers from the essential yet not the main part of application development, which are authentication and authorization as well as file transfer and search. Thus, the Globus platform enables developers to provide robust file transfer, sharing and search capabilities within their own research data applications and services, while leveraging advanced identity management, single sign-on, and authorization capabilities.

The main advantages of this functionality are:

- It lets researchers focus on core features;
- It provides consistent UI through storage systems;
- It uses well-documented open RESTful APIs;
- It includer secure, scalable search using custom metadata.

Same OAuth and OpenID Connect for authentication/authorization, and uses trusted protocols such as GridFTP and HTTPS are used to enhance reliability.

Implementation: demo

For the trial of the Globus Cloud Services, data transfer functionality was under the focus, since it is available for free. Windows machine and Ubuntu virtual machine were used to install Globus Connect Personal, therefore both web and command line UI were utilized and tested. Instructions from Globus official website are clear and easy to follow. Endpoints were configured upon reception of the setup key and its submission to the web service. Furthermore, the directories for Globus access were configured. Unfortunately, special subscription, which is proof of the identity and belonging to certain research communities, was required to actually perform the file transfer.

Nevertheless, if such subscription had been available, with the interface shown in Figure 4 it would be possible to transfer files in several mouse clicks, which is convenient for the end users.

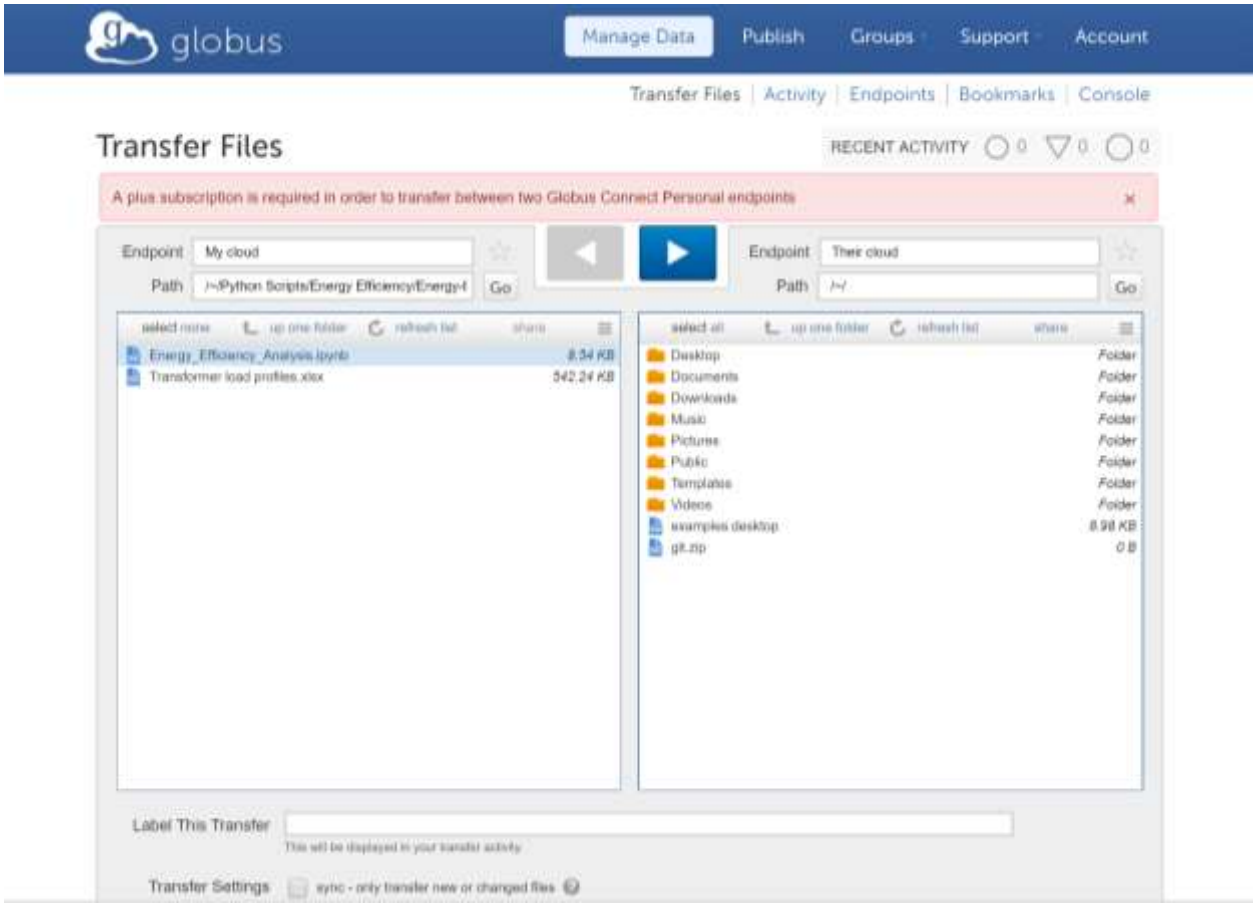

Figure 4. Demo. Globus File Transfer Functionality

Conclusion

The report covers Globus development genesis which is divided by two stages of Globus Toolkit and Globus Cloud Service development. Main functionalities of both products are described to give an overview of options which the products provide for end users. The report finishes with a demo of one of the Globus Cloud Service features.

Overall, Globus is a set of software components which are aimed at grid usage. Main functionalities cover data transfer, sharing, publishing and fast set up of new application services. Globus is aimed primarily at research communities and is currently widely used by communities which focus on Fundamental Physics, Fluid Dynamics and other. Being a reliable research data management service, Globus connects collaborators all over the world and simplifies their research with files when it comes to work for common goals.