计算机科学2004Vol. 31№. 12

# AdpReplica：自适应副本管理机制*)

魏青松　卢显良　侯孟书

（电子科技大学计算机学院　成都610054）（同济大学软件学院　上海200331）

**摘　要**　本文提出一种自适应副本管理机制—AdpReplica。AdpReplica 将期望可用性和一致性维护开销看成副本数的函数，并建立自适应模型。该模型动态管理副本，使文件副本数维持在一个合理的水平，既满足用户的期望可用性，提高访问效率和平衡负载，又减少带宽消耗，保证系统稳定，为用户提供满意的存储 QOS。

**关键词**　副本，自适应，期望可用性，一致性

## AdpReplica：Adaptive Replica Management Mechanism

WEI Qing-Song　LU Xian-Liang　HOU Meng-Shu

(Department of Computer Science of UEST of China, Chengdu 610054)

(School of Software Engineering, Tongji University, Shanghai 200331)

**Abstract**　In this paper, we present an Adaptive Replica Management Mechanism called AdpReplica. The AdpReplica constructs adaptive model regarding user anticipant availability and consistency maintenance overload as a function of a file's replica number. The AdpReplica model manages replica dynamically and maintains a rational replica number, not only satisfying the user anticipant availability, improving access efficiency and balancing overload, but also reducing bandwidth requirement, keeping the system stable and proving satisfied storage QOS.

**Keywords**　Replica, Adaptive, Anticipant availability, Consistency

## 1　引言

基于 P2P 的分布式存储系统通过将地理分布的节点相联，在不可靠和动态的底层网络环境中提供稳定、低代价、高性能的存储服务。为了提高数据可靠性，常用而有效的办法就是在不同的节点中放置副本。多个副本保存在多个独立的节点中将有效提高获得至少一个数据拷贝的机会。副本数越多，数据的可靠性越高。但是，由于文件性质和重要性的不同，用户对不同文件的可靠性要求是不同的，对所有文件应用统一的可靠性保证策略，将浪费大量的存储资源和网络带宽，是不必要和低效的。

文件保存在广泛分布的节点上，数据访问延迟较大，严重影响数据的访问效率。将文件的多个副本合理分布在各节点中，当用户访问某个文件时，底层的路由算法将用户请求传送到距离用户最近的保存该文件副本的节点，从而有效降低访问延迟，大大提高数据访问效率，同时，避免出现负荷失衡。

副本能带来以上好处，但为了保持系统中多个副本的一致性，文件被修改后，必须更新所有的副本。这将带来可观的带宽消耗。随着副本数的增加，维护数据一致性的开销将随之增加[5,6]。

现在的分布式存储系统如 Freenet[1]、Napster[2]等基于文件受欢迎程度(Popularity)复制副本，以提高可靠性和访问效率，并没有考虑副本增加的合理性和代价，没有建立一种评价机制或模型来指导添加副本的必要性。为了维护系统的整体性能，我们必须在可靠性、访问效率和一致维护开销之间找到平衡。

为此，本文提出一种自适应的副本管理机制—AdpReplica(Adaptive Replica Management Mechanism)。AdpReplica 以一致性维护开销为约束条件，考虑数据可靠性和访问效率，在一个合理的范围内动态管理副本数目，既满足数据可靠性需求，提高访问效率和平衡负载，又减少带宽消耗，为用户提供满意的存储 QOS。

本文第2节简介 AdpReplica 的基本策略；第3节是 AdpReplica 的模型分析；具体实现在第4节阐述；第5节进行性能测试；最后是全文的结束语。

## 2　AdpReplica 的基本策略

副本能有效提高文件可靠性和访问效率。但随着副本数的增多，维护各副本数据一致性所消耗的网络带宽就越大，数据不一致的可能性就越大。本文认为：文件访问效率、可靠性和数据一致性是一对矛盾。

### 2.1　可靠性

面向 Internet 的 P2P 存储系统由地理分布的节点构成。系统中的任何单个节点的平均可靠性比较低，节点的退出、网络延迟或节点死机都将导致节点不可到达，从而影响数据可用性(Availability)。

为了提高数据可用性，有效的办法就是在不同的节点中放置多个副本。多个副本保存在多个独立的节点中将有效提高获得至少一个数据拷贝的机会。副本数越多，数据的可靠性越高。

虽然副本能有效提高数据可用性，但是需要付出客观的代价，比如磁盘空间、增加副本过程中的网络传输开销，如果

*)本文受国家95重点攻关项目支持。魏青松　博士研究生，主要研究方向：计算机网络、网络存储、分布式存储等。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。侯孟书　主要研究方向：计算机网络、操作系统、分布式存储等。

文件是非只读文件，文件的一致性维护也需要消耗宝贵的网络带宽。因此并不是副本越多越好，而是必须有个度，而这个度就是满足文件可用性需要的最小代价。因此我们需要一种机制来决定在给定期望可用性的情况下是否需要增加副本。

由于文件性质的不同，用户对文件的可靠性要求是不同的。对所有文件应用统一的可靠性策略，将浪费大量的存储资源和网络带宽，是不必要和低效的。因此，引入区分可靠性机制，为每个文件引入一个文件可靠性参数 $A_{except}$，用户根据文件重要性设置文件的可靠性参数 $A_{except}$，系统根据可靠性计算模型 $A_{except}=Func1(N)$ 计算出维持该可用性 $A_{except}$ 所需要的最小副本数，这样，既维护了可靠性，也节约了资源。

### 2.2 数据一致性

增加文件副本数量是提高文件访问效率的有效方法。但是，单纯地增加文件副本并不一定带来访问效率的提高，相反，副本数达到一定程度后，访问效率会随着副本数的增加而下降。这是因为随着系统中副本数的增多，维护数据一致性的开销急剧增大，影响到整个系统的性能，进而表现访问延迟并不再随副本的增加而减少。也就是说，当副本数在一定范围内增加时，访问效率随之明显增加；当副本数超过一定范围时，访问效率并不随副本数的增加而增加，反而会减少[7]。

系统为维护文件数据一致性而消耗的网络带宽为 $C$，我们以 $C$ 为结合点，实现数据一致性和访问效率的统一协调。文件访问效率越高、数据一致性要求越高，$C$ 就越大。而实际网络能够提供的网络带宽是有限的，$C$ 的增高势必会影响整个网络的性能。如果 $C$ 无限制地增高，就会引起整个网络的拥塞，反过来会降低文件访问效率和数据一致性质量，因此，必须将 $C$ 控制在一定的范围之内。

因此，可以将一致性维护的网络开销作为复制副本的约束条件。当 Hot-Spot 产生时，采用一致性维护开销模型 $C=Func2(N)$ 计算增加一个副本后维护全部副本所需的网络开销，如果满足约束条件，就复制一个副本到距离用户最近的节点上；否则，将访问频率最小的副本向最靠近用户的节点迁移，这样既避免了 Hot-Spot 的进一步扩散，提高了访问效率，负载了平衡，又没有增加一致性维护的网络开销，保证整个系统在资源相对有限的情况下，获得稳定的性能。

综上所述，系统将以 $A_{except}=Func1(N)$ 和 $C=Func2(N)$ 为计算模型，以 $A_{except}$ 和 $C$ 为约束条件，对文件副本数进行约束，在文件可靠性、访问效率和数据一致性之间做折衷，动态管理副本数目，既满足可用性需求 $A_{except}$ 和提高数据访问效率，又使得 $C$ 不超过下层网络能提供的带宽资源，保证系统的整体性能。

AdpReplica 采用如下的基本策略：

1、通过期望可用性 $A_{except}$ 和 $N$ 的函数关系 $A_{except}=Func1(N)$ 找到 $N$ 的最小值 $N_{min}$；

2、通过数据一致性维护开销 $C$ 和 $N$ 的函数关系 $C=Func2(N)$ 找到 $N$ 的最大值 $N_{max}$；

3、如果当前副本数小于 $N_{min}$ 时，添加副本，以满足可用性需求；

4、当某个文件访问过热，根据当前副本数 $N$ 和 $N_{max}$ 来按以下策略决定是添加副本还是迁移副本：

a) $N<N_{max}$ 时，添加副本到距离用户访问最近的节点；

b) $N<N_{max}$ 时，采用数据迁移的策略，将访问频率小的副本向最靠近用户的节点迁移，提高访问效率和负载平衡。

## 3 AdpReplica 模型分析

### 3.1 文件的最小副本数分析

基于 P2P 的分布式存储系统中的任何节点都会因为节点的退出、网络延迟或节点死机等因素导致节点不可到达。保存副本的节点的在线和离线，将直接影响文件的可用性。即便保存副本的节点在线，路由定位服务如果无法准确定位该节点，也会影响该副本的获得。

因此，节点在线的概率和定位服务的准确性是影响文件可用性的关键因素[4]。我们用节点在线的概率和定位服务的准确度为参数来表达文件可用性。为了推导方便，设置如下参数：

$N$：文件 $F$ 的副本总数；

$p$：文件在线的概率；

$R$：副本定位的准确度；

$A_{except}$：文件 $F$ 的期望可用性。

显然，文件 $F$ 的 $N$ 个副本不可用的概率为：

$(1-p)^N$

因此，文件 $F$ 的至少一个副本可用的概率为：

$1-(1-p)^N$

考虑到对象定位服务的准确度 $R$，假设 $R$ 和文件可用性相互独立，文件 $F$ 的一个副本被发现的概率为：

$R\times[1-(1-p)^N]$

因此，对文件 $F$，为了获得期望中的文件可用性，必须满足：

$$R\times[1-(1-p)^N]\geqslant A_{except} \quad (1)$$

通过式(1)，我们可以计算出文件要达到可用性期望值需要维护的最小副本数。例如：对于一个系统，其节点在线几率 $p=40\%$，期望可用性 $A_{except}=80\%$，定位服务的准确度为 85%，则根据式(1)可计算出最小副本数 $N_{min}=6$；如果节点在线几率 $p=10\%$，则最小副本数 $N_{min}=27$。

一旦获得文件的最小副本数 $N_{min}$，就可以根据当前副本数来决定该不该添加副本。如果当前副本数 $N_{exist}<N_{min}$，就必须增加 $N_{min}-N_{exist}$ 个副本，分布到不同节点，保证该文件期望可用性 $A_{except}$ 的获得。

### 3.2 一致性维护的网络开销分析

增加一个副本的过程：向目标节点传输副本和目录更新。因此：

增加副本的开销＝副本复制开销＋目录更新开销

考虑到文件一致性维护的开销，系统总体开销为：

总体开销＝增加副本的开销＋维护一致性的开销

增加一个副本的开销是瞬间的，一致性维护的开销是长期的。这里主要计算一致性维护开销。如果当前副本数为 $N-1$，维护一致性的开销为增加一个副本后维护 $N$ 个副本一致性的开销[3]。为了计算方便，设置如下变量：

$F_u$：文件更新的频率；

$N$：一个文件在系统中的副本总数；

$F_{th}$：文件访问频率阈值；

$B_{req}$：为单位时间内，维护一个文件所有副本数据一致性所耗费的网络带宽。

当一个节点中的副本被更新后，触发数据一致性维护机制，节点将对文件的更改马上通知原文件所在的节点，更新原文件和根节点中的目录项，然后更新分布在各节点的副本及相关的目录 Cache。可以看出，副本数据一致性维护开销分为

三部分：

1. 一个副本文件更新后，更新原文件的开销：

$\alpha \times F_u$　$\alpha$ 为平均更新一个文件的开销

2. 原文件更新后，回写所有节点中的文件副本带来的开销：

$\beta \times (N-1) \times F_u$　$\beta$ 为平均回写一个文件的开销，$\beta \approx \alpha$

3. 目录更新的开销：

$\chi \times N \times F_u$　$\chi$ 为平均更新一条记录的开销

因此，可得单位时间内维护 $N$ 个副本一致性的网络开销：

$$B_{req} = \alpha \times F_u + \beta \times (N-1) \times F_u + \chi \times N \times F_u \quad (2)$$

由于 $\alpha \approx \beta$，将公式(2)化简：

$$B_{req} = (\alpha + \chi) \times N \times F_u \quad (3)$$

以单位时间传输的字节数为网络开销，有：

$\alpha = F_{size}$　$F_{size}$ 为文件的长度

$\chi = L_{msg}$　$L_{msg}$ 为一个消息的平均长度，令其为20

代入式(3)，可得：

$$B_{req} = (F_{size} + 20) \times F_u \times N \quad (4)$$

从式(4)可以发现，对于特定文件，$F_{size}$ 一定，$B_{req}$ 和 $N$ 呈线性关系，随着 $N$ 的增加，维护 $N$ 个副本的网络开销将随之增大。而下层网络能提供的网络带宽是有限的，不能靠无限消耗带宽获得暂时的性能提高，因此必须满足以下条件：

$$B_{req} \leqslant B_{total} \quad (5)$$

将式(4)代入式(5)可得：

$$(F_{size} + 20) \times F_u \times N \leqslant B_{total} \quad (6)$$

综合以上分析，将式(1)和式(6)结合，得到 AdpReplica 模型：

$$\begin{cases} R \times [1-(1-p)^N] \geqslant A_{except} \\ (F_{size} + 20) \times F_u \times N \leqslant B_{total} \end{cases} \quad (7)$$

## 4 具体实现

系统中的各节点运行 AdpReplica 参考模型，各节点通过该模型和文件的属性来维护副本数目，以尽量低的代价维护文件的可用性和系统的高效、稳定运行。为了在系统中实现该模型，为文件引入以下属性：文件期望可用性 $A_{except}$，文件更新频率 $F_u$ 和文件访问频率阈值 $F_a$。

用户在创建文件时，根据文件重要性设置文件属性期望可用性 $A_{except}$，用户在文件创建后，可以修改该属性，以调节期望可用性 $A_{except}$。

为了维护用户期望的可靠性，应该实时检测系统中的副本数目，以决定是否增加副本。一个可选的检测策略是周期性地比较当前副本数和最小需求副本数。根据网络情况和使用模式，检测周期可以动态调整。如果在过去的几次检查中副本数没有变化，就认为当前节点稳定度高，增加检测周期；如果在过去的几次检测中副本持续变化，就认为当前节点极度不稳定，增加检测频率以适应这种变化。周期性探测的方式将周期性发送大量的询问消息和进行计算，将带来很大的网络开销和系统资源。

为了减少系统开销，系统采用事件驱动的检测策略，文件被访问时才进行探测。因为文件的访问频率是不同的，一方面，在一段时间内，有的文件经常被访问，另外的文件访问频率很低；另一方面，同一文件在不同时间段的访问频率是不同的。事件驱动的检测策略不但能实时保证文件的可用性，而且大量减少不必要的检测，从而节约带宽。

只读文件不需要一致性维护，为只读文件进行一致性维护将浪费不必要的网络资源，将只读文件的更新频率设置为0。即便需要更新的文件，其更新的频率也不一样。对不同更新频率的文件应用不同的更新策略不但保证文件的一致性，而且节约大量网络带宽。

由于所有对物理文件的访问请求都是从该物理文件定位信息所在的节点上发送过来的，可以在该物理文件定位信息所在的节点上对该物理文件的使用情况进行监测。AdpReplica 模型对每一个物理文件的使用情况进行监测，监测内容包括访问频率和访问请求来自的区域等信息。对于可更改的文件，保存该文件目录的各节点运行后台进程记录文件在过去单位时间的修改次数，计算历史更新频率，并将历史更新频率记为文件的更新频率 $F_u$。

当某个区域的用户对一个物理文件的访问频率超过某个阈值时，检测当前可用副本数，然后，按以下策略进行：

1. 如果文件只读，说明不需要一致性维护，则添加副本到该区域的某个节点上；

2. 如果文件可更改，根据当前副本数，通过式(7)计算添加一个副本后的一致性维护开销 $B_{req}$，如果 $B_{req} < B_{total}$，将该物理文件添加到该区域的某个节点上；如果 $B_{req} > B_{total}$，则将访问频率最小的节点上的副本迁移到该区域的某个节点上。

## 5 性能测试

### 5.1 容错性测试

以100个 PC 机器作为节点构建一个 P2P 实验网络。测试时采用的方法：在节点在线率一定的情况下，先固定文件副本数，多次测试相应的实际文件可用性，得到一个平均值，然后增加副本数，测试相应的文件可用性，从而得到实际的文件可用性随副本的变化情况。

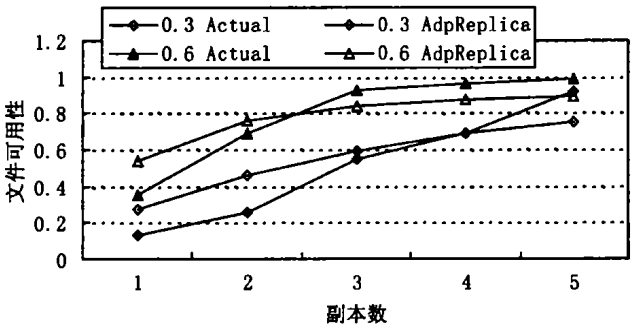


图1　可用性随副本的变化情况

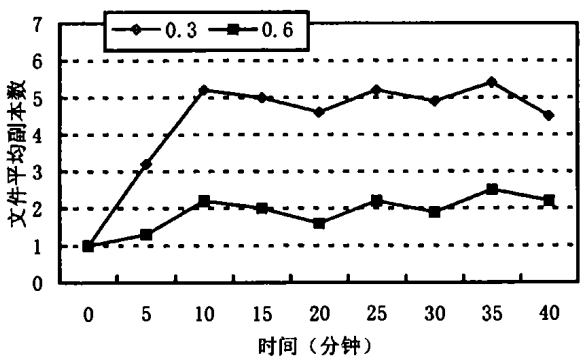


图2　平均副本数随时间的动态变化情况

然后，给定期望文件可用性，根据 AdpReplica 模型计算

（下转第51页）

17 Schnackenberg D, Djahandari K, Strene D. Infrastructure for Intrusion Detection and Response. In: Proc. of DISCEX, Jan. 2000
18 Schnackenberg D, Djahandari K, Strene D, Holiday H, Smith R. Cooperative Intrusion Traceback and Response Architecture (CITRA). In: Proc. of the 2nd DARPA Information Survivability Conf. and Exposition (DISCEXII), June 2001
19 Mahajan R, et al. Controlling High Bandwidth Aggregates in the Network. AT&T Center for Internet Research at ICSI (ACIRI), DRAFT, Feb. 2001
20 Chang H Y, et al. Deciduous: Decentralized Source Identification for Network-based Intrusions. In: 6th IFIP/IEEE International Symposium on Integrated Network Management, IEEE Communications Society Press, May 1999
21 Gil T M, Universiteit V, Poletto M. MULTOPS: A Data-Structure for Bandwidth Attack Detection. In: 10th USENIX Security Symposium, Washington, D.C., USA, Aug. 2001
22 Song D X, Perrig A. Advanced and Authenticated Marking Schemes for IP Traceback. In: Proc. of InfoCom 2001
23 Dean D, Franklin M, Stubblefield A. An algebraic approach to IP traceback. ACM Trans. Information and System Security, May 2002
24 Massey D, Wu C L, Wu S F, Zhang L. On Design and Evaluation of Intention-Driven ICMP Traceback. In: Proc. of IEEE Intl. Conf. on Computer Communications and Networks, 2001
25 Active Network Intrusion Detection and Response project. http://www.pgp.com/research/nailabs/adaptive-network/active-networks.asp, 2001
26 Peng T, Leckie C, Kotagiri R. Adjusted Probabilistic Packet Marking For IP Traceback: [Technical Report of Department of Electrical and Electronic Engineering]. University of Melbourne, Australia, 2001
27 Adler M. Tradeoffs in Probabilistic Packet Marking for IP Traceback. In: Proc. of 34th ACM Symposium on Theory of Computing (STOC) 2002
28 Park K, Lee H. On the effectiveness of probabilistic packet marking for IP traceback under denial of service attack. In: Proc. of IEEE INFOCOM '01, 2001. 338～347
29 Sung M, Xu J. IP Traceback-based Intelligent Packet Filtering: A Novel Technique for Defending Against Internet DDoS Attacks. In: 10th IEEE Intl. Conf. on Network Protocols(ICNP), Paris, France, Nov. 2002
30 Yaar A, Perrig A, Song D. Pi: A Path Identification Mechanism to Defend against DDoS Attacks. In: Proc. of the IEEE Symposium on Security and Privacy 2003
31 Kuznetsov V, Simkin A, Sandström H. An evaluation of different ip traceback approaches. In: Proc. of the 4th Intl. Conf. on Information and Communications Security. Dec. 2002

(上接第36页)

相应的最小副本数，和前面的实际测试结果比较得到图1。从图1可以看出，当节点在线率为0.6时，期望文件可用性为0.76，AdpReplica 模型返回最小副本数为2。当系统保持2个副本时，实际测试的平均文件可用性为0.69，从而验证了模型的正确性。

为了验证 AdpReplica 的动态性，系统初始保持一份副本，为之设定文件期望可用性为0.6，以一定频率访问文件，在不同时刻测试副本数，图2为副本数随时间的变化图。从图2可以看出，系统根据 AdpReplica 模型动态增加副本，当满足用户期望可用性后，将副本维持在一个合理的水平。

### 5.2 性能测试

以100个 PC 机器作为节点构建一个 P2P 实验网络。设定可用带宽 $C=10$MB，系统初始保持一份副本，设定文件访问频率阈值为20，表示当文件的访问频率超过20时触发 AdpReplica 模型，逐渐增加文件的访问频率，测试文件的平均访问延迟和平均副本数，得到图3和4。

图3和图4说明，初始阶段，文件只有一个副本，平均访问延迟较大，随着访问热度的提升，副本随之增加，表现在明显降低访问延迟，但副本数达到一定程度后，副本总数和访问延迟维持在一个稳定的水平，副本没有无谓地增加。

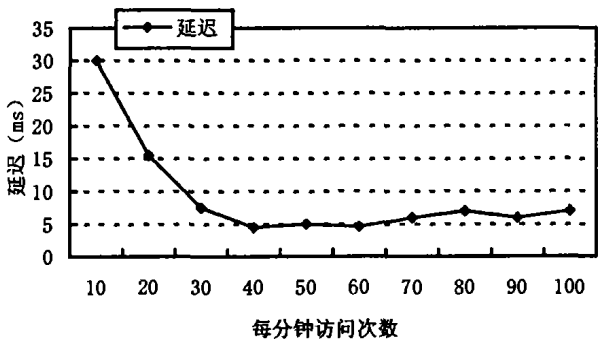


图3 访问延迟随访问频率的变化情况

**结束语** 本文提出一种高效的自适应副本管理机制—AdpReplica。AdpReplica 将期望可用性和一致性维护开销看成副本数目的函数，并建自适应模型，动态管理系统中的副本数，使文件副本数维护在一个合理的水平，既提高了数据可靠性和访问效率、避免 Hot-Spot 的产生、平衡负载，又减少带宽消耗和系统性能的震荡，保证系统的稳定，并且能适应系统的动态扩展，为用户提供满意的存储 QoS。

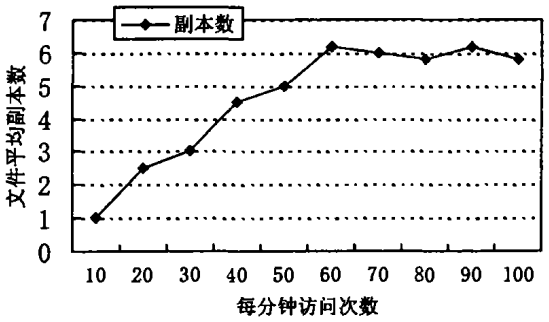


图4 副本数随访问频率的变化情况

## 参 考 文 献

1 FreeNet. http://freenet.sourceforge.net.
2 Napster. http://www.napster.com.
3 Hanhua. Studies on Internet Oriented Distributed Massive File Storage: [PH.D. Dissertation of Peking University]. June 2002
4 Ranganathan K, Iamnitchi A, Foster I. Improving Data Availability through Dynamic Model-Driven Replication in Large Peer-to-Peer Communities. In: Proc. of the Workshop on Global and Peer-to-Peer Computing on Large Scale Distributed Systems, Berlin, May 2002
5 Acharya S, Zdonik S B. An Efficient Scheme for Dynamic Data Replication. Brown University CS-93-43, 1993
6 Wolfson O, Jajodia S, Huang Y. An adaptive data replication algorithm. ACM Transactions on Database Systems, 1997, 22: 255～314
7 Ranganathan K, Foster I. Identifying Dynamic Replication Strategies for a High Performance Data Grid. In: Intl. Workshop on Grid Computing, Denver, CO, 2001