高能物理实验中数据传输系统的研究与实现

曾 珊[1] 齐法制[1] 王 萌[2]

(中国科学院高能物理研究所计算中心 北京 100049)[1] (山东大学物理学院 济南 250100)[2]

摘 要 高能物理实验每天会产生大量的实验数据,由于高能物理实验本身的跨地域的建设特性,这些实验数据需要传输到远程的数据和计算中心进行离线分析。如何将这些数据实时、可靠、高效地传输到远程的数据和计算中心则是目前高能物理实验中需要解决的一个重要问题。介绍了一种高能物理环境下支持大批量数据传输的实时系统。系统提供了与数据产生系统和数据管理系统的接口,并实现实验数据从实验现场到数据中心和计算中心的可靠传输。目前该系统具有的功能包括:多路径源数据扫描、数据传输、数据缓冲区自动释放、传输过程管理、数据传输过程和性能监视、传输过程日志记录等。为了解决广域网上高延迟的特性,提高网络传输效率,系统应支持多流并发传输,同时支持数据中继服务,从而解决传输过程中由于网络或者某一传输节点失效造成的单点故障问题,提高系统自身的健壮性和可靠性。实验表明,该系统在高能物理实验数据的传输过程中具有良好的效果。

关键词 大批量数据,实时传输,可靠传输,多流并发,数据中继

中图法分类号 TP393.07 **文献标识码** A

Study and Implementation of Data Transfer System in Experiment of High Energy Physics

ZENG Shan[1] QI Fa-zhi[1] WANG Meng[2]

(Computing Center, Institute of High Energy Physics, Chinese Academy of Sciences, Beijing 100049, China)[1]

(School of Physics, Shandong University, Jinan 250100, China)[2]

Abstract In the circumstance of high energy physics experiments, usually a large amount of experimental data is produced every day. Also the high energy physics experiments are regional-crossing. Because of this feature, the experimental data needs to be transferred to the remote data center and the computing center for data analysis. So how to transfer the data to the remote data center and computing center reliably and efficiently in real time is an important problem which needs to be solved for carrying on a high energy physics experiment successfully. This paper presented a mass data real-time transfer system under the environment of high energy physics. This system provides the interfaces with the data generation system and the data management system, and can transmit the experiment data from onsite to the remote data center and computing center reliably. The functionalities of the system include: multipath source data scanning, data transmitting, data buffer automatic releasing, data management interface, data transmitting performance monitoring and logging and etc. The system also provides multi-flow concurrent transmission in order to solve the high delay in WAN and improve the transmission efficiency. In addition, to solve the single point problems caused by the network status or failure of the transmission point, the system also supports data relaying to make the system work robustly and reliably. Experiments show that the system performs well in data transmission process under the environment of high energy physics.

Keywords Mass data, Real-time transmission, Reliable transmission, Multi-flow concurrency, Data relay

1 引言

目前,每年高能物理的实验数据量已经达到100PB,随着高能物理实验规模的不断扩大和实验复杂性的不断增加,会产生越来越多的实验数据,由于高能物理实验本身的跨地域的建设特性,数据存储和分析资源部署在统一的数据中心和计算中心,而现存的数据传输工具不能有效地满足高能物理实验数据传输需求,因此研究一种能够将这些越来越庞大的实验数据实时、可靠、高效地传输到远程的存储和计算中心的数据传输系统,对保证物理实验实现其目标具有重要意义。

本文第1节介绍了设计高能物理环境下数据传输系统采用的关键技术;第2节对系统的总体框架进行简单介绍;第3节针对系统的主要功能,对其重要模块的实现进行详细分析。该系统已经在大亚湾中微子实验中进行部署,并完成了4次测试数据的传输、1次正式实验数据的传输,具有良好的应用效果。

本文受科技部973项目(2006CB808104)资助。

曾 珊(1987—),女,硕士,研究实习员,主要研究领域为网络系统研究、开发与网络安全技术研究,E-mail:zengshan@ihep.ac.cn;齐法制(1978—),男,硕士,高级工程师;王 萌(1970—),男,博士,教授。

2 相关技术

2.1 GridFTP

GridFTP[1]是 Globus 项目组开发的一个新的数据传输协议，它基于规范的 FTP 协议，并对其进行全面扩展，旨在为网格上分离的存储系统间的互操作提供一个通用的、可扩展的底层数据传输协议。GridFTP 不但支持 Kerberos 安全机制和 GSI 安全机制，还支持完整性检查、安全鉴别、可靠数据传输和容错传输；并且在 GridFTP 中我们可通过自动调整 TCP buffer/Window 大小来有效地提高数据传输性能[2]。以上特点使得 GridFTP 更安全、快速和高效。

2.2 PostgreSQL

PostgreSQL[3]是由加州大学伯克利分校计算机系开发的开源的对象关系型数据库管理系统(ORDBMS)，它采用客户端/服务器模式，可以通过 SSH(Secure Shell)和 SSL(Secure Socket Layer)连接方式提高访问的安全性。PostgreSQL 支持大部分 SQL 标准并且提供了许多其他现代特性：复杂查询、外键、触发器、视图、事务完整性、多版本并发控制。同样，PostgreSQL 可以用许多方法扩展，比如，增加新的数据类型、函数、操作符、聚集函数、索引方法、过程语言。同时，PostgreSQL 能够比较方便地迁移到 Oracle、Sybase 或者 MSSQL 等商业数据库。

2.3 JBoss

JBoss[4]是一个基于 J2EE 的开放源代码的应用服务器，同时也是企业级 Java 中间件系统，用于实现基于 SOA 的企业应用和服务。JBoss 的一个重要特性是：它不仅能够在一台运行 Java 的机器上部署，同时能够部署 Java 的 J2EE 部分。由于是基于 Java 的，JBoss 应用服务器能够跨平台运行，能够在任何支持 Java 的操作系统上运行，另外一个主要特性是：JBoss 应用服务器以 JMX(Java Management Extensions，即 Java 管理扩展)为微内核，各个模块以管理构件(Managed Bean，简称 MBean)的形式提供相应的服务。JMX 是一个为应用程序、设备、系统等植入管理功能的框架。JMX 可以跨越一系列异构操作系统平台、系统体系结构和网络传输协议，灵活地开发无缝集成的系统、网络和服务管理应用[5]。在各种 J2EE 应用服务器中，JBoss 是最受欢迎而且功能最为强大的应用服务器。

3 设计与实现

3.1 原理

为了将实验数据实时、可靠、高效地传输到远程的存储和计算中心，本文介绍的数据传输系统在 GridFTP 传输数据文件的基础上采用了 PostgreSQL 数据库保存和记录系统的一些配置信息和数据传输过程中的具体信息，包括：传输文件名、数据传输开始时间戳、结束时间戳、文件大小、传输过程中的网络路由信息、传输文件的类型信息(在高能物理试验中，根据不同的实验需求会产生不同类型的文件，这些文件类型信息需要先在系统中进行注册，并生成相应信息保存到数据库中，该类型的文件才能被传输)、传输过程中的报警信息等。同时，通过 JBoss 的 MBean 提供的页面管理服务，图形化地管理系统的各个子模块，并对相应的配置信息进行页面化设置[6,7]。

3.2 系统总体框架

根据物理实验对数据传输系统的要求，系统按照其功能模块划分为：多路径源数据扫描模块、数据传输模块、缓冲区数据清理模块、传输日志管理模块、传输系统监视模块和配置管理模块，系统总体架构如图 1 所示。

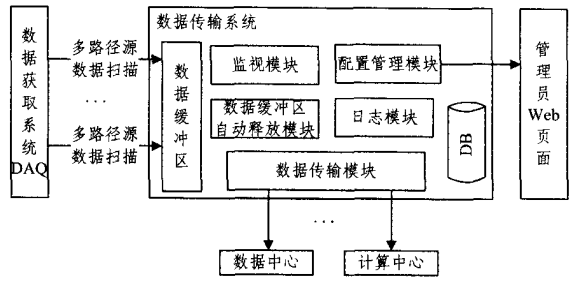

图 1 系统总体架构图

数据获取系统(以下简称 DAQ)位于物理实验现场，会实时获取物理实验产生的大量实验数据，这些实验数据被保存在多个磁盘阵列中。数据传输系统通过多路径源数据扫描模块将保存在多个磁盘阵列的数据直接映射到其数据缓冲区中。数据传输模块则实时扫描数据缓冲区，并将新产生的实验数据传输到目的数据中心和计算中心。在传输过程中，传输系统监视模块会对传输过程的性能进行分析，并以图形化模式展示数据传输系统的状态。日志模块则采用文本形式记录传输的整个过程，方便数据传输过程中日志信息的实时查看，便于较快地进行问题排查和故障诊断。由于数据缓冲区具有一定的容量，为了保证 DAQ 数据能够正常地接收和传输到远程的数据中心和计算中心，数据缓冲区自动释放模块会定时地对数据缓冲区可用空间大小进行检测，并根据设定的阈值，按照一定的规则将数据缓冲区的数据进行清理。配置管理管理模块为管理员提供了管理系统的 Web 页面接口，在 Web 页面上，管理员可以设置 DAQ 数据的磁盘阵列与系统缓冲区的映射关系、远程数据中心和计算中心的机器名以及数据存放路径、数据文件的重传、传输的数据文件类型的注册等。传输过程中的每个文件的传输过程包括：传输文件名、数据传输开始时间戳、结束时间戳、文件大小、传输过程中的网络路由信息、传输文件的类型信息、传输过程中的报警信息等都会被记录到 PostgreSQL 数据库中，便于日后追踪问题和数据分析。

3.3 多路径源数据扫描

物理实验的过程，也是数据产生的过程，由于原始数据不可再生，DAQ 获取到的原始数据需要在实验现场的数据缓冲区中保留一定时间范围(例如两周)以避免意外事故造成的原始数据丢失[4]。原始数据在实验现场会分别存储到不同的磁盘阵列和磁盘目录中，这就要求数据传输系统可以自动识别新产生的数据位于 DAQ 磁盘阵列中的具体位置并支持多源路径输入。系统通过与 DAQ 数据库交互并调用 DAQ 数据库中有关数据存放的物理位置的信息实现源数据多路径扫描，具体的流程如图 2 所示。

系统通过轮询 DAQ 数据库中新增加的数据文件，找到新产生的原始数据被存放的磁盘阵列的位置，根据预设的 DAQ 磁盘阵列和系统缓冲区的目录结构的映射关系，将原始数据映射到系统的缓冲区，实现源数据多路径扫描。

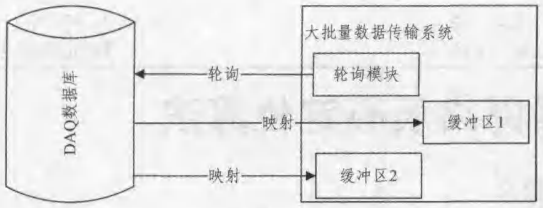

图 2　多路径源数据扫描

3.4　数据传输模块

从 DAQ 数据库中扫描到数据缓冲区中的原始数据通过数据传输模块传输到远程的数据中心和计算中心。具体传输流程如图 3 所示。

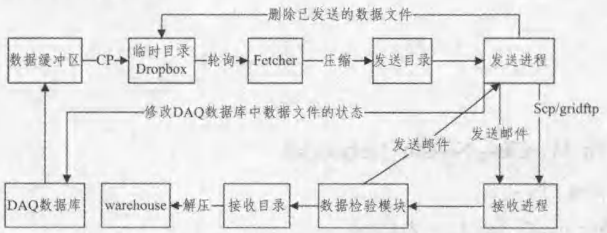

图 3　数据传输模块流程图

数据缓冲区接收到 DAQ 获取的原始实验数据文件后，数据传输系统会将该数据拷贝到其临时区域(Dropbox)中。此时系统的一个守护进程(Fetcher)会对这个临时目录进行查询，一旦发现有新增加的数据文件，就将其进行压缩并转移到发送目录，数据发送进程将发送目录中的压缩文件通过 SCP 或者 Gridftp 方式发送到远程计算中心或者数据中心的接收目录中(Gridftp 由于不加密和多流传输，效率会高些)，同时发送邮件到对应的接收邮件箱，表示数据文件传输开始，当接收模块接收到数据后，数据检验模块会对接收到的数据进行完整性校验，一致后回复邮件给发送模块，表示数据接收成功，并将数据解压到对应的目的磁盘(warehouse)中，发送进程接收到接收确认邮件后，会将临时区域的文件删除，并修改 DAQ 数据库中的数据文件的状态为传输成功(transferred)状态[4]。这样一个数据文件就成功传输完成了，数据文件在目的端则会被存放到根据文件产生的时间而自动创建的目录结构中(其目录结构为年/月/日/文件名)。若发送进程在规定的时间内没有接收到接收方返回的确认邮件，则启动自动重传机制，重试指定次数后，发送进程还没有收到接收方返回的确认邮件，则表示文件传输失败，该文件会被放到有问题的文件目录(problem_files)中，管理员可以进行手动的重传，从而保证数据的正常传输[5]。为了解决广域网上高延迟的问题，提高网络传输效率，该系统应支持多流并发传输，如图 1 所示，该数据传输系统能够同时向远程的多个计算中心和数据中心传输数据而彼此不受任何影响。

3.5　数据缓冲区自动释放模块

为了保证数据正常传输，数据缓冲区必须具有足够的空间接收 DAQ 数据库映射过来的原始数据。然而，数据缓冲区存放于物理磁盘中，具有特定的容量。为避免数据缓冲区满而导致的原始数据无法映射到数据缓冲区中，从而影响大批量数据传输系统的传输效率，该系统中部署了数据缓冲区自动释放模块，通过设置上水位和下水位(阈值)，并按照一定的规则(例如时间顺序、文件重要性指标等)对缓冲区中的数据进行清除，当数据缓冲区占用的磁盘空间大于上水位时，自动释放模块会对数据缓冲区中的已经传输到远程的计算中心或者数据中心的数据按照时间先后顺序进行删除，直到数据缓冲区占用的磁盘空间小于下水位为止。为了保证数据缓冲区中的数据能够始终保持一定时间范围(例如两周)，在进行数据删除操作的时候只对符合时间要求的数据进行操作。具体流程如图 4 所示。

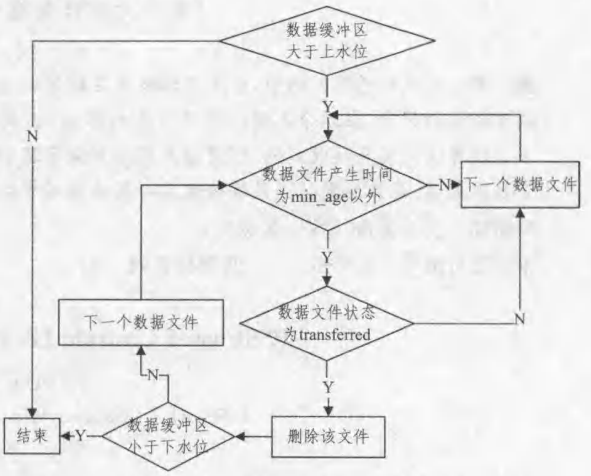

图 4　数据缓冲区自动释放流程图

3.6　数据中继服务

为了解决传输过程中由于网络或者某一传输节点失效造成的单点故障问题，系统支持数据中继服务，从而提高了系统自身的健壮性和可靠性。如 5 图所示，数据传输系统除了提供多流并发功能，还提供数据中继服务。

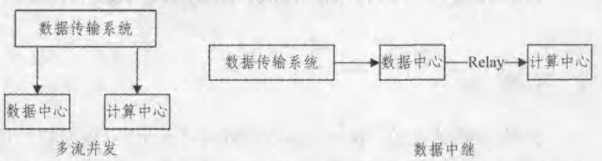

图 5　多流并发和数据中继示意图

3.7　日志模块

图 6　数据文件传输过程中日志信息

日志模块则采用文本形式记录传输的整个过程，方便数据传输过程中的实时查看日志信息，从而较快地进行问题排查和故障诊断。日志文本中的内容包括：数据文件从 DAQ 数据库映射到数据缓冲区的时间、Fetcher 进程开始检索临时目录 Dropbox 和结束的时间、数据文件压缩开始和结束的时

(下转第 108 页)

[14] Son S-W, Jeong H, Noh J D. Random field Ising model and community structure in complex networks[J]. Eur. Phys. J, 2006, B50(3):431
[15] Hughes B D. Random Walks and Random Environments: Random Walks[M]. Oxford, UK: Clarendon Press, 1995
[16] Zachary W W, Anthropol J. Res. 1977, 33:452
[17] Kuramoto Y, Oscillations C. Waves and Turbulence [M]. Berlin, Germany: Springer-Verlag, 1984
[18] Arenas A, Diaz-Guilera C, Perez-Vicente J. Synchronization Reveals Topological Scales in Complex Networks[J]. Phys. Rev. Lett, 2006, 96(11):114102
[19] Fortunato S. Community detection in graphs [J/OL]. arXiv: 0906.0612v2[physics. soc-ph], 2010
[20] Karrer B, Newman M E J. Stochastic blockmodels and community structure in networks[J/OL]. arXiv: 1008.3926v1[physics. soc-ph], 2010
[21] 淦文燕, 赫南, 李德毅, 等. 一种基于拓扑势的网络社区发现方法[J]. 软件学报, 2009, 20:8

(上接第 95 页)

间、数据文件传输开始和结束的时间以及数据文件在临时目录 Dropbox 中被删除的时间等。通过这些日志信息的记录，可以实时地观察到某个时刻数据传输的阶段，同时也可以根据错误信息较快地进行问题排查和故障诊断。图 6 是一个数据文件传输过程中日志信息的截图。

3.8 监视模块

为了更加形象化地监视大批量数据传输系统传输过程中的效果，监视模块在日志模块采集日志的基础上，采用 Web 页面图形化的方法实时显示每个时刻的传输结果，包括文件传输的个数、文件的大小(byte 单位)等，具体如图 7 所示。

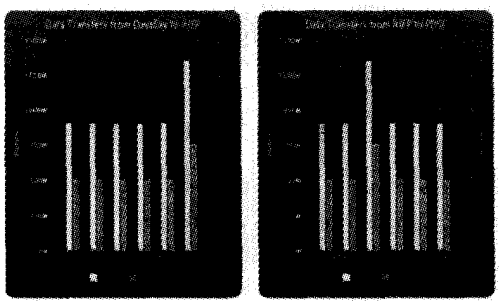

图 7 监视模块中数据文件传输过程效果图

此外，监视模块会对每天传输到各个数据中心和计算中心的文件进行统计，并在当天晚上 12 点发邮件通知管理员，以便管理员进行数据统计和排查问题。

3.9 配置管理模块

配置管理模块利用通过 JBoss 的 MBean 提供的页面管理服务，图形化地管理系统的各个子模块。在 Web 页面上，管理员可以设置 DAQ 数据的磁盘阵列与系统缓冲区的映射关系、远程数据中心和计算中心的机器名、Fetcher 程序轮询 Dropbox 的时间间隔、发送模块和接收模块用于通信的邮件名、数据存放路径、传输超时时限、数据文件的自动重传次数、传输的数据文件类型的注册、传输失败文件存放目录等。

4 系统应用

本文介绍的大批量数据传输系统已经在大亚湾中微子实验中进行部署，并完成了 4 次测试数据的传输，已经完成了 14.1TB 的数据传输，取得了良好的效果。部署图如图 8 所示，分别在大亚湾现场、高能所和高能所的合作单位部署了大批量数据传输系统，采用数据中继的方式，现场的数据传输系统从现场数据磁盘缓冲区中获取原始数据后，将其发送到高能所的数据传输系统，高能所的数据传输系统再将其中继到其合作单位，合作单位的数据传输系统接收数据，并进行线下分析。

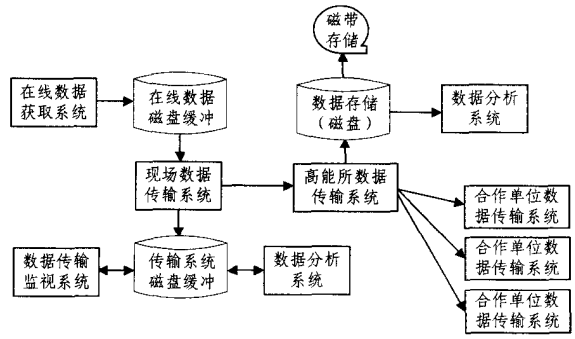

图 8 大批量数据传输系统在大亚湾中微子实验中的部署图

结束语 本文描述的大批量数据传输系统已经应用于大亚湾与高能所的物理实验数据传输过程中，取到了良好的应用效果。但是 DAQ 获取数据速率较大时，使用数据中继方式传输数据，数据接收的时延会比较长，从而影响远程的数据中心分析人员对数据进行分析的实时性。经过分析，产生该问题的原因有两个方面，一是系统部署的问题，具体指大批量数据传输系统的数据处理和数据存储使用的磁盘是同一块磁盘，由于磁盘自身的 IO 问题导致互相影响，目前已经将数据处理和数据存储使用的磁盘分离，以对时延的缩短产生良好的效果。二是系统源代码本身的冗余问题，这是今后改进的方向。

参 考 文 献

[1] Allcock W, Bresnahan J, Kettimuthu R, et al. The Globus Striped GridFTP Framework and Server[C]// Proceedings of the 2005 ACM/IEEE conference on Supercomputing (SC'05). Washington, DC, USA: IEEE Computer Society, 2005:54-54
[2] Allcock W, Bresnahan J, Bresnahan A, et al. GridFTP: Protocol Extension to FTP for the Grid[C]//Grid Forum Internet-Draft, 2001. Washington, DC, USA: IEEE Computer Society, 2001:21-24
[3] Momjian B. PostgreSQL: introduction and concepts. Bruce Momjian, PostgreSQL: introduction and concepts [M]. Boston, MA, Addison-Wesley Longman Publishing Co., Inc., 2001:47-58
[4] Stark S. JBoss Administration and Development[M]. The JBoss Group, USA, 2003:145-159
[5] Kounev S, Weis B. Buchmann A. PerformanceTuning and Optimization of J2EE Applications on the JBoss Platform[J]. Journal of Computer Resource Management, 2004, 2(113):129-135
[6] 聂思敏, 张吉龙, 谭有恒. 羊八井宇宙线观测数据实时传输及处理系统[J]. 核电子学与探测技术, 2007, 2(1):14-17
[7] Patton S. Spade and Data Movement[EB/OL]. http://dayabay.ihep.ac.cn/DocDB/0055/005598/001/DataMovement.pdf, 2010