数据网格下副本一致性问题的研究

The Research of Replica Consistency for Data Grids

梁 鸿 张春明 高元涛 （中国石油大学(华东)计算机与通信工程学院 山东东营 257061）

摘 要： 分析了目前数据网格环境下的副本一致性研究现状，并提出了一种基于优先级与时间戳的副本一致性模型(RCPTM，Replica Consistency based Priority and Timestamp Model)及算法，并在模拟环境 OptorSim 下与其他两种传统的算法进行比较。通过实验模拟表明，本文的模型算法要明显的比传统算法更适合于网格环境中保持副本的一致性。

关键词： 数据网格 副本一致性 优先级 时间戳

1 引言

数据网格[1]通过将 Internet 上存在着的大量分散的、独立的、异构的储存系统组织成一个逻辑意义上的整体，为用户提供高效的、可靠的、可扩展的、海量的存储资源. 但这些数据经常存放在不同的存储系统中和不同的位置上，这让用户使用起来很不方便。副本机制的引入可以带来减少数据访问延迟、减少网络带宽的消耗、提高可用性等优点，但同时也带来了一些诸如副本创建的时机、副本选择的额外开销以及副本一致性等问题。

在数据网格领域，存在三种类型的一致性问题：元数据副本间的一致性、元数据与应用数据间的一致性以及应用数据副本间的一致性。本文主要讨论应用数据副本间的一致性问题，并在目前副本一致性的研究现状的基础上，提出一种基于优先级的和时间戳的副本一致性模型(RCPTM)及其算法，并在 OpterSim 环境中与其他两种传统算法进行实验模拟，实验结果表明，本文提出的模型及算法更适合在网格环境下保持副本的一致性。

2 相关研究

副本一致性是副本管理中的重要组成部分，直接影响到副本管理的性能和正确性。目前，在传统的分布式数据库及分布对象等领域，已经进行了深入的研究。但是数据网格跟数据库、分布对象等领域相比，要维护副本间的一致性，需要充分考虑数据网格系统的以下特点：

- 数据网格系统的副本分布于广域网环境，其访问延迟大，且经常存在网络失效；
- 数据网格环境具有动态性，网格节点可能会动态加入和退出系统；
- 数据网格系统在运行时刻根据需要动态创建和删除副本；
- 数据网格环境中副本的数量较大。

因此传统的如事务控制法、复制控制法以及消息队列法等算法，都有一定的适用局限性，并不适合在网格环境下使用。

欧洲数据网格项目[4]提出了副本一致性服务(RCS，Replica Consistency Service)，RCS[3]负责维护存在的副本的一致性。其中，本地一致性服务运行在每一个存储单元，以确保本地数据的一致性。RCS 同样也考虑了关于副本元数据一致性的问题，RCS 提出了一种文件锁机制来控制文件存取。但是分布锁机制在执行用户写操作时，需要封锁所有副本，直到所有副本都达到一致状态后，才可以对其他用户请求进行响应。由于数据网格系统基于广域网环境，且请求数量和副本数量都很大，通过分布锁机制保证数据网格操作的一致性，实现效率很低，将会在很大程度上增大数据网格系统的数据响应时间。

在国内，中科院的孙毓忠和徐志伟提出了数据网络中的懒惰拷贝和积极拷贝[5]两个一致性协议。懒惰拷贝的协议是副本仅在访问的时候才去更新，它可以节省网络带宽而不用当副本变化时传输实时数据，但

是懒惰拷贝在数据更新的时候，就存在了访问延迟。而积极拷贝，则是副本在源文件改变时，实时的更新所有副本，但是会浪费很大的网络带宽。

3 一致性模型及算法

3.1 副本一致性模型架构

在本节中，主要介绍了副本一致性的模型结构，我们利用网络节点的区域性[2]来扩展本结构。在物理上邻近的一些节点构成一个网格区域（GR：Grid Region），例如一个校园内的节点构成一个网格区域。每个区域内存在一个本地索引节点，用来存放本区域内所有副本的相关信息；区域间的通信通过全局索引节点来进行，全局索引节点采用分布式的结构，节点上存放的是副本的全部映射信息，即全局索引节点到本地索引节点的映射，每个文件可以选择某个全局索引节点来存放自己的所有的副本信息。

在每个网格区域内，对于同一物理文件的多个副本，则至多存在一个主副本，主副本可以被用户直接修改；而剩下的副本则为从副本，从副本没有数量的限制，从副本不能被用户直接修改，只能根据主副本的变化来更新。

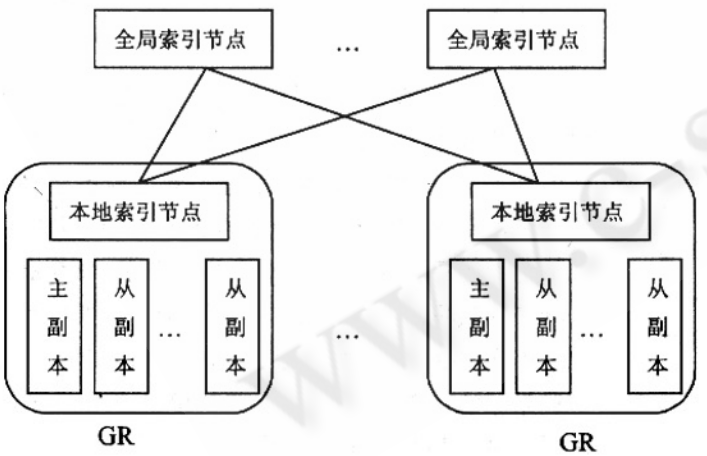

图 1　副本一致性模型架构

3.2 副本一致性的设计实现

在传统的数据库等领域内，副本的一致性基本都采用了锁机制，但是由于网格环境的广域性等原因，在网格环境下的锁机制实现的效率太低，而且会在很大程度上增大数据网格系统的数据响应时间。因此本文提出一种基于优先级和时间戳的副本一致性模型及算法，来减少响应时间。

首先，文件在创建的时候，根据系统当时的时间创建一个时间戳的属性，并同时给其分配一个优先级，优先级可以根据需要修改。优先级分为 5 级，其中，优先级为 1 级的文件更新时间最短，5 级的文件为只读文件，默认级别为 3 级。1 级的更新时间为 1 毫秒，2 级为 2 毫秒，3 级为 8 毫秒，4 级为 16 毫秒。

在为文件创建副本的时候，同时复制时间戳属性和优先级属性。在本地索引节点上，除了存放每个副本的基本信息外，还存放副本的时间戳属性，而在全局索引节点上，也存放时间戳属性。

要想更新副本，首先需要定位副本。在网格环境下，本文采用三级分布式的结构，由于从副本根据本地的主副本更新，所以本文的更新算法分为主副本更新算法和从副本更新算法；主副本更新主要在各网格区域之间进行，从副本更新算法在本网格区域内进行。下面主要介绍本文所采用的定位算法和更新算法。本文所采用的副本定位算法思想如下：

(1) 根据给定的逻辑文件名，查找本地索引节点的信息，然后定位索引信息所在的全局索引节点；

(2) 查找全局索引节点的信息，定位各个网格区域的本地索引节点，然后查找对应主副本的位置信息；

(3) 返回各个网格区域的主副本的位置信息；

(4) 算法结束。

本文采用的主副本的更新算法思想如下：

(1) 主副本的文件修改完毕，提交修改，启动一致性服务。首先查看该文件的级别号，如果级别为 5，则拒绝修改，退出服务；否则，继续执行；

(2) 根据系统时间更新本地主副本的时间戳属性，并更新本地索引节点的时间戳属性；并在本地索引节点上建立相关的更新队列；

(3) 调用副本定位算法，更新全局索引节点的时间戳属性，并将返回的各网格区域的主副本信息添加到更新队列；

(4) 根据更新队列来更新各网格区域的主副本文件,如果成功,则更新网格区域的时间戳,并删除更新队列中的对应的项,否则移至队列的尾部;

(5) 继续执行更新,直至队列为空,更新结束。

本文采用的从副本的更新算法思想如下:

(1) 首先取得从副本文件的级别号及时间戳属性,并根据级别号来确定更新的时间间隔;

(2) 根据更新的时间间隔访问本地索引节点;

(3) 查看时间戳是否与自身的一致,如果一致,则转到(2);

(4) 否则根据本地索引节点所查询到的主副本更新从副本。

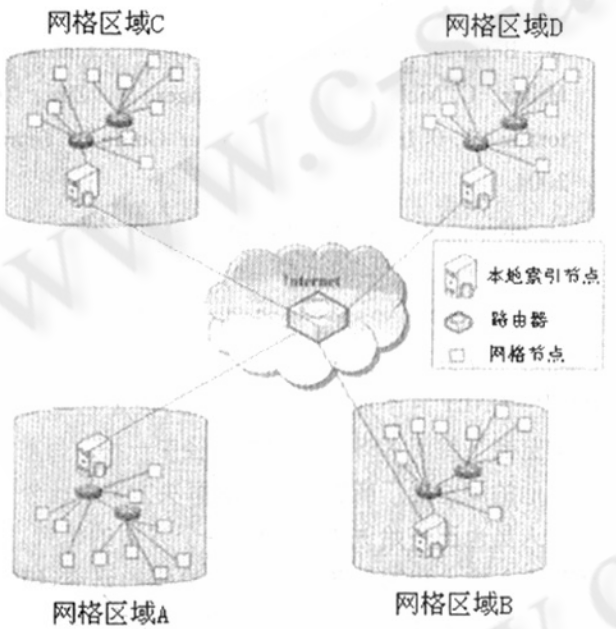

图 2　模拟环境拓扑图

3.3 一致性冲突的解决方法

副本一致性的冲突是指多个用户在不同的网格区域中同时更新同一个文件的主副本而造成的冲突。例如:用户 A 与用户 B 在不同的网格区域中同时更新 FileA 文件的主副本,则会造成 FileA 文件的冲突。解决此问题的传统方式是锁机制,但是分布锁机制在执行用户写操作时,需要封锁所有副本,直到所有副本都达到一致状态后,才可以对其他用户请求进行响应。由于数据网格系统基于广域网环境,且请求数量和副本数量都很大,通过分布锁机制保证数据网格操作的一致性,实现效率很低,将会在很大程度上增大数据网格系统的数据响应时间。

本文采用冲突避让策略解决冲突,如前所述,每个逻辑文件可以选择某个全局索引节点来存放自己的所有的副本信息。因此,在更新全局索引节点的时间戳属性时,如果检测到冲突,则冲突的更新节点撤销本次更新,延迟一定的时间后再执行更新。

采用冲突避让策略,则比锁机制更灵活,而且可以有效减少响应时间。

4 仿真及性能分析

本文采用 OptorSim 来模拟所提出的算法。OptorSim 是由欧洲数据网格项目开发的工具。它是一个基于 EU DataGrid 体系结构的网格仿真工具,主要用于对大规模广域分布式数据网格中的各种数据复制算法进行评估。OptorSim 假定每个网格节点可同时包括计算单元和存储单元,分别提供计算和数据存储服务。各个节点通过网络链路连接,每条链路有相应的带宽。计算单元在处理作业时需要使用存储单元中的数据文件,作业到计算单元的调度由资源代理(Resource Broker)负责。数据文件副本的选择、创建和删除由复制管理器(Replica Manager)中的复制优化服务(Replica Optimisation Service)负责。资源代理通过对各个候选调度方案的性能进行评估来决定如何调度。更多关于 OptorSim 的信息请查看文献[6,7]。

4.1 模拟环境

我们的模拟环境如图 2 所示。环境共包括四个网格区域,每个网格区域内有 11 个网格节点,其中一个节点为本地索引节点。网格区域之间通过 Internet 连接,区域内部的带宽为 1000Mb/秒,区域之间的带宽为 500Mb/秒。

表 1 的数据是本模拟实验中的配置参数。我们每个实验提交 200 个作业,作业间隔是两个作业提交的时间间隔,最大 CE 队列数目是所接受的作业的等待的最大容量,文件处理时间是指文件访问及处理所需要的时间。

4.2 性能分析

本文算法与其他两种算法的模拟实验结果如图 3 所示,从实验可知,本文提出的算法要比其他的两种算法更适合于网格环境中。这主要是因为本文算法减少

了网络的延迟时间,以及文件访问速度等。

表 1　模拟实验参数

参数	值
作业数目	200
作业间隔(ms)	20000
最大 CE 队列长度	100
文件处理时间(ms)	100000
实验数目	50
单文件大小(MB)	500
副本修改的数目	50

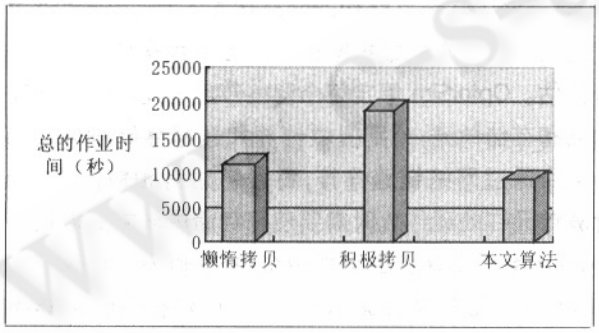

图 3　模拟环境下三种算法的作业总时间

5　总结及展望

本文提出了一种基于数据网格环境的副本一致性模型及算法,并在 OptorSim 的模拟环境下与其他两种算法进行比较,模拟结果表明本文的算法要明显的比其他算法更适合于网格环境。下一步的工作主要集中在副本定位的准确度及响应时间,还有数据传输的速度等问题,并将本算法应用于实际的网格环境下。

参考文献

1　Foster I , Kesselman C. The Grid 2 : blueprint for a new computing infrastructure [M] . [s. l.] : Morgan Kaufmann , 2004.

2　Sang - Min Park, Jai - Hoon Kim and Young - Bae Ko," Dynamic Grid Replication Strategy based on Internet Hierarchy", The second International Workshop on Grid and cooperative Computing (GCC2003), Shanghai, China, December 2003, pp. 838 - 846.

3　Andrea Domenici, Flavia Donno, Gianni Pucciani, Heinz Stockinger, Kurt Stockinger, "Replica consistency in a Data Grid", IX International Workshop on Advanced Computing and Analysis Techniques in Physics Research (ACAT03), Tsukuba, Japan, December 2003.

4　The EU Data Grid Project, http://www.eu - datagrid.org/.

5　Yuzhong Sun and Zhiwei Xu, "Grid Replication Coherence Protocol", The 18th International Parallel and Distributed Processing Symposium, Santa Fe, USA, April 2004, pp. 232 - 239.

6　W. Bell, D. Cameron, R. Carvajal - Schiaffino, P. Millar, C. Nicholson, K. Stockinger, F. Zini, "OptorSim v1. 0 Installation and User Guide", February 2004.

7　OptorSim, http://edg - wp2. web. cern. ch/.