

一种基于网关的网格数据互操作框架

颜秉珩¹, 钱德沛^{1,2}

(1. 西安交通大学电子与信息工程学院, 710049, 西安; 2. 北京航空航天大学计算机学院, 100083, 北京)

摘要: 针对网格数据互操作中存在的存储资源异构、作业需求多样、故障多发等问题, 提出了一种基于网关的数据互操作框架——数据网关. 它是专门的数据互操作解决方案, 可以独立处理数据互操作作业, 促进网格互操作任务的并行化处理. 数据网关采用了一种基于 ClassAD 的作业描述语言, 可以描述任意的服务、数据或者约束条件, 并引入了插件式的扩展机制, 大大提高了数据网关对异构存储资源的扩展性. 数据网关的动态协议选择机制和故障恢复机制可以为上层应用提供一个良好的运行环境, 由此增强了系统的可用性. 实验表明, 数据网关可以快速实现异构网格设施之间的数据互操作, 在无人值守的情况下自动处理各种底层网络和服务故障, 有效地完成各种数据互操作任务.

关键词: 数据互操作; 协议转换; 数据网关

中图分类号: TP393 **文献标志码:** A **文章编号:** 0253-987X(2009)02-0020-05

A Gateway-Based Framework for Data Interoperation between Grids

YAN Bingheng¹, QIAN Depei^{1,2}

(1. School of Electronics and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China;

2. School of Computer Science, Beihang University, Beijing 100083, China)

Abstract: A data-gateway, a gateway-based framework, is proposed to solve the problems in grid data interoperation, such as heterogeneous storage resources, different job requirements, and numerous failures. The data-gateway is a professional solver for data interoperation that can make the data interoperation as an independent and general modular separated from job inter-operation, and promote parallel processing of the interoperation tasks. A job description language based on ClassAD is introduced in the data-gateway, which can provide a very flexible and extensible data model, and can be used to describe arbitrary services and constraints, as well as the scale mechanism in a plug-in style so that the scalability of the data-gateway for heterogeneous storage resources is improved well. Dynamic protocol selecting mechanism and failure recovering mechanism used in the date-gateway can provide a nice environment like failure-free network and storage devices, and enhance the system's availability. Experiment results show that the data-gateway can implement data interoperation between heterogeneous grid infrastructures rapidly, recover failures automatically without any human interaction, and complete all kinds of data interoperation tasks efficiently.

Keywords: data interoperation; protocol translation; data gateway

在过去的若干年, 不同网格的独立发展导致了多种技术的产生, 这使得异构网格之间无法进行有

效的互操作. 如何将现有的网格无缝地融合起来, 实现更大范围内的共享与协作, 是网格互操作所要解

决的主要问题. 一般来说, 网格互操作主要涉及 4 个方面: 作业互操作、数据互操作、信息服务以及与安全相关的内容, 其中数据互操作是作业互操作的基础, 是网格数据服务的延伸.

目前的网格互操作方案是将数据互操作与作业互操作耦合在一起处理, 不利于二者的并行化执行. 将计算互操作任务与数据互操作任务区分对待是合理的, 因为它们有着不同的语义和特性^[1]. 现有的计算互操作任务的调度模块显然不能很好地理解数据互操作任务的语义.

那么, 如何将数据互操作从作业互操作中分离出来, 形成一个通用的模块, 就需要将数据互操作任务由专门调度、管理这种作业的网关来处理, 而计算互操作任务仍由现有的网关处理, 从而实现网格互操作任务并行化处理. 基于这个目的, 本文提出了一种基于网关的数据互操作或称数据网关(DG)框架.

1 数据互操作面临的问题

在异构的网格设施之间实现数据互操作, 将面临以下问题.

(1) 异构地存储资源. 网格应用所需的数据分散在不同的存储系统之中, 执行应用需要与各种各样的存储系统通过交互来完成. 那么, 借助于一个自适应的中间系统来访问不同种类的存储系统, 充分利用不同的底层中间件/传输协议是十分必要的.

(2) 建立多元的安全机制. 一个完整的网格应用需要访问多个存储系统, 完成多种不同的认证过程, 跨网格的应用更是如此. 这就需要为上层应用提供一个统一的认证入口, 以屏蔽底层不同的安全界面, 使应用专注于业务逻辑, 摒弃繁杂的安全细节.

(3) 解决无处不在的系统故障. 网络连接错误、传输性能不稳定、客户端/服务端/存储系统崩溃等故障, 随时都会在网格应用中出现. 不能期望网格应用能预防和处理多种类型的错误, 必须由系统将故障屏蔽起来, 由此为应用提供良好的运行环境.

(4) 满足不同的作业需求. 网格互操作的特点决定了作业的多样性, 仅有全局性的策略是不够的, 每个作业还需要制定自己特定的策略, 以满足具体的需求. 调度模块依赖这些信息, 所以一种灵活的作业描述方式是不可缺少的.

2 数据网关设计

数据网关的主要目标是, 独立处理网格数据互操作作业, 提供一个良好的作业运行环境.

2.1 数据网关的架构

为了提高系统的扩展性和吞吐率, 数据网关采用了 Master/Worker 的架构模式. Master 是整个数据网关的入口, Worker 是一系列的工作单元, 负责完成数据网关中的具体作业. 图 1 展示了数据网关的架构. Master 是一个全局的调度中心、信息中心和日志仓库, 它负责将作业分发到一个合适的 Worker 上, 同时收集所有注册到它上面的 Worker 的信息, 从而为 Worker 进一步调度作业提供必要的信息. 为了提高系统的可用性, 可以增加备份 Master. Worker 是横跨不同存储系统的桥梁, 它利用一个统一的接口来访问异构的网络资源, 完成具体的数据互操作任务.

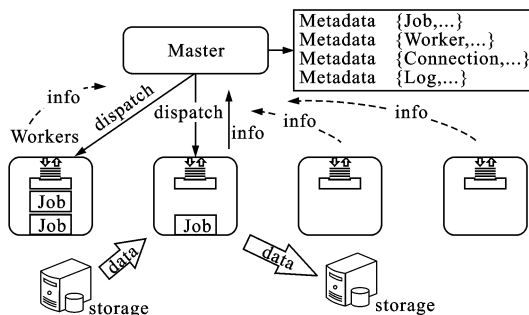


图 1 数据网关的架构

2.2 异构存储资源的访问

为了与网络中的异构资源交互, 我们采用了 Apache Commons 虚拟文件系统(VFS)作为 Worker 访问异构存储系统的统一接口. Apache VFS 为访问不同的文件系统提供了单一的应用程序接口(API). 得益于模块化的设计, VFS 可以很容易地进行扩展, 以支持其他的存储系统或者传输协议. 这一点对于工作在异构网格环境中的系统来说尤为重要, 因为我们不期望数据网关与所有的存储系统、传输协议直接交互.

VFS 屏蔽了底层的异构特性, 使得数据网关可以用统一的接口去访问这些异构资源. VFS 的模块化设计允许用户添加插件来支持不同的存储系统/中间件或者协议, 这也从整体上提高了数据网关对不同传输协议的扩展性.

基于 VFS 提供的统一接口, Worker 可以对各种存储系统进行读/写操作. 数据互操作作业到达 Worker 以后, Worker 启动读写线程, 利用 VFS 完成协议转换, 读线程负责从源存储系统到数据网关的数据传输, 而写线程负责从数据网关到目标存储系统的数据传输.

2.3 多策略支持

为了满足数据互操作环境下的各种作业需求,比如不同的作业类型、不同的安全认证模式、不同的作业策略等等,除了数据网关功能,本文还设计了一种基于 ClassAD^[2]的作业描述语言(JDL),用来提供灵活的、可扩展的数据模型,描述任意的服务、数据或者约束条件。

图 2 所示的 JDL 文件描述了一个典型的数据互操作作业,在 2 个不同的存储系统之间传输数据,作业类型是拷贝目录,采用续传模式,并指定了可选的传输协议以及必要的认证信息.其他的作业类型有创建文件/目录、删除文件/目录,这些作业类型都是根据数据互操作的流程需求准备的。

```
[
  JobType="cpdir";
  Arguments="append";

  SrcHost="euchina 07.buaa.edu.cn";
  SrcPath="/tmp/slc306";
  SrcSchemas={ [ Protocol="GSIFTP";
    User = ":globus-mapping:";
    Password = "file://src_certificate_file_path";
  ],
  [ Protocol = "FTP";
    User = "*****"
  ] };

  DstHost="buaa01.ihep.ac.cn";
  DstPath="/tmp";
  DstSchemas={ [ Protocol="GSIFTP";
    User = ":globus-mapping:";
    Password = "file://dst_certificate_file_path";
  ] };
]
```

图 2 一个复制作业的描述文件

认证方式依赖于具体使用的协议/存储系统.在网格中认证方式有两类,一类是类似于 FTP 的用户/密码认证类型,另一类是类似于 GridFTP 的证书认证类型,前者可以直接在 JDL 中声明,而后者需要说明如何获取相应的代理证书文件,而其他认证方式可由此扩展.本文的数据网关采用一个四元组 (Protocol, User, Password, Port) 来描述协议与认证信息等数据,四元组中的 User 与 Password 的含义随着 Protocol 的不同而有所改变,Worker 在读取这些信息之后,通过恰当的配置 VFS 来完成存储系统的认证过程。

在数据网关中策略分为全局策略和作业级策略,以满足不同作业的需求.通过设置全局策略,可以指定一些通用的策略.用户也可以通过设置 JDL 的作业级策略来覆盖全局策略,满足自己特定的需求.在 JDL 中设置的作业级策略实例如图 3 所示。

在这个例子中,用户将作业的优先级设置为“高”(在数据网关中有高、中、低 3 种优先级),并采

```
[
  JobType="cpfile";
  :
  Priority="high";
  BufferChunks=16;
  RetryCount=3;
  ExpireTime="10 h";
]
```

图 3 JDL 中设置的作业级策略

用 16 个缓冲区来平滑读写之间的速度差异,如果作业失败或者超过 10 h 仍未完成的话,将重启作业,但重试次数不超过 3 次。

2.4 运行时支持

在执行作业时,数据网关通过测试相关协议的工作状态,动态地决定采用何种传输协议来完成任务.用户也可以给出存储系统所支持的协议列表.如果使用某种协议传输失败,数据网关将会尝试使用列表中的其他协议继续传输数据,如图 4 所示。

```
[
  :
  SrcSchemas={ [ Protocol="GSIFTP";...],
    [ Protocol="FTP";... ] };
  DstSchemas={ [ Protocol="GRIAS";...],
    [ Protocol="HTTPS";... ] };
  :
]
```

图 4 协议列表

在上面的例子中,用户给定了源存储系统和目标存储系统支持的传输协议列表.如果使用 GSIFTP 协议访问源存储系统失败,数据网关会尝试用 FTP 协议来读取源数据.同样地,数据网关会选择一个可用的协议将数据写入目标存储系统。

2.5 错误恢复

数据网关的设计目标之一便是屏蔽所有网络、存储系统、中间件或者硬件的错误,提供一个良好的运行环境.为此,数据网关设计了两种错误恢复机制:一种是故障重启,即当数据互操作作业遇到故障而失败的时候重启这个作业;另外一种是超时重启,允许用户为他们的数据互操作作业设置一个可以接受的最大运行时间,一旦作业的运行时间超过了这个限制,数据网关就会终止并重启这个作业,这种机制可以克服不少系统的 Bug,尤其是那些导致传输操作挂起而又没有响应的情况.这里的“重启”是指处理那些没有完成的数据。

2.6 本地数据迁移

为了支持工作节点与存储系统之间的数据互操作,我们开发了一个轻量级的“LOCAL”协议,它可将工作节点转化为一个临时的存储节点,而工作节

点就可以像其他普通的存储节点一样与数据网关交互,在任务完成之后自动关闭“LOCAL”服务,以保证安全。

“LOCAL”协议的用法如图5所示,执行这个作业会把数据下载到提交作业的节点上,它与普通的JDL并没有太大的区别。

```
[
  JobType="cpfile";
  :
  DstPath="C:/data";
  DstSchemas={ [Protocol="LOCAL";...]; };
  :
]
```

图5 “LOCAL”协议的用法

2.7 数据网关的安全机制

数据网关可保证用户与数据网关之间的安全通信、作业与作业之间的安全隔离。数据网关到存储系统的安全访问由相应的访问协议自行保证,因为我们无法改变存储系统固有的安全机制,比如数据网关访问使用 GridFTP 协议的存储系统要比使用 FTP 协议的存储系统更安全。

数据网关用 SSL(安全套接字层)连接来保证数据网关与用户之间的安全通信,确保 JDL 的相关信息(如 FTP 的认证信息)不被泄露,并用签发代理证书机制来保证作业代表用户执行操作的合法性。用户在提交一个作业之后,作业会要求用户为它签发一个代理证书,以便代表用户完成存储系统的操作。对于采用证书认证方式的存储系统,作业可以直接使用这个代理证书进行访问,也可以使用用户指定的其他证书。数据网关还能有效隔离作业的私有数据空间,保证作业相关信息(如私钥/证书)的安全。

3 实验与分析

通过3个实验可以说明数据网关是如何适应网络特性并顺利完成数据互操作任务的,实验环境如图6所示,包括实验涉及的中间件,相关站点支持的传输协议、平台信息和物理位置。数据网关与 IHEP(中科院高能物理所)节点的传输速度大约是 5 MB/s,在 JSI(中德软件所)内部是百兆位局域网,便携电脑(Windows XP, Service Pack 2)位于 JSI 内部。在这个环境中我们验证了数据网关处理网格互操作任务的功能,包括不同存储系统之间的数据传输、错误恢复、动态协议选择和本地数据迁移。

3.1 数据传输与错误恢复

由于 gLite 和 GOS 不能直接交互,所以需要使用数据网关桥接二者的存储节点。实验结果表明,数

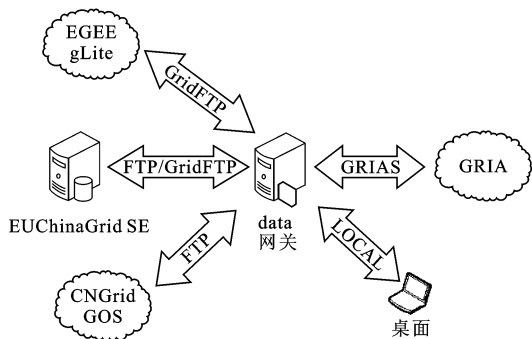


图6 实验环境

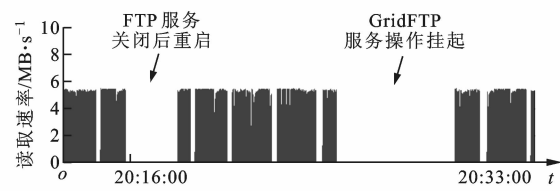
据网关最终在2个异构的存储系统之间成功地建立了一个数据通道,并在没有人工干预的情况下完成了大约 120 GB 的数据传输任务。

从 gLite 到 GOS 的数据传输过程中,数据网关遇到了很多故障,包括软件、硬件方面的网络连接失败、读写操作挂起、拒绝读写数据、服务器崩溃等。无论哪种错误,数据网关都会自动终止当前作业并启动故障恢复机制而无需人工干预。图7是作业执行过程中传输速度变化的一个片段。首先 GOS FTP 服务重启,接着数据网关到 GridFTP 的操作很长时间没有响应,直到 GridFTP 服务恢复正常,期间还夹杂着一些网络故障。但是,数据网关的错误恢复机制能很好地处理这些故障。

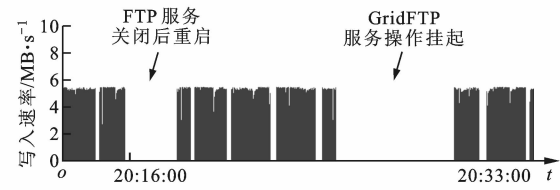
3.2 动态协议选择

在 EUChinaGrid^[3] 项目中,提供了一个支持 FTP 和 GridFTP 传输协议的存储节点 EUChinaGrid SE。网格中间件 GRIA^[4] 拥有自己独特的数据服务(GRIAS),如图6所示,它们不能直接交互完成数据互操作。

本实验中测试了数据网关的动态协议选择功能,其中提交了100个从 EUChinaGrid SE 到 GRIA 的数据拷贝作业。实验开始的时候,SE 上面的 GridFTP 服务与 FTP 服务都是可用的。数据网关的读操作选择了 GridFTP 协议,写操作选择了 GRIAS 协议。在作业执行了一段时间之后,关闭了 SE 上面的 GridFTP 服务,数据网关很快切换到 FTP 服务继续传输数据,切换之后数据传输速度有所下降,如图8所示,这是因为 FTP 服务的传输速率被限制在 2 MB/s 以内。数据网关采用 FTP 协议经过一段时间之后,重新启动了 SE 上面的 GridFTP 服务,由于 GridFTP 协议的优先级高于 FTP 协议,所以数据网关重新采用 GridFTP 协议传输数据,传输速度也逐渐恢复到了原来的水平。这个操作重复几次,观察到的结果均与图8相似。



(a)从 gLite 到网关



(b)从网关到 GOS

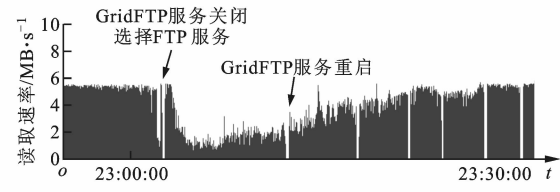
图 7 错误恢复机制自动处理系统故障

实验结果表明,如果存储节点支持多种传输协议,将大大增加系统的可用性,数据互操作也可以充分利用这些服务,在较高性能的传输服务发生故障后,采用其他的传输服务继续执行作业。

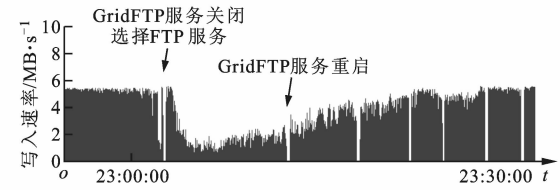
3.3 本地数据迁移

实验提交了一个从 EUChinaGrid SE 到本地(图 3 中的便携电脑)文件系统的拷贝作业。

实验开始时,SE 的 2 种传输服务都是可用的,本地与数据网关的连接也是正常的.经过一段时间后,本地与数据网关之间进行了 3 次断开/恢复操作,如图 9 所示.从中可以看到作业中断了 3 次,待到网络恢复正常时,数据网关会继续执行作业,并不需要人工干预.最后,重启 SE 上面的 GridFTP 服务来验证数据网关的动态协议选择功能.实验表明,本地文件系统与其他存储系统的数据互操作可以从数据网关中获益。

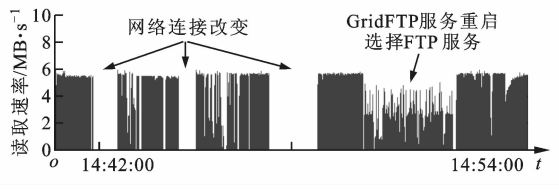


(a)从 EUChinaGrid SE 到网关

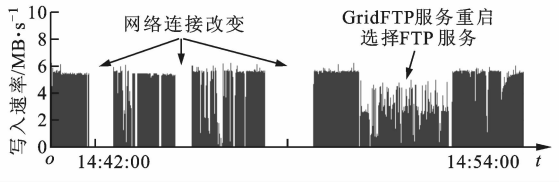


(b)从网关到 GRIA

图 8 动态协议选择



(a)从 EUChinaGrid SE 到本地文件系统



(b)从网关到本地文件系统

图 9 本地数据迁移、错误恢复和动态协议选择

4 结束语

本文的数据网关是一种具有良好通用性、扩展性的数据互操作框架,对各种网格数据互操作环境均有很好的适应能力,可以有效解决网格数据互操作中的各种难题.它将数据互操作从作业互操作中分离出来,促进了网格互操作的并行化处理。

参考文献:

[1] KOSAR T, LIVNY M. Stork: making data placement a first class citizen in the grid [C] // Proceedings of 24th IEEE International Conference on Distributed Computing Systems. Piscataway, NJ, USA: IEEE, 2004: 342-349.

[2] RAMAN R, LIVNY M, SOLOMON M. Matchmaking: distributed resource management for high throughput computing [C] // Proceedings of the 7th IEEE International Symposium on High Performance Distributed Computing. Piscataway, NJ, USA: IEEE, 1998: 28-31.

[3] YAN Bingheng, WU Zhongxin, YANG Dongbo, et al. Experiences with the EUChinaGrid project-implementing interoperability between gLite and GOS [C] // The 2nd IEEE Asia-Pacific Service Computing Conference. Piscataway, NJ, USA: IEEE, 2007: 224-231.

[4] SURRIDGE M, TAYLOR S, ROURE D D, et al. Experiences with GRIA: industrial applications on a web services grid [C] // 1st International Conference on e-Science and Grid Computing. Piscataway, NJ, USA: IEEE, 2005: 98-105.

(编辑 苗凌)