维普资讯 http://www.cqvip.com

一种基于Grid的多媒体内容分布框架*)

李春洪　顾铁成　李文中　陆桑璐　陈道蓄
(南京大学计算机软件新技术国家重点实验室　南京大学计算机科学与技术系　南京210093)

摘　要　基于互联网的大规模多媒体内容分发系统中,广泛地采用了数据复制技术来提高系统的性能。网格技术在构建大规模分布式信息系统方面,有着广阔的前景。本文提出了一种基于网格的多媒体内容分布框架,在网格基础服务的基础上,可以实现适用于流媒体应用的副本管理、传输和查找定位机制。此外,本文还对广域分布式信息系统中,副本的放置算法和用户请求调度机制等关键技术进行了深入探讨。

关键词　复制,网格,副本放置,流媒体

A Grid-Based Multimedia Content Distribution Framework

LI Chun-Hong　GU Tie-Cheng　LI Wen-Zhong　LU Sang-Lu　CHEN Dao-Xu
(State Key Laboratory for Novel Software Technology, Department of Computer Science and Technology, Nanjing University, Nanjing 210093)

Abstract　Replication techniques are widely employed to improve the performance of large-scale content distribution systems. There is a big future for Grid technology in building large-scale distributed information systems. Based on basic Grid services, we propose a Grid-based multimedia content distribution framework, which offers a set of key functions to meet the requirements of continuous multimedia applications, such as replication management, replica transferring and discovering etc. Besides, this paper also discusses some key techniques in detail, such as replica placement algorithms and request scheduling mechanism.

Keywords　Replication, Grid, Replica placement, Continuous media

1　引言

设计和构造分布式系统时,对服务和数据进行复制是一种常用的手段[1]。在以内容服务为中心的大规模分布式系统中,如Web缓存和预取系统、内容分发网络、分布式数据库和P2P文件共享应用等,在网络中创建和分布内容副本可以极大地提高系统的总体性能[2~4]。一般来说,内容复制从以下几个方面增强系统的性能:内容更加接近客户的访问点,用户请求的响应速率得到提高;多副本的存在,使系统可以更好地容忍因网络分割和节点失效等形成的资源间断,系统的可用性得到增强;通过合理配置副本,分散了服务负载,系统的总体吞吐量和资源利用率得到优化,进而增强了系统的可扩展性。

网格计算[5](Grid computing)是一种新型计算模型,对属于不同虚拟组织的异构和动态资源进行协同共享和问题求解,是大规模资源共享和服务协同的有效机制。除了在高性能计算领域得到了成功的应用之外,网格在数据处理、信息共享、服务协同等领域得到了广泛的关注。我们也在自己的研究工作中,针对网上流媒体服务问题,考虑利用网格技术将地域上分布的、异构的流媒体服务器组织起来,统一和协同地为用户提供流媒体服务[7]。和普通Web内容服务不同,流媒体服务持续时间长、带宽要求高。当服务的客户增多时,会产生大量的网络负载和服务器负载。特别是热门媒体服务器和附近的网络处于过载状态,导致高的请求拒绝率。因此,如果在流媒体服务网格中引入动态内容分布机制,即依据服务需求的分布状况,通过内容复制技术将热门影片"推向"多个服务节点,则可以使得流媒体服务负载在网络中均衡分担,从而提高系统的服务能力和服务质量。

本文提出了一种基于Grid的多媒体内容分布框架,为大规模环境下多媒体内容副本的表达、查找、分布和请求调度等提供了统一支持。文章第2节对相关工作进行了介绍;第3节提出了基于Grid的多媒体内容分布框架及其主要组成部分;第4节对副本的放置问题和用户请求调度机制等关键技术进行了深入的探讨;最后对全文进行了总结。

2　相关工作

近年来,利用复制技术进行互联网上有效的内容分布得到较广泛的研究[2~4]。根据复制的实体的不同,可以进一步划分为服务器副本的放置问题和内容对象副本的放置问题。服务器的放置问题以镜像服务器/代理服务器在互联网上的放置为研究目的,可以用K-Median问题[3]来对它进行刻画:在N个节点中选择K个($N \geq K$)节点放置镜像/代理服务器,且每个节点最多放置一个服务器,使得总的网络开销最小。K-Median问题已经被证明是NP完全问题[10],文[11,12]提出了实用的近似算法。由于每个镜像服务器/代理服务器包括了站点的所有内容,因而服务器的放置只能实现粗粒度的复制。以内容文件为复制对象可以实现更加精确的复制:在一定网络资源的约束下,考虑内容对象的工作量(流行度)的差异,确定副本的数量和在网络中的放置。内容副本的放置问题可以

*)基金项目:国家"973"项目(编号:2002CB312002)、国家"863"项目(编号:2001AA113050)。李春洪　博士研究生,研究方向为分布式计算、多媒体网络的QoS;陈道蓄　教授,博士生导师,研究方向为分布式系统、Internet计算、计算机支持的协同等。

维普资讯 http://www.cqvip.com

看成是经典的文件分配问题(FAP: File Allocation Problem)[13]在基于互联网的广域系统上的扩展。文[13]证明了FAP是NP完全问题,文[14]降低了FAP问题对节点存储能力的限制,并提出了一种有效的叠代算法。对于这一问题,在实际应用中,通常通过利用启发式算法寻求具有多项式复杂度的次优解。文[4]对内容副本放置问题中的启发因子进行了深入的研究,文[15]等对CDN中常用的启发式算法进行了实验比较。

网格技术对不同管理域内的异构和动态资源进行协同共享和问题求解,其终极目标是像电力网格等公用设施一样,无缝地为用户提供计算资源和服务[5]。虽然Globus[6]是面向高性能计算的网格中间件,但仍然认为"对分布式数据的访问和访问分布式计算型资源同样重要"。在这种理念的支持下,从Globus ToolKit 2.0开始就推出了数据管理模块,其主要组成部分包括:全局二级存储服务GASS、GridFTP和Globus复制目录等。Globus的数据管理模块为网格环境下数据副本的创建、传输和查找提供了有力的支持。文[16]提出了数据网格环境中在线创建数据副本的算法,该算法通过一个开销/收益模型来进行复制决策:计算在某个节点上建立数据副本的开销和带来的性能收益,以收益是否大于开销作为是否在该节点建立数据副本的依据。Sudharshan等针对数据密集型网格中大体积数据块(bulk data)的传输问题,提出了基于性能预测的副本选择与协同分配和调度机制[17]。

3 基于Grid的多媒体内容分布框架

我们基于Globus建立了多媒体内容的分布框架,如图1所示。系统主要组成部分有:

·多媒体内容元信息目录服务:存放系统中各多媒体内容对象的逻辑ID和描述信息(如文件大小、版本信息、编码格式、传输速率、图像质量等),并且提供外部的查询接口;

·内容副本目录:建立内容副本物理存储位置和逻辑ID的映射,并提供外部查询接口;

·内容复制管理器:负责内容副本的创建、删除和网络传输过程的监控和协调。在副本创建时,内容复制管理器向内容副本目录注册副本的信息,并对副本的传输过程进行监控。如果传输过程失败,实施回滚(roll back)操作,维持副本目录的一致性。在删除某个内容副本时,内容复制管理器负责将副本信息从副本目录中删除;

·副本选择服务:按照特定的性能指标,从多个候选副本中选择一个或者多个满足要求的副本;

·网格信息服务:是网络基础设施的一部分,提供网络中各资源的基本信息(如节点的硬件配置、所使用的操作系统类型和版本、资源管理系统的类型和版本等)、负载信息(如可用存储空间、CPU时间和网络带宽)以及用户自定义的定制信息等;

·访问请求统计分析服务:为了对访问工作量的分布情况进行分析和预测,访问请求统计分析服务记录所有用户的历史请求信息,包括请求者、所属虚拟组织、所请求内容的逻辑ID等;

·副本放置决策:确定当前工作量分布状况下的副本放置策略,其结果提交给内容复制管理器实施。

·副本的创建和传输服务:实施副本的传输和创建,其过程受到内容复制管理器的监控。

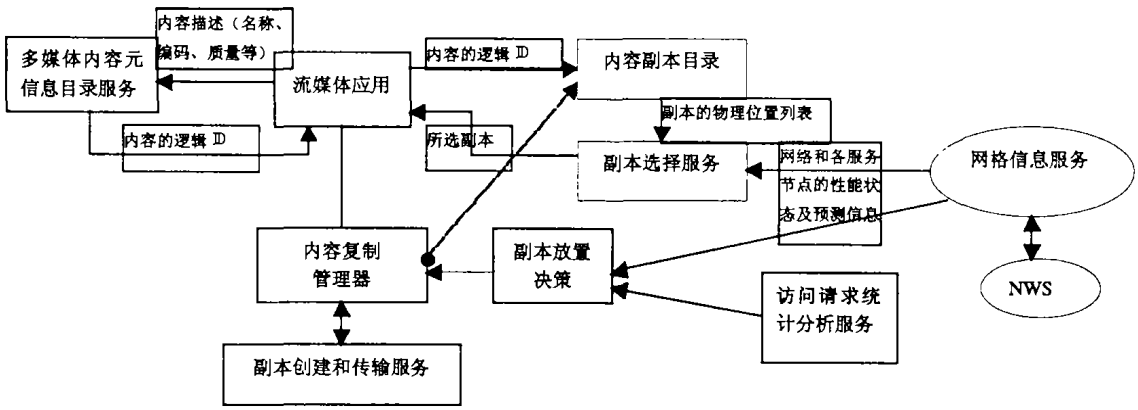

图1 基于Grid的多媒体内容动态分布框架

3.1 应用场景

基于上述多媒体内容分布框架,一个典型的应用场景可以是这样的:

(1)用户/流媒体应用向多媒体内容元信息目录服务发出查找请求,提交所需内容的描述信息,如影片名、编码格式、分辨率等;

(2)多媒体内容元信息目录服务根据描述信息进行匹配查找,返回所请求内容的全局唯一的逻辑ID;

(3)用户/流媒体应用利用内容的逻辑ID查询副本目录服务,获得该内容在系统中所有副本的物理位置信息列表;

(4)利用副本选择服务,对内容的各副本进行评估,从中选择最优的副本,并将其位置信息返回给用户/流媒体应用;

(5)用户/流媒体应用直接建立与内容副本的访问会话。

经过一段时间运行后,系统根据"访问请求统计分析"服务对访问需求重新评估。副本放置决策模块计算新的内容副本分布策略,内容复制管理器调用"副本创建和传输"服务实施从内容的主拷贝向副本所在节点的传输。

3.2 副本的定位

在流媒体服务网格中,一个多媒体内容对象可能在网络中的多处建立了副本。为了使用户能够透明地访问该对象,可以用一个全局的、唯一的逻辑ID来识别该对象及其所有的副本。我们基于Globus的副本目录结构(Replica Catalog),建立了多媒体内容副本目录,以实现逻辑ID到副本物理存储位置的映射。

当新建一个内容副本的时候,复制管理器需要向副本目录的位置项(Location)和逻辑文件项(Logical file)注册信息:

·逻辑文件项(Logical file):逻辑文件项包含多媒体内容的全局唯一的逻辑ID。

维普资讯 http://www.cqvip.com

·位置项(Location):位置项主要包含访问内容副本的URL表示。此外还包括了副本所在物理存储系统的信息(如主机名、端口号和协议等)。

同样,在副本被删除的时候,复制管理器也应该对相应的注册项进行删除操作。

在查询时,副本目录对查询输入进行匹配操作,并返回所有匹配的条目。类似于Globus MDS,Globus副本目录也建立在LDAP(Lightweight Directory Access Protocol)之上。在以前的工作[7]中,我们针对流媒体应用需求对Globus MDS的查询机制进行了改进,利用缓存查询结果和构建Hash查找表等方法,极大地提高了查询的响应速度。在多媒体内容副本目录服务的实现中,也引入了这种设计思想,以提高内容副本定位的效率。

4 关键问题与技术

4.1 副本的优化放置问题

由于网络中的带宽和服务器资源有限,因此内容副本的放置是一种带约束条件的优化问题,可以这样来定义:如何在网络资源和服务器计算及存储资源受限的条件下,动态确定副本的数量和放置的位置,以满足系统的性能和QoS的需求。

4.1.1 *形式化描述* 通常可用一个无向图$G(V,E)$来对网络建模(如图2所示),其中$V=\{v_1,v_2,\cdots,v_I\}$为服务节点的集合,$|V|=I$为服务节点的数量;$E\subseteq V\times V$为G中的边,代表服务节点之间的连接链路,并且被赋以一定的开销值,该数值可以为网络链路的传输延迟、经过的网络跳数或者其它被考察的性能指标。

针对副本的放置,做如下假设:系统中各服务节点具有不同的存储空间,其中节点v_i的存储能力为$S_i(i\in(1,I))$;系统中的内容对象总数为J,内容对象j所需的存储空间为b_j,其流行度为p_j,其中$j\in(1,J)$。

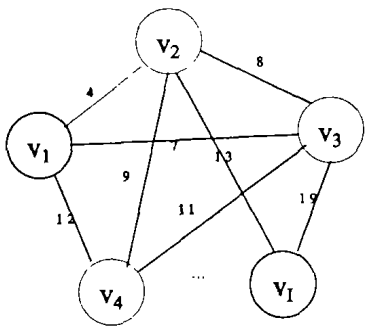

图2 内容副本放置的网络模型

用一个布尔值x_{ij}来表示对象j在节点v_i的放置情况。如果x_{ij}取true,表示节点v_i上存在对象j的副本;反之,如果取值false,则表示节点v_i中不存在对象j的副本。因此,某一时刻网络中副本的分布快照可以用一个$I\times J$的矩阵表示:

$$X=\begin{vmatrix} X_{11} & \cdots & X_{1J} \\ \cdots & \ddots & \cdots \\ X_{I1} & \cdots & X_{IJ} \end{vmatrix}$$

为了进行副本的优化放置决策,定义一个开销模型(cost model)来描述优化的目标。令$d_{ij}(X)$为在某个副本配置X下,始于节点v_i的对内容j的访问请求到最近的副本访问开销。假定存在一个请求调度机制将用户的请求重定向到最近的副本,则所有来自节点v_i的请求的访问开销为:

$$C_i(X)=\sum_{j=1}^{J}p_jd_{ij}(X) \tag{1}$$

设发起自节点v_i的请求率为λ,则总的系统访问开销为:

$$C(X)=\sum_{i=1}^{I}\lambda C_i(X)=\sum_{i=1}^{I}\sum_{j=1}^{J}\lambda d_{ij}(X)$$

至此,副本的放置问题可以形式化描述为:求解副本的优化配置X,使得在满足约束条件$\sum_{j=1}^{J}b_jx_{ij}\leqslant S_i,i=1,\cdots,I$的前提下,$C(X)$取最小值。

4.1.2 *启发式算法* 以上描述的副本放置问题是NP完全问题,无法在多项式时间和空间复杂度内求得最优解。在实际系统中,通常采用启发式方法来寻求次优解[3,4,15]。常用的启发式算法有随机放置算法、基于流行度的放置算法和贪婪算法等:

·随机放置算法:随机放置算法的基本思想是在节点存储能力约束下,随机放置内容的副本。一个对象可以在多个节点上创建副本,但是一个服务节点最多只能拥有某个对象的一个副本。

·基于流行度的放置算法:节点通过统计历史请求记录来估计各对象的流行度(popularity),且将对象按照其流行度的降序排列。在存储能力的约束下,各节点存储流行度最高的前n个对象。

·Greedy Single算法:算法的基本思想是,在初始对象分布X_0下,每个节点i为所有对象j计算访问开销$C_{ij}=p_jd_{ij}(X_0)$。这代表了在初始配置的情况下,每个对象对式(1)的贡献。节点i根据C_{ij}的降序将对象排列,并且在其存储能力的约束下,尽可能存储最多的对象。

·Greedy Global算法:算法的基本思想是,在初始内容分布X_0下,每个节点计算访问各对象的开销$C_{ij}=p_jd_{ij}(X_0)$,选择具有最高访问开销的"节点-对象"对,并在该节点上建立此对象的副本。此操作形成新的副本放置X_1。在X_1下,重复上面的操作,直至所有的节点的存储空间已满。

4.1.3 *讨论* 以上算法中,随机算法具有最小的复杂度,但由于它没有考虑不同内容间带来访问工作量的差异,该算法只是在内容的流行度分布均匀的时候有比较好的效果。基于流行度的放置算法考虑了内容的流行度的差异,只依据本地信息进行复制决策,易于分布式实现。但是该算法没有考虑同一内容在其它节点上的副本放置情况,可能造成某些内容在系统中存在过多的副本;Greedy single算法对基于流行度的算法作了改进,按照式(1)计算每个内容在初始内容放置X_0下的访问开销,以此作为在本地建立副本的依据。但是,为了估计C_{ij}值,系统必须了解有关当前网络拓扑的信息;Greedy Global算法采用了叠代算法,每次决策的结果作为下一次计算的输入,其得出的副本放置解也更加接近全局的最优解。但是,由于采用了多次叠代,Greedy Global算法具有更大的计算开销。

4.2 请求调度机制

假定存在一个重定向机制,它可以保证将用户的访问请求调度到最近的副本去接收服务。系统在实际运行时,网络链路和节点的可用资源会受到许多不确定因素的影响,和用户邻近的服务副本往往无法提供满意的服务质量。因而,用户请求调度机制必须对副本进行动态评估和选择[18]。此外,流媒体服务通常需要端到端的QoS保证(如传输速率、延迟、抖动和差错率等),服务节点、通信链路以及用户端设备等网络部

维普资讯 http://www.cqvip.com

件都必须有足够的可用资源保证用户感知播放质量。因此，在存在多个内容副本的流媒体服务系统中，在选择服务节点时，还必须综合考察网络链路能力和节点负载等多种指标[19~21]。

通过网格信息服务，可以获取网格环境下各资源的动态性信息，如服务器的可用 CPU 时间、内存以及出口带宽等。此外，网格信息服务还提供了扩展接口，允许驻留在各资源的自定义信息提供者向网格信息服务提供定制的信息。为了对多个候选副本进行选择，我们为每个副本所在节点定义了一个聚合的指标 Φ，并选择 Φ 值最大的副本作为请求的服务节点。假设为了满足一定的 QoS 等级，所需的节点资源为 $\vec{R}=\{r_1, r_2, \cdots, r_m\}$，网络资源为 B。某时刻副本所在服务节点 j 的可用资源为 $\vec{RA}=\{ra_{j1}, ra_{j2}, \cdots, ra_{jm}\}$，可用网络资源为 Ba_j，则聚合指标 Φ_j 定义如下：

$$\Phi_j=\sum_{i=1}^{m} w_i \cdot \frac{ra_{ji}}{r_i}+w_{m+1}\cdot\frac{Ba_j}{B}$$

其中，$w_i(1\leqslant i\leqslant(m+1))$为非负数值，且满足 $\sum_{i=1}^{m+1} w_i=1$。w_i 代表了各资源在副本选择中的重要性。

由于流媒体内容在编码格式、编码速率、解析度等方面存在很大的差异，在实践中，我们发现很难确定某内容对象对各类服务器资源的具体需求。为此，正在建立一种基于测量的节点性能预测机制，抽象出服务节点针对不同 QoS 等级的流媒体服务的接入概率 Pa_j[8]，并通过自定义信息提供者向网格信息服务注册。因此，副本的选择问题可以这样来简化，即依据节点接入概率和链路可用带宽两类资源的聚合指标进行选择。

总结 复制技术在大规模内容分发系统中得到了广泛的使用。通过在网络的多个位置部署内容的副本，降低了访问延迟和网络的流量，同时在多个服务器中分布了访问负载，系统的可扩展性也得到了提高。然而，大规模环境下的内容复制存在很多热点问题有待深入探讨，如副本的管理、副本的优化放置和副本的选择等。本文提出了一种基于 Grid 的内容分布框架，利用基本网格服务建立了适合于流媒体内容分发的副本管理、传输和查找定位机制，并对内容副本的优化放置算法和请求调度机制进行了深入的探讨。

参 考 文 献

1 Wu Jie. Distributed System Design. CRC Press, 1999

2 Akamai. http://www.akamai.com, 2002

3 Loukopoulos T, et al. An Overview of Data Replication on the Internet. In: Intl. Symposium on Parallel Architectures, Algorithms and Networks (ISPAN '02), 2002

4 Kangasharju J, et al. Object Replication Strategies in Content Distribution Networks. In: 6th Intl. Web Caching Workshop and Content Delivery Workshop, Boston, MA, 2001

5 Foster I, Kesselman C, Tuecke S. The Anatomy of the Grid: Enabling Scalable Virtual Organizations. Intl. J. Supercomputer Applications, 2001, 15(3)

6 Foster I, Kesselman C. Globus: A Metacomputing Infrastructure Toolkit. Intl J. Supercomputer Applications, 1997, 11(2): 115~128

7 视频服务网格系统结构设计：[南京大学计算机科学与技术系技术报告]. 2003

8 许平. 流媒体系统接纳控制研究：[南京大学硕士学位论文]. 2004

9 Allcock W, Chervenak A, Foster I, et al. The Data Grid: Towards An Architecture for the Distributed Management and Analysis of Large Scientific Datasets. Network and Computer Applications, 2002

10 Mirchandani P, Francis R. Discrete Location Theory. John Wiley and Sons, 1990

11 Li B, et al. On the Optimal Placement of Web Proxies in the Internet. In: Proc. of the IEEE INFOCOM'00 Conf. March 2000

12 Vigneron A, Gao L, Golin M, et al. An Algorithm for Finding a k-median in a Directed Tree. Information Processing Letters, 2000, 74: 81~88

13 Dowdy L W, Foster D V. Comparative Models of the File Assignment problem. ACM Computing Surveys, 1982, 14(2)

14 Mahmoud S, Riordon J. Optimal Allocation of Resources in Distributed Information Networks. ACM Trans. Database Systems, 1976, 1(1): 66~78

15 Karlsson M, Karamanolis C. Choosing Replica Placement Heuristics for Wide-Area Systems. In: Proc. of the Intl. Conf. on Distributed Computing Systems (ICDCS), March 2004

16 Lamehamedi H, et al. Data Replication Strategies in Grid Environments. In: Proc. 5th Intl. Conf. on Algorithms and Architecture for Parallel Processing, Bejing, China, Oct. 2002

17 Vazhkudai S. Distributed Downloads of Bulk, Replicated Grid Data. Journal of Grid Computing, 2004 (to appear)

18 Wang Limin, Pai V, Peterson L. The Effectiveness of Request Redirection on CDN Robustness. In: Proc. of the Fifth Symposium on Operating Systems Design and Implementation, Boston, MA, Dec. 2002

19 Huang An-Cheng, Steenkiste P. Network-Sensitive Service Discovery. Journal of Grid Computing, 2004 (to appear)

20 Gu Xiaohui, Nahrstedt K. A Scalable QoS-Aware Service Aggregation Model for Peer-to-Peer Computing Grids. In: Proc. of the 11th Intl. Symposium on High Performance Distributed Computing HPDC-11 2002

21 Fu Zhenghua, Venkatasubramanian N. Combined Path and Server Selection in Dynamic Multimedia Environments. In: Proc. of ACM Multimedia '99, Orlando, FL, Nov. 1999