Integration of Globus Online with the ATLAS PanDA Workload Management System

The ATLAS PanDA system

PanDA relies on sophisticated dynamic data movement facilities developed in ATLAS, which automate various tasks related to data staging, transfer and cataloging. In certain scenarios, such as small research teams in ATLAS “Tier-3” sites and non-ATLAS Virtual Organizations, the overhead of installation and operation of these components makes their use not very cost effective. Globus Online is an emerging new tool from the Globus Alliance, which already proved popular within the research community. It provides the users with fast and robust file transfer capabilities that can also be managed from a Web interface, and in addition to grid sites, can have individual workstations and laptops serving as data transmission endpoints.

The main component of the Globus Online system is a Web Service that automates data movement across sites and/or systems, called “endpoints” in this context. It is equipped with a user-friendly Web interface that allows the users to initiate and manage data transfer and access logs. In addition, it provides facilities for GUI and REST APIs. The former has syntax similar to sftp and can be scripted for automation purposes if needed. The latter presents a foundation for integration of Globus Online into workflow distributed over Grid sites or in Cloud Computing environment. This "RESTful" interface has Python bindings, which makes it especially well suited for use in the PanDA framework, which is itself heavily Python-based.

User Actions:
• Cache user proxy on an instance of MyProxy service
• Define endpoints via the Web Interface or from an application
• Submit a job via the common CLI interface to PanDA

Globus Online – it works

Globus Online Web Service

Globus Online – how it works

Consider a general case:
• The input data is located on the User’s desktop
• A remote Grid site managed by PanDA will be used for computation
• The output data must arrive back to the user’s desktop

In this case of Globus Online application, the following needs to happen:
• The user activates a Globus Connect endpoint on their desktop
• The user defines the GridFTP server present at the processing site as another endpoint via the Globus Online web interface (a "MYProxy"

One of possible ATLAS use cases:
• The input data is located on a storage element on an ATLAS-managed site, and is properly accounted for in the file catalog
• A remote Grid site managed by PanDA will be used for computation

In addition to the storage element normally used in this particular analysis workflow, the output data must arrive to an additional location outside of DDM sphere of management, such as the user's desktop or a private GridFTP server run by the research team.

Role of X.509 credentials

Client software used by Globus Online requires the use of X.509 credentials, such as the end-user’s proxy, in order to get authenticated to the Web Service and be authorized to initiate data transfer. It is necessary therefore for the Pilot process running on the remote Grid site to obtain the user’s proxy at runtime. We opted to utilize an instance of the MyProxy service deployed at CERN in order to securely cache and retrieve the users’ X.509 proxies. When caching the proxy, an option is used to authorize its later retrieval by a trusted agent, in this case the PanDA Pilot, whose identity is recorded in an encrypted form in the DN of its own proxy, which is used to authenticate and get authorized by MyProxy.

Job lifecycle and data movement in PanDA enhanced with Globus Online capability