文章编号:1005-0523(2000)01-0051-05

# 具有副本透明性的分布式文件系统模型的讨论

陈晓宇,　苏中义

(上海交通大学信息与控制工程系,上海　201100)

**摘要**:在所有典型的分布式文件系统中,很多时候一个文件都有备份,也可称为副本,副本(冗余)是分布式系统的一大特色,一些问题,诸如文件内容的一致性,副本如何管理等都是需要解决的.本文提出一种完全独特的方法,一种基于不同的文件视图的方法,同时操作系统透明地管理文件副本,并且实现文件内容一致性,还兼顾了效率问题.

**关　键　词**:副本;原件;分布式文件系统;文件服务;文件操作方式;文件内容一致性

**中图分类号**:TP316.4　　**文献标识码**:A

## 1　引　言

讨论是对于以下典型的分布式系统来进行的:系统中有 N 台独立运行的计算机,计算机之间的关系是平等的,即不存在固定的客户机/服务器,也不存在主从之分.但是由于分布式操作系统和网络的作用,这些计算机才属于一个分布式系统,在系统中,每台计算机都有一个在全系统范围内唯一的地址.一台计算机可以通过地址向别的计算机申请文件服务.在这样一个分布式文件系统中,有下列一些主要问题需要解决:

1) 为了提高可靠性,需要实行同一文件的多机存放,称为备份,也称副本.如何实现?

2) 对用户来说,最好是系统自动找到相应文件的副本进行访问,用户不必指出副本的具体位置,用户也不必管理数目众多的文件副本,这称为副本透明性.如何实现?

3) 如何透明地实现各个副本内容的一致性?

4) 用户通过何种手段控制文件,系统如何操作文件?

分布式文件系统中,文件备份是一个很重要的内容,因为这可以提高系统的可靠性.一个文件损坏了,可以在其它计算机上取到它的副本.而现在的一些分布式系统对文件副本的管理不外乎是两种方式:集中式和分散式.集中式管理有实现简单的优点,但如果管理机崩溃则整个系统瘫痪.另外,管理机也可能成为整个系统的瓶颈.而分散式管理实现复杂,不容易保持文件内容的一致性.现代操作系统大多是基于微核结构的,如果在一个已有的分布式文件系统中在基于微核的基础上加入若干操作系统的功能,则可以扩充系统的功能且具有向下兼容的优点.本文将对"副本"这个看似简单的概念作出重新定义,在这个定义的基础上提出一种区别对待文件的方法,以便透明地管理文件的副本,这个方法具有集中+分散的特点,具有两者优点.一般地,我们假设文件在各台计算机上以目录结构存放,每台计算机以地址+盘符

**收稿日期**:1999-12-23;**修订日期**:2000-02-18



**作者简介**:陈晓宇(1975-),上海人,上海交通大学在读硕士研究生.

+目录名+文件名来远程访问别的计算机上的文件,其中,目录名可能是多级的.我们不妨把机器地址+盘符+目录名称为路径名.因为副本也是一种文件,所以首先需要对副本作出定义.

## 1 副本的说明

首先要澄清一个概念:即什么是副本.一个普通用户把一个文件简单地拷贝到本地机上使用,使用之后便删除了,或者丢弃不管.这个"拷贝"也能称为原文件的一个副本吗? 副本应该是文件的可靠备份.这里我们要把普通用户的简单的复制行为和系统或系统管理员的复制行为区分开.我们把系统自动对文件的复制和系统管理员对文件的复制行为统称为系统行为.规定只有系统管理员有权限使用特殊的副本创建命令在另外一些机器上创建某个文件的副本.使用普通的复制命令只能简单地复制一个文件,不被系统看成是原文件的副本,而看成一个新的文件.有时系统为了平衡访问负载,自动地在别处建立相同的文件,把对原文件的访问分散到各处以消除瓶颈,这些被系统行为自动建立的相同文件是原文件的副本.这个工作也可以由系统管理员手工来做.所以在本文的定义中,由系统行为复制出来的文件的拷贝才称为副本.本文的系统中,文件分为两种类型:原文件和副本文件.原文件就是通常意义上的文件,而副本文件是系统中特有的.每台计算机上有一块特定的外存空间用于存放别的计算机上的文件的副本,这块空间称为副本空间.只有系统或系统管理员意识到副本空间的存在,普通用户是"看"不到这个"空间"的,换句话说,对于系统或系统管理员和普通用户来说,他们看到不同的文件视图.普通用户只能看到通常意义上的"原文件"视图,不妨称为原文件空间,而系统管理员则两个文件视图都能看到.一个普通意义上的文件称为原文件( 或称原件) ,它可以在其它若干台计算机上备有副本,也可以没有任何副本,为了方便,这时我们仍然称该文件为原文件.对于一个"原文件"来说,它除了一些通常意义上的文件信息( 诸如文件大小,存取权限,在该机上的位置等) 以外,还有一个专门的副本信息段指出该文件是否有副本,如果有,它的所有副本的位置,操作系统可以通过访问文件原件的这个信息段了解该文件的副本所在地.当然,这个信息段是由系统管理的,对于普通用户来说也是不可见的.对于副本空间中的副本文件来说,每个副本文件复制一份原文件除了副本信息段以外的文件信息,还要记录下该副本文件的原文件所在地.普通用户只能管理原件,操作系统在后台透明地管理副本.系统管理员通过操作系统的帮助可以管理两个文件空间.

## 2 文件操作方式

既然对于系统管理员和普通用户来说系统具有不同的文件视图,当然两者的操作方式也会不同.系统管理员具有普通用户的一切操作方式,另外还可以管理副本.如果用户要访问的文件位于其它机器上,那么该机器的操作系统通过网络找到另外的机器上的文件.这里就涉及到如何找到另外的机器的问题.这本来是个网络中的问题.网络中一台主机找到另一台主机称为路由.在分布式系统中,各台计算机是由网络连接起来的,所以如何找到另外的计算机也可以看成是一个路由问题.在计算机网络中,有专门的路由算法.在参考中,提出过一种距离

－向量路由算法( Distance－Vector Routing) 19它可以使主机找总是沿着较好的路径找到另外的主机,以减少响应时间19具体路由算法见参考文献[1]19如果用户要操作一个文件,系统自动识别是何种文件操作,如果操作不会改变文件内容( 只读) ,系统自动地给文件加上一个 R 锁,如果操作会改变文件的内容,系统给文件加上 X 锁19如果一个文件没有被加锁,那么对它加 R 锁和 X 锁都能成功19如果一个文件已经加了 R 锁,那么可以再对它加 R 锁( 重复读) ,但不能对它加 X 锁19如果一个文件被加了 X 锁,那么以后任何对它加锁的请求都被拒绝19这样做的目的是为了不使一个文件同时被多个用户读写,但可以同时被多个用户读,以保证文件数据的一致性19.当然,在正式操作一个文件之前还要进行权限检查等等常规工作,以后,除非特别说明,我们都是假定已经由操作系统做了这些工作19.下面我们对一些常用的文件操作命令作讨论19.

1) 普通用户的文件操作方式系统采用用户提出要求,获得文件服务的模式19.普通用户使用路径名＋文件名的方式来远程访问文件19.(当然,如果文件位于本地机上,路径名中的机器地址可以省略) 普通用户必须首先联系文件原件所在地19.( 普通用户也只能这样做,因为副本位置对他们来说是透明的) 不妨把提出文件服务要求的计算机称为客户机,把提供文件服务的计算机称为服务机19我们这种访问方式并不要求一定要把文件下载到本地机上使用,这可以使分布式系统中的用户都可以使用同一种方式操作文件( 提出要求,获得服务) 19.

**a**. Read 命令

客户机提出文件访问要求后,服务机总是先根据客户机的要求试探是否能给该文件原件加上 S 锁,如果不能加上 S 锁,则返回出错信息让客户机自己处理19.否则服务机返回确认信息和这个文件所有副本的位置( 如果该文件有副本的话) ,然后给该文件原件加上 S 锁19客户机操作系统得到该文件所有副本位置后,同原件位置一起,根据路由算法选择一个对该客户机来说的最佳位置,然后由本机操作系统向该位置所在地发出文件操作请求19.以后,客户机对文件的任何操作都和该台计算机联系19如果客户机操作系统选择了原服务机,那么文件操作仍然和文件原件所在的计算机联系19如果读操作在一台副本计算机上进行,客户机在访问文件结束后副本所在的计算机通知原件所在计算机解除一个 S 锁19.( 注意:只是一个 S 锁)

**b**. Write 命令

对于写文件的命令,获得文件服务的过程和读命令相似,只不过文件原件被加上了 X 锁19.客户机在访问文件结束后,很可能已经改变了文件( 副本) 的内容,这时,先前被进行文件访问的计算机如果不是文件原件所在的计算机,那么它要与文件原件所在计算机联系,把整个文件副本传送过去代替原件,当然,在传送过程中,要访问该文件的原件也是不容许的19在传送结束后,再由文件原件所在计算机通知除了上述这个副本所在计算机之外的所有该文件的其它副本所在的计算机作废( 删除) 它们机器上的副本19.然后原件所在的计算机考虑是重新发送文件还是仅保留一个有效副本,发送文件可以在计算机空闲时进行,也可以立刻进行,取决于具体的实现,当然,要对原件的副本信息段作相应修改19.

**c**. Delete 命令

如果客户机要删除一个文件,则由文件原件所在地通知所有的副本所在地删除副本,然后再删除原件,再返回确认消息给客户机19.这里可以作一个优化处理19.即服务机在收到命令后,立即给客户机发回确认信息以便让客户机程序能继续运行下去,然后再与各个文件副本所在地联系以协同一致地删除文件原件和所有副本19.

**d**. Create 命令

这个命令很简单，只要建立一个文件原件即可19.以后如果要建立一些副本，那也是系统或系统管理员的事19.

**e**. Move 命令

如果用户要把一个文件从一台计算机移到另一台计算机上，则先由文件原件所在地把文件按要求移出去，再把该文件的所有副本的位置也发送给新的计算机，在给客户发回确认信息，再通知所有副本所在地改变原件位置，最后，新的计算机建立文件原件和副本信息段19.

其它一些文件操作命令也可类似地讨论19.

2) 系统管理员的操作方式

访问文件原件的方式与普通用户完全相同19.所不同的是，系统管理员可以对副本进行操作19.

**a**. 增加副本

在指出了原件和目的计算机后，系统在目的计算机上的副本空间内加上该文件的一个副本，该文件原件的副本信息段内添加上该目的计算机的位置19.

**b**. 删除副本

由副本所在的计算机与文件原件所在的计算机联系，在删除了副本后，对原件的副本信息段的与之相关的内容作相应删除19.

**c**. 移动副本

可以完全类似地讨论19要注意的是，副本只能在两台计算机的副本空间之间移动19同时需要文件原件所在地的副本信息段作些修改19.

**d**. 查看副本信息

系统管理员可以通过原件的副本信息段的内容来察看副本，也可以通过任意一个副本来找寻原件19换句话说，系统管理员可以象管理普通文件那样管理副本19系统可以使用特定的模型来对性能进行评估，可以在此基础上自动地建立一些文件副本以分散流量或提高可靠性19.

## 3 一些说明

1) 本文讨论的文件模型属于原件＋副本的创新模式，文件访问类似于客户机/服务器模式，副本管理实现透明性，系统把文件和它们的副本区别对待，文件访问采用集中＋分布的方式19.这是本文方法的最大的特点19.可以在操作系统上增加该功能，符合操作系统设计的"可扩充性原则"，并且实现了文件内容的一致性19.

2) 如果用户申请文件访问失败而需等待的话，有可能造成死锁19.可以采取一般操作系统中普遍采用的方法来解决，例如申请不成功就剥夺该用户进程，一次只能申请一个文件等19.

3) 本文的方法有一定的容错性19.例如，一台计算机突然崩溃了，所有文件全部损坏，在它恢复运行后，可以向系统发送一个广播信息，信息中包括该机的地址，以便向其它计算机查询，可以恢复以前该机在其它计算机上备有副本的所有文件原件，还可以恢复副本19.

4) 系统中一台计算机可以同时为客户机和服务机，换句话说，它可以同时获得文件服务和为其它计算机提供文件服务19.

5) 本文中的方法使得用户不管在哪里都可以用同一种方式来访问文件(向原件申请服务),系统透明地决定如何给予和给予何种服务,符合分布式操作系统的要求[9].

## [ 参 考 文 献 ]

[1] Andrew S Tanenbaun, Computer Networks·, Third Edition, Prentice Hall·
[2] Andrew S Tanenbaun, Distributed Operating Systems·, Prentice Hall·
[4] 萨师煊,王珊[9].数据库系统概论[M][9].第 2 版,北京:高等教育出版社[9].
[5] 鞠九滨[9].分布计算系统[M][9].北京:高等教育出版社[9].
[6] 李学干,苏东庄[9].计算机系统结构(第 2 版)[M][9].西安:西安电子科技大学出版社[9].
[7] [美]Kai Hwang, 高等计算机系统结构[M][9].北京: 清华大学出版社[9].

# A Discussion about the Transparent Management of Backups in Distributed File Systems

**CHENG Xiao-yu, SHU Zhong-yi**

(Dept·of information and Engineering, Shanghai Jiaotong University, Shanghai 201100 China)

**Abstract:** A solution to the problem that how backups are implemented will be proposed in this essay· How to keep file consistency will be also discussed·

**Key words:** Backup, Original file; File service; File consistency; File operation