数据网格中一种基于副本和缓存的元数据管理系统

石 柯 王庆春 吴 松
(华中科技大学计算机科学与技术学院 武汉 430074)
(keshi@hust.edu.cn)

摘 要 元数据管理是数据网格的关键技术之一.对全局分布式存储系统GDSS(global distributed storage system)中的元数据管理进行了改进,提出了一种基于副本和缓存的分布式元数据管理系统RCMMS(replication and cache based distributed metadata management system),缓存设置在GDSS系统中的存储服务点SSP(storage service point)端.还讨论了RCMMS的设计、实现以及测试.RCMMS提供了动态管理元数据副本的有效算法.分析和测试表明,副本结合缓存的元数据管理方案在性能上超过了GDSS现有的元数据管理系统,有着较好的可靠性.

关键词 元数据管理;数据网格;全局分布式存储系统;存储服务点;可靠性

中图法分类号 TP316.4

A Replication and Cache Based Distributed Metadata Management System for Data Grid

SHI Ke, WANG Qing-Chun, and WU Song
(*School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan* 430074)

Abstract Metadata management is a key technique of data grid. In this paper a replication and cache based metadata management system (RCMMS) is presented to improve metadata management of a global distributed storage system (GDSS). Storage Service Provider (SSP) of GDSS has dedicated buffer to cache metadata. The design, implementation, and evaluation of RCMMS are discussed. RCMMS provides efficient algorithms to manage highly dynamic metadata replicas. The evaluation demonstrates that the replica plus cache metadata management system outperforms the existing metadata management system of GDSS and has good reliability.

Key words metadata management; data grid; global distributed storage system(GDSS); storage service point(SSP); reliability

1 引 言

元数据是描述数据的数据,数据网格[1~3]中的所有元数据构成了元数据目录[4],它采用统一的结构来描述元数据.本文提出了一种基于副本和缓存的分布式元数据管理系统RCMMS,它借鉴了数据副本[5~6]和Web缓存的思想,引入了元数据副本和元数据缓存的概念,不仅支持多个元数据副本,而且支持元数据副本的动态加入或删除.RCMMS通过稀疏强连通图来描述元数据副本的结构以及它们之间的关系,来实现以上功能.

2 相关工作

数据网格技术的发展非常迅速,对元数据管理技术的研究成为众多项目的研究目标.美国圣地亚哥超级计算中心SDSC的SRB[7]是用途较广的数据

收稿日期:2004-09-21
基金项目:国家"八六三"高技术研究发展计划基金项目(2001AA111011);国家杰出青年科学基金项目(60125208);国家"九七三"重点基础研究发展规划基金项目(2003CB316903)

网格软件之一. Griddaen[8]是国防科学技术大学设计的一个数据网格系统. SRB 和 Griddaen 的元数据信息都存放在关系型数据库中.GDSS 是华中科技大学集群与网格计算实验室开发的一个存储虚拟化项目,它实现了数据网格的基本功能. GDSS 的元数据信息存放在 LDAP 数据库中. 以上提到的项目中,如果元数据目录在一次操作中出错并且它不存在副本的话,那么元数据服务将终止. 而在 RCMMS 中,元数据目录被复制来获得高可用性.

3　RCMMS 系统设计

为了提高网格环境下元数据的访问性能及可用性,本文改进了 GDSS 系统中的元数据管理模型 GNS,为每个域中的元数据创建副本,在 SSP 端创建缓存,借鉴 Griddaen 引入一个全局元数据服务器. 然而,在 Griddaen 中,每个域中只有一个本地元数据服务器,而在 RCMMS 中每个域有多个本地元数据服务器,它们之间以层次或对等方式交互.

3.1　相关概念

RCMMS 以分区的粒度来复制元数据,元数据副本即分区副本. 分区是 LDAP 服务器中目录树的子树,如图 1 所示,当创建元数据副本时,也就是创建分区 *P-A*,*P-B*,*P-C* 或 *P-D* 的副本.

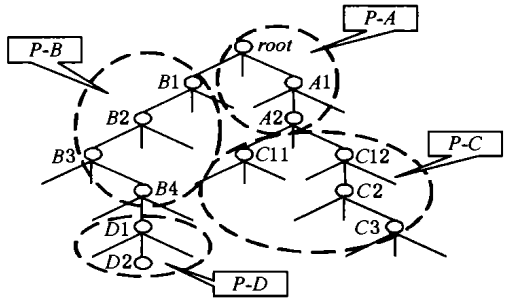


图 1　目录树-分区的关系

用一个稀疏强连通图来表示这些副本的结构以及它们彼此之间的关系. 如图 2 所示,图中的每个节点表示一个分区副本. RCMMS 将副本分成两种类型:主副本和从副本,图中带阴影的节点为主副本,其他节点为从副本. 在当前 RCMMS 系统中,原始分区是没有创建任何副本的分区,被指定为主副本,系统中始终只有一个主副本,并且位置固定.

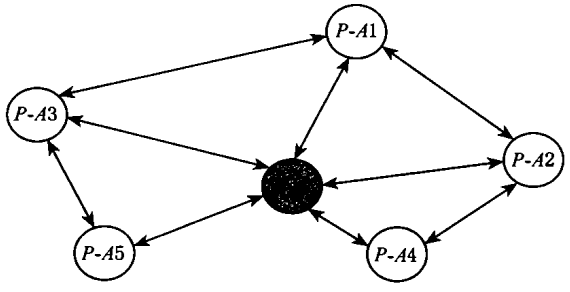


图 2　分区 *A* 的副本稀疏强连通图

3.2　RCMMS 体系结构

RCMMS 系统结构如图 3 所示,其组成结构包括全局元数据服务器(GMS)、域元数据服务器(LMS)、元数据副本服务器(RMS)和 SSP 缓存(SC).

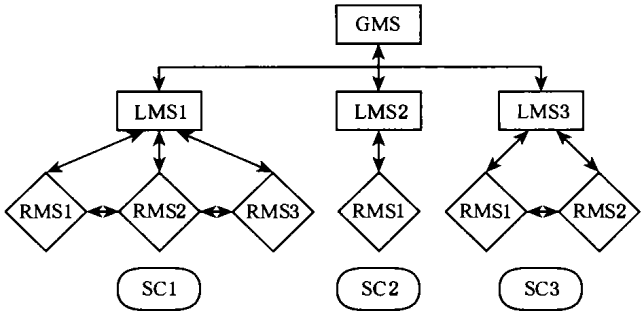


图 3　RCMMS 的体系结构

GMS 的任务是创建和维护 LMS 的索引信息. 索引信息表由两列组成,第 1 列是 LMS 所在的域名,第 2 列是 LMS 的 IP 地址.

RCMMS 本地元数据服务器包括 LMS 和 RMS,由 3 个主要模块构成:

① SSP-RCMMS 协议处理器. 接收来自 SSP 的请求,更新本地元数据副本,并产生复制请求转发给复制引擎.

② 复制引擎. 接收来自 SSP-RCMMS 协议处理器以及其他复制引擎发来的请求,创建、修改或删除副本,并在必要的时候将请求转发给其他节点. 这是本地元数据服务器最关键的部分.

③ 成员信息模块. 维护其他节点的状态信息,包括系统活性、可用磁盘空间和相应副本的位置.

SC 保存热点元数据. 它可以提高元数据读写效率,只占用较少的空间.

4　元数据副本及缓存管理

在 RCMMS 中,通过副本创建算法在 RMS 上创建的元数据副本为从副本,当一个节点的磁盘空间耗尽时,可以删除某些从副本来增加磁盘空间. 如下是一个副本的关键属性:

struct *Replica*

pid: PartitionID

ts: TimeStamp /* 物理时钟,IP 地址 */

vv: VersionVector /* IP 地址↔时间戳 */

masterPeers: Set(NodeID)/* IP 地址集 */

时间戳(*ts*)和版本矩阵(*vv*)记录副本最后一次被修改的时间. *MasterPeers* 是从副本指向该分区主副本的单向连接.

Peers 是从副本指向它的相邻从副本的连接.

元数据副本创建策略包括如何选择要创建的副本、在何时创建副本和在何地放置副本 3 个方面的问题.

首先讨论如何决定要复制的分区. 该算法动态地将一棵目录树按照条目数划分得到多个分区,在该服务器上建立关于分区信息的索引表. T_{resp}为分区响应时间,$T_{threshold}$为响应时间阈值,当 $T_{resp} > T_{threshold}$时,这个分区就是要创建从副本的分区,并且它被系统指定为主副本.

然后,根据 PQ 参数原则[9]触发副本创建进程. 设置参数 P,$Q(P<Q)$,P 值要求系统可以对剧烈增长的 SSP 访问请求做出及时响应,Q 值则对元数据副本创建更为谨慎,其允许元数据副本及所在服务器性能暂时波动,判断触发时机步骤如下:

(1) $T_{resp} < T_{threshold}$,放弃元数据副本创建,否则转(2);

(2) 在 P 时段内,且 $\forall et \in P$,$T_{resp} > T_{threshold}$且 $\frac{dT_{resp}}{dt} > 0$,则转(4),否则转(3);

(3) 在 Q 时段内,若 $\forall et \in Q$,$T_{resp} > T_{threshold}$,则转(4),否则放弃元数据副本创建;

(4) 触发副本创建进程,创建从副本.

确定副本创建时机后,LMS 根据客户端的访问模式选择一个不存在该元数据副本的 RMS 来放置副本. 比较了不同访问模式下 fastspread[10] 和 cascading[10]两种策略. fastspread 策略在 SSP 元数据访问路径上的所有服务器上创建副本;cascading 策略则只在下一级服务器上创建副本,并沿着层次结构扩展.

新创建的从副本必须要加入到相应的副本连通图中,以保证同一分区不同副本之间的一致性. 因此,复制引擎模块将选择 x 个已知副本,使它们指向新创建的从副本. 这 x 个副本必须满足以下要求:

① 要包括域内该分区的主副本;

② 要包括距离新创建从副本最近的一个从副本;

③ 尽可能选择任意 $x-2$ 从副本,使图的不连通概率尽可能低. 为了删除元数据副本,服务器发送通知给该副本在连通图中的邻居. 每个邻居从主副本开始遍历,重新寻找一个可用副本,建立一条新的边,以保证图的强连通性.

为了提高元数据读写效率,RCMMS 模型中采用了缓存技术. 缓存模块存放在 SSP 服务器上,它保存热点元数据和正在被执行写操作的元数据,根据访问的局部性原则把元数据服务器上的元数据复制到 SSP 服务器的缓存中. 由于只保留少量元数据,缓存并不占用太大空间. 需要注意的是,与副本元数据不同,缓存中的元数据是不完整的.

5 系统测试

本节评估 RCMMS 的系统性能和可靠性. 我们研究了单域固定结构 RCMMS(如图 4 所示)的基本性能和故障恢复. 实验结果显示 RCMMS 比 GNS 更加适合数据网格环境.

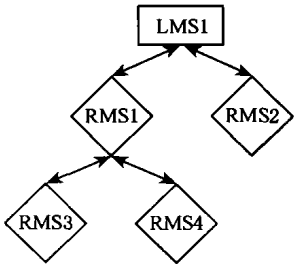


图 4 单域固定结构 RCMMS

5.1 测试环境

本系统的测试环境为 16 个节点的集群系统,每个节点的配置为 Xeon 1 GHz 的处理器,512 MB 的内存,40GB 的硬盘,节点间以 100 Mbps 以太网相连接.每个节点安装 Red Hat 7.3,内核为 2.4.9 版 . 测试用集群除了进行测试之外,还执行日常事务,如 FTP 服务、Web 服务及一些开发任务.

5.2 性能评估

下面的实验中 SSP 端没有设置缓存. 当访问模式为集中模式时,即 SSP 对分区 A 的访问集中在 RMS3,fastspread 将在 RMS1 和 RMS3 上创建副本;而 cascading 则只在 RMS1 上创建副本. fastspread 比 cascading 的响应时间快了 18%,前者的存储空间消耗和副本更新时间都接近后者的 2 倍.

当访问模式为均匀模式时,即 SSP 对分区 A 的

访问均匀分布在 RMS3 和 RMS4,fastspread 将在 RMS1,RMS3 和 RMS4 上创建副本;而 cascading 则只在 RMS1 上创建副本. fastspread 比 cascading 的响应时间快了 21%,前者的存储空间消耗和副本更新时间都接近后者的 3 倍.

当访问模式为随机模式时,即 SSP 对分区 A 的访问是随机的,fastspread 将在 RMS1,RMS2,RMS3 和 RMS4 上创建副本;而 cascading 则只在 RMS1 和 RMS2 上创建副本. fastspread 比 cascading 的响应时间快了 8%,而前者的存储空间消耗和副本更新时间都接近后者的 2 倍.

在 3 种情况下,采用 fastspread 策略所得到的响应时间都比 cascading 少,但是前者是以存储空间消耗和副本更新为代价的. 因此,集中模式下 RCMMS 采用 fastspread 策略,均匀或随机模式下,RCMMS 采用 cascading 策略. 根据这样的策略选择,我们将 RCMMS 和 GNS 在性能上做了一个比较(如图 5 所示). 由此我们可以得出结论,虽然 RCMMS 以空间消耗和副本更新为代价,但是获得的响应时间提高是显著的,所以在总体性能上比 GNS 有了明显的提高.

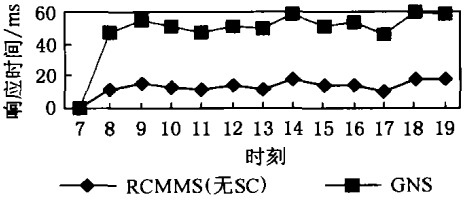


图 5　RCMMS(无 SC)和 GNS 响应时间比较

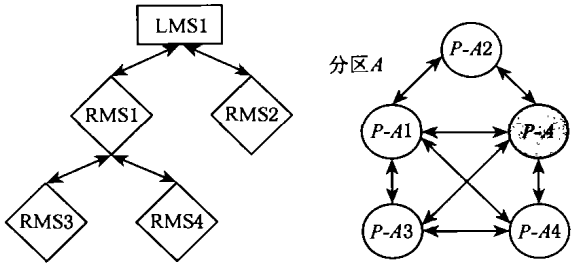
进一步,我们在 SSP 端设置缓存,RCMMS 的系统性能又提高了将近 20%.

5.3　可靠性评估

在 RCMMS 系统中,当某个 RMS 长期不可用时(比如两周),它上面的元数据副本将迁移到一台或多台 RMS 上,重新构造副本强连通图,以替换失效的元数据副本.

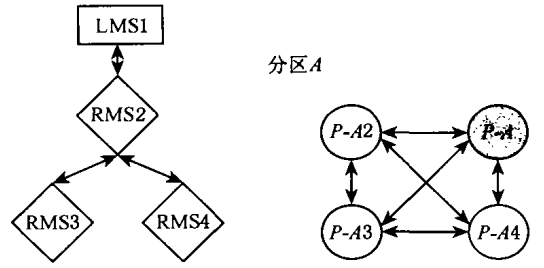
下面的实验中,我们仍然使用性能测试中的单域结构,并且系统在 SSP 端设置了缓存. 系统的初始状态如图 6 所示,其中 *P-A* 为分区 *A* 的主副本,位于 LMS1 上,*P-A*1～*P-A*4 为分区 *A* 的从副本,分别位于对应的元数据副本服务器 RMS1～RMS4.

当 RMS1 不可用时,RMS 服务器和 LMS 服务器中的成员信息模块将启动,RMS3 和 RMS4 将从 LMS1 获取得到 RMS2 的信息,因为 RMS2 是距离它们最近的元数据副本服务器,因此 RMS3 和 RMS4 更新本地的成员信息表,把父节点更新为 RMS2,同时告知 LMS1. LMS1 更新本地成员信息表的同时,通知 RMS2,让它把自己的子节点更新为 RMS3 和 RMS4. 然后各服务器启动复制引擎,重新构造分区 *A* 的强连通图,如图 7 所示. 当 RMS1 上的元数据量为 6MB(包括元数据条目 33900 个)左右的时候,整个过程需花费 7 分钟左右的时间.



(a) 副本服务器拓扑结构　　(b) 分区*A*的副本强连通图

图 6　RCMMS 的系统初态



(a) 副本服务器拓扑结构　　(b) 分区*A*的副本强连通图

图 7　RMS1 失效后 RCMMS 的修复状态

6　结束语

综上所述,RCMMS 和 GNS 相比,提高了网格中元数据的访问速度和可靠性. 然而我们的测试仅仅局限于单域固定结构,并没有在真正的广域多变环境下运行,多域环境的复杂性对于系统性能和可靠性都是一个很大的挑战. 在测试中我们还发现 LDAP 服务器的稳定性不是很好,当系统突然断电再重启之后,LDAP 数据库往往不能正常工作. 因此我们也在考虑将现有的系统向关系型数据库上移植. 另外,如何利用元数据副本稀疏强连通图进行元数据副本一致性的维护,也是系统的一个难点,关于这方面的讨论将在今后的研究中进行.

参　考　文　献

1　I Foster. The grid: A new infrastructure for 21st century science. Physics Today, 2002, 55(2): 42～47

2 B Allcock, J Bester, J Bresnahn, *et al*. Data management and transfer in high performance computational grid environments. Parallel Computing Journal, 2002, 28(5): 749～771

3 I Foster, C Kesselman, S Tuecke. The anatomy of the grid: Enabling scalable virtual organizations. International Journal of Supercomputer Applications, 2001, 15(3): 200～222

4 Ian Foster, Carl Kesselman. Computational grids. In: The Grid-Blueprint for a New Computing Infrastructure. San Francisco: Morgan Kaufmann, 1999. 15～51

5 Corina Ferdean, Mesaac Makpangou. A scalable replica selection strategy based on flexible contracts. The 3rd IEEE Workshop on Internet Applications, San Jose, California, 2003

6 H Lamehamedi, B Szymanski. Data replication strategies in grid environments. The 5th Int'l Conf on Algorithms and Architectures for Parallel Processing, Beijing, 2002

7 Michael Wan, Arcot Rajasekar, Wayne Schroeder. An overview of the SRB 3.0: The federated MCAT. http://www.npaci.edu/DICE/SRB/FedMcat.html, 2003-09

8 李东升，李春江，肖侬，等. 数据网格环境下一种动态自适应的副本定位方法. 计算机研究与发展，2003，40(12)：1775～1780
(Li Dongsheng, Li Chunjiang, Xiao Nong, *et al*. Dynamic self-adaptive replica location method in data grids. Journal of Computer Research and Development(in Chinese), 2003, 40(12): 1775～1780)

9 Byoung-Dai Lee, Jon B Weissman. An adaptive service grid architecture using dynamic replica management. The 2nd Int'l Workshop on Grid Computing, Denver, Colorado, 2001

10 Kavitha Ranganathan, Ian Foster. Identifying dynamic replication strategies for a high-performance data grid. In: Proc of the Int'l Workshop on Grid Computing. Berlin: Springer-Verlag, 2001

石 柯 男，1973 年生，博士，副教授，主要研究方向为网格与集群计算、移动计算、嵌入式系统等.

王庆春 女，1980 年生，硕士研究生，主要研究方向为并行分布式系统、集群与网格计算.

吴 松 男，1975 年生，博士，讲师，主要研究方向为分布式存储系统、集群计算、并行多媒体服务.