第32卷 第18期 计 算 机 工 程 2006年9月
Vol.32 № 18 Computer Engineering September 2006
·网络与通信· 文章编号：1000—3428(2006)18—0111—03 文献标识码：A 中图分类号：TP311

基于负载均衡的网格数据移动服务

梁 英[1,2]，胡志刚[1]
（1. 中南大学信息科学与工程学院，长沙 410000；2. 湖南商学院计算机与电子工程系，长沙 410205）

摘 要：网格是一种新兴的基于 Internet 的并行和分布式计算框架。该文提出了一种网格环境下基于负载均衡的数据移动服务。对网格环境下的数据移动需求及其产生的问题进行了详细的分析，并对数据移动服务的概要设计和详细设计进行了具体的分析。
关键词：网格；数据管理；数据移动

Data Movement Service Based on Load Balance in Grid

LIANG Ying[1,2], HU Zhigang[1]
(1. College of Information Science and Engineering of CSU, Changsha 410000;
2. Department of Computer and Electronic Engineering, Hunan Business College, Changsha 410205)

【Abstract】Grid emerges as a new infrastructure for Internet-based parallel and distributed computing. This paper proposes the data movement service based on load balance in grid. The demand and the matter of data movement in grid are analyzed. The general design and the particular project are introduced concretely.
【Key words】Grid; Data management; Data movement

网格(Grid)技术是将地理上分布、系统异构的多种资源通过高速网络连接起来，以获得复杂问题的求解能力[1]。网格数据管理作为网格环境中一个单独的模块，要向用户提供透明地访问和使用网格上存储资源和数据资源的手段，使用户能够容易地实现网格中的数据共享。负载平衡是目前并行计算和分布式计算领域主要的研究内容之一，也是将来网格计算中需要解决的主要问题之一。尽管目前有关网格数据管理和负载均衡方面的研究有许多，但在网格环境下以负载均衡为目标的数据移动服务却缺乏一整套完整的实现策略。因此，对网格环境下基于负载均衡的数据移动服务作深入的研究是非常必要的。

1 基于负载均衡的数据移动服务分析

网格中的数据在使用过程中，随着时间的推移，某个数据集的用户群可能会发生变化，对数据的密集请求在不同的时间段来自不同的地理区域。此时，为避免大量数据的远程传输，减轻单个服务节点的负载压力和由于频繁远程访问对网络带宽的消耗，满足网格用户所需的服务质量，就必须考虑数据的移动。

1.1 数据移动需求分析

在网格环境下，当一个数据拥有者发布一个数据服务时，并不能预先准确地判断数据的潜在使用者有哪些，也无法把自己的数据部署到一个比较合适的位置。随着时间的推移，一个数据的用户群可能发生变化，密集的请求在不同的时间段来自不同的地理区域。

当网格上的某个数据集需要移动时，表明该数据文件在这个节点上的存在已不能满足给网格上的用户提供所需要的服务质量，或者造成了传输资源的浪费。在下列情况下，网格上的数据集有移动的需求：(1)存储该数据集的节点上的数据访问请求负担过重，无法及时响应请求者的访问要求。(2)请求该数据时的平均传输距离超过了网格的半径。(3)请求一个数据集的用户群发生了位置上的偏移。

1.2 数据移动所产生的问题及解决途径

当网格环境中产生数据移动需求时，可以考虑进行数据的移动，但在数据的移动过程中及移动之后会产生一些问题。数据的移动可能影响正在使用的数据，使得用户在数据的移动过程中无法请求使用数据。为避免这种情况的发生，采用对源数据集创建副本的方式进行数据的移动可以较好地解决问题。

数据移动之后应该让原来访问该数据的请求在不做任何改变的情况下仍然能够访问到该数据。这一问题的解决可以通过副本的定位服务来完成。即数据的逻辑名字中不包含与数据的具体位置有关的信息，每个数据都有一个全局唯一的逻辑数据名。当请求者通过数据文件的名字来访问数据时，就可以做到数据的移动不影响用户原来的访问。

1.3 负载均衡分析

由于网格中的各节点机器具有高度的自治性，同时网格系统也是一个典型的分布式系统，因此在网格环境下可能发生某些节点负载很重，节点上的数据访问请求得不到及时响应，而另一些节点却几乎处于空闲状态，这就是负载不平衡的状况。负载均衡就是将重负载节点上的访问请求转移到轻负载的节点上，使得网格中各节点数据的访问负载趋向平衡，以缩短用户访问请求的响应时间，从而改善整个系统的性能。一个动态负载均衡算法通常包含以下 3 个组成部分[2,3]：

(1)转移策略：用于决定在什么条件下一个节点需要向另一个节点迁移部分或全部数据集。在数据移动服务中采用阈值策略来判定数据是否需要转移。

基金项目：国家自然科学基金资助重点项目(60433020)
作者简介：梁 英(1977—)，女，讲师、硕士，主研方向：网格计算；胡志刚，教授、博导
收稿日期：2005-10-08 **E-mail：**yingl@126.com

(2)定位策略：负责寻找一个结点的搭档结点。当转移策略确定了某结点为发送者时，它负责寻找接收者。在数据移动服务中它负责决定将数据集的副本移动到合适的节点上。

(3)信息交换策略：用于确定何时、到何处去搜集有关系统中其它结点的状态信息。这一部分的工作可以由网格监控服务完成，本文不作具体的讨论。

2 数据移动服务概要设计

数据移动服务主要存在于网格体系结构中的汇聚层，如图 1 所示。数据的移动服务独立于存储系统的存储技术和数据在网络中的传输协议。

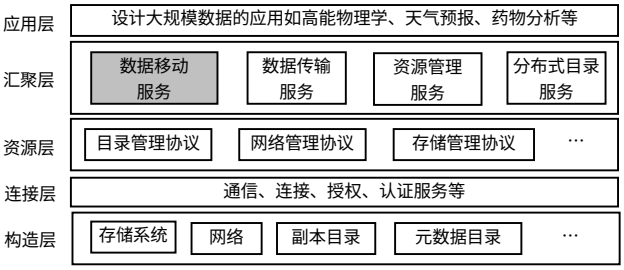

图 1 数据移动服务在网格体系中的位置

在数据移动服务中主要采用建立副本的方式进行数据的移动，服务中的主要构成部分包括：数据副本的管理和副本目录的维护。数据副本管理主要负责副本的创建、数据一致性维护和副本的删除工作。副本目录的维护主要是在副本创建后必须及时修改相关的副本信息，从而将用户对数据访问引导到新建立的副本上去。数据移动服务的组成以及该服务与其它网格服务的关系如图 2 所示。

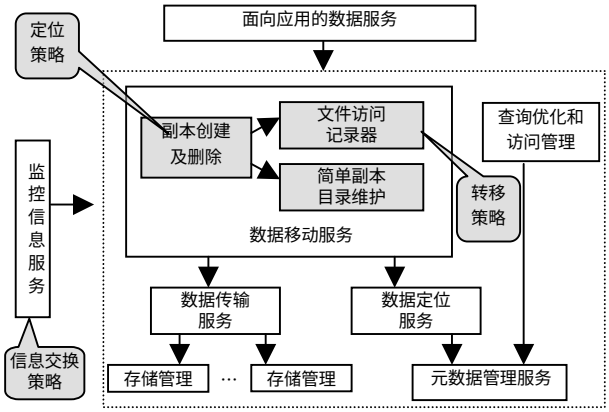

图 2 数据移动服务的组成及与其他服务的关系

3 数据移动服务详细设计

3.1 副本数据创建

副本创建时应考虑的因素有：运行系统的访问负载，网络状况、数据副本的大小、用户的访问特征等。数据移动服务中副本创建的主要目标是减轻现有节点的访问负载，实现系统内的负载平衡，考虑的主要因素是系统运行的负载情况。

3.1.1 副本创建触发

数据移动服务接收到网格监控系统发来的网格内任务的执行状况和主机的运行信息后，按照一定的系统负载预测方法获取存储节点的负载值L[4]，再结合系统内用户对数据文件的访问热度，由数据移动服务根据PQ参数原则触发副本创建进程。

设系统内有N个可用存储节点，其中含有同一数据文件(包括元数据文件及其副本)的存储节点有$S_1,S_2,\ldots,S_m$(其中m<n)，节点对应的负载值为$L_1,L_2,\ldots,L_m$。

设置参数 P、Q(P<Q)，P 值要求系统可以对剧烈增长的用户请求作出及时响应，Q 值则对副本创建更为谨慎，其允许数据文件及副本主机服务器性能暂时波动；T 为系统内用户对数据文件在单位时间内的访问次数阈值；L 为系统内存储了同一数据文件的节点的负载平均阈值。

判断触发时机步骤如下：

(1)若系统中的用户单位时间内访问数据文件的次数小于 T，则放弃副本的创建；

(2)在 P 时段内，若对于 P 时段内的任意时间 t

$(\sum_{i=1}^{m} L_i /m)_t \geq L$

则转(4)；

(3)在 Q 时段内，若对于 Q 时段内的某一时间 t

$(\sum_{i=1}^{m} L_i /m)_t < L$

则放弃副本创建；

(4)触发副本创建服务。

3.1.2 副本创建策略

副本创建服务触发后，必须根据一定的策略进行副本的创建工作，常见的副本创建策略有以下几种：最优客户端策略，串联策略，快速广播策略[5]，T-Value域间副本扩展策略和Economic Model[6]策略等。

以上的副本创建策略在进行副本创建时所考虑的主要指标是网络带宽和访问延迟。但在基于负载均衡的数据移动服务中，所考虑的主要指标是系统的负载情况，而以上策略均不能较好地满足负载平衡的要求。因此，提出了一种新的基于负载均衡的副本创建策略。此策略分为域内副本创建和域间副本创建两部分。

域内副本创建策略描述如下：设本地域内有N个可用存储节点，若这些节点中存在单个节点在单位时间内对某一数据文件的访问次数超过了设定的阈值T，则直接在此节点上创建该数据文件的副本。否则，从N个节点中随机选取节点上未含有频繁访问数据文件(包括源数据文件及其副本)的存储节点$W_1,W_2,\ldots,W_k$(其中k<n)。设这些节点的负载分别为$L_1,L_2,\ldots,L_k$；计算节点对应的概率$P_1,P_2,\ldots,P_k$，取概率值最大的服务器W_i，在此节点上创建副本，其中

$$P_i = X_i \Big/ \sum_{j=1}^{k} X_j$$

$$L_{sum} = \sum_{i=1}^{k} L_i$$

$$X_i = \frac{L_{sum} - L_i}{L_{sum}} \quad , i=1,2,\ldots,k$$

域间副本创建策略为：若频繁的访问请求来自远程域，则将数据文件副本在远程域的管理节点上进行创建，然后由远程域的管理节点根据文件的访问情况进行域内副本创建。

3.2 相关的目录设计

根据网格环境局部自治性的特点，网格上的服务主机拥有对本机上所有资源的管理权，它可以决定自己的数据文件可供外部用户使用还是只对本地用户开放。按照网格环境中层次化的管理结构，可以将数据移动服务中用到的目录分为两种：本地副本目录和全局副本目录。本地副本目录中存储的信息是关于本地节点上存储的所有副本的映射信息，提供本地副本的查询能力。全局副本目录中存储的信息是上层网格中全部副本的索引信息。由于上层网格中只保存索引信息，因此可减少存储和更新的开销。

3.3 数据一致性维护

数据移动服务是采用生成副本的方式进行的，那么当用户对副本或源文件进行了写操作时，必须考虑数据的一致性维护工作。如果要实现完整语义的数据一致性，就要考虑采用复杂的文件加锁机制或者协商机制等方法。由于数据的强一致性维护实现太过复杂，代价也很高，因此可考虑只进行简单的弱一致性维护。在分层结构中，将用户对数据的修改以异步的方式通知到所有的备份。

3.4 文件访问记录器的设计

文件访问记录器是为满足数据移动服务的需要而设计的。该记录器的主要功能是记录本地或远程用户对文件的访问情况，如访问的用户、用户所在管理域、访问的次数等。

3.5 副本的删除

在数据移动服务中，如果某个网格节点要接收一个新的远程数据文件的副本，但存储空间不足，这时就必须选择一些利用率相对较低的其他数据文件的副本进行删除。在副本删除时应先将副本目录中的有关信息删除，然后再将副本数据删除。

4 性能分析

用户对文件的访问大多数时候遵循一定的规律。在模拟网格环境下主要采用两种访问模型研究数据移动服务的性能：(1)时间连续性访问模型，在这个模型中，如果一个用户访问了某个文件一次，则在不久的将来，该用户很可能再次访问此文件。(2)地理连续性访问模型。在这个模型中，如果一个用户访问了某个文件，那么该用户的邻居用户也很有可能去访问该文件。在这两种访问模型下数据移动服务可以有效地减少远程访问所造成的响应时间延迟，同时也可以使整个模拟网格的数据访问不再集中在某一个节点上，从而达到负载平衡的目的。并且，随着访问情况的时间连续性和地理连续性越明显，数据移动服务的效果就越好。

5 结束语

网格数据管理是网格中的关键技术之一。网格中的数据在使用过程中，随着时间的推移，对数据的密集请求也会发生迁移。此时，为避免大量数据的远程传输，减轻单个服务节点的负载压力和由于频繁远程访问对网络带宽的消耗，满足网格用户所需的服务质量，就必须考虑数据的移动。当数据产生移动需求时，要考虑在适当的时候从多个候选节点中选择合适的节点作为移动数据的目的节点。良好的数据移动策略可实现网格节点的负载均衡并可提高数据请求的响应速度，进而提高整个网格的运行效率和服务质量。

参考文献

1 Foster I, Kesselman C. The Grid: Blueprint for a New Computing Infrastructure[M]. Morgan Kaufmann Publishers, 1999: 6-8.

2 薛　军, 李增智, 王云岚. 负载均衡技术的发展[J]. 小型微型计算机系统, 2003, 24(12): 2100-2103.

3 Watts J, Taylor S. A Practical Approach to Dynamic Load Balancing[J]. IEEE Transactions on Parallel and Distributed Systems, 1998, 9(3): 235-248.

4 李庆华, 郭志鑫. 一种面向工作站网络的系统负载预测方法[J]. 华中科技大学学报, 2002, 30(6): 49-51.

5 Ranganathan K, Foster I. Identifying Dynamic Replication Strategies for a High-performance Data Grid[C]. Proceedings of International Workshop on Grid Computing, Denver, CO, 2002-12.

6 Lamehamedi H,　Szymanski B, Shentu Z. Data Replication Strategies in Grid Environments[C]. Proc. of the 5th Intl. Conf. on Algorithms and Architectures for Parallel Processing, 2002: 378-383.

(上接第 110 页)

2.2 一个实例

根据表 1 提供的信息，可举例说明算法的工作过程。约定符号意义：网桥 i 用 Bi 表示，网桥 i 上的端口 j 用 Bi,j 表示；转发链路用()表示，冗余链路用[]表示。

上述算法扫描每个网桥的每个端口得到如下连接：

(B1,2 , B2,1) ; (B1,1 , B3,1) ; [B2,2 , B3,2] ; (B1,1 , B4,1); [B2,2 , B4,2]

经过统计发现 B1,1 和 B2,2 分别与多个端口之间存在链路，所以所有以这两个端口为端点的链路通过共享网段，根据算法拆分链路得到最终的连接状况：

(B1,2, B2,1) ; (B1,1, LAN A) ; (B3,1 , LAN A) ; (B2,2, LAN B) ; [B3,2, LAN B] ; (B1,1, LAN A) ; (B4,1, LAN A) ; (B2,2, LAN B) ; [B4,2, LAN B]

该算法所得到的是与图 1 吻合的正确的网络拓扑。

3 结论

本文提出的算法只需要对每个端口的生成树信息进行分析就可以计算出网络的拓扑。假设网络中有M个网桥，第i个网桥的端口数是N_i个，那么算法的时间复杂度为$O(n) = O(\Sigma N_i)$。由于端口的生成树信息相比端口的地址转发表信息要少的多，因此本算法相比文献[1]和文献[2]中的算法更简单，效率更高，并且能够发现它们所无法发现的冗余链路。然而由于算法基础的限制，无法直接发现网桥和主机、服务器、路由器等网络设备的连接情况，但在本算法的基础上，可以再利用端口的地址转发表信息发现网桥和其它网络设备的连接关系。

IETF 没有定义 IEEE 802.1S 多生成树的网桥 MIB，运用本算法的思想无法发现交换网络中 VLAN 的拓扑情况，这一方面需要进一步的研究。

参考文献

1 Lowekamp B, O'Hallaron D R, Thomas R. Gross Topology Discovery for Large Ethernet Networks[C]. Proc. of SIGCOMM'01, San Diego, California, USA, 2001-08-27.

2 Bejerano Y, Breitbart Y, Garofalakis M, et al. Physical Topology Discovery for Large Multi-subnet Networks[Z]. Bell Labs Tech. Memorandum, 2002-07.

3 Decker E, Langille P, Rijsinghani A, et al. Definitions of Managed Objects for Bridges[S]. RFC 1493, 1993-07.

4 Case J, Fedor M, Schoffstall M, et al. Simple Network Management Protocol[S]. RFC 1157, 1990-05.

5 IEEE 802.1D. Standard for Local and Metropolitan Area Networks: Media Access Control (MAC) Bridges[S]. 1998.

6 IEEE 802.1W. Standard for Local and Metropolitan Area Networks: Media Access Control (MAC)Bridges——Amendment 2: Rapid Reconfiguration. 2001.

7 Comer D E. 林　瑶, 蒋　慧, 杜蔚轩等译. 用 TCP/IP 进行网际互联(第 1 卷): 原理、协议与结构(第 4 版)[M]. 北京: 电子工业出版社, 2003.