P2P 分布式存储系统*)

田荣华[1] 卢显良[1] 侯孟书[1,2] 王晓斌[1,2]

(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 成都 611731)[2]

摘 要 设计了一个基于 P2P 的分布式存储系统。该存储系统采用高可扩展的 P2P 体系结构,将大量分散的节点组织成一个逻辑网络,充分利用原先被忽视的端系统资源,构建大规模分布式存储系统。该存储系统采用高效的结构化 P2P 路由机制、动态自适应的副本管理、信任机制和激励机制为用户提供高效、可靠的分布式存储服务。

关键词 对等网络,分布式存储系统,覆盖网络

P2P-based Distributed Storage System

TIAN Rong-Hua[1] LU Xian-Liang[1] HOU Meng-Shu[1,2] WANG Xiao-Bin[1,2]

(School of Computer Science and Engineering, UESTC, Chengdu 610054)[1] (Chengdu College, UESTC, Chengdu 611731)[2]

Abstract This paper designs a P2P-based distributed storage system: PeerStore. Based on a highly scalable P2P architecture, PeerStore organizes large numbers of nodes distributed in Internet into a united overlay network and builds P2P-based distributed storage system by utilizing the end-systems resources ignored in the past. Featuring with effective routing mechanism, adaptive replication management, reputation mechanism and incentive mechanism, PeerStore provides users with efficient, reliable, massive storage service.

Keywords Peer-to-peer network, Distributed Storage system, Overlay network

随着 Internet 技术和计算机技术的不断发展,分布式存储技术取得了长足的进步,然而用户和数据的不断增加,系统规模的不断扩大,对分布式存储技术提出了更高的要求,另外,随着网络带宽的大幅增加和计算机能力的迅速增强,在传统的客户机/服务器模式中被忽视的客户机成为一种宝贵的资源。微软公司对 4801 台个人计算机,10,568 个文件系统,共 10.5TB 的数据进行了跟踪试验,分析并总结了这些机器的使用情况[1]。实验表明,个人计算机平均只有 50%的存储资源得到了利用,大量的存储空间处于空闲状态。因此"Harness the edge of Internet"成为当前一个新的研究和应用目标。

P2P 存储系统就是充分利用计算机的空闲计算资源、存储资源和带宽资源,构建高可扩展、高可靠、高性能的分布式存储系统。

1 相关工作

自 1999 年以来,由 Napster 点燃的 P2P 计算模式正在逐渐成为研究和应用的热点。P2P 计算模式的兴起得益于 Internet 的广泛普及、网络带宽的大幅增加以及基于 Internet 端系统计算能力的迅速增强。上述因素促使原先在其它计算模式中被忽视的端系统成为一种宝贵的计算资源。到目前为止,P2P 研究已经涉及非常广泛的方面,主要包括:分布式数据存储、大规模并行计算、即时通讯等。P2P 分布式存储系统的目的就是通过互联网将端系统闲散的网络资源整合起来,实现大规模的文件共享和存储。现有的 P2P 分布式存储系统中比较著名的系统有 Napster[2], Gnutella[3], Kazaa[4], Freenet[5], Chord[6], CAN[7], PAST[8], OceanStore[9]等。

根据节点集中程度和网络拓扑结构这两个特征,大致可以将现有主流的 P2P 存储系统分类如表 1。

表 1 P2P 存储系统分类

	结构化	非结构化
混合非集中式		Napster, BitTorrent
完全非集中式	Chord, CAN, OceanStore, PAST, CFS	Gnutella
部分集中式		Kazaa, Morpheus

本文的目标是设计一个基于 P2P 的分布式存储系统 PeerStore(Peer-to-Peer based Distributed Storage System)。

2 PeerStore 分布式存储系统设计

PeerStore 的基本思想是利用 P2P 技术,通过 P2P 网络将 Internet 范围内零散的计算机连接起来,整合这些计算机上的空闲存储资源,形成一个高可扩展、高可靠、高性能、廉价的分布式存储系统。

PeerStore 由地理分布的多个节点构成,每个节点都是拥有存储空间的独立计算机,节点之间以 P2P 覆盖网络的方式组织,采用结构化的路由算法实现节点定位及就近访问,文件以副本的形式分布在系统的多个节点中,从而提高存储的可靠性,并通过信任机制保证系统节点从高可信的节点取得服务,采用激励机制抑制"搭便车"和"公共悲剧"现象。

从系统功能的角度可以把 PeerStore 分为五层:应用层、

*)电子信息产业发展基金资助项目,编号:[2002]1106。田荣华 博士研究生,主要研究生,主要研究方向:分布式存储,分布式计算;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。侯孟书 博士,主要研究方向:分布式存储系统,P2P 计算;王晓斌 副教授,主要研究方向:程序设计语言与编译。

会话层、数据层、路由层和物理层：

·应用层：系统用户通过用户界面直接与应用层交互。通过应用层提供的文件服务接口，用户看到的将是一个虚拟的存储空间，用户可以上传、下载、共享自己的文件，也可以访问由其它用户共享出来的文件。由于应用层屏蔽了下层路由、复制、传输等技术细节，用户可以像使用本地存储系统一样访问分布式存储空间。在应用层中，可以利用系统下层提供的文件存储共享功能，开发各种应用。

·会话层：该层实现了用户管理和目录管理等功能。用户的登录信息在会话层得到处理，会话层为每个用户提供独立的目录空间。用户登录后，节点负责检查目录中的共享文件所在的节点是否在线，如果节点不在线，则启动搜索机制，在系统中查找可用的共享文件。

·数据层：用户文件以多个副本的形式分散存放在系统的多个节点之上，冗余的副本不仅提供数据容错，而且起到平衡负载、提高访问效率的作用。数据层使用动态副本管理机制，根据文件的受欢迎程度，动态的调整文件副本的数量和存放位置，以降低系统热点，平衡负载，提高文件的可用性。通过信任机制，节点可以选择可信度高的节点提供的服务，而从可信度低的节点获取服务的可能性要小得多。通过激励机制，系统可以有效地抑制“搭便车”和“公共悲剧”的现象，从而确保系统拥有一定数量的可用存储资源。

·路由层：采用一种基于 P-Grid 改进的 P2P 路由算法 PNS-PGrid。通过该路由机制，可以将系统中松散的节点结合到一起，形成一个结构化的分布式 P2P 网络。PNS-PGrid 算法充分考虑节点间在底层的物理拓扑，使 P2P 网络拓扑尽可能反映节点的底层物理拓扑，从而减少搜索延迟，提高搜索效率，并最终提高建立在 PNS-PGrid 算法之上的存储系统的性能。

·物理层：物理层由地理分布的具有存储空间和计算能力的计算机即系统节点以及连接它们之间的底层网络部件构成。各节点贡献自己的存储空间和计算资源，是构成 P2P 存储系统的基本元素，是文件存储的实体，是路由转发的中间节点。物理层是整个系统的物理基础。

PeerStore 具有以下特点：

·充分利用 Internet 上的多个节点的闲散资源，构建一个持久的、高可用的分布式存储系统。节点可能属于不同的地域、不同的部门、不同的组织等，这就避免了存储介质的运输，节省了共享数据的费用，而且可以充分利用各个节点贡献的存储和带宽，使得 PeerStore 提供的存储资源超过任何单个节点所能提供的存储资源。

·有效的路由机制 PNS-PGrid 确保用户的服务请求，路由到在底层物理网络上距离自己较近的节点，以获得较低的文件访问延迟，提高构建在其上的应用的系统性能。

·PeerStore 提供的持久存储服务在文件的访问语义上，有别于传统的文件系统。PeerStore 对于每个文件指定一个全局唯一的文件标识符 fileID，对于共享模式的文件，一旦插入 PeerStore 系统，其内容和 fileID 将不再改变，这就避免了文件一致性协议，简化了系统设计的复杂性。

·PeerStore 提供了信任机制。由于 P2P 存储系统的开放、匿名等特征，节点不为自身的行为负责，导致恶意节点滥用 P2P 存储资源，甚至传播病毒等文件。PeerStore 根据确定性理论的不确定推理建立了信任模型，以确保 P2P 存储系统的良性发展。

·PeerStore 提供了激励机制。由于节点的自兴趣性和理性，导致了“搭便车”和“公共悲剧”等问题，PeerStore 将节点的贡献和其占用的公共资源挂钩，鼓励节点贡献其存储资源。

PeerStore 的体系结构如图 1 所示。

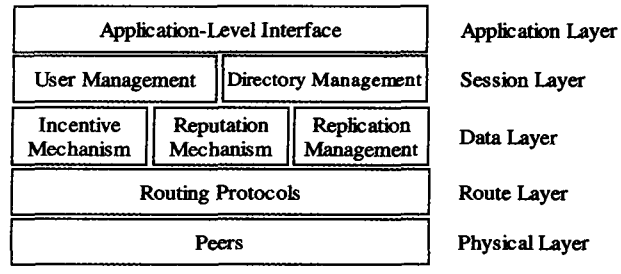

图 1　PeerStore 体系结构

小结　PeerStore 由地理分布的多个节点构成，每个节点都是拥有存储空间的独立计算机，节点之间以 P2P 网络的方式组织，采用结构化的路由算法实现节点定位及就近访问，文件以副本的形式分布在系统的多个节点中，从而提高了存储的可靠性，并通过信任机制保证系统节点从高可信的节点获取服务，采用激励机制鼓励节点贡献其存储资源。

参 考 文 献

1　Douceur J R, Bolosky W J. A large-Scale Study of file-system Contents. SIGMETRICS'99, 1999, 27(1): 59～70

2　Napster. http://www.napster.com. 1999

3　Gnutella: To the Bandwidth Barrier and Beyond. http://lambda.cs.yale.edu/cs425/doc/ gnutella.html. 2001

4　Kazaa home page. http://www.Kazaa.com. 2000

5　Clarke I, et al. Freenet: A Distributed Anonymous Information Storage and Retrieval System. ICSIWorkshop on Design Issues in Anonymity and Unobservability, July 2000

6　Stoica, Morris R, Karger D, et al. Chord: A scalable peer-to-peer lookup service for internet applications. In: Proceedings of the ACM SIGCOMM '01 Conference, San Diego, California, August 2001

7　Ratnasamy S, Francis P, Handley M, Karp R. A scalable content-addressable network. In: Proceedings of SIGCOMM 2001, August 2001

8　Druschel P, Rowstron A. PAST: A large-scale, persistent peer-to-peer storage utility. HotOS VIII, Schoss Elmau, Germany, May 2001. 75～80

9　Kubiatowicz, Bindel D, Chen Y, et al. OceanStore: An Architecture for Global-Scale Persistent Storage. In: Proceedings of the Ninth International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS 2000), November 2000