

P2P 环境下存储系统的构架与可靠性分析

夏磊^{1,2,3}, 刘鹏², 袁致晓³, 赵梦欣³

(1. 解放军理工大学 气象学院, 江苏 南京 211101;

2. 解放军理工大学 网络技术研究中心, 江苏 南京 210007;

3. 94600 部队气象台, 河南 郑州 450046)

摘 要:随着气象业务数据量的日益膨胀,海量数据存储已成为了各气象台站迫切需要解决的问题,同时存储设备出现物理故障后数据的丢失,也常常给气象保障业务带来巨大的损失。为此,文中设计了基于 P2P 的存储系统的整体构架,通过对完全副本和纠删码两种冗余方式的比较,确定了采用基于 Reed-Solomon 算法的数据冗余方案,并提出了基于节点可靠性的分层数据存储策略,将一部分数据编码存储在可靠度相对较高的节点上,保证了数据的高可靠性。由于采用 P2P 技术,系统具有很好的扩展性,相同构架和方法可以复制到其他台站。

关键词:P2P;可靠性;冗余;完全副本;纠删码

中图分类号:TP302

文献标识码:A

文章编号:1673-629X(2009)09-0079-04

Architecture and Reliability Analysis of Storage System in P2P Environment

XIA Lei^{1,2,3}, LIU Peng², YUAN Zhi-xiao³, ZHAO Meng-xin³

(1. Institute of Meteorology, PLA Univ. of Sci. & Tech., Nanjing 211101, China;

2. Research Center for Grid Technology, PLA Univ. of Sci. & Tech., Nanjing 210007, China;

3. Meteorological Observatory of 94600 Troops, Zhengzhou 450046, China)

Abstract: With the inflation business data in meteorological observatory day by day, data storage become the problem needed solving urgently in meteorological observatory, meanwhile, the data lose after store equipment break down, often bring enormous loss to meteorological guarantee business too. For this reason, designed the whole framework of storage system based on P2P, through the comparison of two kinds of redundant ways that the complete copy and erasure code, confirmed the scheme of data storage based on Reed-Solomon algorithm, and proposed that the tactics of data storage based on node dependability, store some data coding on the relatively higher reliability node, have guaranteed the high dependability of the data. The system has a good scalability because of the use P2P technology, the same structure and practices can be copied to other meteorological observatory.

Key words: P2P; reliability; redundant; complete copy; erasure code

0 引言

随着地面观测,卫星云图,VSAT 等技术的大量应用和改进,各种气象业务数据呈指数地快速增长,海量数据存储已成为了气象业务发展迫切需要解决的问题;同时数据的可靠性也成为关注的热点问题。传统上,数据存储采用“信息存储中心+镜像备份中心”的集中式信息存储模式,在这种模式下,数据镜像备份虽然可以在一定程度上保证信息存储的可靠性,但通常

需要进行 1:1 或 1:n 的备份,存在大量存储空间浪费;更为严重的是,如果数据中心和备份数据同时出现故障,数据将无法恢复,重要数据的丢失必然对气象业务的正常开展产生极大的影响。当采用传统的信息存储模式建立存储系统后,随着系统规模的增加,需要不断增加备份数据才能保证系统的可靠性。

为解决这些问题,需要利用分布式存储技术,通过完善保证数据存储可靠性的算法,建设一种可以提供海量存储能力、具有高安全性和可靠性的信息存储平台,给气象业务的健康运转提供强有力的保障。

1 相关研究

P2P 是一种分布式网络,可以定义为:网络的参与

收稿日期:2008-12-24;修回日期:2009-03-10

基金项目:国家自然科学基金资助项目(60403043)

作者简介:夏磊(1981-),男,湖南郴州人,硕士研究生,研究方向为存储网络,P2P 存储;刘鹏,教授,博士,研究方向为网络计算。

者共享他们所拥有的一部分硬件资源(处理能力、存储能力、网络连接能力、打印机等),这些共享资源通过网络提供服务,能被其它对等节点(Peer)直接访问而无需经过中间实体。在此网络中的参与者既是资源(服务和内容)提供者(Server),又是资源获取者(Client)^[1]。P2P 打破了传统的 C/S 结构模式,网络中每个节点的地位都是同等的,每个节点既可以作服务器,为其它节点提供服务,同时也可以作客户机,享用其它节点提供的服务^[2]。

基于 P2P 的存储系统,也即一种基于对等模式的存储系统,其存储节点以一种功能对等的方式组成。与传统的存储系统相比,存储系统有如下优点:

(1)无中心服务器,系统具有高扩展性,可利用廉价机群搭建大规模高性能存储服务,不存在单点性能瓶颈问题。

(2)各个节点功能对等,整个系统在缺失任意节点后仍能正常工作,有高容错性。

(3)系统性能的起伏随用户规模的变化不大,相比 C/S 结构,具有极小的系统总开销。

(4)系统具有海量数据存储能力。

(5)每个节点将可以利用带宽资源存储数据,极大地提高了传输速度。

目前 P2P 存储技术已经引起了人们的广泛关注和深入研究,知名的存储系统有: Berkeley 的 OceanStore, MIT 的 CFS 和清华大学的 Granary。

OceanStore^[3]系统构建在较为稳定的由服务商提供的节点集合上,节点间通过协议保证互相提供连续的服务;系统假设每个节点都可能不可信,但其系统整体上又是可信的;系统中的数据是不断演化的,因此系统能够自调整;系统中的数据是可以共享和全局可访问的,系统既保证数据私密性又保证其完整性;系统提供一定的数据一致性保证^[4]。OceanStore 底层采用 Berkeley 提出的 Tapestry 路由算法,冗余策略上, OceanStore 为减小空间和带宽消耗使用纠删码存储归档的数据,同时使用完整副本来提高数据访问的效率。

CFS(Cooperative File System)^[5]是一种只读存储系统,系统主要分为三个层次: Chord 做底层覆盖网络; DHash 为客户从服务器上获取分块,将分块分散到各个服务器上去,并维护缓存和副本;文件系统层(FS)负责提供文件系统接口与数据块之间的转换。CFS 用复制及缓存的方法实现了负载均衡: 对大文件实施分块复制,对小文件进行缓存。CFS 认为存储空间不是稀缺资源,因此不采用纠删码技术。

Granary^[6]系统的设计目标是能够自适应地支持高动态系统和稳定系统,它以对象为存储单位。系统

底层采用 Tourist 路由算法,冗余策略上采用完全副本方式提高可靠性和访问效率,冗余后的数据为防止相关性错误,被均匀的分发到网络中的节点上,并通过概率复制算法保证复本之间的一致性。

上述存储系统,在冗余方案上采取副本,数据分块或副本加编码的方式,可见在冗余方法的意见还不一致。因此,文中将通过比较副本和纠删码两种冗余方案,选择合适的冗余方式和数据存储策略组织机房中较为稳定的内部存储节点,总的目标为:用廉价节点构建设计简单,高可靠性的存储系统。

2 可扩展的系统构架

2.1 体系结构

存储系统从整体上可分为四个层次:资源层、通信网络层、P2P 组件层和综合服务层,如图 1 所示。

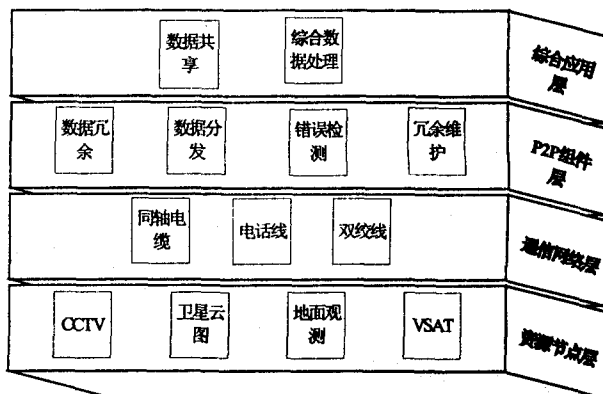


图 1 体系结构图

资源节点层包括各种软硬件资源和信息接收设备,主要有 CCTV 报文接收机、WT-8 卫星云图接收机、地面气象实时观测系统以及各种气象信息系统及相关的数据库。这些资源是形成存储系统的基础设施。通信网络层主要完成底层各资源节点的物理连接,主要包括同轴电缆、电话线以及双绞线等。P2P 组件层的主要作用是对数据资源进行必要编码运算和冗余备份,实现数据资源的分布式存储,为上层的应用服务提供高可靠性的数据。P2P 组件层主要提供数据冗余、分发和错误检测、冗余数据维护等功能集。综合服务层向各种应用提供统一的数据服务。

2.2 基于 Reed-Solomon 算法的数据冗余方案

P2P 是一个动态的环境,对等点可以随时加入 P2P 网络,也可以随时退出^[7]。为保证有节点失效的情况下系统的可靠性,必须对要存储的数据做一定的冗余。如果没有冗余的数据,在某一节点失效后,数据将必然无法恢复。系统中主要使用两种冗余方法:完全副本冗余和纠删码冗余。

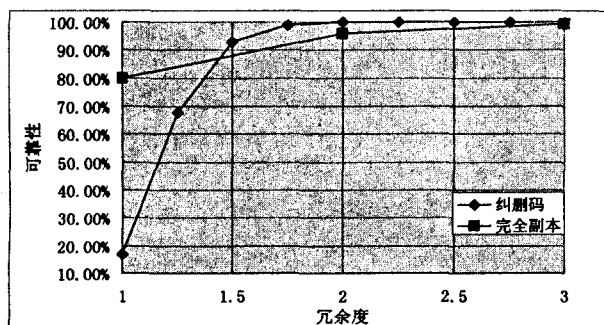
完全副本冗余,即保存多个要存储的数据的完整副本。设系统可靠性为 R , 节点可靠度为 p , 数据冗余倍数为 r , m 为编码碎片数。对于完全副本模式的冗余存储,在不同的节点上保存数据的多个完整副本,系统中任何节点有效即可满足系统有效,满足系统结构的可靠性关系:

$$R = 1 - (1 - p)^r$$

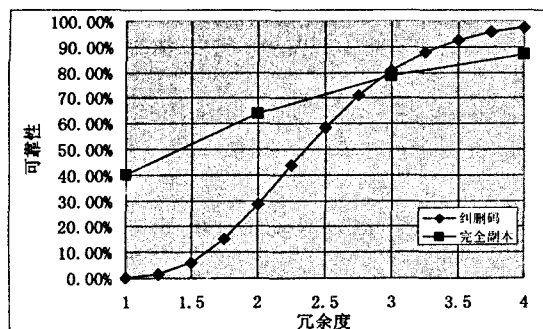
纠删码则是指将要存储的数据先切分为 m 个部分,然后通过编码算法变换为 n ($n > m$) 个部分,其中任意 t ($t \geq m$) 个部分可以用来恢复原始数据,即满足系统结构的可靠性关系:

$$R = \sum_{i=0}^{mr-m} \binom{mr}{i} p^{mr-i} (1-p)^i$$

考察 $p = 0.8$ 和 $p = 0.4$ 时,完全副本冗余与纠删码冗余(编码碎片数 $m = 8$) 模式下的系统可靠性随冗余度变化的情况。



(a) $P=0.8$



(b) $P=0.4$

图2 可靠性比较

从图2的曲线变化可看出:纠删码并不适应低冗余度的存储环境,但随着冗余度的逐渐增大,纠删码开始出现优于完全副本的可靠性效果;因此对于本系统,冗余方案采用纠删码方式。Reed-Solomon算法是一种前向错误校正算法,实现基于数据块的错误纠正,主要用于数据通信和存储应用中。Reed-Solomon码是BCH码的一个子集,也被称为RS(n, k),这里 n 代表编码后共有 n 个码子, k 代表产生 k 个校验码,每个码

子长度为 s 个比特。在接受端,只要在 n 个码子中有任何 $n - k$ 个是正确的就可以恢复原始数据。相对于完全副本方式,Reed-Solomon造成了一定的计算量,也增加了系统设计和实现的复杂度;但它能在相同冗余数据的情况下提供比副本方式更高的系统可靠性。

2.3 基于节点可靠度的数据存储策略

在实际应用中构建一个P2P存储系统时,所面对的大部分是低可靠度的节点。而根据图2,纠删码模式在 p 值较小的情况下,在开始的很长一段冗余度区域内都会出现不如完全副本冗余的情况,随着冗余度的增加整个系统的可靠性提高并不明显;且文献[8]中提到,如果网络中的节点以随机的方式构建,那么在网络中相邻的两个节点可能在实际的网络环境中相距很远,导致路由开销过大,传输性能不理想。针对这种情况,提出一种基于节点可靠度将数据分层存储策略。通过一段时间的统计,按照可靠度的相对大小,将节点分为高可靠度节点层和低可靠度节点层的双层结构,如图3所示。在数据编码后,优先将一部分数据编码块存储在高可靠度节点上。

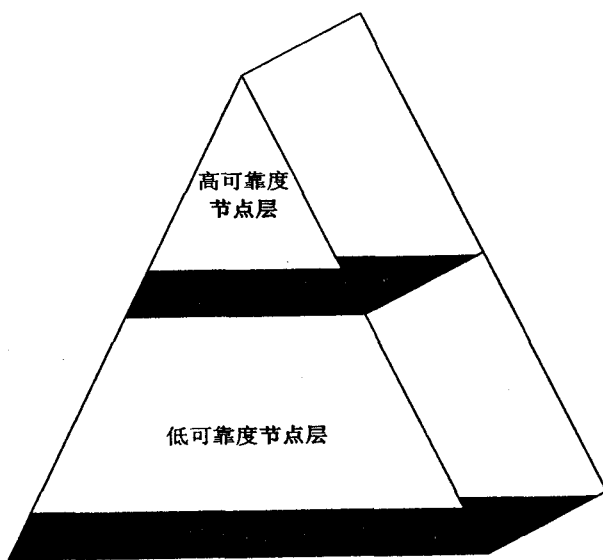


图3 节点组织结构

3 系统可靠性分析

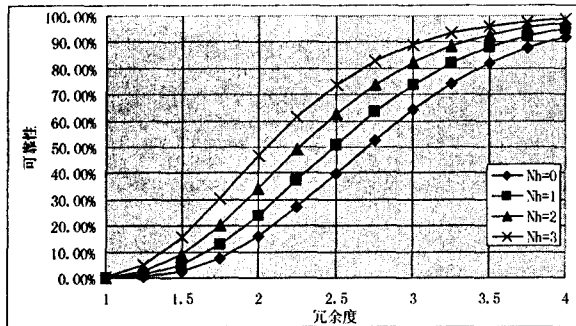
设高可靠度节点平均可靠度为 p_H , 低可靠度节点平均可靠度为 p_L , 碎片数为 m , 数据冗余倍数为 r , 高可靠度节点数为 n_H 。系统可靠性:

$$R = \sum_{i=0}^{n_H} \binom{n_H}{i} p_H^{n_H-i} (1-p_H)^i \left(\sum_{j=0}^{mr-m-i} \binom{mr}{j} p_L^{mr-j} (1-p_L)^j \right) \quad (0 < n_H < mr - m)$$

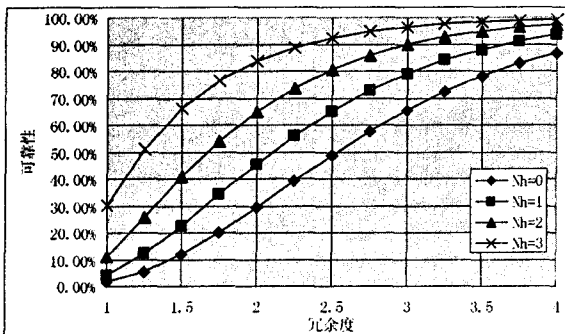
当 $n_H = 0$ 时即所有数据都存储在低可靠度节点上,在这种情况下即

$$R = \sum_{j=0}^{mr-m} \binom{mr}{j} p_L^{mr-j} (1-p_L)^j$$

考察 $p_H = 0.95, p_L = 0.4, m = 8$, 不同 n_H 的取值对系统可靠性的影响。



$m=8$



$m=4$

图 4 系统可靠性分析

从图 4 的实验分析结果曲线可以得到结论,采用文中提出的基于节点可靠度的分层存储策略,高可靠度节点的利用率得到最大限度的利用,在相同的冗余条件下,随着编码块落在高可靠性节点上的数目的增加,系统可靠性也得到了进一步的提高。特别的,在冗余度为 1.5~3.4 范围内提高更为明显,这说明,当系统遭受一定的损坏后,节点数较少,系统经过一定的冗余恢复后,若处在相对较低的冗余情况,系统可靠性也

能得到有效的保证。

4 结束语

目前,如何管理好日益增长的气象数据已经成为顺利有效开展气象业务的关键点,针对机房内部较稳定节点及系统设计足够简单的目标,提出的基于节点可靠度的数据存储策略,在一定程度上提高了系统的可靠性。文中研究可以作为探索中的试验田,通过实际应用充分体现 P2P 应用于存储的优势,发现并研究该方向建设过程中面临的问题,为其他气象台站建立统一的海量数据存储系统探索一条有效的解决途径。

参考文献:

- [1] 罗杰文. Peer-To-Peer 综述[DB/OL]. 2006. <http://www.intsci.ac.cn/users/luojw/P2P/ch01.html>.
- [2] 江武汉,叶从欢,孙世新. P2P-Grid 结构模型研究与设计[J]. 计算机技术与发展, 2006, 16(2): 135-138.
- [3] Wells C. The oceanstore archive: Goals, structures, and self-repair[R]. UC Berkeley Masters Report. California, USA: [s. n.], 2002.
- [4] 田敬,代亚非. P2P 持久存储研究综述[J]. 软件学报, 2007, 18: 2481-2494.
- [5] Dabek F, Kaashoek M F, Karger D, et al. Wide-Area cooperative storage with CFS[C]//In: Proceedings of the 18th ACM Symposium on Operating Systems Principles (SOSP 2001). Banff: Chateau Lake Louise, 2001.
- [6] Zheng W, Hu J, Li M. Granary: architecture of object oriented Internet storage service. E-Commerce Technology for Dynamic E-Business, 2004[C]//IEEE International Conference. Shanghai, China: [s. n.], 2004: 294-297.
- [7] 袁卫东,战守义. 一种分组 P2P 网络模型[J]. 微机发展(现更名: 计算机与技术发展), 2005, 15(8): 50-52.
- [8] 杨文俊. P2P 网络系统中节点自组织管理机制[J]. 计算机技术与发展, 2006, 16(7): 57-60.

(上接第 78 页)

网络资源调度算法的研究提供了很好的参考。

参考文献:

- [1] 陈宇寒. 网络计算技术研究[J]. 计算机技术与发展, 2008, 18(5): 82-85.
- [2] Moreno R A. Job scheduling and Resource Management Techniques in Dynamic Grid Environments[C]//in: 2003 annual Crossgrid Project Workshop & 1st European Across Grids Conference. Santiago de Compostela, Spain: [s. n.], 2003.
- [3] Buyya R, Abramson D, Giddy J. An economy driven resource management architecture for global computational power grids [C]//Int'l Conf on Parallel and Distributed Processing Tech-

niques and Applications. Las Vegas: [s. n.], 2000.

- [4] Di Martino V. Scheduling in a grid computing environment using genetic algorithms[C]//Mililotti M. the 16th Int'l Parallel and Distributed Processing Symp (IPDPS2002). Florida, USA: [s. n.], 2002.
- [5] Abraham A, Buyya R. Nature's heuristics for scheduling jobs on computational grids[C]//The 8th Int'l Conf on Advanced Computing and Communications (ADCOM 2000). Cochin, India: [s. n.], 2000.
- [6] 胡自林,徐云,毛涛. 基于效益最优的网格资源调度[J]. 计算机工程与应用, 2005(7): 69-71.
- [7] 林剑柠,吴慧中. 基于遗传算法的网格资源调度算法[J]. 计算机研究与发展, 2004, 41(12): 2195-2199.