文章编号:1001 - 9081(2006)10 - 2282 - 03

一种基于安全的网格数据副本策略模型

李 静[1,2],陈蜀宇[3],吴长泽[1]
(1. 重庆大学 计算机学院,重庆 400044;2. 重庆教育学院 计算机与现代教育技术系,重庆 400067;
3. 重庆大学 软件学院,重庆 400044)
(li_jing1@163.com)

摘 要:数据网格中数据副本的存在是为了获得对数据的更好的访问性能,同时也是为解决容错问题而采用的一种冗余技术,但系统复杂程度的增加会带来一系列不可预测的安全隐患。安全与容错是既相互统一,又相互矛盾的关系,应将它们综合起来加以研究。为此将数据副本机制与信息安全结合起来,提出一种决定数据资源副本数量的数学模型。该模型综合考虑服务提供者一方经济利益和声誉度,在合理的假设下对一个两目标的优化问题作了简化处理,通过数值计算和分析确定了数据副本数量的最佳限额。

关键词:网格; 数据副本; 安全; 容错
中图分类号: TP393.03 **文献标识码**:A

Model of data replication strategy based on security in grid

LI Jing[1,2], CHEN Shu-yu[3], WU Chang-ze[1]
(1. *College of Computer Science, Chongqing University, Chongqing* 400044, *China;*
2. *Department of Computer and Modern Education Technology, Chongqing Education College, Chongqing* 400067, *China;*
3. *College of Software, Chongqing University, Chongqing* 400044, *China*)

Abstract: Data replication in data grid is an important technique, by which, better access performance can be got. It is also a fault-tolerant approach, but at the same time brings some hidden trouble for security. Therefore, security and fault-tolerance should be colligated to research. Combining data replication with security, we proposed a mathematic model which can determine data replication quantity. In this model, taking economic interest and reputation index of service supplier into account, an optimization problem of two targets was simplified under reasonable hypothesis. Numeric calculation and analysis show that there is an optimal quantity for the data replication.

Key words: grid; data replication; security; fault-tolerance

0 引言

数据网格是对广域范围内大规模的数据集进行分布式管理、分析和使用的一个综合的体系结构[1],实现安全、可靠和有效的网格环境中的数据传输以及访问、复制等操作,并提供到不同存储系统的统一的接口,从而使得数据密集型的高性能计算和大量的共享数据密集型的事务处理及科学研究成为可能,在电子商务、数据仓库和数据挖掘、高能物理、生物信息科学等商业和科学工程领域起到了至关重要的作用。

数据复制在数据网格中是一个关键的问题,其目的是为了获得对数据的更好的访问性能,包括较短的访问时间和实现容错。数据复制技术已被广泛使用在分布式系统和数据库管理技术中,但不能满足严格的一致性、安全和高速的广域网间大量数据的传送,因此不适应数据网格应用的需要。复制应完成这样一些功能:生成新的完整的或部分的数据拷贝;把这些新的拷贝,也就是副本注册进一个复制目录中;允许用户和应用去查询目录以发现所有现存的部分和全部文件的副本;基于负载平衡等策略选择一个“最好的”副本用于访问[2]。

数据副本的存在不仅是为了提高服务质量,同时也是为解决容错问题而采用的一种冗余技术。冗余技术可部分解决部件的失效、硬件故障、软件错误、程序员的失误甚至安全协议的漏洞等问题,但毫无疑问,系统复杂程度的增加会带来一系列不可预测的安全隐患。在网格的多个节点上存放完整或部分数据库的冗余副本,从容错的观点看,这对于提高系统的容错性能是有利的;然而从安全的观点看,数据库冗余度越大,给攻击者提供的攻击点就越多,泄密的可能性就越大,这对系统的安全将是十分有害的。因此,可以说安全与容错是既相互统一,又相互矛盾的关系。这样的关系要求人们在解决这两个问题时应将它们综合起来加以研究。

本文将数据副本机制与信息安全结合起来,提出一种决定数据资源副本数量的策略模型。该模型不仅因为复制足够的数据副本而为网格服务消费者一方保证了一定的服务质量,而且综合考虑服务提供者一方经济利益和声誉度,确定了

收稿日期:2006 - 04 - 29;**修订日期**:2006 - 06 - 19
基金项目:教育部跨世纪优秀人才支持计划(NCET-04-0843);重庆市自然科学基金资助项目(2005BB2192)
作者简介:李静(1974 -),女,重庆人,博士研究生,主要研究方向:网格、网络安全; 陈蜀宇(1963 -),男,重庆人,教授,博士生导师,主要研究方向:网格、网络安全; 吴长泽(1980 -),男,四川人,博士研究生,主要研究方向:网格.

数据副本数量的最佳限额。

1　相关工作

Globus 数据网格的复制管理体系结构是一个分层的体系结构。用于复制文件的工具主要是复制目录和 GridFTP。在最底层是一个复制目录层,采用维持从逻辑文件名到物理位置映像的方法跟踪一个单一逻辑文件的多个物理拷贝。复制目录包括三种类型的对象:一组逻辑文件名、一个位置对象和一个逻辑文件表目。用户请求不会直接送到数据访问器,而是由复制管理器转发。通过网格缓存提供高级访问服务和实现优化。复制管理器是一个智能化的服务,通过分析用户访问模式以发现在何地和如何以一种优化的方法访问文件。作为这些访问模式分析的结果是在远程网站进行复制的生成和清除[2]。

GDMP 是一个基于 Globus Toolkit 的多线程客户机—服务器系统,主要解决快速和有效的点到点的文件复制,目前实现的复制策略是具有签名模式的异步复制机制[3]。GDMP 服务器是一个守护进程,运行在一个产生和输出数据的节点上。服务器本身使用通讯模块,用于接受来自于应用客户机的请求。一个线程池同时处理多个客户机请求, 每个客户机使用一个线程。在 GDMP 中数据复制经过多个步骤[4]:首先在数据源使用数据拷贝工具把所要复制的数据拷贝到一个新的文件中;其次使用广域文件复制的方法把该数据移动到目的网站中;最后把该源站的新文件删除掉。

Globus 项目组开发了一个统一的网格数据传输和访问协议 GridFTP[5],它提供广域网环境中安全的和有效的数据传送。GridFTP 是对 FTP 协议的一个扩展,是目前所使用的各种网格数据存储系统所提供的特性的一个超集。在传送和访问数据时,GridFTP 提供了健壮、灵活的认证,以及完整性和机密性。作为一个公用的数据访问协议,对于网格数据提供者和用户来说,GridFTP 都具有优越的性能。

以上关于数据资源的副本技术涉及到了数据复制的优化、快速性和有效性以及安全性等问题,但都没有从冗余和安全的矛盾性这个角度来考虑副本的制定策略。

2　问题分析及模型假设

在介绍数据副本策略模型之前,先定义以下几个角色:

1) 网格货币(Grid Currency,GC)。它是网格环境中的一般等价物,可以是网格环境中的一种紧俏资源,对所有网格资源的计量都可按比例地转换为对这种网格货币的计量。网格货币与现实生活中的货币有一定的对应关系,最终可转换成现实生活中的货币[6]。

2) 数据网格服务提供者(Grid Resource Provider,GRP)。它是通过出售数据资源,允许他人使用自己的服务,从而获得网格货币的计算机。

3) 数据网格服务消费者(Grid Resource Consumer,GRC)。它是通过付出网格货币,购买网格数据资源,获得网格服务的计算机。

如果把数据网格看做一个市场,在激烈的市场竞争中,服务提供者为争取更多的客源而希望尽量提高服务质量,而单一数据资源会造成访问瓶颈和单点失效,因此服务提供者会给出多个数据副本。现在的问题是如何确定数据副本的数量。

服务提供者首先根据市场需求估计所需副本量,另外还需要考虑黑客的恶意攻击会造成部分数据副本失效(不能访问或被篡改);如果不限制副本的数量,那么获得的利润将不足以支付维护费用。如果消费者访问到失效的副本,可能会对其造成一定的损失,导致服务提供者声誉受损和一定的经济损失,如客源减少等,因此服务提供者可以付给消费者一定的赔偿金。综上服务提供者需要综合考虑经济利益和声誉度(Reputation Index,RI),确定数据资源副本数量的最佳限额。

服务提供者的经济利益可以用数据服务出售收入扣除使用成本和赔偿金后的利益来衡量,声誉度可以用被黑客攻陷或其他原因而失效的数据副本限制在一定数量为标准。这个问题的关键因素,即黑客对数据资源的攻击是随机的,所以经济利益和声誉度两个指标都应该在平均意义下衡量,这是个两目标的优化问题,决策变量是数据资源副本数量的限额。

在提出数据副本策略模型之前,我们作出以下假设:

假设 1　根据市场需求估计的所需副本量为常数 n,消费者购买数据服务价格为常数 r,由于消费者数量 t 与副本量 n 成正比,因此总销售额 $t \times r$ 可以表示为 $n \times w$,称 w 为理想数据副本单价。数据资源成本为常数 c,副本维护成本忽略不计。理想数据副本单价 $w = c/\lambda n$,其中 λ 是利润调节因子。对计算资源和存储资源等网格资源来说,多一份资源就多一份成本;而数据资源与它们不一样,虽然有多个副本,但实质内容一样,因此假设数据资源成本为常数 c 有合理性的。

假设 2　为容错而实际发布的副本数量为 m,$m > n$,每个副本失效的概率为 p,各副本被攻击或其他原因而失效是独立的。

假设 3　每个失效数据副本造成的经济损失,也就是给消费者的赔偿金为常数 b。

3　数据副本模型设计

服务提供者的经济利益可以用平均利润 S 来衡量。每轮服务的利润 s 为从服务单价收入中减去服务成本和可能发生的赔偿金。当 m 个数据副本中有 k 个失效时:

$$s = \begin{cases} (m-k)w - c, & m-k \leqslant n \\ nw - c - kb, & m-k > n \end{cases} \tag{1}$$

由假设 2,失效的副本数量 K 服从二项分布,于是:

$$p_k = P(K=k) = C_m^k p^k q^{m-k},\ q = 1-p \tag{2}$$

平均利润 S,即 s 的数学期望为:

$$S(m) = \sum_{k=0}^{m-n-1} [(nw-c) - kb] p_k + \sum_{k=m-n}^{m} [(m-k)w - c] p_k \tag{3}$$

化简(3)式,并注意到 $\sum_{k=0}^{m} k p_k = mp$,可得:

$$S(m) = qmw - c + (w-b) \sum_{k=0}^{m-n-1} k p_k + (n-m) \sum_{k=0}^{m-n-1} p_k \tag{4}$$

当 n,w,c,p 给定后可以求 m,使 $S(m)$ 最大。

服务提供者从声誉度和经济利益两方面考虑,应该要求

失效的副本数量不要太多,而由于失效副本的数量是随机的,可以用失效的副本超过若干人的概率为度量指标。失效的副本数超过 i 人的概率为 $P_i(m)$,所以:

$$P_i(m) = \sum_{k=i}^{m} p_k \tag{5}$$

对于给定的 n,i,显然当 m 变大时 $P_i(m)$ 单调增加。

综上所述,$S(m)$ 和 $P_i(m)$ 虽然是这个优化问题的两个目标,但可以将 $P_i(m)$ 不超过某个给定值作为约束条件,以 $S(m)$ 为单目标函数来求解。

为减少 $S(m)$ 中的参数,取 $S(m)$ 除以服务成本为新的目标函数 $H(m)$,其含义是单位成本获得的平均利润,注意到假设 1 中有 $w = c/\lambda n$,由(4)式可得:

$$H(m) = S(m)/mc = \frac{1}{\lambda n}\left[qm + (1 - b/w)\sum_{k=0}^{m-n-1} kp_k + (n-m)\sum_{k=0}^{m-n-1} p_k\right] - 1 \tag{6}$$

其中 b/w 是赔偿金占理想数据副本单价的比例,问题化为 $\lambda,n,p,b/w$,求 m 使 $H(m)$ 最大,而约束条件为:

$$P_i(m) = \sum_{k=i}^{m} p_k \leqslant \alpha \tag{7}$$

其中 α 是小于 1 的正数。

4 模型求解及分析

由于模型(6)、(7)无法解析地求解,可以设定几组数据,用 Matlab 软件作数值计算。

设 $n = 300, \lambda = 0.6, p = 0.02$,当 m 从 300 到 330 变化,b/w 分别取 0.2 和 0.4 时,计算 $H(m)$ 的值。结果如图 1 所示。

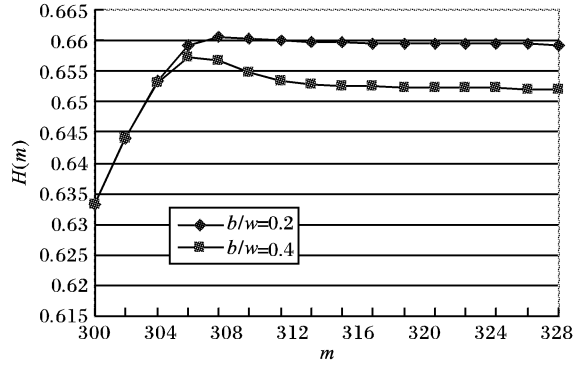

图 1 $n = 300$ 时 $H(m)$ 随 m 的变化

当 m 从 300 到 330 变化,b/w 取 0.2 时,计算各有 10 个和 15 个数据副本失效的概率,分别用 $P_{10}(m)$ 和 $P_{15}(m)$ 来表示。结果如图 2 所示。

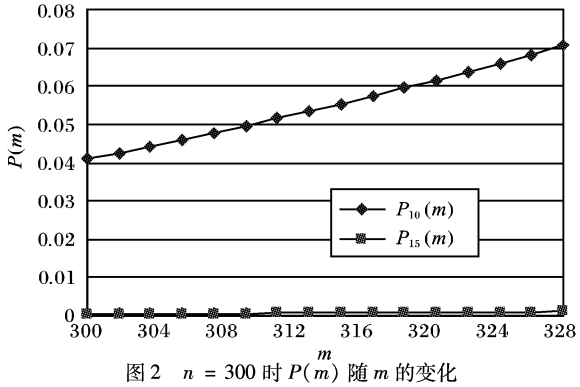

图 2 $n = 300$ 时 $P(m)$ 随 m 的变化

另外取 $n = 400$,其他参数不变时,$H(m)$ 随 m 的变化和 $P(m)$ 随 m 的变化规律相似。

我们对实验结果进行分析如下:

1)对于所取的各个 $n,p,b/w$,平均利润 $H(m)$ 随着 m 的变大都是先增加再减少,但是在最大值附近变化很小,而访问到失效数据副本的消费者超过 10 个或 15 个的概率 $P_{10}(m)$ 和 $P_{15}(m)$ 增加得相对较快,所以应该参考 $H(m)$ 的最大值、给定约束条件(7)式中可以接受的 α,确定合适的 m。

2)对于一定 n,p,当 b/w 由 0.2 增加到 0.4 时,$H(m)$ 的减少不超过 2%,所以不妨付给访问到失效数据副本的消费者以较多的赔偿金,提高声誉度。

3)综合考虑经济效益和声誉度,可给定 $P_{10}(m) < 0.05$,$P_{15}(m) < 0.001$,由图 1、图 2 可知,对于 $n = 300$,若估计 $p = 0.02$,可取 $m = 308$。

5 结语

在数据网格环境中,已有的数据副本机制有其局限性。本文将数据副本机制与信息安全结合起来,提出一种决定数据资源副本数量的策略模型。该模型从安全与容错对立统一的关系出发来研究解决问题,它不仅为网格服务消费者一方保证了一定的服务质量,而且综合考虑服务提供者一方经济利益和声誉度,确定了数据副本数量的最佳限额。我们在基本合理的假设下对一个两目标的优化问题作了简化处理,虽然得到的模型无法解析地求解,但数值计算的结果已经满足我们对问题进行分析的需要。分析结果表明,将访问到失效数据副本的消费者限制在一定范围内时,存在一个数据副本数量的最佳值,使平均利润最大。

参考文献:

[1] FOSTER I, KESSELMAN C, TUECKE S. The anatomy of the grid: enabling scalable virtual organizations [J]. International Journal of High Performance Computing Applications. 2001, 15(3): 200 - 222.

[2] VAZHKUDAI S, TUECKE S, FOSTER I. Replica selection in the globus data grid [Z]. Present International Work shop on Data Models and Databases on Clusters and the Grid (DataGrid 2001). New York: ACM Press, 2001.

[3] CHERVENAK A, FOSTER I, KESSELMAN C, *et al.* The data grid: towards an architecture for the distributed management and analysis of large scientific data sets [J]. Network and Computer Applications. 2001, 26(5): 187 - 200.

[4] GULLAPALLI S, CZAJKOWSKI K, KESSELMAN C, *et al.* The grid notification framework [EB/OL]. Grid Forum Working Draft GWD-GIS-019. http: //www. gridforum. org, 2001.

[5] ALLCOCK W, FOSTER I, NEFEDORA V. High-performance remote access to climate simulation data: a challenge problem for data grid technologies[R]. Technical Report, New York, Argonne National Laboratory, 2001.

[6] MEDVINSKY G, NEUMAN C. NetCash: A Design for Practical Electronic Currency on the Internet [A]. Proceedings of 1st the ACM Conference on Computer and Communication Security[C]. Brighton: Springer-Verlag, 1993. 100 - 103.