

遥感数据异地归档方法研究与实现

李斌斌^{①②}, 冯钟葵^①, 唐梦辉^①

(^① 中国科学院对地观测与数字地球科学中心, 北京 100094; ^② 中国科学院研究生院, 北京 100049)

摘要: 遥感数据异地备份是在数据中心和备份端各自保存一份遥感数据, 并提供数据恢复手段的容灾技术。遥感数据异地归档是遥感数据异地备份的基本组成部分, 是实现遥感数据异地备份的关键。为实现遥感数据异地备份中异地归档功能, 本文针对异地归档过程中的难点和问题, 提出了遥感数据异地归档方法, 并从技术角度出发提出解决方案。

关键词: 遥感数据; 异地; 归档; 备份

doi: 10.3969/j.issn.1000-3177.2011.02.015

中图分类号: TP79 **文献标识码:** A **文章编号:** 1000-3177(2011)114-0086-05

Research and Implementation of Remote Archiving Method of Remote Sensing Data

LI Bin bin^{①②}, FENG Zhong kui^①, TANG Meng hui^①

(^① The Center for Earth Observation and Digital Earth, Beijing 100086;

(^② Graduate University of the Chinese Academy of Sciences, Beijing 100049)

Abstract: The remote sensing data remote backup(RSDRB) is a kind of disaster recovery technology, which saves the remote sensing data both in the data center and backup station and provides a means of data recovery. The remote archiving method of remote sensing data is an essential component of RSDRB and key to implement RSDRB. To achieve remote archiving function of RSDRB, this paper presents the remote archiving method of remote sensing data. Then this paper analyses the method and puts forward solutions of the difficulties and problems in the process of remote archiving.

Key words: remote sensing data; remote; archiving; backup

1 引言

当前, 数据安全问题已经成为国民经济各部门日益关注的问题, 而数据异地备份是解决这一问题普遍采用的技术。数据异地备份是指在两个相距一定距离的空间, 各自保存一份关键业务数据, 并提供数据恢复手段的数据安全存储方式。两地之间的距离常根据工作条件、数据的重要性、以及行政区划等因素统筹考虑^[1]。

遥感数据具有种类多, 覆盖面广、保存时间长等特点, 无论是遥感原始数据还是遥感数据成果都是

极其珍贵的遥感数据资料。如何有效地保存和利用这些珍贵的遥感数据, 确保数据的安全性和可访问能力, 是国内外遥感领域面临的重要课题。美国的 USGS、法国 CNES 等遥感数据接收和运行机构已经建立了数据安全存储体系, 国内一些机构也在积极规划采用数据异地备份的方式进行数据的安全存储与管理。遥感数据异地备份是在数据中心和备份端各自保存一份遥感数据, 在任何一端数据出现损坏时, 能快速恢复数据, 从而增强遥感数据安全性。遥感数据异地备份由数据的异地归档和数据的恢复两个基本部分组成, 其中数据异地归档是指在备份

收稿日期: 2010-01-29 修订日期: 2010-03-11

作者简介: 李斌斌(1984-), 男, 研究方向: 遥感信息处理。

E mail: constantheart@163.com

端复制存储一份数据中心的遥感数据,并在两端存储遥感数据的编目信息;数据恢复是指在任何一端发生灾难或数据损坏后,从另一端获取相应的遥感数据,然后重新对本数据进行编目登记工作,更新数据库信息。异地归档是遥感数据异地备份的基础,为实现数据恢复功能提供了数据和信息支撑。

为实现将遥感数据在异地的安全存储与管理,本文针对异地归档过程中的难点和问题,提出了遥感数据的异地归档方法,并从技术角度给出具体的解决方案。

2 遥感数据异地归档的特点

遥感数据的归档包含了对数据进行整理、质量评估、登记或编目以及存储数据等多个环节,是形成完整的数据归档、保存的工作过程^[2]。遥感数据的异地归档是从数据异地备份技术中衍生出来的方法,既涉及到计算机领域的数据备份等技术手段,又涉及到遥感领域的归档工作方法,是利用计算机手段实现遥感数据的异地存储和编目的工作方法。由于遥感数据与其他业务数据相比,具有数据量大,种类多等特点,当前,遥感数据的异地备份与银行、电信等业务的备份系统有所不同,还未提出严格的RPO(Recovery Point Object)、RTO(Recovery Time Object)等指标要求。遥感数据的异地归档主要关注数据中心和备份端两端系统的协调、数据在异地的存储和编目等问题,其具有以下几方面的特点:

(1) 遥感数据的数据量巨大、种类繁多,共享资源存在使用竞争,归档策略难以确定。

在进行异地归档工作时,数据中心一般已经存有一定量的数据。尤其对遥感数据来说,其数据量可能是海量的,部分数据文件可达几GB,从而需要占有大量的传输和存储资源。另一方面,遥感数据种类繁多,不同的遥感卫星有不同类型的数据,同一颗卫星的遥感数据也会因波段、分辨率的不同而有差异^[3]。因此,大量的遥感数据的异地归档是无法一次完成的,只能采取一定措施,有步骤、有计划地分别归档。此外,遥感数据的异地归档工作需要读取数据中心的遥感数据,并且利用传输网络进行数据传输,这些都有可能产生对共享资源的使用冲突。因此,在部署异地归档工作时,既要考虑如何利用资源快速、有效地实现遥感数据的异地归档工作,又要避免对共享资源使用冲突。

(2) 数据中心和备份端系统相互独立,异地归档过程复杂,工作协调困难。

遥感数据异地归档过程复杂,涉及到数据的提取、编目、传输、存储等多个动作,而且需要数据中心和备份端两地共同协作才能完成。但是,数据中心和备份端的系统是相互独立的,不是从属关系,无法由一端独立完成遥感数据异地归档工作。因此,在如何协调两地工作、完整有序地完成各个工作内容方面存在着困难。

(3) 在异地归档过程中,遥感数据存在质量隐患。

在异地归档过程中,需要进行数据复制、数据传输、数据异地存储等。在此期间都有可能出现问题,导致数据出现质量问题,如数据不完整,数据错误等。数据的质量问题将直接影响到异地归档的结果,最终可能导致整个异地备份工作失去意义。因此,在数据归档过程中,需要保证遥感数据质量的可靠性。

3 遥感数据异地归档方法研究

遥感数据异地归档方法是按一定的归档策略和控制方法,协调数据中心和备份端两端的工作,完成对遥感数据的异地存储和在两端对遥感数据的数据编目工作。针对遥感数据异地归档过程中的难点,本文将从归档策略、质量校验和控制方法三个方面来分析,并给出解决方案。

3.1 归档策略

遥感数据的异地归档策略是指对遥感数据进行备份的工作方法,包括了备份窗口、备份次序和备份方式三个方面^[4]。通过对归档策略的研究,可解决因遥感数据量大、种类繁多、共享资源使用竞争等问题给异地归档带来的困难。

(1) 备份窗口

备份窗口是指用于备份的时间段。确定备份窗口的目的,主要是避免与其他业务系统争用共享资源,如存储数据的磁带库和传输数据的网络资源等。备份窗口的大小,可根据对数据的实际操作特性来设定,例如,如果数据从早上8点到午夜被使用,则从午夜到次日早上8点之间的时间段可用于备份。在可用备份时间段内,备份窗口由要备份数据的总量和读取、传输数据的速度等因素决定。备份窗口的计算可由以下公式表达:

$$W = S/V \quad (1)$$

其中: W 是需要的备份窗口, 该值不能大于系统所给定的备份时间段; S 是需要备份的数据量; V 是由传输网络, 存储结构等因素决定的备份速度。

备份窗口因不同的应用要求和工作平台而不同, 需要对实际的应用系统进行充分的测试和调研来确定。

(2) 备份次序

备份次序是指按一定次序, 将遥感数据进行归档备份, 确保备份端对数据中心所有数据进行了备份, 保持两端数据的一致性。一般情况下, 备份系统的建立要滞后于业务系统, 在建设异地备份系统时, 数据中心可能已经存储大量的业务数据, 尤其对遥感数据的异地备份来说, 遥感数据的数据量大, 种类繁多, 因此备份次序就显得尤为重要了。遥感数据的备份次序有多种方式, 如按照数据存入存储系统的时间、数据的种类、数据的存储路径等。可以根据遥感数据的实际存储情况, 按照简单、可靠的原则做出选择。例如, 若遥感数据是按照数据类型存储在不同的存储空间, 此时可以按数据存储时间和数据种类相结合的方式, 依次将所有数据进行异地归档, 具体方法如下:

设集合 $D = \{D_i | 1 \leq i \leq N\}$ 表示所有遥感数据的集合, 其中 $D_i = \{d_1, d_2, \dots, d_j, \dots\}$ 表示同一种类遥感数据的集合, d_j 表示某条遥感数据, t 为 d_j 上的时间属性, 半开半闭区间 $T = (t_1, t_2]$ 表示时间段。

首先, 按实际需要按对 D 中元素进行优先级排序, 如 $D = \{D_1, D_2, \dots, D_N\}$;

然后, 设定初始值: $i = 1, t_1 = -\infty, t_2$ 根据实际情况设定。

最后, 利用公式(2)选出 D_i 中在 T 时间段内的所有需要归档的遥感数据 D_a 。

$$D_a = \{d_j | d_j \in D_i \text{ 且 } t[d_j] \in T\} \quad (2)$$

如果 $i < N$, 则令 $i = i + 1$, 继续利用公式②选择归档数据; 如果 $i \geq N$, 则令 $i = 1, t_1 = t_2, t_2$ 根据实际情况另行设定, 然后继续利用公式(2)选择归档数据。

(3) 备份方式

在确定备份窗口和备份次序后, 需要根据实际情况来制定备份方式。较为常见的备份方式有完全备份、增量备份和差异备份三种。

完全备份忽略两次备份操作之间数据的变化情况, 是一种每次都对备份目标执行完整备份的方式,

在备份后, 每个文件都被标记为已备份。增量备份只备份上次备份后增加或修改的部分, 每次备份都基于上次备份进行, 增量备份过程中, 只备份那些未备份的数据, 备份后标记为已备份。差异备份同增量备份类似, 也是只备份两次备份之间改变的部分, 与增量备份的区别是, 差异备份在备份后不将已经备份的数据标记为已备份。

遥感数据异地归档的对象是遥感数据, 其目的是在异地保存一份遥感数据的拷贝。从这个角度出发, 在初次异地归档时, 采用完全备份策略, 即将数据中心中的所有遥感数据拷贝到备份端。当然, 数据中心的数据可能是海量的, 其拷贝过程需要花费大量的时间, 这就需要考虑备份次序和备份窗口等问题。当数据中心所有的数据在备份端完成拷贝后, 采用增量备份的策略, 及时将数据中心不断接收的新数据拷贝到备份端。

3.2 质量校验

质量校验是异地归档过程中的重要内容, 它关系到数据的安全存储, 并影响着后续工作的开展。因此, 在遥感数据存入备份端存储系统前, 必须保证遥感数据的质量是可靠的, 本文利用信息摘要实现了对遥感数据的质量校验。信息摘要是把任意长度的输入信息揉和产生长度固定的摘要的算法, 有如下特点^[5]:

①无论输入的信息有多长, 计算出来的信息摘要的长度总是固定的;

②一般地, 只要输入的信息不同, 对其进行摘要以后产生的摘要信息也必不相同; 但相同的输入必会产生相同的输出;

③信息摘要函数是单向函数, 即只能进行正向的信息摘要, 而无法从摘要中恢复出任何的信息, 甚至根本就找不到任何与原信息相关的信息。

本文利用信息摘要, 对遥感数据执行质量检查, 确定遥感数据的可靠性, 算法为:

①在数据中心使用信息摘要算法对遥感数据生成信息摘要 R_1 ;

②数据中心把遥感数据及其信息摘要 R_1 一起发送至备份端;

③备份端通过使用与数据中心相同的信息摘要算法, 对接收的遥感数据生成新的信息摘要 R_2 ;

④对比 R_1 与 R_2 , 以确认遥感数据是否被修改过。如果 $R_1 = R_2$, 说明遥感数据质量是可靠的, 否则说明遥感数据在传输过程中发生了改变。

3.3 控制方法

遥感数据的异地归档包括本地编目、数据传输、异地存储和异地编目四个环节。

本地编目: 在数据中心, 通过一定方式提取遥感数据的元数据, 建立本地信息库; 数据传输: 采用一定的传输技术手段, 将数据中心已经完成本地编目的遥感数据拷贝到备份端; 异地存储: 按一定的存储方式, 将遥感数据存储到备份端的存储系统中; 异地编目: 在备份端对所存储的遥感数据进行编目, 获取所需要的元数据, 建立异地信息库。

通过对控制方法的研究, 采用合适的控制技术和协调机制, 协调数据中心和备份端的工作, 最终完成数据在备份端的存储和归档信息的建立。

(1) 控制技术^[6]

控制技术分为同步复制和异步复制。同步复制技术是指将数据中心数据以完全同步的方式复制到备份端, 每一次数据中心的 I/O 事务均需等待远程复制的完成确认信息, 方予以释放。利用同步传输方式进行备份端归档, 可以保证在一端出现灾难时, 另一端保存一份与数据中心数据完全一致的数据备份, 但该方法对原有的系统性能影响较大, 对硬件条件要求高, 实现技术复杂。

异步复制技术的基本思想是, 首先在数据中心完成基本 I/O 操作, 然后由数据中心发送“I/O 操作完成”确认信息给备份端, 最后更新备份端存储。相对于同步复制技术, 异步复制技术具有对数据中心系统性能影响小, 对网络带宽要求小等特点。其缺点是, 备份端数据的存储时间要滞后于数据中心的数据存储, 当数据中心发生灾难时, 可能造成部分数据无法恢复, 但是滞后时间是可预知和可控的, 根据实际条件的不同, 滞后时间从几小时到几天不等。

如果采用同步复制技术, 由于遥感数据的数据量巨大, 需要高速的数据读写和传输技术, 不仅代价高而且实现技术复杂。特别是利用原有的业务系统进行扩展完成遥感数据的异地归档时, 同步复制技术对原有业务系统的影响是明显的。而且, 一般情况下, 在数据中心遥感数据需要经过一系列的整理、质量评估、登记或编目等处理过程, 然后进入数据中心存储系统, 此过程是自动进行的, 一般无法人工干预, 以至无法实施同步复制技术。

异步复制技术在数据中心发生灾难时有损失数

据的风险, 但损失的数据是可控的, 而且它克服了同步复制技术的缺点, 因此, 遥感数据的异地归档可采用异步复制技术, 具体为:

①在数据中心完成本地编目;

②通知备份端, 本地编目工作完成;

③备份端向数据中心发送确认信息, 准备接收遥感数据;

④传输数据, 将相关的遥感数据及其元数据信息传输到备份端;

⑤异地存储, 确定遥感数据在备份端存储路径, 完成数据在备份端的存储;

⑥在备份端完成异地编目。

(2) 基于消息队列的工作协调机制

消息队列技术是分布式应用间交换信息的一种技术。基于消息队列的存储-转发机制, 使得消息的发送方和接收方松耦合, 进行异步消息传递, 从而实现各应用模块间的数据交换和事务协调^[7]。

消息队列为应用模块之间的进程间通信提供了强大灵活的机制。与模块间的直接调用相比, 它们具有若干优点, 其中包括:

稳定性: 应用模块失败对消息的影响程度远小于模块间的直接调用, 因为消息存储在队列中并一直留在那里, 直到被适当地处理;

消息优先级: 更紧急或更重要的消息可在相对不重要的消息之前接收, 因此可以为关键的应用程序保证足够的响应时间;

脱机能力: 发送消息时, 它们被发送到队列中并一直留在那里, 直到被成功地传递。当因任何原因对所需队列的访问不可用时, 用户可以继续执行操作。同时, 其他操作可以继续执行, 如同消息已经得到了处理一样, 这是因为网络连接恢复时消息传递是有保证的。

通过消息队列, 在数据中心和备份端内部依次分步执行各个工作模块, 并且使得数据中心和备份端两个控制系统建立联系, 以异步的工作模式协调两端的工作, 最终由相对独立的数据中心系统和备份端系统协调完成遥感数据异地归档工作。消息队列协调机制的工作原理如图 1 所示。

工作协调方法如下:

①数据中心控制系统向队列 1 发送启动命令, 启动本地编目工作; 然后从队列 2 中读取本地编目结果, 如果本地编目成功则继续, 否则退出;

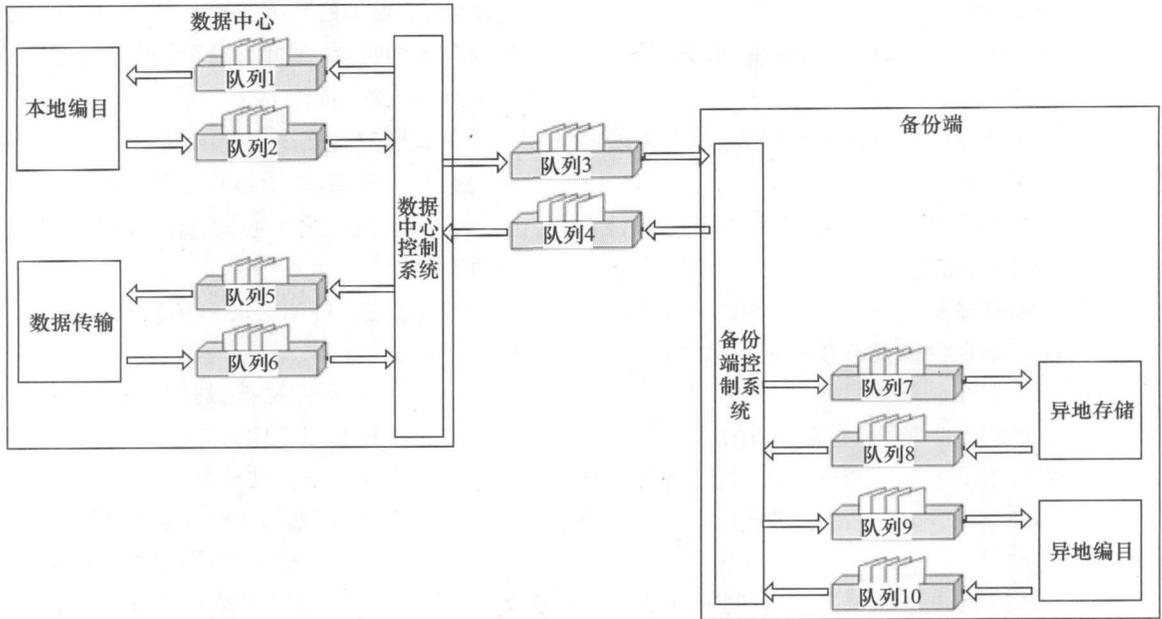


图1 消息队列协调机制的工作原理图

②数据中心控制系统向队列3发送通知消息,通知备份端准备接收数据;然后从队列4中读取备份端的反馈信息,如果备份端同意归档则继续,否则退出;

③数据中心控制系统向队列5发送启动命令,启动数据传输工作;然后从队列6中读取数据传输结果,如果成功则表明遥感数据异地归档在数据中心的工作完成;

④备份端控制系统在接收到遥感数据后,向队列7发送启动命令,启动异地存储工作;然后从队列8中读取异地存储的结果,如果失败则退出;

⑤备份端控制系统向队列9发送启动命令,启动异地编目工作;然后从队列10中读取异地编目的结果,如果异地编目成功则表明遥感数据异地归档工作完成。

4 遥感数据异地归档方法实现

(1) 采用 NcFtp 技术传输数据

NcFtp 是基于 Ftp 协议的传输工具,它具有高的传输速度和可靠性,并且支持目录文件传输,使用方便。

(2) 采用 MD5 信息摘要算法校验遥感数据传输质量^[8]

信息摘要常用算法有 MD5、SHA-1、REPEMD-160。对于数据传输的应用来讲,它们都具有良好的可靠性,但在处理速度上有明显的差别,表1是在同一平台下,采用不同摘要算法对 10GB 数据的处理时间对比。

表1 信息摘要方法处理速度对比

算法	时间(s)	试验速度(MBps)
MD5	255	40.2
SHA-1	300	34.1
REPEMD-160	325	31.5

由上表可知 MD5 具有最快的运行速度,因此在实现过程中,本文选择 MD5。

(3) 使用消息中间件实现消息队列协调机制

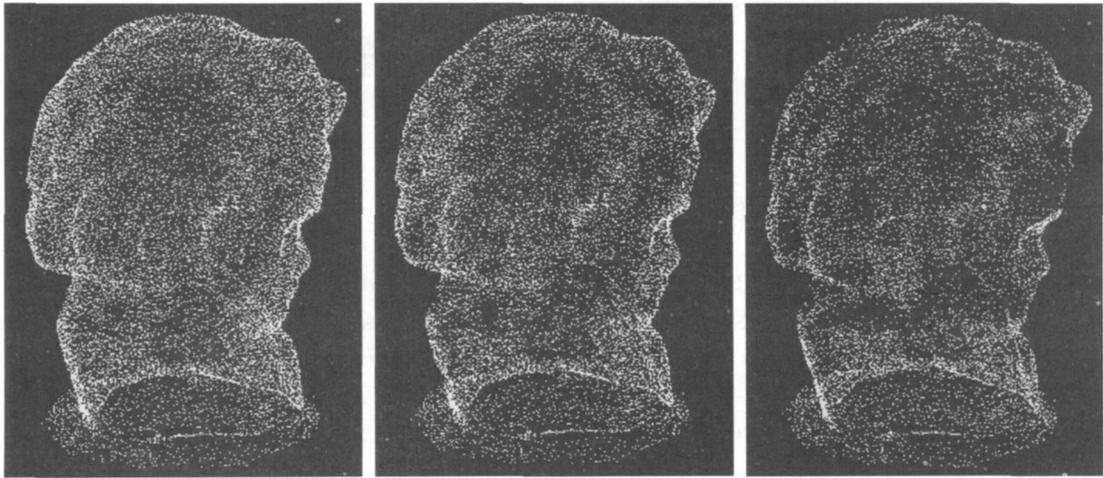
消息中间件是一个独立运行的通信环境,有统一的编程调用接口,可以跨平台、跨协议,它的地位在操作系统之上,在应用系统之下,所以被称为消息中间件。本文选择了应用广泛,稳定可靠的 IBM WebSphere MQ 作为消息中间件。

按照第3章提出的解决方案,并通过以上技术,有效地解决了遥感数据异地归档过程中的关键问题,协调了数据中心和备份端两端的工作,实现了遥感数据的稳定传输,并保证了备份数据的可靠性。

5 结束语

遥感数据异地归档是实现遥感数据异地备份不可或缺的重要组成部分,也是解决遥感数据异地备份过程中数据异地归档问题和难点的技术保障。本文针对遥感数据异地归档过程中的难点,从归档策略、质量控制、控制方法三个方面分析遥感数据异地归档方法,并给出了解决方案。本文的分析和介绍将对数据异地备份工作,尤其是遥感数据的异地备份工作起到积极意义。

(下转第 105 页)



(a)压缩率为 62.6%的点云图

(b)压缩率为 70.9%的点云图

(c)压缩率为 78.9%的点云图

图 6 塑像压缩后的点云图

6 结束语

针对点云数据量大的特点,本文提出了基于不规则格网的点云缩减算法。采用球面投影的方法先建立规则格网,然后依据格网内顶点法矢的标准差大小作为细分准则构建出不规则格网,这样压缩的结果就会在曲率变化大的位置保留较多的点,而在平缓的位置保留较少的点,从而在数据压缩的同时保留了点云数据的几何信息,使得采用压缩后的点进行模型重建具有很高的保真度,也即保留了点云

数据的真实几何信息。同时,采用球面投影的方法建立网格不仅适用于单站数据,对于已经拼接到一起的多站数据只要找到合适的物体中心,同样可以应用。本文实验中的塑像数据为一个封闭的整体,找到封闭体的中心,以此作为球心即可应用本方法进行数据缩减;而对于单站点云数据,则直接以坐标原点作为球心即可。本文所述的压缩算法特别适用于表面不规则的物体,而在现实世界中,表面不规则的物体占有很大比重,所以这种压缩方法会有很大的应用前景。

参考文献

- [1] 郑德华. 三维激光扫描数据处理的理论与方法[D]. 上海: 同济大学, 2005.
- [2] 官云兰. 地面三维激光扫描数据处理中的若干问题研究[D]. 上海: 同济大学, 2008.
- [3] 周晓云. 散乱数据几何造型技术研究[D]. 北京: 北京航空航天大学, 1993.
- [4] 朱心雄, 等. 自由曲线曲面造型技术[M]. 北京: 科学出版社, 2000: 215- 230.
- [5] 顾孝烈, 鲍峰, 程效军. 测量学[M]. 上海: 同济大学出版社, 2006: 112.
- [6] 武汉测绘科技大学测量平差教研室. 测量平差基础[M]. 北京: 测绘出版社, 1996: 66.
- [7] 金涛, 董水光, 等. 逆向工程技术[M]. 北京: 机械工业出版社, 2003: 128- 179.

(上接第 90 页)

参考文献

- [1] 田永飞. 区域数据中心重要业务系统的异地备份方法研究[J]. 华南金融电脑, 2009(5): 10- 12.
- [2] 冯钟葵, 张洪群, 王万玉, 等. 遥感卫星数据获取与处理关键技术概述[J]. 遥感信息, 2008(4): 91- 97.
- [3] 汪超亮. 遥感卫星数据编目存档系统的研制[J]. 地球信息科学, 2005, 7(2): 116- 119.
- [4] 杨鹏锐, 张延园, 牟虹燕. 数据容灾系统的原型设计及实现[J]. 微电子学与计算机, 2005, 25(5): 46- 50.
- [5] 张亚玲, 王尚平, 王育民, 等. 信息摘要算法 MDA-192[J]. 计算机工程与应用, 2002(24): 86- 93.
- [6] 郑国昌. 同步、异步数据容灾的性能分析[J]. 现代计算机, 2003(4): 85- 87.
- [7] 嵇智辉, 倪宏, 刘磊. 动态消息队列负载均衡策略的研究与应用[J]. 计算机工程, 2009, 35(8): 31- 37.
- [8] 张振权, 罗新民, 齐春. 数字签名算法 MD5 和 SHA-1 的比较及其 AVR 优化实现[J]. 网络安全技术与应用, 2005(7): 64- 67.