面向高能物理计算的网格文件系统*)

程耀东　汪　璐　刘爱贵　陈　刚
(中国科学院高能物理研究所计算中心　北京 100049)

摘　要　针对网格环境下高能物理计算的需求，基于现有的网格中间件，提出并设计了一个网格文件系统 HEP-GridFS，目的是将异构的、动态变化的、大规模的网格存储资源虚拟成单一的、稳定的文件系统视图，并根据用户实际需求提供不同的服务质量。描述了它的体系结构以及关键技术，主要包括名字服务、存储资源管理、数据复制、用户访问接口、工作流程等。

关键词　高能物理，网格，网格文件系统，数据存储

Grid File System for Computing in High Energy Physics

CHENG Yao-dong　WANG Lu　LIU Ai-gui　CHEN Gang
(Computing Center, Institute of High Energy Physics, CAS, Beijing 100049, China)

Abstract　Based on the existing grid middle wares, a grid file system for computing in high energy physics was proposed and implemented. The goal was to integrate heterogeneous, dynamic, large-scale grid storage resources into a single and stable file system view and provide different QoS according to user's requirements. Its architecture and key technologies were described mainly including name service, storage resource management, data replication, user interface, and work flows.

Keywords　High energy physics, Grid computing, Grid file system, Data storage

1　引言

近年来，网格技术[1]发展突飞猛进。高能物理是网格发展的驱动者与受益者，一直走在网格技术研究和应用的前列，这主要是由高能物理的需求决定的。

首先，高能物理的数据量巨大。随着新一代高能物理加速器的建设完成，比如欧洲核子中心 CERN 的 LHC、北京正负电子对撞机 BECPII 等，实验规模不断扩大，实验复杂性也不断增加，会产生越来越多的数据。预计到 2010 年，全世界高能物理的实验数据将达到 100PB，并在以后的几年中超过 1000PB。

其次，数据需求全球共享。比如，LHC 在投入运行以后，每年将产生大约 15PB 的实验数据，全世界 6000 多位物理学家需要对此进行分析。BEPCII 上的 BESIII 实验也是一个大型国际合作项目，有着同样的需求。

最后，高能物理中很多数据需要长期保存。比如，LHC 和 BESIII 实验的数据必须保证在 10 年以上的生命期内可被利用。

本文针对这些需求，在现有高能物理网格体系结构的基础上，提出和设计了网格文件系统 HEP-GridFS。在后面的章节中，本文将描述它的体系结构和关键技术，主要包括名字服务、数据复制、存储资源管理、用户访问接口以及工作流程等。

2　相关工作

2.1　高能物理网格的体系结构

全球大型强子对撞机(LHC)计算网格，称为 WLCG[2](Worldwide LHC Computing Grid)，将全球的高性能计算中心整合到一个平台中，来共同应对大型强子对撞机 LHC 实验带来的数据挑战。到 2008 年底，WLCG 将覆盖 50 多个国家的 500 多个研究机构，把超过 20 万台计算机的计算资源整合在一起。

WLCG 按分级机构来组织计算环境[3]，分为 5 层：第 1 层，以 CERN 为中心的 Tier-0；第 2 层，以国家/地区为中心的 Tier-1；第 3 层，以国家/地区为中心的 Tier-2；第 4 层，以国家研究机构为中心的 Tier-3；第 5 层，以物理学家的分析平台。

WLCG 的各级功能有所不同，Tier-0 主要负责原始数据记录、数据重建和向 Tier-1 中心分发数据；Tier-1 负责数据分析、数据存储管理、再处理和地区支持中心；Tier-2 负责磁盘存储管理、物理模拟计算、终端用户分析和并行交互分析。

2.2　网络文件系统的相关工作

传统的分布式文件系统的主要功能是支持用户在一定网络范围内，对一定数量的分布文件进行透明访问，典型的系统有：NFS(Network File System)，WebNFS，AFS(Andrew File System)，DFS 等。这些文件系统很难适应大规模的动态网格环境。P2P(peer-to-peer)存储系统是近年来的一个研究热点，像 Oceanstore，Farsite 等。但是，P2P 中提供存储的节点随时都可能暂时甚至永久离开系统，给持久存储[4]带来很大困难，当前在这方面的研究并没有重大突破，因此目前不适合用于高能物理网格的存储。

针对当前高能物理网格的体系结构，本文提出将 Tier0，Tier1 和 Tier2 中的永久存储资源整合起来，提供一个高可靠的、易管理的、易扩展的、高性能的以及易于使用的文件系统服务。

*)基金项目：国家自然科学基金委资助项目(90412017)。**程耀东**　博士后，主要研究方向为海量存储与网格计算。

3 HEP-GridFS 关键技术

3.1 体系结构

围绕网格文件系统的需求和设计目标，按照 Globus 五层沙漏模型，HEP-GridFS 采用完全分布式、高度模块化以及协议分层的方案，图 1 显示了 HEP-GridFS 的体系结构。

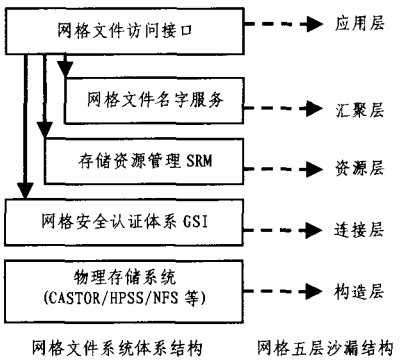


图 1 HEP-GridFS 的体系结构

物理存储系统位于构造层，是网格文件系统的物理基础。网格安全认证体系 GSI(Grid Security Infrastructure)位于连接层，是网格文件系统的安全基础。网格文件系统要整合不同的存储系统，必须要能够屏蔽不同的认证/授权方式。网格安全认证体系 GSI 在不改变原有的安全机制的基础上，可在虚拟组织中的不同安全机制上建立信任关系，因此，GSI 为 HEP-GridFS 的实现提供了安全保证。存储资源管理 SRM[5](Storage Resource Management)位于资源层。其功能是在网格中提供动态空间分配和共享文件管理功能。网格文件名字服务位于汇聚层，主要的功能是提供网格文件的逻辑名和物理存储系统上的一个或者多个文件标识之间的映射，支持副本管理功能的实现，最重要的是，它还提供一个全局透明的统一树状名字空间。网格文件访问接口位于应用层，主要为用户提供统一、方便的接口。

3.2 名字服务

高能物理的数据分布存储于全球不同的站点，名字服务提供了统一透明的命名空间，允许用户在分布式的异构环境下快速定位所需的数据。

首先，名字服务提供一个统一透明的逻辑名字空间，便于用户记忆和使用。一般采用类似于传统文件系统的树形目录结构，比如/grid/user/yao/f1。不管用户在任何时候、任何地理位置、任何主机上登录，他所看到的目录结构都是一致的，不会因为物理文件的移动或者登录位置的不同而看到不同的名字空间。

其次，名字服务器记录物理文件的实际位置，并维持逻辑文件名到物理文件名的映射，同时还保证逻辑文件与其物理实体的一致性。第三，名字服务器中记录网格文件的元信息，比如文件大小、所有者、访问时间、checksum 类型以及 checksum 值等。名字服务还提供访问权限控制，包括两种基本的认证方式：网格安全认证 GSI 与传统的基于用户的访问权限控制。

由于集中名字服务器结构存在性能、单点故障和扩展性的问题，HEP-GridFS 引入了分布式名字服务结构。在系统中，建立多个名字服务器，客户端只需要和其中一个(本站点或本区域的)名字服务器相连。每个名字服务器主要负责本区域的名字服务(注册、查询、更改等)工作。当一个名字服务器无法满足客户端的请求，比如请求的文件名字不在该名字服务器中，它会根据一定的策略向其它的名字服务器转发请求。如果下一个名字服务器中仍没有该文件的信息，就接着向下一台服务器转发请求，直到该请求得到满足。

3.3 存储资源管理

存储资源管理包括对存储空间和数据两个方面的管理。传统的海量存储系统，比如 CASTOR，dCache，EnStore，HPSS 等对本地的存储资源(磁盘阵列、磁带库等)能进行很好的管理与使用。由于各个存储系统的接口各不相同，国际网格论坛组织 OGF 制定了网格存储管理的标准规范 SRM[5](Storage Resource Manager)。各个存储系统开发与 SRM 的接口，最终对网格用户提供统一的调用方式。同时，SRM 还支持数据缓存、预取、加锁、空间预留等高级功能，大大方便了网格存储资源的管理。

为了对用户提供不同的存储服务质量，HEP-GridFS 将多个具有类似特点的存储系统，组成不同的“网格存储池”，比如高性能存储池、高可靠性存储池等。在引入“网格存储池”以后，HEP-GridFS 的基本存储单元从单个存储系统变成了网格存储池。对于某个具体的存储系统，可以属于多个存储池。名字服务器中的逻辑目录也可以对应于单个或多个网格存储池(如图 2)。

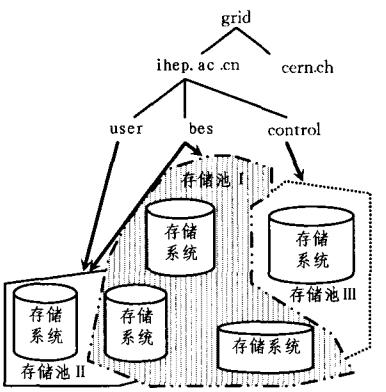


图 2 网格文件命名空间与网格存储池

3.4 数据复制

为了获得更好的数据访问性能和实现容错等，在 HEP-GridFS 中广泛采用数据复制。数据复制完成这样一些功能：生成新的完整的或部分的数据副本；把这些新的副本注册到名字服务目录中；允许用户和应用去查询目录以发现所有现存的部分或全部文件的副本；基于存储和由网格信息服务所提供的网络性能预测功能选择“最好的”的副本用于访问。

虽然高能物理网格采用了分级的组织方式，但是从技术的角度来看，各个站点之间是平等的，可以自由交换数据。同时，高能物理的大部分数据是 WORM(Write Only, Read many)访问模式。因此，实现复制的手段时比较灵活。一种常见的且简单有效的方式是由管理员定义复制策略，比如哪些数据需要复制、复制到哪些站点、何时删除等。根据管理员定义的策略，系统自动维护副本一致性。

3.5 用户访问接口

HEP-GridFS 提供了 3 种基本的用户访问接口：专用 SHELL 方式、应用程序开发库以及虚拟文件系统接口。

专用 SHELL 方式，是一种命令行的接口，提供了多个类似于 Linux 文件操作的命令，比如 gls，其参数和显示方式同传统的 ls 相同。类似地，实现的还有 gcat(显示网格文件内容)、gmkdir(创建目录)、grm(删除文件)、grmdir(删除目录)、grename(重命名)、gcp(拷贝)、gchown(更改所有者)、gchmod

(更改属性)等。

HEP-GridFS还提供了应用程序开发库,其API函数在参数和格式上与传统的文件系统基本相同,函数名称采用传统的函数名加上前缀"hgfs_",比如open调用变成了"hgfs_open"。类似的还有"hgfs_read","hgfs_lseek","hgfs_stat","hgfs_readdir"等等。在编译的时候,将包含函数原型的头文件"hgfs_api.h"加入到源文件中,然后同库文件libhgfs_api.so一起进行编译。编译后的程序就可以透明使用网格上的数据。通过调用API的方式,用户可以快速开发基于网格文件系统的应用。

网格上的数据分布在不同的存储系统中,而不同的存储系统,其访问方式也不尽相同,如GridFTP,RFIO(Remote File IO),dCache等。对于不同的文件访问协议,客户端需要使用相应的命令、应用程序或API来访问。如GridFTP使用globus-url-copy命令,RFIO使用rfcp,dCache使用dccp。这对于用户来说是非常不便的。为此,HEP-GridFS设计和实现了一种可扩展的网格文件访问接口[6]GFAI(Grid File Access Interface),为多种网格文件访问/协议提供一个统一的接口。这样,用户就不用关心数据具体存放在什么存储系统中,或者要用什么传输协议来访问。

通过实现虚拟文件系统,HEP-GridFS成为本地文件系统的一部分,可以让用户像使用传统的文件系统一样透明访问网格文件。其实现如图3所示,分为两个模块:一个在内核区,一个在用户区。所有来自VFS的内核调用请求被传递给内核区的模块,然后内核区的模块把请求传递给用户区的模块,接着用户区模块调用相应的HEP-GridFS的应用程序API来完成操作。HEP-GridFS的内核区模块基于FUSE[7](Filesystem in Userspace)文件系统进行开发。

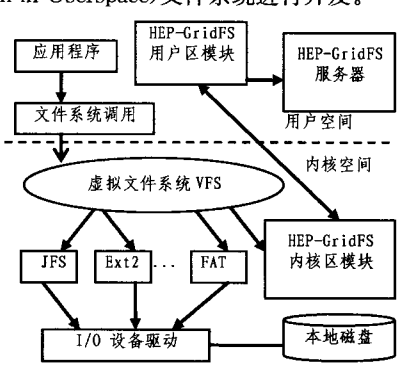


图3 HEP-GridFS虚拟文件系统的实现

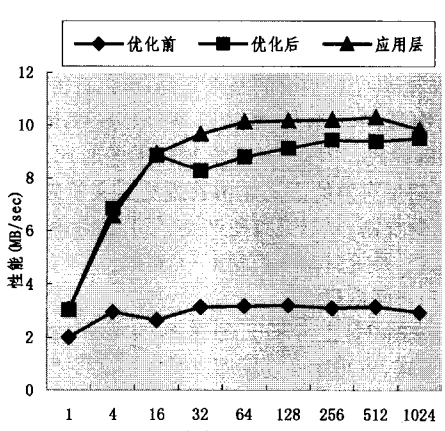


图4 HEP-GridFS的写性能

虚拟文件系统方便了用户的使用,但是如果数据访问的性能大大降低,就失去了意义。因此,在实现过程中,我们采取了预取、缓存、延迟写等多种方式,对性能进行优化。图4显示了一个测试结果。测试环境由两台机器组成,一台作为网格名字服务器和存储服务器,一台作为客户端,并用来安装虚拟文件系统。两台机器之间采用百兆以太网连接。机器的配置大致为CPU:Intel P4 2.80GHz;内存:512MB;内核:2.4.21-32.0.1.EL.cernsmp。本地文件系统采用ext3。测试工具采用Iozone 3.217[8]。

从图4可以看出,应用层情况下的性能最好,基本上达到网络饱和带宽。因为数据直接从存储服务器上获取,不通过内核层转换。优化后性能非常接近在应用层情况下的性能。

3.6 HEP-GridFS的工作流程

图5描述了HEP-GridFS的工作流程[9]。当用户需要访问一个网格文件时,通过以下步骤:①调用用户访问接口(SHELL命令、应用程序开发库或者是虚拟文件系统);②用户访问接口查询网格文件名字服务器;③名字服务器根据逻辑文件名字返回一系列的物理文件名;④用户访问接口将这一系列的物理文件名发送给副本选择器;⑤副本选择根据历史以及预测信息,再根据一定的算法在这一系列的物理文件名中选择一个最优的副本,并将这个副本地址返回给用户访问接口;⑥用户访问接口根据副本地址调用网格访问接口GFAI向相应的网格存储系统发出请求,协商好数据访问的协议;⑦网格存储系统通过协商好的网格文件访问协议向用户访问接口传送数据;⑧用户访问接口将网格存储系统传输回来的数据根据用户的需求进行适当的可视化,返回给用户。其中,复制管理器定期读取与其相关的网格存储服务器的数据访问信息,统计与分析用户的访问行为,再根据网格名字服务器中记录的信息,按照管理员定义的复制策略,进行数据的自动复制。副本选择器使用网络信息服务(Network Weather Service)对可用网络带宽进行预测,然后选出其中最优的副本。

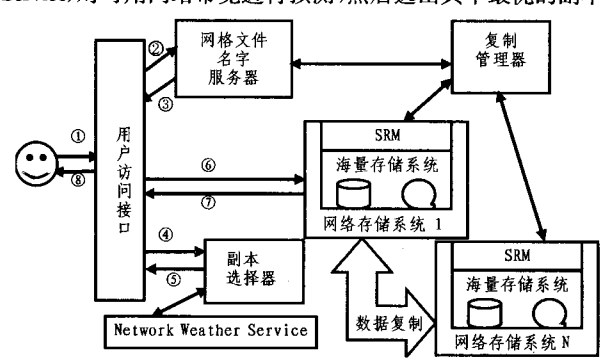


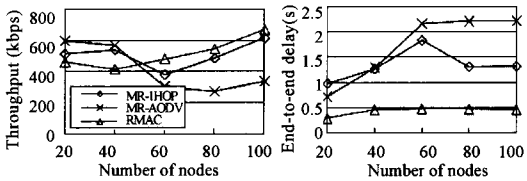
图5 HEP-GridFS的工作流程

结束语 高能物理计算是典型的数据密集型应用,对数据的访问性能、可管理性、可扩展性以及易用性有着极高的要求,特别是在分布式的网格环境下,数据管理和共享方面具有更多的挑战。本文提出并设计的网格文件系统HEP-GridFS,以"分布存储、统一管理、优化服务"为基本设计理念,将分布的存储资源虚拟成单一的视图,极大方便了存储管理与用户使用。同时,它也可用于其它大规模数据存储领域,在不改变原有本地自治管理的前提下,将分散的存储资源整合起来进行统一管理与使用。随着计算需求的不断变化,还需要对网格文件系统进行更加深入的研究,比如数据安全性、副本一致性、自动复制、完备的文件系统语义等。文件系统是网

(下转第87页)

的路径传输，不至于重建路由。这种交叉层优化既能发挥物理层提供的多速率优势，又在一定程度上保存了网络层路由维护的灵活性，同时在 MAC 层增加了对动态环境的及时响应。

固定分组发生速率为每秒 40 个，节点数量在 20～100 之间变化时，模拟的结果如图 11。在节点数量大于 50 之后，RMAC 在吞吐率方面的优势开始显现。其余同之前的模拟有类似的结论。



(a)吞吐率　　(b)端到端延迟

图 11　不同节点数量下动态随拓扑模拟结果

结束语　本文讨论了多速率环境下移动自组织网络中交叉层优化对网络性能的影响。分别设计实现了位于网络层和 MAC 层的两种多速率交叉层路由方案 MR-AODV 和 RMAC，并详细分析了简单拓扑、静态线性拓扑和动态随机拓扑等不同场景下，在不同层次上实现多速率路由功能时的吞吐率、端到端延时、分组递交率和协议控制开销等性能特征。在网络层制定路由策略，可以综合物理层多速率以及 MAC 层信道状态等交叉信息进行路由的全局优化，故能有效地提高网络吞吐率，降低端到端延迟，保证路由的准确性和易维护性。但是要达到这一目标，网络层需要更为准确的路由度量指标，并且由于网络层本身存在着迟滞性，会影响到路由的灵活性。而在 MAC 层通过添加中继，对网络层路由信息做进一步局部优化，能够保证对动态环境的及时响应，减小转发分组在层间的滞留时间，因此特别适用于节点密度高、负载较大的动态网络环境。

后续工作包括更好地实现网络层和 MAC 层在建立路由方面的跨层合作，网络层需要综合更多的下层提供的信道、拥塞状况等信息，以提供更准确的路由选择。MAC 层同样需要更为准确的判据，以保证及时从上层提供的路由中作出最佳选择。

参 考 文 献

[1] Srivastava V, Motani M. Cross Layer Design: A survey and the road ahead[J]. IEEE Communications Magazine, 2005, 43(12): 112-119

[2] Carneiro G, Ruela J, Ricardo M. Cross Layer Design in 4G Wireless Terminals[J]. IEEE Wireless Communications Magazine, 2004, 11(2): 7-13

[3] Holland G, Vaidya N H, Bahl P. A Rate-adaptive MAC Protocol for Multi-Hop Wireless Networks[C]// ACM MOBICOM'01. Rome, 2001

[4] Li Z, Das A, Gupta A K, et al. Full Auto Rate MAC Protocol for Wireless Ad hoc Networks[J]. IEEE Proceedings Communications, 2005, 152(3): 311-319

[5] Awerbuch E, Holmet D, Rubens H. High Throughput Route Selection in Multi-Rate Ad Hoc Wireless Networks[G]. Wireless on Demand Network Systems. Berlin: Springer, 2004: 253-270

[6] Yang G, Xiao M, Chen H, et al. A Novel Cross-layer Routing Scheme of Ad hoc Networks with Multi-rate Mechanism[C]// Proc. of International Conference on Wireless Communications, Networking and Mobile Computing. 2005

[7] 王炫，李建东，张文柱. 支持多速率传输的动态 ad hoc 路由协议[J]. 电子与信息学报，2006，28(10)：1907-1911

[8] Yongho S, Jaewoo P, Yanghee C. Multi-rate Aware Routing Protocol for Mobile Ad hoc Networks[C]// Proceedings of IEEE Vehicular Technology Conference. 2003: 1749-1753

[9] Zhu H, Cao G. rDCF: A Relay-enabled Medium Access Control Protocol for Wireless Ad Hoc Networks[J]. IEEE Transactions on Mobile Computing, 2006, 5(9): 1201-1204

[10] Zeng W, Tan H, Suda T. A Relay Based MAC Protocol to Support Multi-rate Feature in Mobile Ad hoc Networks[C]// Proc. of The Second Annual International Conference on Mobile and Ubiquitous Systems: Networking and Services. 2005: 145-154

[11] Perkins C, Royer E B, Das S. Ad hoc on-demand distance vector (AODV) routing[S]. RFC3561. 2003

[12] Goldsmith A J, Chua S G. Variable rate variable power M-QAM for fading channels. IEEE Transactions on Communications, 1997, 45(10): 1218-1230

[13] Takai H, Kleinrock L. Optimal Transmission Ranges for Randomly Distributed Packet Radio Terminals[J]. IEEE Transactions on Communications, 1984, 32(3): 246-257.

[14] 李东生，向勇，史美林，等. 基于交叉层设计的能量负载平衡自组网路由技术[J]. 清华大学学报：自然科学版，2006，46(10)：1771-1775

[15] Yang Ning, Sankar P, Lee Jungsik. Improving ad hoc network performance using cross-layer information[C]// Proc. of IEEE International Conference on Communications. 2005: 2764-2768

[16] Barrett C, Marathe A, Marathe M V, et al. Characterizing the Interaction between Routing and MAC Protocols in Ad-hoc Networks[C]// Proc. of the 3rd ACM International Symposium on Mobile Ad hoc Networking & Computing. 2002: 92-103

(上接第 38 页)

格中非常重要的一部分。对于网格文件系统的深入研究必将对网格技术的完善与发展起着极大的推动作用。

参 考 文 献

[1] Foster I, Kesslman C. The Grid 2[M]. 北京：电子工业出版社，2004：224-262

[2] WLCG website. http://lcg.web.cern.ch/LCG/

[3] 于传松. 高能物理与网格计算. 核电子学与探测技术，2004，24(6)：563-567

[4] 田敬，代亚非. P2P 持久存储研究. 软件学报，2007，18(6)：1379-1399

[5] Abadie L, Badino P, et al. Storage Resource Managers: Recent International Experience on Requirements and Multiple Co-Operating Implementations // 24th IEEE Conference on Mass Storage Systems and Technologies (MSST 2007). 2007: 47-59

[6] 刘爱贵，程耀东，许冬，等. 可扩展网格文件访问接口的设计与实现. 计算机工程，2007，33(20)：259-261

[7] Szeredi M. Filesystem in Userspace. http://fuse.sourceforge.net

[8] Norcott W, Capps D. IOzone Filesystem Benchmark. URL: http://www.iozone.org/

[9] 程耀东，刘爱贵，陈刚，等. 高能物理网格数据管理关键技术研究. 计算机应用研究，2007，24(10)：20-22