P2P 分布式存储系统中冗余策略研究 *

董 辉 , 雷大军

(湘南学院物理与电子信息工程系,郴州 423000)

摘 要:由于 P2P 系统具有高动态性,为了提高存储的可靠性,必须采用冗余策略,使数据文件以副本的形式分布在系统的多个节点中。阐述 P2P 分布式存储系统中使用的冗余策略,并分析它们对文件可用性的影响以及在真实 P2P 系统中的应用。

关键词:P2P 系统;复制;冗余;纠删码;数据可用性

0 引 言

Peer-to-Peer(P2P)存储系统,也即对等存储系统,是指存储节点以一种功能对等的方式组成的存储网络,这种结构与传统的客户/服务器的集中控制模式相对应。在一个 P2P 覆盖网络中,一般假设系统有海量的节点且每个节点可以自由加入和退出系统,因此,临时和永久的节点失效与传统系统相比将会多很多。在不能确保每个节点可用的情况下,要实现数据文件的高可用性,一个 P2P 存储系统必须要用数据冗余策略屏蔽这些节点失效,保证数据的持久存储。冗余策略的基本思想是在满足一定的性能要求下,维护尽量少的副本数量以减小系统维护的代价,同时要求各个副本保持负载平衡。

1 冗余策略简介

P2P 系统具有高度的动态性和异构性,节点之间是异构的、独立的,每个节点可以提供的资源有限,并且以不同的频率动态加入或离开网络。传统静态复制技术没有考虑 P2P 系统的这些特点,因此需要开发新的解决方案。关于 P2P 存储系统的研究机构很多,例如 Berkeley、MIT、MSR 和 USD 等,主要的系统包括 OceanStore、CFS、OverCite、PAST、FarSit、BiVault 和 Total Recall 等,它们都是 Internet 上基于 P2P 结构的分布存储应用,目的是向用户提供强持久性、高可用性、可扩展性和安全性的服务。其中,这些系统的冗余策略主要有复制(Replication)和纠删码(Erasure Coding)两种方式,而复制策略主要又可分为全文件复制和分块复制两类。

2 数据可用性分析

数据可用性定义为数据在时间 t 能被访问到的概率。在 P2P 系统中,节点暂时离开可能造成数据暂时不能被访问,会降低数据可用性。因此,在 P2P 存储系统中,数据可用性是非常重要的衡量指标。

2.1 可用性概率模型

系统可用性分析需要的变量定义如下:

P:数据文件可用的概率

n:系统中的节点数目

c:复制倍数

b:文件块数

Y:h 个随机选取节点中可用节点数目

Aj, j>=1 and j<=f:文件 j 被需求的可用性

h:系统中任意选取的节点个数

μ:节点可用的概率均值

系统节点总数为 n,从中任意取 h 个节点,其中可用的节点个数为 y 的概率会满足二项式分布概率公式:

$$P(y)=\binom{h}{y}\mu^{y}(1-\mu)^{(h-y)} \quad (1)$$

2.2 全文件复制数据可用性分析

在全文件复制中,假如文件有 c 个副本分布在不同的节点上,c 个副本中至少应该有一个副本有效才能用于恢复数据。因此,全文件复制的可用性定义为:

* 基金项目:湖南省教育厅 2008 年优秀青年基金资助项目(No.08B073)

收稿日期:2009-07-02 修稿日期:2009-09-01

作者简介:董辉(1978-),女,硕士研究生,研究方向为软件容错

$$P(y\geq 1)=1-(1-\mu)^c \quad \text{所需复制倍数 } c=\frac{\log(1-P)}{\log(1-\mu)} \quad (2)$$

根据(2)可以推算出：当节点可用概率均值μ为一固定值时，文件副本数目 c 越大，文件可用性 P 就越高；不同 P2P 系统想达到相同的文件可用性 P，节点可用概率均值μ越低则所需副本数越多。显然，当文件很大，节点可用概率均值μ低会导致需要较大的副本数目才能保证一定的文件可用性，因此，以全文件的形式复制和处理大文件是耗时和麻烦的。

2.3 分块复制数据可用性分析

分块复制将文件分割成小块，文件存储负担可由系统中的节点均匀地承担，而且有利于并行下载和负载平衡。如果将文件划分成 b 块，并对每块生成 c 个副本，文件可用性可定义为：

$$P(y\geq 1)=(1-(1-\mu)^c)^b \quad (3)$$

将(2)与(3)进行比较，可发现，在节点可用概率均值μ相同的情况下，如果要实现相同的文件可用性，分块复制所需的副本数目 c 会大于全文件复制。

2.4 纠删码数据可用性分析

根据纠删码的实现思想，将数据对象分割成 b 片，再编码成 n 片(n>b)。就是说纠删码的冗余因子为 c=n/b。数据可用性可由(1)推出以下计算公式定义：

$$P(y\geq b)=\sum_{j=b}^{cb}\binom{cb}{j}\mu^j(1-\mu)^{(cb-j)} \quad (4)$$

当 cb 的值足够大时，(4)的二项式分布近似为正态分布，利用这个近似，通过代数化简和正态逼近，可以解出 c 的值，如(5)所示，具体推导过程可见文献[5]。

所需复制倍数

$$c=\left(\frac{\sigma\sqrt{\frac{\mu(1-\mu)}{b}}+\sqrt{\frac{\sigma^2\mu(1-\mu)}{b}+4\mu}}{2\mu}\right)^2 \quad (5)$$

其中 σ 为对应数据可用性的正态分布的标准差。表 1 列出了不同的文件可用性对应的正态分布标准差。例如 σ=3.3 对应 3 个 9 的可用性：

表 1 文件可用性对应的正态分布标准差

P	σ
0.800	1.29
0.900	1.65
0.990	2.58
0.995	2.81
0.999	3.30

从(5)看出，所需复制倍数 c 与 n 无关，唯一决定 c 的因素就是 b 值，当 b 增大，复制倍数降低，但会导致系统复杂度增加且下载延迟增大。

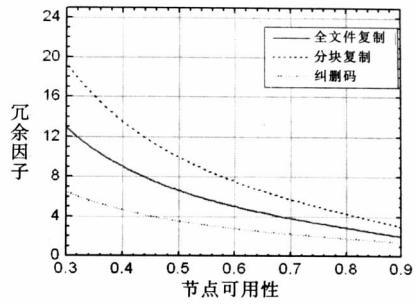

图 1 (b=10)

图 1 展示了全文件复制、分块复制和纠删码冗余策略为达到文件可用性为 0.99 所需冗余因子理论值随节点平均可用性变化的情况，其曲线分别由(2)、(3)和(5)决定。由图可见，当节点可用性一定时，纠删码冗余策略的冗余因子最小，全文件复制次之，分块复制所需的冗余因子最大。当进行写操作时，复制所需网络带宽等于数据大小与冗余因子的乘积，而纠删码所需网络带宽等于编码后的数据大小，显然，在文件可用性和节点可用性相同的情况下，纠删码可以降低数据冗余因子，同时减少了节点对网络带宽的占用。

3 冗余策略在 P2P 存储系统中的应用

Internet 上基于 P2P 结构的分布存储应用有很多，各个系统所设计的目标和工作环境是不同的，根据不同的工作环境，所采取的冗余策略也不尽相同。全文件复制和分块复制的设计和实现较简单，所需冗余因子明显高于纠删码冗余策略，但当节点平均可用性较低，即动态性较高的网络环境中，纠删码会带来设计和实现上的复杂度；在高动态性的网络环境中，纠删码更节省网络带宽；而当节点平均可用性较高，即动态性较低的网络环境中，全文件复制和分块复制比纠删码更节省网络带宽。

Total Recall[6]是圣地亚哥加州州立大学(UCSD)设计的 P2P 存储系统，设计目标是自动配置系统所需要的各种参数，以避免繁琐且困难的人工设置。Total Recall 应用的是一种多变的复制机制，针对不同类型数据分别使用复制或纠删码方式或混合方式。系统会根据可用性的需求，基于文件大小、等级以及文件读写请求的访问模式等来自动选择最有效的复制方案和冗余度。复制方式所需的冗余度采用(2)的计算公式，纠删码需要的冗余度采用(5)所示的计算公式。Total Recall 对小文件采用复制策略和积极修复法，对

大文件采用纠删码和懒惰修复法。通过模拟发现，Total Recall 系统可以获得的数据可用性远远高于设计的目标可用性，并发现使用纠删码能够减小修复带宽。

伯克利大学(Berkeley)开发的 OceanStore 系统[7]使用了复制和纠删码两种复制策略，一方面使用纠删码存储归档的数据以减小空间和带宽消耗，另一方面使用复制来提高数据访问的效率。OceanStore 把冗余的数据碎片存放在网络中无错误相关性的节点集合上，然后将碎片的位置信息保存在文件 ID 对应的根节点处，当有客户请求数据时，先根据数据的 ID 联系到数据的根节点，进而获取每个数据碎片。系统研究者通过建模和分析发现，纠删码能极大地节约系统的存储空间和维护带宽，且有利于提高系统可靠性。

CFS[8]是一种只读存储系统，实现了文件存取的安全、高效、健壮及负载平衡。CFS 系统认为存储空间不是稀缺资源，因此不采用纠删码，而只采用复制方式冗余：将数据实施分块复制，然后用 DHT 直接分发的方式将副本按顺序放置在数据 ID 对应的负责节点及后继节点上，当底层覆盖网络检测到有一个副本丢失时，数据的主节点负责立即再修复出一个副本。

总结上述典型 P2P 存储系统的冗余策略，不难发现，各个系统根据不同的工作环境，所采取的冗余策略也不尽相同，对复制和纠删码的冗余方式各有偏好。

4 结 语

P2P 分布式存储系统能充分利用计算机空闲的存储资源和带宽资源，但由于系统中的节点具有高动态性，极大地影响了系统的可靠性。合理的冗余策略会在多个节点存放数据文件的副本，能有效地提高数据文件可用性。本文系统地分析了各类冗余策略的特点及数据可用性，以及各类冗余策略的适用的环境，并综述了 Internet 上不同的 P2P 分布式存储应用所采取的冗余策略。显然，在 P2P 系统分布式存储应用中，应该根据系统应用要求及网络环境来选择合适的冗余策略，增强系统可靠性。

参考文献

[1]L Rizzo. Effective Erasure Codes for Reliable Computer Communication Protocols[J]. ACM Computer Communication, Review, 1997, 27(2):24~36

[2]M. Mitzenmacher. Digital Fountains: A Survey and Look Forward[J]. In 2004 IEEE Information Theory Workshop, October 2004:271~276

[3]Plank J. A Tutorial on Reed-Solomon Coding for Fault-Tolerance in RAID-Like Systems.[J] Software Practice and Experience, 1997,27(9): 995~1012

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

[5]Bhagwan R, Savage S, Voelker G. Replication Strategies for Highly Available Peer-to-Peer Storage System[J]. Technical Report, CS2002-0726, UCSD, 2002

[6]Bhagwan R., Tati K., Cheng Y., Savage S., and Voelker G.. Total Recall: System Support for Automated Availability Management[J]. In Proceedings of NSDI, San Francisco, USA, 2004

[7]Kubiatowicz J., Bindel D., Chen Y., Czerwinski S., Eaton P., Geels D., Gummadi R., Rhea S., Weatherspoon H., Weimer W., Wells C., and Zhao B.. Oceanstore: an Architecture for Globalscale Persistent Storage[J]. In Proceedings of ASPLOS, Cambridge, MA, USA, 2000

[8]Dabek F., Kaashoek M. F., Karger D., Morris R., and Stoica I.. Wide-area Cooperative Storage with CFS[J]. In Proceedings of ACM SOSP, Banff, Canda, 2001

Research on Redundancy Policies of P2P Distributed Storage System

DONG Hui , LEI Da-jun

(Department of Physics and Electronic Information Engineering, Xiangnan University, Chenzhou 423000)

Abstract: Storage redundancy must been employed by dynamic P2P systems to improve reliability. Distributes the replication of files over many peers in network. Introduces redundancy policies in P2P distributed storage system, analyzes the efficiency of redundancy policies to file availability, and then introduces the application of redundancy policies in real P2P systems.

Keywords: P2P System; Replication; Redundancy; Erasure Coding; Data Availability