基于随机 Petri 网的 P2P 存储系统可靠性模型和分析

刘志明[1,2] 沙基昌[1] 阳小华[2] 万亚平[2]
(国防科技大学信息系统与管理学院 长沙 410073)[1] (南华大学计算机应用研究所 衡阳 421001)[2]

摘 要 可靠性是可信性研究的基本内涵之一,也是 P2P(Peer-to-Peer)存储系统研究的关键。P2P 存储系统结点具有高动态特征,结点频繁地上下线给系统的可靠性分析带来困难。结点的暂时离线会增加副本数量,从而导致系统不必要的损耗。较多的副本数量会增加系统的可靠性,但是同时会导致系统的一致性维护成本增加。而较少的副本数量又会导致系统的可靠性显著降低。针对副本的数量和可靠性之间的关系,采用随机 Petri 建立了 P2P 存储系统的可靠性模型并加以分析,围绕数据可靠性度量要求和数据副本的数量建立关系模型,从理论上指明研究系统可靠性的目的和基本原则,这可以在系统组建之初帮助优化系统的设计。

关键词 随机 Petri 网,P2P 存储系统,可靠性,副本冗余
中图法分类号 TP302.7 **文献标识码** A

Model and Analysis of Peer-to-Peer Storage System Reliability Based on Stochastic Petri Net

LIU Zhi-ming[1,2] SHA Ji-chang[1] YANG Xiao-hua[2] WAN Ya-ping[2]
(College of Information System and Management, National University of Defense Technology, Changsha 410073, China)[1]
(Institute of Computer Applications, University of South China, Hengyang 421001, China)[2]

Abstract Reliability is one of the basic connotations of dependability, and it is also the key of P2P Storage Systems study. P2P storage system nodes are highly dynamic. That nodes are from online to offline frequently will lead difficulty to system reliability analysis. If increasing the number of copies because nodes are temporarily offline, it will cause system to unnecessary outage. The larger number of copies will increase system reliability, but it will also cause the system to increase the cost of consistency maintenance. And the fewer number of copies will result in a significant reduction in reliability of the system. To address the problem of the quantity of copy and system reliability, a reliability model of P2P storage system which revolved around relationships between the requirements of data reliability measurement and the number of copies was built by stochastic Petri nets(SPN) and used to analysis. It identified the purpose and basic principles of studying system reliability from theory and can help to optimize the system design on early stage.

Keywords Stochastic Petri nets(SPN), P2P storage system, Reliability, Copy redundancy

1 引言

P2P 存储系统是一个基于对等网络的数据存储系统,可以提供高效率、鲁棒、容错和负载平衡的文件存取和共享功能。它把分散的结点以一种行为对等的模式组成一个提供统一存储服务的系统,能够为用户提供统一和一致的逻辑数据视图和透明的数据存储服务。结点的组成既可以是一般的用户个人电脑或是服务器,也可以是它们的混杂。P2P 存储系统可以充分利用用户闲散存储资源形成一个互助、自组织的存储网络,解决了服务器的性能瓶颈,使用户对数据的访问速度得到极大提高。

可靠性是系统性能中的重要组成部分。对于存储系统而言,数据服务的可靠性是可信性研究的核心。存储系统可靠性描述了系统有效提供数据服务的能力,即系统能正常提供服务的概率。P2P 存储系统结点的动态性远高于传统存储系统。在 P2P 网络上搭建可靠的存储服务需要更多努力,用户随机行为对系统可靠性造成一定困扰。通过分析 P2P 存储系统的可靠性,可以帮助系统的设计者和用户制定高可靠的数据分发策略以及高效率的数据搜索算法。因此,必须加强对 P2P 存储系统可靠性的研究。正如文献[1]所述,P2P 存储系统可以在动态性、可靠性和用户行为三者之间保持一种平衡的关系,以保证系统在高动态性前提下仍然能够较好地运行。

P2P 存储系统的可靠性研究通常涉及到 3 个问题:数据冗余、故障检测以及修复机制。副本冗余是提高分布式系统可靠性的关键技术,较多的副本还可以使得用户较快获得数据搜索和查询结果。P2P 存储系统一般通过文件或者数据的冗余来提高系统的可靠性。因此,数据的冗余度测量、数据副

到稿日期:2010-05-15 返修日期:2010-09-05 本文受国家自然科学基金项目(70671051),国防基础研究课题(A3720060121)资助。
刘志明(1972—),男,博士生,副教授,主要研究方向为可信计算、知识管理,E-mail:nhdxlzm@foxmail.com;沙基昌(1945—),男,博士,教授,博士生导师,主要研究方向为信息系统管理科学、军事运筹学;阳小华(1963—),男,博士,教授,博士生导师,主要研究方向为智能信息处理、可信计算与核安全;万亚平(1973—),男,博士生,主要研究方向为网络存储技术、海量存储技术。

本的一致性维护，都是问题研究的关键。较大的数据副本数量通常会带来很大的一致性维护成本，因此需要在系统的可靠性和一致性维护成本之间取得一个折衷，以满足各种用户和环境体验的要求。

研究 P2P 存储系统可靠性，必须考虑实际的系统特征。副本冗余作为提高系统可靠性的关键因素，设计之初就必须考虑采用合理的数据副本数量和副本的放置策略。Adamic 等人的研究表明[2]，文件在网络中被访问的频率表现出幂率特征，少数一些文件被访问的频率很高，而大多数的文件被访问的频率很低。因此需要针对网络中文件被访问的不同频率特征设计合理的副本冗余措施。访问频率高的文件在设计的时候考虑使用较多的副本数量，这还可以在一定程度上带来系统性能的提升。而对访问频率不是很高的文件，则使用较少的副本冗余数量，以节省系统的网络带宽和存储资源。

比如文献[3]提出的拥有者复制策略(OR，owner replication)，当对某个数据对象提出查询请求并查询成功以后，只在数据查询请求结点处创建副本，才可以保证系统拥有较少的副本数。Forster 等人[4]为了提高 P2P 系统的可用性提出了一个副本计算模型，规定针对每个对等结点计算当前的可用性，并通过对系统状态的收集决定应该保持的副本数量。对一些类型的数据和服务，这种方法被证明是保证系统高可用性的一个有效途径。

研究副本冗余对 P2P 存储系统可靠性的影响，需要考虑各种实际使用的环境和用户需求。而建立系统可靠性模型，也必须充分考虑不同的冗余方式对可靠性带来的影响，比如从双副本到多副本。如前所述，已有的一些 P2P 系统在设计的时候，针对某些较少访问请求的数据提供较少副本数量。

为了更好地比较副本冗余和可靠性的关系，本文使用随机 Petri 网建立了系统的可靠性模型并加以分析，通过考虑 P2P 系统中不同结点的不同动态特征，使得模型可以较好地反映 P2P 存储系统的可靠性目标。

2 随机 Petri 网

描述可靠性的一个重要模型工具就是随机 Petri 网(Stochastic Petri Net，SPN)，它能对复杂系统做可靠性分析和研究。

一般的 P2P 存储系统是一种非可控系统。其基本特征和思想在于：整个系统处于一种无核心状态，每个结点都能以功能对等的方式提供各种服务，对于存储系统而言提供的是数据存储服务；理论上说，这种系统的结点一般不受其它结点控制，每个结点可以自由地加入或者退出系统。因此，对此种系统的可靠性分析如果仅仅只采用简单的无状态方法，很难描述完备。随机 Petri 网作为一种很好的工具得到了普遍的使用。Petri 网[5]源于西德科学家 C. A. Petri 发表的博士论文，它提供了一种描述并发、竞争和同步系统的建模方法。Petri 网作为一种数学图形工具，其应用非常广泛，不仅能描述系统的静态行为，而且能刻画系统的动态特性，尤其适合描述含有并行成分的复杂系统。它能从组织结构、控制和管理的角度，精确描述系统中事件(变迁)之间的依赖(顺序)和不依赖(并发)关系。

2.1 多副本可用度分析

P2P 存储系统通过连接物理位置分散的结点，在动态不可靠的广域网环境下为用户提供性价比高、可靠的存储服务。这种服务带给我们的是平等和自由，用户可以获得自己需要的存储服务和数据请求。同样，为了提高 P2P 存储系统的可靠性，必然存在多个副本的冗余设计，使得只要系统至少还保存一个副本就可以保证数据不被丢失。对用户而言，在法律许可的范围内，只要能够找到自己所需的数据，不用管其来自哪里，由谁提供，只要能够得到即可。

对同一数据复制多个副本保存，使得当某一个或某一些副本丢失的时候能够恢复原始数据。这些副本保存在不同的结点集合上面，并且要尽量避免这些结点之间错误相关(比如说处于同一个局域网内)。不同的数据根据用户需求设定不同的数据冗余度(即保存的数据副本数)。多个数据构成了 P2P 存储系统的存储服务。

假设系统保存了 s 个数据，其中每个数据 D_i($i=1,2,\cdots,s$)存在 n 个副本，通过 P2P 网络可以访问任意一个副本。该数据的一个或者多个副本失效都不会引起数据 D_i 丢失。只有所有副本都失效时，该数据才彻底丢失。

为分析简单，假设在同一时刻没有两个或两个以上的副本失效，且任一时刻也不会有两个或多个副本被修复，即某一时刻只会有一个副本得到修复。定义 P2P 存储系统的可靠性指标，即系统的可用度。可用度定义为 P2P 存储系统在规定条件下，任意某一时刻系统处于正常可用状态的概率。

设 $A(t)$ 为数据 D_i 的状态空间，且有：

$$A(t)=\begin{cases} i, & (i=0,1,\cdots,n-1)，数据\ D_i\ 没有丢失，但有\ i\ 个副本失效 \\ n, & 数据\ D_i\ 丢失，该数据的所有副本都已经失效 \end{cases}$$

从上述状态空间可以看出，当 D_i 处于状态 n 时，数据不可用，数据丢失。而当处于其它状态时数据是可用的，并且根据副本失效的数量确定相应的修复策略。

图 1 所示为数据 D_i 的随机 Petri 网模型。位置 P_i($i=0,1,2,\cdots,n-1$)表示数据处于正常状态；位置 P_n 表示数据丢失，系统发生故障。初始记号表示起始时刻所有副本处于正常状态，没有副本失效。

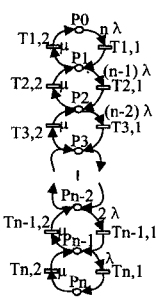

图 1 数据 D_i 的 SPN 模型

其中各个符号含义如下：

位置 P_0、P_1 到 P_n 分别表示数据 D_i 的副本 0 个丢失、1 个丢失，直至数据 D_i 所有副本全部丢失。

$T_{i,1}$($i=1,2,\cdots,n$)表示不同数量副本丢失变迁，$T_{j,2}$($j=1,2,\cdots,n$)表示数据恢复变迁。

λ 表示数据副本发生故障的概率，μ 表示副本发生故障后的修复概率。这里假定不同结点的副本故障概率和修复概率一样。

如同上小节所示，由图 1 中 SPN 模型和变迁实施规则可以得到与之同构的马尔可夫链的状态转移图。根据马氏理论有线性方程组：

$$\begin{cases} P \cdot R = 0 \\ \sum_{i=0}^{n} P_i = 1 \end{cases} \tag{1}$$

式中，P_i 为稳态情况下处于位置 P_i 的概率，马尔可夫链中 n 个标记的稳定概率是一个行向量 $P=(P_1,P_2,\cdots,P_n)$。R 为变迁速度矩阵。

元素 r_{ij} 为从位置 P_i 转移到位置 P_j 的所有转移率之和。如果两位置之间不存在直接联系，则 $r_{ij}=0(i\neq j)$。r_{ii} 由 $\sum_j k_{ij}=0$ 决定，R 的阶数和状态数量一致。求得 R 中各个元素，得矩阵 R：

$$R=\begin{bmatrix} -n\lambda & n\lambda & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\ \mu & -(\mu+(n-1)\lambda) & (n-1)\lambda & 0 & \cdots & 0 & 0 & 0 & 0 \\ 0 & \mu & -(\mu+(n-2)\lambda) & (n-2)\lambda & \cdots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & \mu & -(\mu+2\lambda) & 2\lambda & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 & \mu & -(\lambda+\mu) & \lambda \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 & \mu & -\mu \end{bmatrix}$$

根据式(1)可以得到数据 D_i 在位置 $P_i(i=0,1,2,\cdots,n)$ 的概率为：

$$p_0=\left[\sum_{i=0}^{n}\frac{n!\cdot\left(\frac{\lambda}{\mu}\right)^i}{(n-i)!}\right]^{-1} \tag{2}$$

$$p_i=p_0\left[\frac{n!\cdot\left(\frac{\lambda}{\mu}\right)^i}{(n-i)!}\right],i=1,2,\cdots,n \tag{3}$$

由式(2)和式(3)得到数据 D_i 的可用度为：

$$A=\sum_{i=0}^{n-1}p_i=1-p_0\cdot n!\cdot\left(\frac{\lambda}{\mu}\right)^n \tag{4}$$

2.2 P2P 存储系统可靠性分析

对一个完整的 P2P 存储系统来说，存在多个数据，每个数据保有多个副本。为了提高系统的可靠性，每个副本放置在不同的结点上面。副本数量越多，系统的可靠性就越高。同时，副本放置在不同的结点上面，还可以及时地响应用户数据请求，将网络负载分担在不同的结点集上，减少用户对某个数据集中访问所造成的压力，有效提高系统对数据的访问效率。研究这样的系统可靠性才能反映 P2P 存储系统实际情况。

本节研究由多个数据所构成的完整 P2P 存储系统的可靠性表示方法。

假设 P2P 系统存在多个数据 $D_i(i=1,2,\cdots,n)$，每个数据 D_i 同时复制 r 个副本进行冗余容错设计，则数据 D_i 发生丢失的概率为：

$$\chi_i=\frac{r!\cdot\left(\frac{\lambda}{\mu}\right)^r}{\sum_{k=0}^{r}\frac{r!\cdot\left(\frac{\lambda}{\mu}\right)^k}{(r-k)!}} \tag{5}$$

若所有数据 $D_i(i=1,2,\cdots,s)$ 完全丢失的概率相同，则对于一个有 s 个数据、每个数据有 r 个副本的 P2P 存储系统，其可靠性可以表示为：

$$R(\chi_1,\chi_2,\chi_3,\cdots,\chi_s;t)=e^{\left(-\sum_{i=1}^{s}\chi_i t\right)} \tag{6}$$

3 结果分析

根据式(5)和式(6)，假设一个系统，其数据副本故障发生率为 0.001；副本故障发生后数据修复的概率为 0.002；整个系统共有 100 个数据；由不同的副本数得到单个数据的可靠性和多数据的整个系统可靠性如图 2 所示。

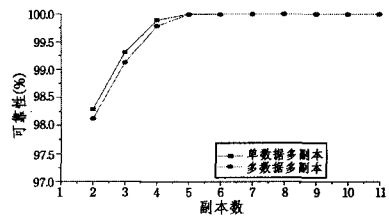

图 2 系统可靠性分析图

图 2 表明，不同的副本数会导致不同的系统可靠性。多副本数据的系统相较于单副本数据的系统，可靠性更接近于 99.999%的高可靠性标准。但是，从图中也可以看出，单纯地增加副本，并不会对可靠性产生多大的影响。当副本数保持在 5 个左右时，可靠性基本达到了系统高可靠性的要求。再多地增加副本数量，并不会对可靠性带来多大的影响，相反会增加副本维护的成本。这可以帮助我们在设计 P2P 存储系统的冗余策略时，根据自己的可靠性要求来设计合理的副本数。

现实世界中用户下层网络的带宽是有限的，如果在数据副本的维护上面消耗太多的带宽，虽然可以提高数据的可靠性，但是过高的数据一致性维护开销将占用系统有限的计算资源并导致系统底层网络的拥塞，进而最终降低系统数据的访问能力。对于各种不同拓扑结构的 P2P 存储系统来说，可以使用本文的方法来对系统可靠性建模，并指导各种复杂的 P2P 网络可靠性设计。

结束语 P2P 存储系统是一种复杂的分布式系统，具有分布式系统的固有特点，如可扩展、透明、异构等；并且作为一种 P2P 网络又必然表现出结点频繁变化的动态性。这些综合因素的影响导致在高动态的网络环境中，存储系统的可靠性问题表现出新的特点和新的问题，从而也构成了 P2P 存储系统独特的研究方向。

研究系统可靠性的一种方法就是随机 Petri 网，它能对复杂系统做可靠性分析和研究。本文对多数据、多副本的复杂 P2P 存储系统建立了相应的随机 Petri 网模型。通过理论分析表明，单纯地增加每个数据的副本数，并不会显著提高系统的可靠性；相反会极大地增加系统的维护成本。为了提高系统的可靠性，需要多副本来屏蔽结点的错误，关键是要考虑如何选择副本的数量。因此，对副本冗余度的设定和对结点状态的估计是 P2P 存储系统设计的难题。

另外还有一个关键问题必须要考虑，这就是为了提高

P2P存储系统的可靠性，通常每个数据需要保留多个副本，导致数据一致性问题也必须得到解决，否则副本保存的冗余数据变成了一堆垃圾，不仅对系统可靠性无益，反而会严重影响存储系统的可用性。数据的一致性问题主要与系统的更新操作有关：更新操作发生时，离线结点因为无法接收到更新操作而导致数据副本间的不一致。P2P网络的结点状态非可控使得一致性协议的开销增大或者导致丢失更新操作，为此需要解决如何用最少的开销保证副本最终的一致状态，以及根据数据的更新频率和故障概率来选择合适的数据修复机制，保证在一致性问题解决的条件下如何高效率地修复数据。

删除冗余数据也会影响一致性并占用存储资源。在对无用数据删除的过程中，由于用户结点的离线会使本应该被删除的无用数据仍然保留在结点上，如果结点此时上线，将发现本地数据和最新数据的版本不一致，使得本地拥有的数据变成了垃圾数据。因此必须解决多个数据副本的一致性问题，保证必须删除的数据都被完全删除。一个可行的解决办法就是采用定时机制，当某个数据副本在结点上的持有时间超过了系统预先设定的时间阈值以后，由系统自动删除该数据，但这会产生一点维护代价。

参考文献

[1] 代亚非. P2P存储在云计算时代的新的机遇[J]. 中国计算机学会通讯，2009，6(5)：54-56

[2] Adamic L A，Lukose R M，et al. Search in Power-law Networks [J]. Physical Review，2001，E 64

[3] Qin L，Pei C，Edith C，et al. Search and replication in unstructured peer-to-peer networks[C]// Proceedings of the 16th International Conference on Supercomputing. New York：ACM Press，2002：84-95

[4] Ripeanu A，Forster I. A decentralized，adaptive replica location mechanism[C]// Proceedings of 11th IEEE International Symposium on High Performance Distributed Computing. 2002：24-26

[5] Lin Chuang. Stochastic Petri Nets and System Performance Evaluation[M]. Beijing：Tsinghua University Press，2000

（上接第47页）

[23] Wong S，Yang H，Lu S，et al. Robust Rate Adaptation for 802.11 Wireless Networks[C]// Proc. of ACM MOBICOM. 2006：146-157

[24] Chen Xi，Qiao Daji，Yu Jeonggyun，et al. Probabilistic-Based Rate Adaptation for IEEE 802.11 WLANs[C]// Proc. of IEEE GLOBECOM. 2007：26-30

[25] Wu S，Biaz S. ERA：Efficient Rate Adaption Algorithm with Fragmentation [R]. CSSE 07-04. Auburn University，June 2007

[26] Biaz S，Wu Shaoen. OTLR：Opportunistic Transmission with Loss Recovery for WLANs[C]// Proc. of WCNC. 2008：1541-1546

[27] Reis C，Mahajan R，Rodrig M，et al. Measurement-based models of delivery and interference in static wireless networks[C]// Proc. of ACM SIGCOMM. 2006

[28] Acharya P A K，Sharma A，Belding E M，et al. Congestion-aware Rate Adaptation in Wireless Networks：A Measurement-driven Approach[C]// Proc. IEEE SECON. June 2008：1-9

[29] Camp J，Knightly E. Modulation Rate Adaptation in Urban and Vehicular Environments：Cross-layer Implementation and Experimental Evaluation[C]// Proc. of the ACM MobiCom. Sept. 2008：315-326

[30] Heusse M，Rousseu F，Berger-Sabbatel G，et al. Performance Anomaly of 802.11b[C]// Proc. IEEE INFOCOM. Mar. 2003：836-843

[31] Pavon I，Choi S. Link Adaptation Strategy for IEEE 802.11 WLAN via Received Signal Strength Measurement[C]// Proc. IEEE ICC. 2003：1108-1123

[32] Zhang J，Tan K，Zhao J，et al. A Practical SNR-guided Rate Adaptation[C]// Proc. of the IEEE INFOCOM. Apr. 2008：2083-2091

[33] Judd G，Wang X，Steenkiste P. Efficient Channel-aware Rate Adaptation in Dynamic Environments[C]// Proc. of the ACM MobiSys Conf. June 2008：118-131

[34] Rappaport T. Wireless Communications：Principles and Practice [M]. Englewood Cliffs. NJ：Prentice-Hall，2002

[35] Vutukuru M，Balakrishnan H，Jamieson K. Cross-layer wireless bit rate adaptation[C]// Proc. of ACM SIGCOMM. 2009：3-14

[36] Jamieson K，Balakrishnan H. PPR：Partial Packet Recovery for Wireless Networks [C] // Proc. of ACM SIGCOMM. August 2007：409-420

[37] Holland G，Vaidya N，Bahl P. A rate-adaptive MAC protocol for multi-hop wireless networks[C]// Proc. of ACM MOBICOM. 2001

[38] Sadeghi B，Kanodia V，Sabharwal A，et al. Opportunistic Media Access for Multirate Ad Hoc Networks[C]// Proc. of ACM MOBICOM. 2002，24-35

[39] Lin C R，Chang Y H J. AAR：An Adaptive Rate Control Protocol for Mobile Ad Hoc Networks[C]// Proc. of the 11th IEEE International Conference on Networks(ICON). Sept. 2003

[40] Wang J，Zhai H，Yuang F Y. Opportunistic Media Access Control and Rate Adaptation for Wireless Ad hoc Networks[C]// Proc. of IEEE ICC. June 2004

[41] Li Z，Das A，Gupta A K，et al. Full auto rate MAC protocol for wireless ad hoc networks[C]// IEEE Proc. on Communication. 2005(3)：311-319

[42] Qiao Daji，Choi Sunghyun. Goodput Enhancement of IEEE 802.11a Wireless LAN via Link Adaptation [C] // Proc. of IEEE ICC. June 2001：1995-2000

[43] Qiao Daji，Choi Sunghyun，Shin K G. Goodput Analysis and Link Adaptation for IEEE 802.11a Wireless LANs[J]. Proc. of IEEE Trans. on Mobile Computing(TMC)，2002，1(4)：278-292

[44] Choudhury S，Gibson J D. Payload Length and Rate Adaptation for ultimedia Communications in Wireless LANs[J]. IEEE Journal on Selected Areas in Communications，2007，25(4)

[45] Chen C，Luo H，Seo E，et al. Rate-adaptive. Framing for Interfered Wireless Networks[C]// Proc. of IEEE. INFOCOM. 2007

[46] Yang X，Vaidya N. On the Physical Carrier Sensing in Wireless Ad Hoc Networks[C]// Proc. of IEEE INFOCOM. 2005：2525-2535

[47] Zhai H，Fang Y. Physical carrier sensing and spatial reuse in multirate and multihop wireless ad hoc networks[C]// Proc. of IEEE INFOCOM. Apr. 2006：1-12

[48] Yang X，Vaidya N. Spatial Backoff Contention Resolution for Wireless Networks [R]. Urbana-Champaign：Univ. of Illinois，2006

[49] Kim T S，Lim H，Hou J C. Improving Spatial Reuse Through Tuning Transmit Power，Carrier Sense Threshold，and Data Rate in Multihop Wireless Networks[C]// Proc. of ACM MobiCom. 2006