NSTL集中备份体系框架设计
——大数据安全实践*

□ 张婧 梁冰 / 中国科学技术信息研究所 北京 100038
胡铁军 / 国家科技图书文献中心 北京 100038

摘要：NSTL海量科技文献数字资源正在以每年20%的速度增长，同时网络应用服务系统产生并累积了大量系统日志。这些宝贵的科技文献资源及具有数据挖掘价值的日志数据还处于分散管理、缺乏有效备份和快速恢复技术手段的状态，为确保国家科技文献信息国家战略保障体系的平稳运行，亟需建立、建成一个应用于NSTL城域网的统一的集中备份体系，以保护这些数据的安全。文章介绍了NSTL集中备份体系架构的基本设想和实现的技术方案。

关键词：数据备份，数据库网络，虚拟化，存储，大数据
DOI：10.3772/j.issn.1673—2286.2012.11.004

1 概述

“大数据”是虚拟技术、云计算和数据中心三者使用率增加后的逻辑衍生物，都能在计算资源的标准化、整合和集中化上发挥杠杆作用，从而实现规模经济，也帮助成本效益的实现。而NSTL拥有国内最大的科技文献实体馆藏，经过十余年的努力，NSTL在资源建设、数据加工和网络服务系统方面取得了长足的进步，并建立了拥有41个服务站、网络覆盖全国的科技文献信息服务体系，为用户提供科技信息检索与文献传递等全方位的文献信息服务，目前正致力于采用虚拟化技术进行数据中心建设，其中面临的问题之一就是如何保障在经过标准化、整合和集中化后形成的“大数据”的安全。

由网络服务系统、国际科技引文检索系统、回溯数据服务系统以及科技信息资源与服务集成揭示系统、嵌入式资源集成服务系统组成的NSTL科技文献信息网络服务系统，历经多次改造升级，已经发展成为集中外文科技期刊、会议文献、学位论文、科技报告、专利、标准和计量规程等文献信息于一身，文摘、引文、题录数据总量达1.8亿条，资源丰富、品种齐全、功能友好、高度开放、全文传递、具有自主知识产权的大型公益性国家科技数字图书馆。仅自建数字资源就以每年3500万条的速度在增长，数据中心存储容量约65TB，数据量约35TB，是国内科技信息行业最大的数据中心之一。面对日益发展的数字化环境，科技文献的出版形式、用户需求和信息服务模式持续变化，对承担科技文献国家战略保障体系建设历史重任的NSTL提出了新的要求和挑战。

十余年来，NSTL积累了海量的数字资源，且正在以每年20%的增速增长，同时还有数个基于互联网的业务系统在24小时提供服务。但这些珍贵的历史数据和宝贵的实时数据以及具有数据分析价值的日志数据还处于分散管理、缺乏有效的备份措施和手段的状态。一旦发生关键存储设备失效或其他意外事故，就会带来不可弥补的损失和风险。同时，随着数据总量日益庞大，已经不能用简单或单一的办法（例如一块硬盘或U盘）来实现有效的完整数据备份，从而保证时间间隔稳定，保障备份的有效性和完整性，实现数据安全和版本控制。因此，一个集中统一的备份系统，对提高国家科技文献IT应用水平的作用是显而易见的。特别是在当今的全面信息化社会，为确保国家战略信息资源的

* 基金项目：十二五科技支撑计划项目技术创新服务平台关键技术研究与应用示范（编号：2011BAH30B01）。

存储应用安全，亟需建立、建成一个覆盖NSTL城域网的、满足NSTL长远发展需要的备份体系。

2 集中备份体系设计的目标

2.1 建立具有可扩展性的集中备份体系、兼顾长远发展需要

截至2012年，NSTL建成并投入运行的网络服务系统包括原文检索服务系统、回溯数据系统、联合联机编目系统、文献综合系统、数据加工系统，这些系统不论是在线服务数据、用户请求数据，抑或是网络系统日志数据，目前都还是各系统独立进行备份和存放，没有采取集中统一的备份措施。即，十余年来，NSTL采购的、加工的、服务系统使用的、各单位认为应该保存的数据资源，以及其他需要NSTL集中保存的数据资源，都应该但还没有进行有效的集中存放。

通过有针对性地设计集中备份体系，对包括在线系统的服务数据（包含文件和数据库）、在线系统的日志数据（文件型）、NSTL采购的各类数据（光盘版、电子版等）、NSTL加工的各类数据（文件型）、各单位单独保存的各类数据资源（光盘版、电子版等）等，依托NSTL城域网使NSTL的十余年来宝贵的战略资源实现集中存放、统一保管、可靠保存，是符合NSTL未来的发展需要的。

从当前数据备份关注的技术发展来看，采取自动备份与物理磁带库和虚拟磁带库的有机结合，或是基于虚拟化、云计算，都是可以选择的方向。根据NSTL各业务管理系统特点和业务逻辑，在考虑适用性的前提下，结合目前虚拟化技术的发展和应用，设计NSTL备份体系，不仅是能够满足现有业务需要，还必须考虑应用新的技术，充分利用这些先进成熟的技术手段或方法，借鉴同业应用的成熟经验，设计并实现NSTL信息系统的集中备份管理。结合当今大数据时代的发展，以超前实用的原则，既能实现NSTL的集中备份，又能够为NSTL下一步异地容灾备份系统建设积累经验和技术，满足未来5到10年业务发展需要。

2.2 采用先进、适用的集中备份技术

网络环境的不确定因素对数字科技文献资源的安全提出了严峻挑战，数字资源长期保存问题如果不能很好地解决，将会严重影响到我国对数字化科技信息资源持久地获取与利用。文献信息服务机构已经逐渐意识到并开始进行数字资源长期可靠保存的研究。就NSTL各类数据资源进行集中备份的需求，结合当前IT技术的最新发展，充分利用先进的技术手段，建成一个架构合理、技术先进、适用性强、扩展性好的NSTL集中备份体系，从而推进国家数字资源长期保存体系的建设，建立数字资源的国家合作保存服务体系具有战略意义。

3 NSTL备份体系框架设计

3.1 NSTL数据备份的现状和需求

NSTL拥有众多业务系统，一方面通过因特网提供7×24在线文献信息服务，另一方面对内部工作网用户提供业务信息处理支持。同时，各成员单位既有购置的各类数据库资源，还有个性化的业务软件，为用户提供定制数据。目前采用以下几种备份手段：

（1）特别重要的网络服务系统，采取双系统备份方式。提供特定目的的服务系统，仅采用在磁盘阵列上采取镜像的方式防止硬盘损坏导致数据丢失。数据库系统和文件类数据，除原始数据存放在阵列上以外，中间数据和加载数据均不保存。有关用户访问行为的日志数据仅存放在运行系统中，不导出另行存放。

（2）各独立的业务系统中的各种类型数据，原始数据存放在阵列上。而由于部分系统数据量很大，数据资源保存的完整性和有效性不能得到保证。

（3）自行采购的数据库资源基本上是以原始介质（光盘）方式存放；对自行加工的原始数据资源，各单位保存方式各不相同。

综上所述，NSTL自行采购的、自主加工的、提供服务的各个业务系统的以及其他需要NSTL集中保存的数据资源，目前还处于分散保存、缺乏集中管理的状态。

因此，亟需建立一套统一的集中备份体系，从而将分散保存的数据有效、完整地保存，在系统或数据出现问题时能够及时有效地进行快速恢复，确保服务的连续性。

NSTL需要保存的数据类型可分为以下几类：

（1）关系型数据库的在线数据，包括SQL

Server、Oracle等关系型数据库；

（2）业务服务系统的在线数据文件，包括AIX、Linux和Windows等系统；

（3）业务服务系统的原始数据、中间文件和加载文件；

（4）业务服务系统的日志文件，包含WebLogic、负载均衡、Apache和Tomcat；

（5）网络设备和网络安全设备的日志文件，包含华三、思科等各类设备；

（6）NSTL采购的数据库资源（光盘版）；

（7）NSTL采购的文件型资源（文件型，光盘版）；

（8）NSTL加工的电子版文件型数据；

（9）其他各单位认为应该保存的电子版文件型数据。

根据上述不同类型的数据特点和应用方式，我们提出了一套可行的集中备份体系设计方案，既便于数据备份和恢复，又方便提交数据保存的用户自行查找和下载，同时还能满足集中保存数据的安全、有效和完整性。

3.2 备份和存储技术的现状和发展趋势

3.2.1 备份技术的现状和发展趋势

如今，备份技术的成熟度已经达到可以大规模普及的程度，简化部署这些产品的新技术也已经出现。

（1）重复数据删除技术

重复数据删除技术的出现改变了传统的数据备份方式，称得上是数据存储影响力最大的技术变革。它的优点是大大地节省备份磁盘空间，在当今大数据时代，这种技术显得尤为重要。

（2）DPM数据保护管理

DPM功能能提供远非简单的备份报告（告诉用户哪些备份已经完成，哪些没有）。除了具备基本一般的备份软件都有的报告功能，涉及趋势预测、容量规划、跨产品报告和跨产品问题分析等超出了传统备份产品的功能就需要有专业化的DPM软件实现。这一功能对于复杂的业务系统数据备份是非常重要的。

（3）虚拟服务器备份

服务器虚拟化技术的应用让许多数据中心用户从中获益。太多的应用程序要求使用“专用的服务器”，不过实际上虚拟机也可以被认为是“专用的服务器”。大多数应用对CPU和I/O的要求并不高，在服务器虚拟化产品的帮助下，共享的硬件资源就可以满足性能要求。

3.2.2 存储技术的现状和发展趋势

（1）高速I/O技术

存储高速I/O技术包括PCI-E（PCI Express）、SAS（Serial Attached SCSI）、iSCSI（Internet SCSI）、FCoE（Fibre Channel over Ethernet）。

目前PCI-E 3.0是8.0Gbps；SAS 2.0是6.0Gbps；Ethernet已经达到了10Gbps；FC（Fibre Channel）当前速率只有8Gbps，并且后续没有看到明确的目标，因此iSCSI和FCoE借助40G/100G Ethernet的标准化，必将成为主流。

（2）硬盘技术

SATA硬盘继续朝大容量方向发展，目前已经有单盘3TB的硬盘量产，相信后续还会有更大容量的硬盘商用；SAS硬盘逐步夺取原来FC硬盘的市场，主流的硬盘厂家已经不再发展FC硬盘，取而代之的就是SAS硬盘，2.5英寸的SAS硬盘会逐步成为主流，SAS硬盘的容量也在逐步增大；SSD盘发展迅速，高可靠和高性能是其显著的特点，虽然目前价格还比较高而容量还比较小，随着技术的发展，其价格会逐步下降，同时容量也会大幅度提高。

（3）硬盘节能技术

在传统硬盘方面，同等容量的2.5英寸硬盘的功耗是3.5英寸硬盘的一半；SSD盘由于没有机械部分，其工作和空闲时的功耗都得到了大幅度的降低；在存储设备系统方面，主流厂商都逐步采用Spin down技术，使得长时间没有I/O访问的硬盘进入到Spin down（磁盘低功耗运行）状态，从而大大降低系统的功耗，这在备份应用领域尤其适用。

（4）非对称双活控制器（Asymmetric Active/Active Controller）

在新的SCSI协议标准里，增加了ALUA（Asymmetric Logic Unit Access）特性，即允许多个控制器访问同一个逻辑单元。最新的OS版本里已经开始支持这个特性，如Windows 2008、HP-UX 11iv3、Solaris 9/10、Linux、AIX 6.1和VMware。这就要求在双控制器的存储系统中，要支持两个控制器非对称地访问同一个逻辑单元。

（5）数据迁移

目前存储硬盘有SSD盘、SAS盘或FC盘以及SATA盘，SSD盘的特点是性能高、功耗低，但是容量小、价格高；SAS盘和FC盘的特点是性能中等、容量也中等，价格适中，功耗大；SATA盘的特点是容量大、价格低，但是性能低、功耗大。在一个存储系统里可以同时存在这三种介质，分成Tier0（SSD）、Tier1（SAS/FC）和Tier2（SATA）三个层次，存储系统可以根据主机访问的不同，调整数据在不同的Tier之间进行迁移，从而满足主机对存储性能的要求。

（6）自动精简配置

存储系统的资源规划是一件非常难以确定的事情，很难做到准确的资源规划，经常是规划多了或者少了，这样导致设备要经常调整资源，对系统影响比较大。针对这个问题，目前存储系统开始支持自动精简配置，用户在初始配置资源时可以配置得比较大，而实际上存储系统会根据实际需要来分配资源，在资源不够时可以自动增加资源，从而简化了设备的配置，提高资源的利用率。

3.3 NSTL备份体系设计

NSTL需要备份的数字资源分散在中心和各分中心，且格式多样，有在线数据库、各种类型的文件、光盘、系统运行实时生成的日志等。中心系统与各分中心组成城域网，大部分备份到中心的数字资源还需要提供访问页面，供上传数据的分中心在需要时下载。因此，简单地利用备份软件不能全面实现这些资源备份。根据以上特点，经过对NSTL业务体系的分析和市场调研，确定了NSTL数据备份的技术路线和备份系统框架。

3.3.1 技术路线

（1）对于各单位需自行提交的文件型数据资源，基于WWW技术定制开发集中保存模块，实现数据资源自主上传、关键字检索、按需下载和资源分配统计；

（2）对于在线服务系统的在用数据（在线数据库和在线文件），利用成熟的企业级备份软件和虚拟带库构建集中备份模块，实现关键数据自动地实时备份；

（3）对于设备和软件的实时日志资源，基于syslog-ng搭建日志集中收集模块，实现用户行为记录的集中保存；

（4）基于FCSAN/IPSAN架构、存储间复制技术和集中备份模块，建立NSTL集中的数字资源储存池，集中保存上述各类数据，并定期进行备份，保证数据的安全；

（5）定期对各模块进行数据恢复功能进行测试，保证各类数据可储存、可使用、可恢复。

3.3.2 NSTL备份体系基础架构

根据技术路线要求，NSTL备份体系基础架构包含了4个方面，如图1所示。

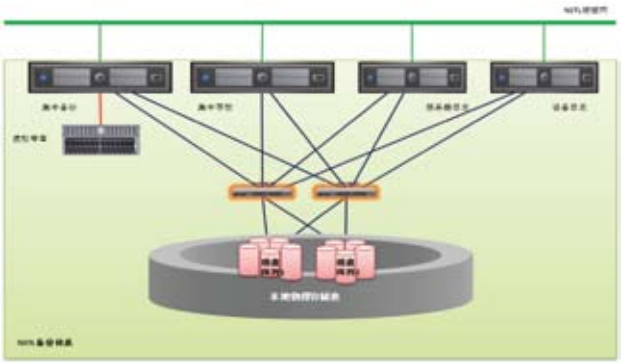

图1 NSTL备份体系基础架构

3.3.3 NSTL备份体系集中备份模块

NSTL文献服务、联机联合编目、引文服务、回溯服务、文献综合管理、联合数据加工等业务系统分布在UNIX、linux、Windows等不同的平台上，数据库以Oracle、TRIP、SQL Server、MySQL为主，对操作系统及业务数据和数据库的数据备份工作处于半自动半人工操作，管理繁琐复杂，且各系统备份数据分散存放，备份数据的安全性和可用性得不到保障。

从当前IT行业实际应用情况看，传统的企业级备份软件已非常成熟，在NSTL虚拟化架构中可直接部署企业级备份软件，即可完成NSTL集中备份系统所承担的整个网络服务系统的集中备份管理工作。具体实施可以以企业级备份管理系统（例如EMC Networker）为核心，采用高性能虚拟带库实现，将备份服务软件构造在虚拟化架构中的逻辑服务器上，形成备份服务器服务端。客户端通过网络将需要备份的数据，根据备份策略，传输到备份服务器上，备份服务器按照备份策略中

的设置，将数据存放到虚拟带库中。选择虚拟带库可使数据读出和写入速度较传统带库提高百倍，因此通过这种方式实现的备份能够完全满足数据恢复的要求。该模块用于在线数据库（Oracle、SQL Server、MySQL）、业务服务系统所属操作系统文件和在线服务文件等。这样通过集中备份管理系统自动灵活地完成每天繁重的数据及系统的备份，可极大地提高业务系统数据的安全性和可用性，为NSTL的数字资源提供安全保障。

3.3.4 NSTL备份体系集中保存模块

大量由NSTL采购的、NSTL加工的、各单位认为应该保存的数据以及其他需要NSTL集中保存的数据在备份到中心系统的同时，还需要提供访问页面便于数据上传单位根据各自的需要进行下载，因此NSTL备份体系集中保存模块主要应用流程设计（上传和下载），如图2所示。

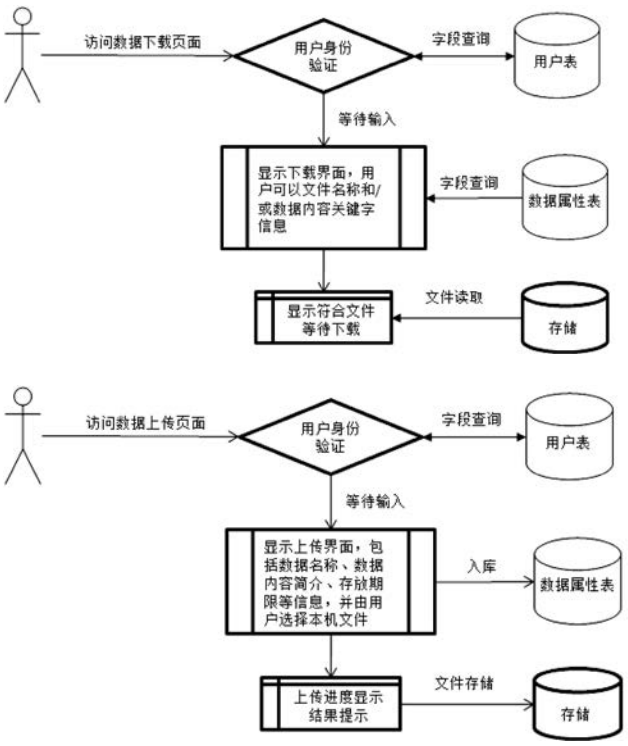

图2 NSTL备份体系集中保存模块主要应用流程

整个模块基于Linux/Tomcat/MySQL实现，包括了以下功能：

- 用户身份验证

只有经过用户身份验证，才可以进行文件的上传和下载，且下载前查询时，仅能查询本帐号曾经上传的文件。

- 文件属性信息收集及入库保存

在文件上传前，由上传人输入文件的相关信息，包括数据名称、数据内容简介（限定200字以内）和计划保存的期限（超过保存期限的文件将被自动删除）。一旦文件或文件集上传成功，这些信息会被存放到数据库中，可供下载和数据维护时使用。

- 文件选择及上传

通过易传文件同步传输控制软件实现文件的上传，包括断点续传、自动重传、带宽优化、加密传输、传输调度、全程监控等功能。

- 文件信息查询

网站提供基于关系型数据库的字符匹配查询的简单页面（不含组合查询）。这些匹配会查询包括数据名称、数据内容简介中的文字、数据文件名称等内容。

- 文件下载

一旦查询到符合匹配条件的文件，则在页面中直接显示这些文件/文件集，并由用户主动点击下载，下载时仅提供单线程服务。

- 集中存放数据的安全

被集中存放的数据，首先是基于用户进行访问权限控制，非本用户帐号或者管理员不能访问这些文件。其次，上传完成后，文件按账户被存放在特定目录中。这些目录不能被浏览器直接浏览遍历。最后，被上传的数据在服务器上以文件形式存放，而此服务器安装有NSTL集中备份子系统的客户端，被定期按照全备、增量等形式进行全磁盘文件的备份。若是以后异地数据中心建成，这些数据块还将被复制到异地数据中心。

3.3.5 NSTL备份体系日志收集模块

经过十余年的信息化建设，NSTL城域网存在多种运行日志，而日志可以分析对公众服务的网站的访问历史和趋势，可以分析设备是否正常、网络是否健康和网络安全运行态势等。因此，NSTL备份体系日志收集模块使运维管理人员可以有根据地面对问题，分析问题产生原因，及时解决问题。同时，详细地分析这些日志也会对NSTL的业务重点及业务针对性进行决策支持，特别是面向互联网提供业务服务的网络服务系统和回溯系统的访问日志，存在大量的日志产生环节，其设备或软件数量约有40个。

而各类系统生成的syslog格式日志只能通过用户指定facility/priority的方式把消息发到不同的地方。系统预先定义了12+8个（mail、news、auth等）facility和8个不同的优先级（alert到debug）。这存在一个大问题，大量的程序使用同样的facility（daemon），使日志都保存到一个文件中（messages），即使它们是毫无关联的信息，这样就会造成日志是一个庞大的、很难使用的宝库。第二个问题是，大多数的程序无法改变日志配置，只能修改软件或设备的源代码，而syslog-ng作为syslog的替代工具，一个设计原则就是建立更好的消息过滤粒度，syslog-ng能够进行基于内容和优先权/facility的过滤。另一个设计原则是更容易进行不同防火墙网段的信息转发，它支持主机链，即使日志消息经过了许多计算机的转发，也可以找出原发主机地址和整个转发链。最后一个设计原则就是尽量使配置文件强大和简洁。其框架如图3所示。

图3 syslog−ng架构

3.3.6 数据恢复功能测试模块

该模块是针对NSTL备份体系中集中备份模块和集中保存模块需要进行定期的数据恢复功能测试，以确保被备份和保存的文件是完整的、可靠的。这两个模块由于采用商业化软件的备份软件和易传文件同步控制软件，从功能上具备了可靠的数据恢复机制，因此，定期测试既可以使操作人员熟悉相关步骤，也可以达到完善管理操作手册、使运维管理人员能够在需要使用时有据可查、有章可循的目的。

4 结语

毋庸置疑，随着时间的推移，NSTL的数据积累会越来越庞大，增速会越来越快。同时，各类科技文献资源电子化存储的实现和NSTL应用数据中心的形成，必然形成“大数据”格局，大数据的安全必须未雨绸缪。本文从NSTL各业务系统的实际需求出发，结合当前各类数据资源现状和发展趋势，提出了科学、实用、系统的NSTL备份体系框架，同时也充分考虑了NSTL异地数据中心的发展需要，在备份架构和异地容灾方面预留了扩展，既能够解决当前大量已有数字资源集中保存、保留极为珍贵的用户行为记录，又能够满足NSTL在信息化方面不断发展对信息基础设施的迫切需求，从而极大地推进了NSTL在国家数字资源长期保存体系方面的建设步伐，具有一定的社会影响力。

参考文献

[1] 贺德方.建设国家科技图书文献中心共铸国家科技文献资源共享体系[J].数字图书馆论坛,2010(10):7-10.
[2] 乔晓东.从NSTL战略定位到最新进展及未来发展规划[J].数字图书馆论坛,2010(10):11-17.
[3] 热点备份技术现状分析：重复数据删除[OL]. [2012-09-23]. http://storage.chinabyte.com/57/11565057.shtml.
[4] syslog-ng [OL]. [2012-09-23]. https://wiki.archlinux.org/index.php/Syslog-ng.
[5] The syslog-ng Open Source Edition 3.3 Administrator Guide [OL]. [2012-09-23]. http://www.balabit.com/sites/default/files/documents/syslog-ng-ose-3.3-guides/syslog-ng-ose-v3.3-guide-admin-en.html/index.html-single.html.

作者简介

张婧，中国科学技术信息研究所高级工程师。主要研究方向：计算机网络安全等。E-mail: zhangj@istic.ac.cn
梁冰，哈尔滨工业大学计算机博士。中国科学技术信息研究所高级工程师。主要研究方向：网络信息服务、数字图书馆技术研究等。E-mail: liangb@istic.ac.cn
胡铁军，研究员。主要研究方向：计算机网络建设与管理、文献信息基础理论研究与建设。E-mail: hutj@nstl.gov.cn

The Design of the NSTL Backup System Framework – The Best Practice of the Big Data Security

Zhang Jing, Liang Bing / Institute of Scientific and Technical Information of China, Beijing, 100038
Hu Tiejun / National Science and Technology Library, Beijing, 100038

Abstract: This paper introduces the structure of NSTL data backup system. It aims to establish the NSTL data backup system and to solve the problem of online big data backup.

Keywords: Data backup, Database network, Virtualization, Storage, Big data

（收稿日期：2012-09-22）