一种新的动态副本管理机制*)

侯孟书[1,2] 王晓斌[1,2] 卢显良[1] 任立勇[1]

(电子科技大学计算机科学与工程学院 成都 610054)[1] (电子科技大学成都学院计算机系 成都 611731)[2]

摘 要 提出了一种新的动态副本管理机制。该机制根据文件受欢迎的程度,增加受欢迎程度高的文件副本数量,选择高性能节点存放文件副本,从而使系统自动调整文件副本数量以及副本存放位置,平衡节点负载,提高文件的可用性。

关键词 副本,对等网络,Zipf 分布

A Novel Dynamic Replication Management Mechanism

HOU Meng-Shu[1,2] WANG Xiao-Bin[1,2] LU Xian-Liang[1] REN Li-Yong[1]

(School of Computer Science & Engineering, UESTC, Chengdu 610054)[1]

(Department of Computer Science, Chengdu College, UESTC, Chengdu 611731)[2]

Abstract This paper presents a novel dynamic replication management mechanism called "DynRM". In DynRM, the system distributes the file replication according to the degree of the file popularity, and selects the high capacity node to store the popular replication. The system adjusts the location and amount of replication according to the replication strategy in order to balance the node load and improve the file availability.

Keywords Replication, Peer-to-Peer network, Zipf distribution

1 引言

在P2P网络中,不同的节点,其计算能力、存储能力、网络带宽是不同的,甚至差异很大。通过合理地选择高性能的节点放置副本,可以大大提高系统处理数据的能力,减少访问延迟,提高系统的整体性能。

文[1]提出了一种模型驱动的动态副本管理机制,该机制从文件可用性出发动态管理副本,能较好地适应P2P网络环境,保证文件可用性,但该模型没有考虑访问效率以及维护数据一致性的开销。文[2]分析了在非结构化P2P系统中,如何在系统节点不断变化的情况下,保持系统中副本的高可用性。Gnutella[3]中没有明确的副本策略。FastTrack[4]将高带宽的节点作为超节点(supernode),每个超节点对邻居节点的共享数据建立索引副本。这和本文的副本策略有区别,本文将副本放置在超节点上,而FastTrack将共享文件的索引放置在超节点上。Freenet[5]支持主动副本,在查询命中返回的路径上放置副本,但是Freenet没有考虑节点的异构性。本文的副本策略考虑了系统中节点的异构性,将副本仅仅放置在高性能节点上。

文[6]针对结构化P2P存储系统提出了一个无中心的体系结构,并建立了相应的模型,分析了受欢迎文件的相应时间以及节点有限容量的相关问题。但是该模型存在以下问题:

* 路由算法没有采用邻近度路由技术,使受欢迎的文件副本不能在短时间内进行有效的扩散。

* 没有考虑节点的异构特征,在P2P网络中,节点在处理能力、带宽、存储能力、在线时间等方面呈现多样性。如果将副本放置在带宽、处理能力比较低的节点上,不但影响了放置节点的性能,而且对整个系统的性能提高不多。

本文针对文[6]存在的问题,提出了一种动态副本管理机制DynRM(dynamic replication management),该机制根据副本的受欢迎程度,对受欢迎程度高的文件,建立更多的副本,从而使系统自动调整副本数量,减少人工干预,降低管理负担。在扩散副本时,选择高性能的节点存放副本,以提高系统的处理能力,从而降低数据的访问延迟,平衡负载。

2 DynRM的基本原理

2.1 P2P查询的分布

Zipf分布最初由G. K. Zipf发现于语言学中,表现为如果将语言中的词汇按照出现的频率由高到低排序,那么某个词汇出现的频率与该词汇所对应的序号之间有着如下关系:

$$P(r)=\frac{1}{r^{\alpha}}$$

其中,$P(r)$表示某个词汇出现的频率,r表示该词汇所对应的序号。在某种具体的分布中,α为一个正的常数,称为Zipf指数。文[7]的研究表明,Gnutella的查询分布符合Zipf分布,且zipf指数的取值在[0.63,1.24]之间。

在P2P存储系统中,对某个文件的查询越多,表明该文件越受欢迎。如果对受欢迎程度高的文件建立更多的副本,一方面可以减轻被访问节点的负载,防止系统热点的产生,另一方面可以使节点就近访问,大大减少由于受欢迎文件的访问而产生的网络流量,减轻网络负担,提高系统的响应时间。

2.2 超节点的选取

在P2P网络中,节点在带宽、处理能力、存储能力、在线时间等方面呈现多样性,即节点是异构的。对于那些高带宽、

*)电子信息产业发展基金资助项目,编号:[2004]217。**侯孟书** 博士,讲师,主要研究方向:分布式存储,P2P计算。**王晓斌** 副教授,主要研究方向:计算机网络,数据控掘。**卢显良** 教授,博士生导师,主要研究方向:计算机网络,操作系统,信息安全;**任立勇** 博士。

强处理能力、大存储空间以及不频繁加入和离开系统的节点，我们称为超节点(supernode)。

如果将副本放置在超节点上，不仅能够做到物尽所用，更重要的是缩短了系统的响应时间，提高了副本的可用性。对于超节点的选取，参照文[8]的标准，简要叙述如下：

带宽：节点的访问带宽对系统的响应时间有着重要的影响。文[7]的研究表明在非结构化的 P2P 系统中，有 20～30%的系统节点的访问带宽大于 3Mbps。

计算能力：超节点在处理正常工作之外，还要处理大量的文件访问请求，因此超节点要有较强的计算能力。

存储能力：超节点要有足够大的共享磁盘空间用于存放副本。

在线时间：超节点不能频繁加入和离开系统，否则即使该节点放置了副本，也会由于节点的频繁离线导致副本不可用。根据文[3]对 Gnutella 的研究表明，50%的发起节点能在线 5 小时以上，接近 30%的发起节点能在线 24 小时以上。

2.3 副本选取和放置

本文的副本管理机制 DynRM 就是建立在自主开发的 PNS-PGrid[9]之上，增加受欢迎程度高的文件的副本，并且将副本放置在超节点上，从而达到缩短系统响应时间、平衡节点负载的目的。DynRM 的具体原理如下：

节点为文件访问次数设定一个计数器，计数器随文件被访问的次数而增加，当计数器的值超过预先设定的阈值 r 时，系统就认为该文件为受欢迎文件，启动副本策略。如图 1a)所示，假设文件 F 的索引值为二进制串 000，且存放在节点 1 上，如果对文件 F 的访问次数已经超过了阈值 r，则认为文件 F 为受欢迎文件。接下来，将文件 F 的索引值 000 从左边去掉一位得到二进制串 00，然后再对二进制串 00 取反，得到二进制串 11，那么将文件 F 的副本放在以二进制串 00 和 11 为路径的所有超节点上。如图 1b)所示，则路径 00 对应的节点是节点 1 和节点 2，路径 11 对应的节点是节点 7 和节点 8，由于节点 1 拥有原来的副本，而节点 2 和节点 7 不是超节点，故将文件 F 放置在节点 8 上。从图 1 中可以看出，受欢迎的文件 F 实际上在虚拟二叉树中上升了一层，同理，对文件 F 及其副本重新计数，如果文件 F 及其副本的访问次数再一次超过了阈值 r，则系统将使文件 F 再升一层，即对文件 F 的索引值 000 从左边去掉两位，得到二进制串 0，然后对 0 取反得到二进制串 1，那么将文件 F 的副本放置在以 0 和 1 为路径的所有超节点上。如图 1c)所示，路径 0 对应的节点有 1,2,3 和 4，路径 1 对应的节点有 5,6,7 和 8，由于节点 1 和 8 拥有文件 F 的副本，而节点 2,3 和 7 不是超节点，故将文件 F 的副本放置在超节点 4,5 和 6 上。由图 1 可知经过两次副本放置，节点 1,4,5,6 和 8 拥有了文件 F 及其副本。

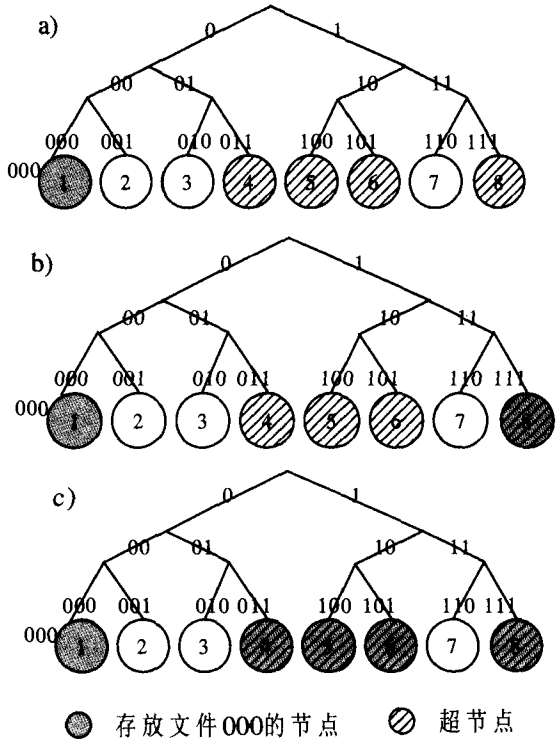

图 1 副本放置

3 DynRM 模型表示

为了简化模型表示，引入以下假设和符号：

假设 P2P 存储系统中有 N 个节点，M 个文件，其中 M 个文件分别为 $f_i(i=1,\cdots,M)$，假设对文件 f_i 的查询次数符合 Zipf 分布，设对文件 f_i 的查询次数为 $q_i=i$，设系统中文件 f_i 的受欢迎程度为 $\rho_i=\frac{q_i}{\mu}$，其中 μ 为节点单位时间内能服务的查询数，ρ_i 越大，说明文件 f_i 越受欢迎，也表示存放文件 f_i 的节点越繁忙。假设用 r 表示阈值，且 $0\leqslant r\leqslant 1$。当 $\rho_i>=r$ 时，则将文件 f_i 上升到 p_i 层，并将其副本放在相应层所对应的超节点上，其中：

$$p_i=\left\lfloor \log_2\left(\frac{qi}{\mu * \gamma}\right)\right\rfloor+1 \tag{1}$$

由虚拟二叉树的性质可知，$1\leqslant p_i\leqslant \lfloor \log_2 n \rfloor$，其中 $i=1,\cdots,M$。

假设超节点占整个系统节点的比例为 α，则处于 p_i 层的文件 f_i 的副本数为 $\alpha 2^{p_i}$。由文件 f_i 的查询次数为 q_i 可知，每个存放文件 f_i 的节点接受到的平均查询次数为 $\overline{q_i}=\frac{q_i}{\alpha 2^{p_i}}$。在以下叙述中，称文件 f_i 的原始文件为原始文件 f_i，称文件 f_i 的副本文件为副本 f_i，而文件 f_i 则是指原始文件 f_i 和副本 f_i。由路由算法 PNS-PGrid 可知，系统按照前缀匹配的原则查找文件，大量的查询仍然会投递到存放原始文件的节点，只有在查找过程中可能会遇到副本，这就导致了存放原始文件的节点接受的查询次数和存放副本的节点接受的查询次数有差异。

设系统中每个文件的原始文件数为 ε，则对于一个存放原始文件 f_i 的节点 W 而言，其存放其它文件 $f_j(f_j\neq f_i)$ 的概率为 $\frac{\alpha 2^{p_j}-\varepsilon}{N-\varepsilon}$，故节点 W 收到的查询总数为：

$$q_{\alpha}=\overline{q_i}+\sum_{j=1,j\neq 1}^{M}\frac{\alpha 2^{p_j}-\varepsilon}{N-\varepsilon} * \overline{q_j} \tag{2}$$

对于存放副本 f_i 的节点 X，设该节点还存放了原始文件 $f_j(f_j\neq f_i)$ 和副本 $f_k(f_k\neq f_i\neq f_j)$，则节点 X 收到的查询总数为：

$$q_{mj}=\overline{q_i}+\overline{q_j}+\sum_{k=1,k\neq i,k\neq j}^{M}\frac{\alpha 2^{p_k}-\varepsilon}{N-\varepsilon} * \overline{q_k} \tag{3}$$

对于存放副本 f_i 的所有节点而言，上述等式成立，则存放副本 f_i 的节点平均收到的查询总数为：

$$q_n=\frac{1}{N-\varepsilon}\sum_{j=1,j\neq 1}^{M}\left(\overline{q_i}+\overline{q_j}+\sum_{k=1,k\neq i,k\neq j}^{M}\frac{\alpha 2^{p_k}-\varepsilon}{N-\varepsilon} * \overline{q_k}\right) \tag{4}$$

(下转第 114 页)

18 Nisan N, London S, et al. Globally distributed computation over the internet: The POPCORN project. Int Conf Distributed Computing Systems, Netherlands, May 1998
19 Waldspurger C, Hogg T, et al. Spawn: A distributed computational economy. IEEE Trans Softw Eng, 1992, 18(2)
20 Amir Y, Awerbuch B, Borgstrom R S. A cost-benefit framework for online management of a meta computing system. In: 1st Int Conf. Information and Computational Economy. Charleston, SC, Oct. 1998
21 Lalis S, Karipidis A. An open marketbased framework for distributed computing over the internet. In: 1st IEEE/ACM Int Workshop Grid Computing, Bangalore, India, Dec. 2000
22 Global Grid Forum(GGF). http://www.ggf.org
23 Buyya R, Abramson D, Venugopal S. The Grid Economy, Special Issue on Grid Computing, 2005, 93(3)698~714
24 Hausheer D, Stiller B. Decentralized Auction-based Pricing with PeerMart. In: The 9th IFIP/IEEE International Symposium on Integrated Network Management (IM 2005), Nice, France, May 2005
25 Golle P, Leyton-Brown K, Mironov I, et al. Incentives for Sharing in Peer-to-Peer Networks. In: Proc. of the 2001 ACM Conference on Electronic Commerce, 2001
26 Tamai M, Shibata N, Yasumoto K, et al. Distributed Market Broker Architecture for Resource Aggregation in Grid Computing Environments. In: Proc. of Cluster Computing and Grid 2005 (CCGrid2005)

(上接第 51 页)

设系统中文件 f_i 的原始文件和副本总和为 T_i，则 $\alpha 2^{p_i} \leqslant T_i \leqslant \alpha 2^{p_i} + \varepsilon$，故对于拥有文件 f_i（可能是 f_i 的原始文件，也可能是 f_i 的副本）的节点而言，其平均收到的查询数为：

$$\overline{q_i} = \frac{\varepsilon}{T_i} * q_{\alpha} + (1 - \frac{\varepsilon}{T_i}) * q_n \tag{5}$$

在系统中，假设查询流的到来符合泊松分布，则由 M/D/1/∞排队系统[10]可知，系统中节点对文件 f_i 的查询的平均响应时间为：

$$\overline{W_i} = (2 - \frac{\overline{q_i}}{\mu}) / 2\mu(1 - \frac{\overline{q_i}}{\mu}) \tag{6}$$

对所有文件查询的平均响应时间为：

$$\overline{W} = (\sum_{i=1}^{M} i * \overline{W_i}) / \sum_{i=1}^{M} i \tag{7}$$

文[6]的模型可以看成上述模型的特例，令 $N=M, \alpha = 1, \varepsilon = 1$ 就可以得到文[6]的模型。如果将 α 看成节点的在线率，则本文提出的模型更具有一般性。

4 DynRM 模型分析

首先将文[6]中的模型(NO-DynRM)和本文的模型(DynRM)进行了对比，如图 2 所示。从图中可以看出，随着阈值 r 的增加，两个模型的平均响应时间都随之而增加。当阈值 $r=0$ 时，NO-DynRM 模型的平均响应时间为 1.517641，而 DynRM 的响应时间为 1.170684，这个时候，系统中的副本数量达到最大数，故此时系统响应时间较快。由于 DynRM 系统中对受欢迎的副本索引缩短位数，然后取反，使得受欢迎的副本不但在本区域扩散，而且扩散到地域较远的其它区域，故 DynRM 模型的系统响应时间较 NO-DynRM 模型的系统响应时间快。当阈值 r 增加到一定程度时，即 r=0.5 时，两个模型的平均响应时间接近 5，为 4.975593。当阈值 r 大于 0.5 以后两个模型的系统平均响应时间处于稳定状态，实际上此时两个模型中的文件都没有增加副本，故响应时间没有变化。

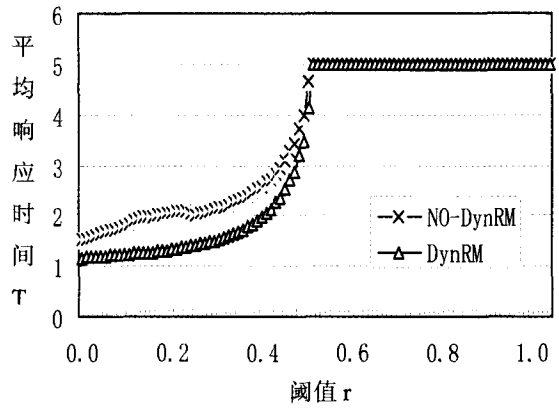

图 2 DynRM 和 NO-DynRM 对比图

结束语 在 P2P 分布式存储系统中，副本能有效提高文件的可靠性，降低访问延迟，避免 Hot Spots 的产生和平衡负载。本文在分析了现有的分布式存储系统动态副本管理机制的基础上，提出了一种动态副本管理机制——DynRM。在 DynRM 中，当文件的受欢迎程度达到一定的阈值时，系统就增加文件的副本，在放置增加的副本时，选择高性能的节点放置副本，以提供更好的系统性能。通过增加文件副本，一方面提高了文件的可靠性，降低了访问延迟；另一方面避免了系统热点的产生，到达了负载平衡的目的。

参 考 文 献

1 Ranganathan K, Iamnitchi A, Foster I. Improving Data Availability through Dynamic Model-Driven Replication in Large Peer-to-Peer Communities. In: The 2nd IEEE/ACM International Symposium on Cluster Computing and the Grid (CCGRID'02) Berlin, Germany, May 2002
2 Cuenca-Acuna F M, Martin R P, Nguyen T D. Autonomous Replication for High Availability in Unstructured P2P Systems[C]. In: the 22nd IEEE Inte Symposium on Reliable Distributed Systems, 2003
3 Gnutella: To the Bandwidth Barrier and Beyond. http://lambda.cs.yale.edu/cs425/doc/gnutella.html, 2001
4 Backx P, Wauters T, Dhoedt B, et al. A comparison of peer-to-peer architectures. In Germany. http://allserv.rug.ac.be/~pbackx/architectures.pdf. 2002
5 Clarke I, et al. Freenet: A Distributed Anonymous Information Storage and Retrieval System. In: ICSI Workshop on Design Issues in Anonymity and Unobservability. July 2000
6 Cudre-Mauroux P, Aberer K. A Decentralized Architecture for Adaptive Media Dissemination
7 Saroiu S, Gummadi P N, Gribble S D. A measurement study of peer-to-peer file sharing systems [C]. In: Proc. of Multimedia Computing and Networking (MMCN), Jan. 2002
8 Xu Z, Hu Y. SBARC: A Supernode Based Peerto-Peer File Sharing System. in(ISCC), Kemer-Antalya, Turkey, June 2003
9 侯孟书. 基于 P2P 的分布式存储研究:[博士论文]. 电子科技大学, 2005
10 唐应辉, 唐小我著. 排队论—基础与应用. 电子科技大学出版社, 2000