基于微服务架构的Globus科研数据管理平台分析*

袁晓明　王美琴
（苏州大学图书馆，苏州 215006）

摘要：科研数据的爆发式增长和远程共享对数据管理提出需求，可实现弹性扩展、高性能传输、云托管的微服务架构数据管理平台为数据的高效管理提供可能。本文调研分析了芝加哥大学阿贡国家实验室研究团队研发的Globus科研数据管理平台。该平台采用分布式微服务架构，包括身份管理、组群服务、数据传输和数据共享4个微服务模块，模块之间相互独立，通过可视化Web界面实现科研数据的传输和共享。该平台已经应用于多个科研项目的数据管理，完成TB数量级的文件传输，其跨区域高效传输、灵活共享的数据管理模式可为我国科研数据管理提供借鉴。

关键词：科研数据管理；Globus；云服务；SaaS；微服务
中图分类号：G250　DOI：10.3772/j.issn.1673-2286.2021.12.004
引文格式：袁晓明，王美琴. 基于微服务架构的Globus科研数据管理平台分析[J]. 数字图书馆论坛，2021(12)：22-27.

随着科学技术的发展、科学仪器的更新和科学研究方法的改变，高性能探测设备和分析仪器应用到科研过程中，随之产生了大量高分辨率图片、音视频等科研数据，一些学科领域的科研数据爆发式增长，科学研究已经进入数据密集型阶段[1]。同时，高质量科研数据的共享、再现对大数据和人工智能跨领域数据分析日益重要，荷兰莱顿数据科学中心Mons[2]指出，科研数据管理不仅是科研工作者的责任，也是科研项目的重要组成部分，其中数据的存储、迁移和利用越来越受到国内外研究机构、高校及科研人员的重视。

数据密集型科研环境下，科研数据管理对科研数据价值的发挥产生了重要影响，国内外机构或图书馆围绕科研数据管理平台的开发和本地部署开展相关实践，以帮助研究者管理科研数据。在美国，如麻省理工学院图书馆的数字存储系统DSpace[3]、哈佛大学的Dataverse[4]、康奈尔大学的Datastar[5]、普渡大学的PURR[6]、宾夕法尼亚州立大学的仓储服务系统Scholar Sphere[7]等。国内机构也尝试采用开源数据管理软件本地化搭建科研数据管理平台，如复旦大学采用Dataverse构建复旦大学社会科学数据平台，武汉大学图书馆基于开源软件DSpace搭建的“大学科学数据共享平台”等。当前，科研数据管理平台主要提供仓储式数据管理服务，只支持本机构用户并依赖机构的管理，无法在机构间或更广范围内共享数据。随着科研数据的急剧增长和跨区域的科研合作共享，仓储式服务的科研数据管理平台面临以下挑战。①数据访问的限制。仓储式科研数据管理平台是一个复杂的Web服务应用程序，通常包含用户信息、数据信息，设置了较复杂的安全防护，限制了平台的访问和可伸缩性，无法满足高性能独立数据通道的高速访问[8]。②高性能数据传输需求。研究数据量的激增，需要数千个文件或者TB级别数据量的高效传输，传输效率对此类科研数据管理平台是一项挑战。③平台部署维护技术和资金支撑。仓储式服务的科研数据管理平台是筒仓式开发部署的，

* 本研究得到2020年度江苏省JALIS数字图书馆专题研究项目“基于社会化精准服务提升馆员专业能力的研究”（编号：2020KT08）资助。

系统独立运行本地的用户管理、身份验证、授权和数据传输，不仅部署需要强大技术支持，还需要管理人员的长期技术跟踪。④数据共享范围有限。各平台数据以不同的分类体系进行组织和元数据主题标引，不同数据组织方式使平台之间技术不可扩展、数据无法迁移，共享范围有限。

仓储式服务的科研数据管理平台虽然在机构范围内的数据管理起到了一定的作用，但随着网络结构性能提升和数据共享全球化的需求，其服务范围受到了根本限制。科研数据管理平台面临应用开发方式、数据存储方式、系统部署和服务功能的挑战，基于微服务架构的平台应用可满足数据高效传输和共享、敏捷开发和动态扩展的需求。

为此，笔者调研分析了芝加哥大学研究开发的基于微服务架构的SaaS模型Globus科研数据管理平台的服务方式和架构，及其应用于科研数据管理的案例，为国内科研机构和科研人员的科研数据管理提供借鉴。

1　Globus科研数据管理平台简介

微服务架构是一种细粒度、自治、协同工作的服务体系[9]，其将大型复杂系统从功能上分解成设计、开发和部署中相互独立自治的小型服务，并通过轻量级机制进行通信，采用标准的API和基于容器的平台来强调松散耦合和高内聚[10]。架构模块具有技术异质性、可独立部署、可弹性扩展等特点，便于系统的技术升级和功能更新，近年来迅速发展并被尝试应用于软件平台、面向服务架构的开发。

Globus是2010年由芝加哥大学阿贡国家实验室研究团队研发的科研数据管理平台，是一种基于微服务架构的软件即服务（SaaS），以Amazon云服务实现数据管理，提供身份认证与授权、数据迁移与复制、数据共享、数据发布与发现等功能[11-12]，并以Web访问服务形式为研究者提供了一套功能强大的科研数据管理功能。在技术方面，该体系结构具有高度的容错性、可弹性扩展、易于部署，且随着负载的增加，服务可动态地分配虚拟机；在数据处理上，云服务具有高效的数据处理效率和弹性计算能力，保证了密集数据的高效、稳定传输；在服务模式上，SaaS支持多租户访问，用户不需要安装或操作任何软件，任何授权用户都可以建立和管理自己的数据发布集合，易于广泛使用[13]。

随着存储介质的更新、存储端点的增加，Globus平台不断更新服务功能、完善数据管理生态系统，包括对安全HTTP数据访问支持、新型存储系统（Amazon S3、HDFS）的兼容、数据端点搜索和管理员管理功能的提升，有效解决了科研人员数据管理过程中的冗余事务。

Globus数据管理包含两个核心组件，即托管服务和代理软件。

Globus实现了第三方传输的托管服务模型，所有微服务模块托管于Amazon云服务器，通过Web应用程序为所有的微服务提供统一接口的协同访问页面。用户发出数据处理指令后，其数据数据处理流程（传输、共享、发布、发现以及身份和凭证管理）都在Amazon云上运行（见图1），此过程中Globus以数据监护方式参与其中，进行用户管理、权限控制和数据流程控制，不传输数据，不保存或记录任何数据。

代理软件（Globus Connect）实现了身份验证和数据访问的机制，包括服务器和个人2个安装版本。Globus Connect Server是一个Linux软件包，部署于存储服务器；Globus Connect personal是一个轻量级的单用户代理，可以部署在Windows、macos和Linux计算机上，使这些系统能够参与Globus文件共享网络。目前，Globus Connect服务器已更新至v5.4版本，可实现多端点共享、超大文件即时传输、兼容多形式存储系统之间的文件转移[14]，且新的版本具有管理控台和发现新端点功能，可监控数据传输状态，支持对多形态存储介质终端的发现和连接。

2　Globus科研数据管理平台的微服务模块

Globus SaaS的微服务模块包括3个关键组件：REST API、一个或多个后端任务工作程序和数据持久层[11]。Globus处理REST API服务模块一般部署在Amazon EC2云服务器上，其处理REST API请求的所有逻辑单元都是同步执行的，在持久存储层中注册所需的活动后即终止任务，由后端任务工作程序进一步处理，所以模块在磁盘和内存中运行非常短暂，具有强大的数据处理能力。Globus管理团队可根据系统负载添加或删除API的服务能力，弹性扩展微服务模块。数据持久层部署在Amazon云存储服务器上，利用其可跨区域复制性能实现系统容错，并定期创建远程快照帮助实现故障恢复。Globus使用了S3和PostgreSQL关系型数据库（RDS），各系统组件封装在虚拟云（VPC）中并互相

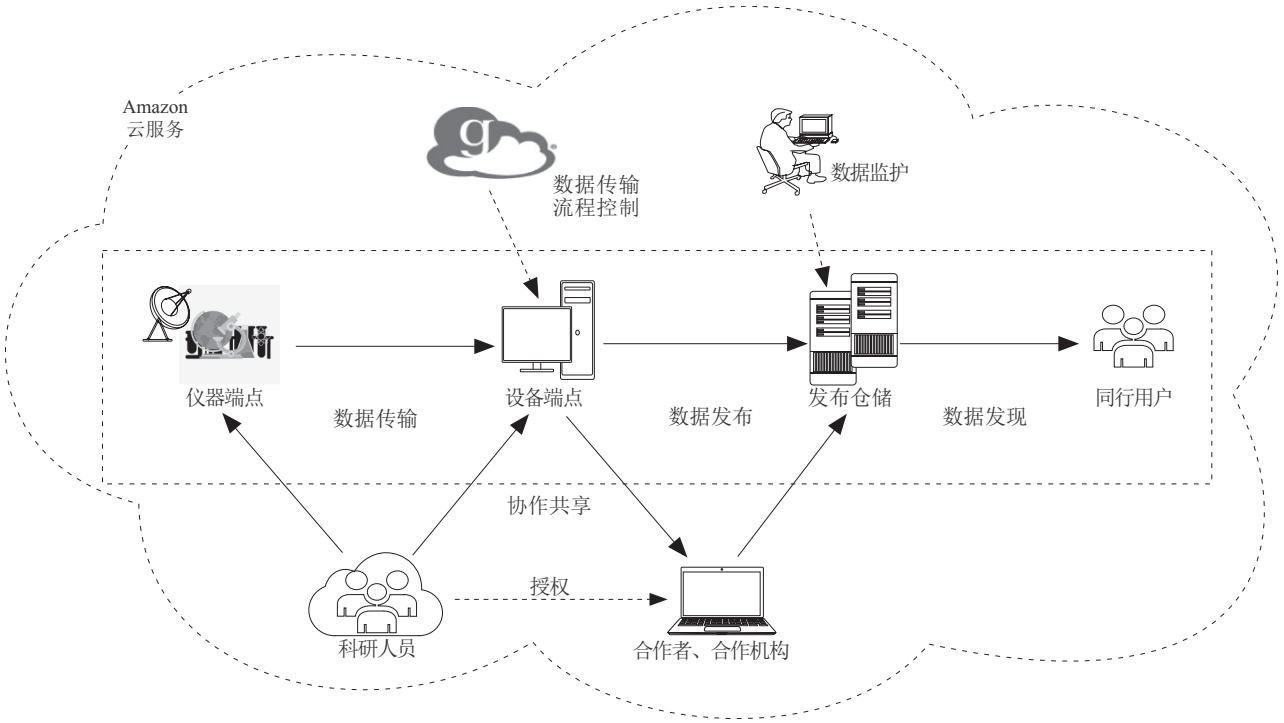

图1 Globus数据管理流程

独立，实现了Amazon云托管服务。

Globus数据管理由Globus Auth、Globus Groups、Globus Transfer、Globus Publication 4个分布式微服务模块组成，可分别实现身份认证服务、群组管理服务、数据传输服务和数据出版服务。

2.1 身份认证服务

Globus Auth模块是其他微服务的基础和安全模型的核心，贯穿数据管理服务全过程。Globus Auth是一个Python Web应用程序，Python应用层管理身份、账户和客户端，通过REST API接口注册和配置客户端、获取用户和令牌的信息以及检索链接的标识，符合标准Web协议OAuth 2和OpenID Connect规范，可与第三方应用程序集成[15]。

Globus Auth代理了终端用户、身份提供、资源服务器和客户端（如Web界面、移动设备、桌面命令行等）之间的身份验证与授权交互，支持用户多重身份（如机构身份、校园ID、Google账户）登录，实现了身份联合模型，将不同身份链接为用户身份集合使用，消除了在使用分布式网络基础设施时经常需要多个账户、身份、凭据的冲突。科研人员数据管理过程中可随时切换身份，用一个凭证身份验证登录，使用另一个身份连接到特定的远程存储资源，以及基于其他身份与合作者共享数据等。此外，Globus Auth提供了临时委托访问令牌，增减客户端的访问权限。

2.2 群组管理服务

Globus Groups模块是采用Web框架实现的一个层次化的组模型，是在Globus Auth基础上实现的团队管理功能。Globus Groups通过评估用户的成员身份授权其可访问粒度，应用于科研团体（组）的授权、角色和共享。该模块提供了一个面向团体用户和团体成员管理的工作流集合，可以让用户自定义存储数据对成员可见性、成员资格、工作流（邀请、接受、暂停等）、成员角色，适合机构对科研人员的数据权限管理或团队项目合作。Globus Groups还利用了其他AWS服务，包括用于电子邮件的SES和用于内部通知的SNS（见表1）。

2.3 数据传输服务

科研协作通常需要对跨区域分布式资源的复杂数据管理，用户需要在采集、存储、分析和归档之间移动

表1 Globus Amazon云服务架构及主要功能

Amazon云服务	为Globus提供的服务
EC2	保障Globus服务的高效运行，提供Web API，运行后台任务，搭建内部基础设施
RDS	存储Globus服务状态
dynamodb	NoSQL数据库服务，实现无缝扩展，存储Globus服务状态
VPC	建立安全虚拟网络的Amazon私有云
ELB	将客户端请求定向执行到可用的服务
Amazon S3	存储运行任务的状态、服务数据备份、静态Web内容
IAM	在Globus中管理对Amazon资源的访问
Cloud Watch	监控Globus资源的状态
SNS	向Globus员工发送通知
SES	向用户发送电子邮件

大量数据。Globus Transfer模块是Globus数据管理服务的核心模块，为科研人员和机构提供了高性能的文件传输和同步服务，简化了两个存储端点之间移动大数据的过程，实现了Globus用户之间的数据安全共享。

Globus利用部署在存储系统上的Globus Connect软件来协调第三方数据的安全传输，数据传输基于虚拟的“共享端点”模型，用户利用Globus Transfer Web、CLI和REST接口在现有端点上的任何文件系统位置创建虚拟的“共享端点”，使用GridFTP协议在端点之间传输数据[6]。GridFTP提供了一个模块化的数据存储接口（DSI），以支持现代网络环境下的不同存储介态，如高性能存储系统（HPSS）、云存储对象及传统存储系统之间的数据传输[16]。

Globus使用两个独立的通信通道，在Globus和端点之间建立控制通道，以启动和管理传输、检索目录列表和建立数据通道，在两个端点（GridFTP服务器）之间直接建立数据通道，用于系统之间的数据传输，Globus服务无法访问数据通道。Globus Transfer是同时使用S3和PostgreSQL RDS数据库的多层存储模型，存储了大量数据传输状态的信息，PostgreSQL RDS管理了安全隐私性信息（如用户、端点等），S3用于存储详细的传输信息，如文件列表和性能标记。Globus Transfer主要实现以下功能。

（1）高性能、可靠的数据传输：保障用户数据传输可靠性和完整性，Globus可对传输控制协议（TCP）缓冲区大小、并发控制通道数量等参数灵活调控，并自行校验传输文件完整性、故障恢复后自动重启传输。

（2）实现跨区域的第三方传输：以第三方管理模式参与两个远程端点数据传输，保障用户数据的隐私性和安全性。

（3）就地数据共享：允许用户使用Globus Connect软件将本地资源公开为Globus端点，并根据文件共享程度授权访问权限。

2.4 数据出版服务

Globus Publication模块支持用户管理发布共享数据[17]。Globus数据发布是在DSpace机构存储库系统的基础上实现的，并采用Globus微服务替换了DSpace内置功能：用户和组管理分别替换为Globus Auth和Globus Groups，使用Globus Transfer处理数据管理和访问策略。该模块管理数据的存储位置、应收集的元数据、应用的持久标识符的形式、使用的管理工作流以及谁可以提交、管理和访问数据的“集合”，通过数据发布服务进行发布，工作流完成后，元数据文件会复制到发布终端。

Globus数据管理平台4个微服务模块可互相调用完成数据管理流程，也可单一与其他应用接口集成，为科研机构和用户提供身份管理、数据传输和共享以及组管理等服务模式。为进一步实现Globus数据管理的智能化，Globus从3个方面着力开发服务模块以适应数据管理的发展需求[18]：①高级数据搜索服务，支持对文件系统元数据和内部文件结构及内容的搜索，以更精细的数据索引粒度从文件中深度索引获得高质量的结果；②构建新的数据收集模型，以灵活的数据共享模型来集成数据共享和数据发布服务，实现用户的数据集合管理；③主动数据管理模块，开发一个模块化的主动数据管理环境，允许用户定义Globus生态系统中的行为规则。

3 Globus科研数据管理平台的服务案例

截至2021年7月，Globus在全球80多个国家和地区拥有12万余个注册用户和3万多个活跃端点，完成了1 244 031TB科研数据的传输和管理，拥有机构订阅用户100多个，其中包括60多所顶尖的研究型大学和DOE实验室[19]。Globus为多个科研机构及大型实验室实现了科研数据管理、同行之间的合作共享、精密仪器与数

据分析中心的数据传输，其服务模块也可被集成到其他数据管理平台。

Globus已被多个机构或研究项目用于科研数据管理的实践。例如，Globus为美国国家大气研究中心（NCAR）的“研究数据档案”（RDA）[20]数据服务提供了高效数据传输和用户认证管理。RDA主要收集气象和海洋观测数据，包含700多个数据集、800万个文件[21]，需要进行大量的数据分析输出，并为用户提供数据浏览和下载服务。RDA数据管理服务集成了Globus Transfer、Globus Auth微服务，实现了高效数据传输和身份管理功能。Globus为用户提供了简单的Web界面，通过专门的软件和GridFTP协议实现数据传输，数据传输过程自动完成，在发生系统故障后能恢复传输，确保数据传输的完整性，用户通过Globus监控数据传输量、时间戳、传输端点及传输文件状态。目前RDA内部端点数据传输速度达10GB/s，RDA与外部端点数据传输速度达2GB/s。同时，RDA采用了Globus身份管理和身份验证功能，用户可以使用GlobusID或Globus集成的其他身份链接登录RDA数据库，支持联合身份认证，改善了用户体验，为用户提供了易于使用、可靠、高性能的数据交付服务。

此外，Globus打破了不同机构之间的数据合作和共享的壁垒、实现了精密仪器与服务器之间的传输。例如：芝加哥大学测序中心与生物医学信息学中心（IBI）跨区域合作DNA测序项目，测序中心技术人员使用Globus移动到测序中心数据库，并通过Globus传输至IBI数据中心，科研人员即可在IBI设施中获取他们所需的DNA测序数据，实现了机构间的数据共享和项目合作；凯斯西储大学（CWRU）采用Globus将高性能Titan Krios透射电子显微镜产生的数据迅速转移到CWRU数据中心库，便于科研人员的数据调用和分析。Globus数据管理服务也被机构用于本地化数据管理平台的开发，加拿大计算机协会、Portagenetwork、加拿大研究图书馆协会合作利用Globus搭建了本地化科研数据管理平台——联邦研究数据存储库FRDR，构建了加拿大科学研究数据的收集、保存、访问和共享平台。

4 结语

在协作共享的大数据环境下，微服务架构和云存储应用于科研数据管理的便利性日益凸显。以管理机构联合科研机构、IT服务商、科研人员等利益相关者推动构建我国微服务架构云存储的科研数据管理平台，以解决目前多区域项目合作中数据传输低效和科研数据孤岛分布的现状，实现跨区域、跨学科的数据传输、共享和利用，对提高国内科研数据管理水平具有重要意义。基于微服务架构的Globus科研数据管理平台的服务模式为我国科研数据管理提供了良好的借鉴。

参考文献

[1] 吴金红，陈勇跃. 面向科研第四范式的科学数据监管体系研究[J]. 图书情报工作，2015，59（16）：11-17.

[2] MONS B. Invest 5% of research funds in ensuring data are reusable[J]. Nature，2020，578：491.

[3] 袁红卫，黄松，刘嫣. 麻省理工学院科学数据管理与共享平台调研及启示[J]. 图书馆学研究，2019（13）：82，95-101.

[4] HARVARD Dataverse[EB/OL]. [2021-11-01]. https://dataverse.harvard.edu/.

[5] Datastar[EB/OL]. [2021-11-01]. http://datastar.mannlib.cornell.edu/.

[6] Research Data Management for Purdue[EB/OL]. [2021-10-22]. https://purr.purdue.edu/.

[7] Scholarsphere[EB/OL]. [2021-10-22]. https://scholarsphere.psu.edu/.

[8] CHARD K，DART E，FOSTER I，et al. The modern research data portal：a design pattern for networked，data-intensive science[J]. Peerj Computer Science，2017，4（6）：e144.

[9] NEWMAN S. Building Microservices[EB/OL]. [2021-11-01]. https://www.oreilly.com/library/view/building-microservices/9781491950340/.

[10] 程秀峰，丁芬，夏立新. 基于微服务架构的文献信息资源保障平台构建研究[J]. 数字图书馆论坛，2021（4）：2-10.

[11] ALLEN B，ANANTHAKRISHNAN R，CHARD K，et al. Globus：A Case Study in Software as a Service for Scientists[C]// ScienceCloud’ 17. Washington：2017.

[12] CHARD K，TUECKE S，FOSTER I. Efficient and secure transfer，synchronization，and sharing of big data[J]. IEEE Cloud Computing，2015，1（3）：46-55.

[13] FOSTER I，VASILIADIS V，TUECKE S. Software as a Service as a path to software sustainability[EB/OL]. [2021-11-17]. https://www.globus.org/sites/default/files/saas-as-a-path-to-sustainable-software-delivery.pdf.

[14] Globus. Globus Connect[EB/OL]. [2021-11-12]. https://www.

globus.org/globus-connect.

[15] TUECKE S，ANANTHAKRISHNAN R，CHARD K，et al. Globus auth：A research identity and access management platform [C] //2016 IEEE 12th International Conference on e-Science（e-Science）. IEEE，2016.

[16] LIU Z，KETTIMUTHU R，CHUNG J，et al. Design and evaluation of a simple data interface for efficient data transfer across diverse storage [J]. ACM Transactions on Modeling and Performance Evaluation of Computing Systems，2021，6（1）：1-25.

[17] CHARD K，PRUYNE J，BLAISZIK B，et al. Globus Data Publication as a Service：Lowering Barriers to Reproducible Science [C] //2015 IEEE 11th International Conference on eScience. IEEE，2015.

[18] CHARD K，TUECKE S，FOSTER I. Globus：Recent Enhancements and Future Plans [C] //the XSEDE16. ACM，2016.

[19] Globus. The Globus Research Data Management Universe [EB/OL]. [2021-10-20]. https://www.globus.org/file/globus-research-data-management-universe.

[20] NCAR. Research Data Archive [EB/OL]. [2021-10-20]. https://rda.ucar.edu/.

[21] CRAM T. Globus integration in the NCAR RDA data portal：Recent enhancements [C] //In Globusworld 2018. Western Digital. Chicago：2018.

作者简介

袁晓明，女，1985年生，硕士，馆员，研究方向：信息咨询、学科服务、数据保存，E-mail：yuanxiaoming@suda.edu.cn。
王美琴，女，1975年生，硕士，副研究馆员，研究方向：图书情报、信息服务。

Analysis of Scientific Research Data Management Platform Based on Microservice Architecture

YUAN XiaoMing　WANG MeiQin
(Soochow University Library, Suzhou 215006, P. R. China)

Abstract: The explosive growth of scientific research data and remote collaboration and sharing put forward the demand for data management. The microservice architecture data management platform with easy elastic expansion, high-performance transmission and cloud hosting makes it possible for efficient data management. This paper investigates the Globus data management platform based on cloud storage developed by the Argonne National Laboratory research team of the University of Chicago. Globus scientific research data management platform adopts distributed micro service architecture, including four micro service modules: identity management, group service, data transmission and data sharing. The services are independent of each other, and scientific research data transmission and sharing are realized through visual Web interface. The platform has been applied to data management of multiple scientific research projects, and has completed terabyte file transmission. The data management mode and flexible data sharing of Globus can provide reference for scientific research data management in China.

Keywords: Research Data Management; Globus; Cloud service; SaaS; Microservice

（收稿日期：2021-11-01）