

基于数据网格的教育资源服务系统的构建¹

吴永和¹, 肖君², 王雁林¹

1.华东师范大学网络教育学院及现代远程教育研究中心, 2.上海远程教育集团

摘要: 本文研究应用网格技术来构建一个分布式网络教学资源服务系统, 即构建基于数据网格的教育资源服务系统 ERSDG, 分别从 Globus Toolkit 工具包、系统模型、系统架构、系统功能等方面系统地阐述了如何构建 ERSDG 系统。

关键词: 网络教育资源; Globus Toolkit; 数据网格; 教育资源库

教育资源是网络教育的基础, 国家已加大涵盖各级各类教育的信息资源开发, 形成多层次、多功能、交互式的国家教育资源服务体系, 如北京已投入 1800 多万元、上海首期投入 1600 万元、广东省预算投入 2350 万元, 所进行的基础教育资源库建设。如何将开发好教育资源方便、有效利用呢? 随着技术发展, 资源共享方式从 PC 的人为共享, 到计算机网络的自动共享, 再迈入网格 Grid 的智能化共享, 网格技术成为解决教育资源共享和方便利用重要的技术方法。本文研究应用 Globus Toolkit 工具包构建管理分布式网络教学资源 ERSDG (Education Resource Service DataGrid) 系统, 将上海教育资源从中心分布 (镜像) 到各区县节点上, 并为师生提供高效、高速的资源服务, 更好地共享和利用已建成的上海教育资源库。

1. Globus Toolkit 工具包

网格是继传统因特网、Web 之后的第三个大浪潮, 称之为第三代因特网, 它试图实现互联网上所有资源的全面连通, 包括计算资源、存储资源、通信资源、软件资源、信息资源、知识资源等。其中由全球网格论坛 (GGF) 下属 Globus 项目组成员联合开发的 Globus Toolkit 标准工具包, 已被公认为当前建立网格系统和开发网格软件事实上的参考标准。目前, 美国国家技术网格 NTG、欧洲数据网格、日本的数据农场 Data Farm 等项目都采用了 Globus 系统。

Globus 随着体系结构的变化经历了几次飞跃, 变得越来越完善 [1] [2]。自 1997 年起, Globus Toolkit 的 GT2 (五层沙漏结构) 已成为了网格计算的事实标准。随着 Web Service 技术发展, Globus 逐渐融入 Web Service 技术标准, 采用角色 (服务提供者、服务请求者、服务代理者) 和操作 (发布、查找、绑定) 方式, 其中三个操作采用不同的技术, 发布服务采用 UDDI (统一描述、发现和集成), 查找服务采用 UDDI 和 WSDL (Web Service 描述语言), 绑定服务使采用 WSDL 和 SOAP (简单对象访问协议)。2002 年, Globus 项目组推出了一个全新的网格标准 OGSA——开放网格服务体系, 把 Globus 标准与 Web Services 的标准结合起来, 网格服务统一以服务的方式对外界提供。随后推出符合 OGSA 规范的 Globus Toolkit 3.0 (GT3)。2004 年, 公布了统一网格计算和 Web 服务的新标准“Web 服务资源框架” WSRF 和“Web 服务通知” WSN; WSRF 是 OGSI 的重构和发展, 利用了新的 Web 服务标准, 而 WSN 为 Web 服务提供基于消息发布和预定的能力; WSRF 和 WSN 都是建立在已存在的 Web 服务定义和技术基础上的, 帮助实现了网格计算、系统管理和 Web 服务的统一。2005 年年初, Globus Toolkit 4 (GT4) 发布, 实现了 WSRF 和 WSN 标准。

¹ 本研究受 863 课题 (2004AA1Z2330) 和全国教育科学“十五”规划重点课题 (DKA010352) 支持。

GT4提供API来构建有状态的Web服务，其目标是建立分布式异构计算环境，其体系架构[3] [4]包括安全（Security）、数据管理（Data Management）、实施管理（Execution Management）、信息服务（Information Services）和公共运行环境（Common Runtime）等5个部分组件集，组件种类分为Web服务组件和非Web服务组件两大类。在安全组件集负责安全认证、身份鉴别、证书管理、安全委托和单点登录等，有公共认证CA、授权认证AA、代理Delegation、证书管理CM，CM包括静态的SimpleCA和动态的Myproxy；数据管理组件集负责数据传送和复制，从底层到高层有网格文件传输协议GridFTP、可靠定位服务RLS、可靠文件传输RFT和数据复制服务DRS等；实施管理组件集负责任务调度，主要有网格资源分配管理GRAM和远程控制GTP等；信息服务组件集负责系统系统信息数据收集和监控管理，采用监控发现服务MDS协议，包括web监控发现服务WebMDS、索引Index、触发器Tigger等；公共运行环境组件集是公共基础运行环境，包括Python、Java、C语言的WS Core和CCL等。

由于数据网格应用的需求迫切，Globus系统在原有面向计算网格的基础上增加了数据网格的功能，对数据的高速传输、数据复制、数据复制的选择、元数据管理等进行了研究和实现，成为数据网格应用的开发平台。本研究基于Globus Toolkit 4.0.1构建一个教育资源服务数据网格。

2. 系统模型设计

ERSDG 提供数据信息与资源文件同步数据功能和提供资源服务（访问）功能，其模型如图 1 所示，以三个节点为例来说明。

主节点（中央节点）通过教育资源访问口访问现有资源管理系统中的教育资源数据库，从资源库获取新的教育资源数据，同步到 ERSDG 中，将数据信息存储到数据库（资源信息）或轻量目录访问协议 LADP（用户信息）中。由 GRAM 进行自动的统一调度，数据管理的高层应用 DRS 提供服务，透明地调用 RFT 和 RLS，

再调用 GridFTP，将教育资源文件同步到各节点上，实现教育资源自动分布和同步功能。

提供教育资源访问服务。由 GRAM 进行统一调度，由 SimpleCA 进行安全认证，由 RFT 将分布在各节点上教育资源文件按最佳算法提供给用户，即根据就近原则和系统及网络负载情况，在最佳的节点处复制下载教育资源。利用 MDS、RFS、RLS、GridFTP 等协议和智能工具实现资源下载服务功能。

在监控管理方面，由 MDS 的收集各节点的状态信息，汇集到中央节点，调用 WSMDS 发布相关监控信息，由免费 Java 组件 JFreeChart 工具提供图形显示。

由于资源文件分布在各个节点上，方便有效管理，需要有一个全局的命名空间，对系统中的文件进行统一命名，将文件的物理特征与逻辑视图独立开，使物理层次上的变化与逻辑层上的变化互不影响。系统对数据采用三种命名空间，即用户层面的逻辑文件名 Logical File Name (LFN) 和分布在各节点上的物理文件名 Physical File Name (PFN)以及内部资源数据对

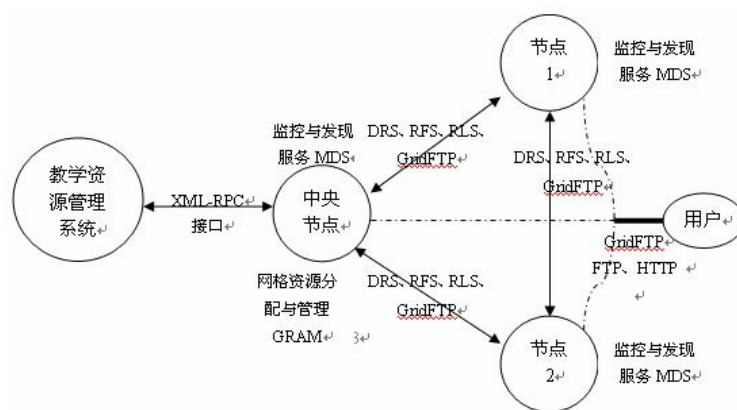


图 1 基于数据网格的教育资源服务系统 ERSDG 模型

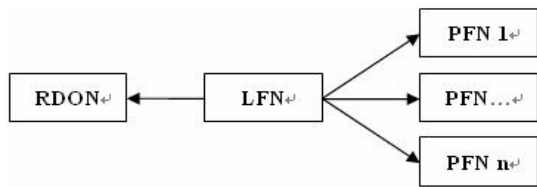


图 2 文件名字空间的组成

象名 Resource Data Object Name (RDON)。用户逻辑文件名 LFN：面向用户的、在用户的逻辑视图中所使用的文件名称。物理文件名 PFN：分布在各节点上文件实体，通过 URL 对该文件进行访问。内部资源数据对象名

RDON：内部资源数据对象的全局唯一标识符，与资源 LOM 属性描述关联，特别是基于教育资源建设规范 CELTS-41 和基础教育教学资源元数据应用规范 CELTS-42 的资源属性，是在系统内部使用的文件名称，在数据实体的整个生命周期中保持不变。它们之间的关系用图 2 来表示。用户从用户逻辑文件名获知用户的逻辑视图中的文件名称；并与 RDON 关联，获取资源属性描述；与 PFN 关联，访问分布在各节点上物理文件，对应的 PFN 有多个。

3. 系统架构设计

ERSDG 将分布在互联网上的教育资源整合成具有统一逻辑视图的教育资源服务系统。整个系统主要包括教育资源服务点 ERSP(Education Resource Service Point)、教育资源数据同步系统 ERDS(Education Resource Data Synchronization)、全局命名服务器 GNS(Global Name Server)、认证服务中心 CAS(Certificate Authority Service)、教育资源管理器 ERM(Education Resource Manager)、教育资源服务代理 ERSA(Education Resource Service Agent)、客户端以及可视化管理 [5][6]，如图 3 所示。

ERSP 是整个系统的入口，都通过它访问系统所有模块，主要提供数据同步 API、FTP、CAS、ERM 和 GNS 等接口；系统中 ERSP 的个数可根据需要动态增加。

ERDS 访问原有教育资源数据库通讯的 API 接口，从资源数据库获取新的数据信息和教育资源，同步到 ERS DG 中。

GNS 负责系统的元数据管理，主要包括元数据操作接口、元数据容错系统、元数据搜索系统。

CAS 包含证书管理系统，主要负责系统的安全性和数据的访问控制。

ERM 包括教育资源调度模块和副本管理模块，其主要负责资源的申请和调度，同时提供透明的副本创建和选择策略，自动地实现教育资源分布和节点间教育资源同步。

ERSA 透明地提供多样性资源访问，为系统提供统一资源访问接口，同时提供了文件操作方式和扩展的 FTP 操作方式，并对文件复制管理操作提供支持，为高效传输提供服务。

客户端目前支持二种形式：通用 FTP 客户端和特制客户端。用户通过系统提供的特制客户端，不但能够具有搜索和共享等功能，还可以获得高性能的服务。

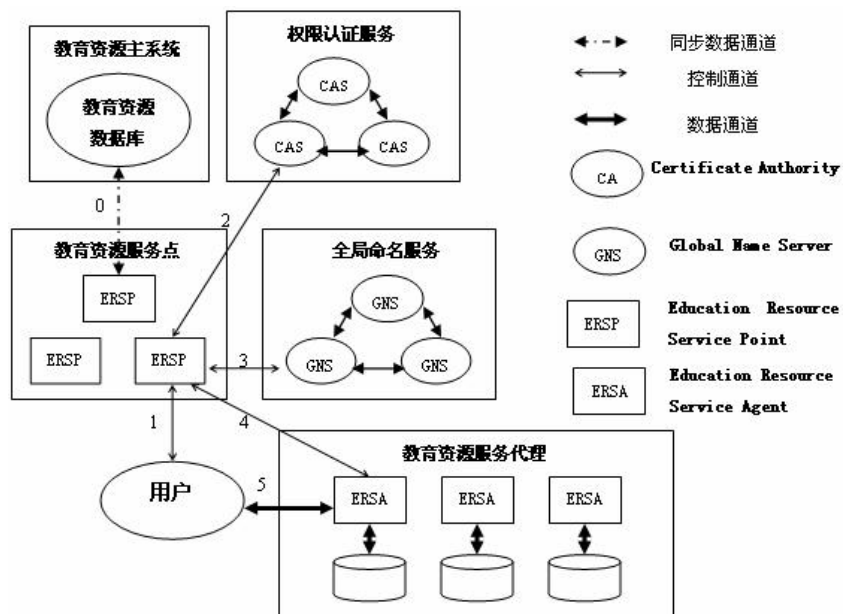


图 3 基于数据网络的教育资源服务系统 ERS DG 架构

其中系统同步数据和用户访问流程如图 4 的三个通道标志所示,即在双箭头线上所标识的数字为该访问步骤。

系统同步数据:ERDS 将现有教育资源数据库的信息和资源文件同步到 ERSDG 系统中(步骤 0)。

用户访问资源流程:用户首先访问整个系统的入口 ERSP(步骤 1),通过 ERSP 访问 CAS,获得认证(步骤 2),认证通过后,由 ERSP 访问 GNS 获得资源信息(步骤 3),根据获得资源信息,再由 ERSP 访问 ERSA(步骤 4),ERSA 给用户 提供资源访问服务(步骤 5)。

4. 系统功能设计

ERSDG 功能包括教育资源服务点、信息服务器、命名服务器、教育资源管理器、资源服务代理、客户端和系统监控等部分 [5][6][7],如表 1 所示。

基于数据网格的教育资源管理应用系统 ERDataGrid																				
教育资源服务点					信息服务器		命名服务器			教育资源管理器		资源服务代理		系统监控			客户端			
教育	用户	文件	终端	GNS	ERM	证书	全局	元数据	元数据	元数据	资源	副本	资源	广域	系统	ERS	ERM	GNS	标准	特定
资源	管理	管理	服务	通信	通信	管理	信息	数据	数据	数据	调度	管理	调度	网传	总控	服务器	服务器	服务器	FTP	客户
数据	模块	模块	模块	模块	模块	模块	管理	服务	容错	搜索	模块	模块	模块	传输	模块	监控	监控	监控	客户端	端
同步							模块	模块	模块	模块				控制		模块	模块	模块	客户端	客户端
接口																				
模块																				

表 1 教育资源数据网格 ERSDG 系统功能模块

教育资源服务点包括教育资源数据同步接口模块、用户管理模块、文件管理模块、终端服务模块、GNS 通信模块和 ERM 通信模块等。利用 Delegation 开发代理功能模块,由 GRAM 进行统一调度。

信息服务器包括证书管理模块和全局信息管理模块,由 SimpleCA 来实现安全认证,利用 MDS 收集信息并汇集到中央节点。命名服务器包括元数据服务模块、元数据容错模块和元数据搜索模块,通过 MDS 协议 Index、Tigger 来实现。

教育资源管理器包括资源调度模块和副本管理模块,利用 GRAM、GridFTP、RLS、RFT 和 DRS 来实现。建立了一个在广域网上的高效数据传输机制,包括分布式合作传输、分片传输、部分数据传输和断点续传等。

资源服务代理包括资源调度模块和广域网传输控制,利用 GRAM、GridFTP、RLS、RFT 来实现。ERSA 实现了资源的虚拟化,分布式的虚拟化和内部共享管理机制,并根据用户的特征而自动变化创建策略,为用户提供高效、灵活性的资源访问服务方式。

客户端包括标准 FTP 客户端和特定客户端,利用 GridFTP 及其客户端(如 UberFTP、globus-url-copy)来实现。

系统监控包括 ERSP 服务器、ERM 服务器、ERM 服务器和 GNS 服务器等监控模块,利用 WebMDS 调用监控信息,并由 JFreeChart 工具显示图形。

5. 总结

本文研究构建 ERS DG 系统, 选择 Red Hat Linux 操作系统, 采用 Web service 资源框架 WSRF 作为网格系统架构, 利用国际主流的 Globus Toolkit 4.0.1 网格平台及相关开发工具包, 用 Java 语言进行编程开发, 按照该系统模型设计、系统架构设计和系统功能设计的要求, 实现数据信息与资源文件等数据同步功能和提供资源服务功能, 具体包括分布式网络教学资源动态管理、节点间教学资源自复制和访问教学资源应用动态的调度等, 提供一个虚拟化网络教学资源分布应用环境; 为上海地区师生提供高性能的教育资源服务, 如在全国建立适当的分布节点, 将能为全国师生提供通畅便捷的资源服务, 实现教育资源全面的共享。

参考文献:

- [1] 网格计算 都志辉, 刘鹏 清华大学出版社, 2002
- [2] 网格服务体系结构的演变 邹德清 金海 计算机用户 第 2 期 2005.1
- [3] Borja Sotomayor The Globus Toolkit 4 Programmer's Tutorial <http://www.globus.org/toolkit/docs/>
- [4] A Globus Primer <http://www.globus.org/toolkit/docs/> 2005.5.8
- [5] 华中科技大学集群与网格计算湖北省重点实验室 信息存储系统教育部重点实验室
虚拟化存储系统 2003 年 12 月
- [6] 肖依 付伟 黄斌 卢锡城 Griddaen 数据网格系统的设计与与关键技术实现
<http://www.chinagrid.com/>
- [7] GT4 Admin Guide <http://www.globus.org/toolkit/docs/4.0/April 2005>