

## 一种适用于并行文件系统的高可用机制

张 虎<sup>1</sup>, 伍卫国<sup>1</sup>, 董小社<sup>1</sup>, 钱德沛<sup>1,2</sup>

(1 .西安交通大学计算机科学与技术系, 710049, 西安; 2 .北京航空航天大学计算机学院, 100083, 北京)

**摘要:** 针对并行文件系统文件数据分布存储的特性所带来的系统可靠性和可用性降低的问题, 提出了一种提高并行文件系统可靠性与可用性的机制. 该机制采用数据镜像技术, 应用逻辑镜像环来指定存储节点间的镜像关系, 使得一个存储节点既存储自身的数据, 也作为其他节点的数据备份节点. 该机制还可以通过调整镜像深度, 以满足用户对文件系统不同级别的可靠性和可用性要求. 同时, 建立了马尔可夫链模型, 以评估所提机制的可靠性和可用性. 数学求解表明, 在镜像深度为 2 时, 应用该机制的并行文件系统发生数据丢失的平均时间大大提高, 约为原系统的 32 倍.

**关键词:** 并行文件系统; 逻辑镜像环; 镜像深度; 马尔可夫链模型

**中图分类号:** TP302.8 **文献标识码:** A **文章编号:** 0253 987X(2005)10 1085 03

### High Availability Mechanism for Parallel File System

Zhang Hu<sup>1</sup>, Wu Weiguó<sup>1</sup>, Dong Xiaoshe<sup>1</sup>, Qian Depei<sup>1,2</sup>

(1 .Department of Computer Science and Technology, Xi an Jiaotong University, Xi an 710049, China;

2 .School of Computer Science and Technology, Beihang University, Beijing 100083, China)

**Abstract:** Parallel file system achieves a high I/O throughput by dividing a file into multiple blocks and storing them on multiple I/O nodes. However, the reliability and availability of the parallel file system is sacrificed for distributing file data over multi I/O nodes. A new mechanism named logic mirror ring was developed to improve the reliability and availability of the parallel file system. A logic mirror ring was built over all I/O nodes to indicate the mirror relationship among the nodes, i.e., each node maintained both its own data and the mirror data of other nodes. Moreover, the mirror depth can be adjusted to different levels based on the requirements of the reliability and availability. The effects of logic mirror ring on the reliability and availability of the parallel file system was evaluated by a Markov chain model. The mathematic solution shows that when the mirror depth is 2, applying the proposed mechanism to parallel file system, the average time of losing data is about 32 times of the original system.

**Keywords:** parallel file system; logic mirror ring; mirror depth; Markov chain model

并行文件系统是将文件数据分条或者分块地存储在多个存储节点上, 以获得并发的文件访问, 从而提供良好的聚合 I/O 性能. 但是, 基于文件分解的存储策略在提供良好 I/O 性能的同时也使得系统的容错能力有所降低, 而且系统的可用性会随着规模的增大而急剧降低. 针对上述问题, IBM 的 GPFS<sup>[1]</sup> 系统通过提供块设备级的数据镜像和文件

层的拷贝来保证文件系统的可靠性, 但需要有专用的存储区域网络和双通道 RAID 控制器等硬件设备来支持. CEFT-PVFS<sup>[2]</sup> 系统则不需额外的硬件设备, 它由 2 套完全对称的 PVFS<sup>[3]</sup> 系统组成, 通过数据的实时同步互为镜像, 从而避免了系统内的单点故障, 但这种方式改变了原有并行文件系统的物理拓扑, 会对原系统的并发性产生一定的影响. 为此,

本文提出了一种在存储节点间建立逻辑镜像环来提高系统可靠性、可用性的机制,不需要额外的硬件设备,也不会改变原有文件系统的拓扑结构。

## 1 基于逻辑镜像环的高可用机制

### 1.1 并行文件系统模型

通过考察几种典型并行文件系统,可以得出本文研究所基于的并行文件系统的拓扑模型,该模型具有以下特征和规定。

(1)并行文件系统中存在多个存储服务器,每个文件的数据被分解存储在多个存储服务器上,不同数据服务器之间的数据不重叠。

(2)并行文件系统的客户端与存储节点通过互连网络连接,每个客户端可以访问任意一个存储节点,同时每个数据服务器均可以直接响应来自客户端的数据访问请求。

### 1.2 逻辑镜像环

定义1 逻辑镜像环是用来表征并行文件系统数据服务器之间数据镜像方向的虚拟有向环,环中的每个节点代表一个数据服务器。对于一个数据服务器数目为  $n$  的并行文件系统,可以标识为  $S_0, S_1, \dots, S_{n-1}$ 。所谓逻辑镜像环是将节点标识组织成一个环,并规定环的方向,该环可以由原系统中的所有数据服务器以任何顺序组成。为了便于描述,可采用同标识顺序一致的逻辑环  $R_0 = \{S_0, S_1, \dots, S_{n-1}\}$ 。

定义2 镜像深度是一个数值  $m$ ,它表示文件系统中每一份数据的拷贝数目。在逻辑镜像环  $R_0 = \{S_0, S_1, \dots, S_{n-1}\}$  内,满足  $0 < m \leq n$ 。如果  $m$  为 2,表明每份文件数据都会存储在 2 个不同的数据服务器节点上。

定义3 邻接距离的概念因逻辑镜像环的存在而存在。在一个逻辑镜像环中存在节点  $S_k$  和  $S_l$ ,如果从  $S_k$  到  $S_l$  之间存在  $i-1$  个节点,以逻辑镜像环的方向为正向,则  $S_l$  与  $S_k$  的邻接距离为  $i$ ,  $S_k$  与  $S_l$  的邻接距离为  $-i$ 。邻接距离可以为负,表示计算距离的方向与环的方向相反。

对于一个镜像深度为  $m$  的逻辑镜像环  $R_0$ ,环中任意一个节点  $S_i$  上的原始数据均会被镜像到与其邻接距离小于  $m$  的节点上。例如,某个逻辑镜像环的镜像深度为 3,是指所有节点上的原有数据被复制到与其邻接距离为 1 和 2 的节点上。也就是说,对于  $R_0$  中的任意节点  $S_k$ ,节点  $S_{(k+1) \bmod n}$  和  $S_{(k+2) \bmod n}$  上存在节点  $S_k$  的数据备份。

定义4 邻接复制是逻辑镜像环中原始数据节点同

其多个数据备份节点之间的数据同步操作。在文件系统的运行期间,对于一个数据的写操作所产生的数据变化都要实时地通过邻接复制操作同步到其备份数据所在的节点上。

### 1.3 容错原理

在镜像深度为  $m$  的逻辑镜像环  $R_0$  中,如果所有节点处于正常状态,那么每个节点只处理面向原始数据的访问请求,并通过邻接复制来保证所有备份节点同原始数据节点的数据同步。当某个数据节点发生故障致使不能被访问时,假设该节点为  $S_k$ ,那么与  $S_k$  邻接距离为 1 的邻接节点  $S_{k+1}$  会首先接管  $S_k$  的原始数据服务,其他  $S_k$  的备份数据节点依然通过邻接复制操作完成数据同步。这时,在  $S_{k+1}$  上要额外处理本该发向  $S_k$  的所有访问请求。因此,对于镜像深度为  $m$  的逻辑镜像环,在最大的可容忍故障情况下,必定存在一个节点,它会同时接收原来连续  $m$  个节点的数据访问请求。也就是说,如果  $S_k \sim S_{k+m-2}$  都发生故障,则  $S_{k+m-1}$  上仍会响应  $S_k, S_{k+1}, \dots, S_{k+m-1}$  的所有访问请求(为了方便讨论,设定  $k+m-1 < n$ )。从上述可见,当一个逻辑镜像环的镜像深度为  $m$  时,可以容忍环中连续  $m-1$  个节点失效,由此可以得到如下定理。

定理 对于一个有  $n$  个存储节点的并行文件系统,在应用了镜像深度为  $m$  的逻辑镜像环后,最多能容忍连续  $m-1$  个节点失效。

## 2 可靠性与可用性分析

### 2.1 可靠性分析

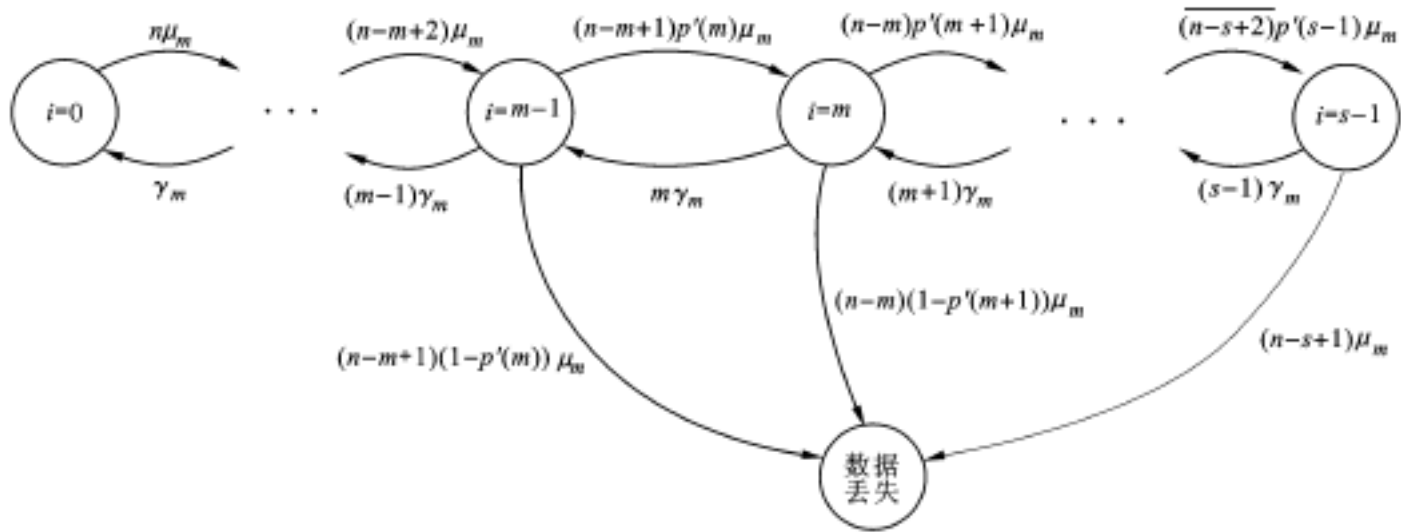
系统可靠性可以用系统平均数据丢失时间  $L$  来度量,同时应用马尔可夫链(Markov Chain, MC)模型来计算系统的  $L$ 。为了便于分析,先作如下假设。

(1)系统内所有节点的状态改变是瞬间完成的。这样考虑可以避免 MC 模型中出现太多对结果影响不大的状态。

(2)不同节点的失效概率和故障修复概率是相互独立的,不会因为某个节点的失效而导致与其相关的节点的失效概率有所增加。

用于系统可靠性分析的 MC 模型如图 1 所示<sup>[2,4,5]</sup>。

在图 1 中,以  $i$  标识的状态代表系统中存在  $i$  个故障节点,但该状态仍然可用。当镜像深度为  $m$  时,只要系统中不存在连续  $m$  个故障节点,则整个系统仍是可用的,所以最大的状态标号  $s-1$  符合



$n$ :逻辑镜像环中节点数目; $m$ :环的镜像深度; $\mu_m$ 、 $\gamma_m$ :分别代表镜像深度为  $m$  时单个节点的失效概率和修复概率

图 1 马尔可夫链模型

$$s - 1 = n - \lceil n/m \rceil$$

$p(i+1)$  为系统在  $i$  状态时, 又有一个节点发生故障, 但系统仍然可用. 依据文献[4]的求解方法, 可以得到系统的平均数据丢失时间  $L$ .

图 2a 给出了系统应用不同的  $n$  和  $m$ , 其可靠性与原有系统 ( $m=1$  时) 的比较结果, 从中可以看出, 该机制可以为系统提供良好的可靠性保证. 在节点数目为 8、镜像深度为 2 时, 从计算的数值可以看出系统的平均数据丢失时间约为原有系统的 3 200%.

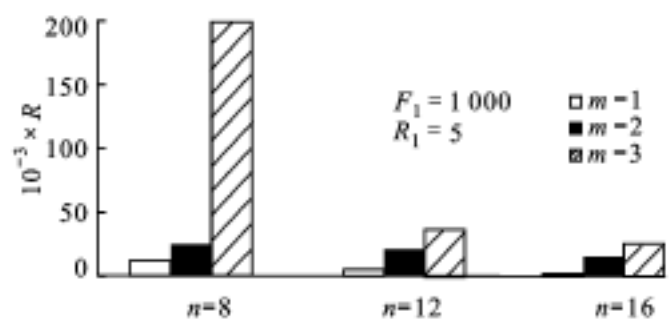
### 2.2 系统可用性

对于应用逻辑镜像环的并行文件系统, 可以认为整个系统的平均故障发生时间  $F$  即为系统的平均数据丢失时间  $L$ , 因为在正常情况下, 只有在出现数据丢失时系统才会停止服务. 因此, 可以认定整个系统的平均故障恢复时间  $R$  为单个节点故障恢复时间  $R_m$  的  $m$  倍, 因为系统停止服务时, 必定在逻辑镜像环中存在  $m$  个连续的失效节点, 这  $m$  个连续节点的恢复必须串行进行, 其他故障节点的恢复则可以与这  $m$  个节点的恢复同步进行. 当镜像深度为  $m$  时, 单个故障节点恢复要与  $m$  个数据备份同步进行, 可以从概率上认为  $R_m$  为镜像深度为 1 的单个节点故障恢复时间的  $m$  倍, 此时根据可用性的评价公式<sup>[6]</sup> 计算系统可用性得

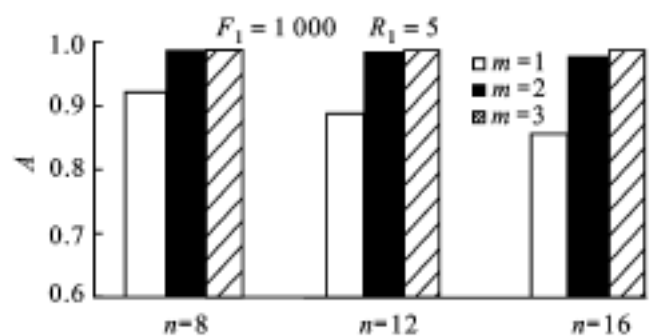
$$A = \frac{F}{F + R} = \frac{L}{L + m^2 R_1} \quad (1)$$

图 2b 给出了系统在采用不同的存储节点数量  $n$  和镜像深度  $m$  时, 系统的可用性比较.

在设计一个并行文件系统时, 对于给定的存储节点数目  $n$  和单个节点的  $R_1$ 、 $F_1$ , 以及对系统可用性的要求  $A$ , 可以根据式(1)获得对平均数据丢失时间  $L$  的要求, 从而根据  $L$  的求解过程<sup>[4]</sup>, 应用数值求解的方式确定合适的  $m$ , 使得系统满足可用性的



(a) 可靠性



(b) 可用性

图 2 系统可靠性与可用性

要求.

### 3 结 论

本文针对并行文件系统存在的可靠性问题, 提出了在并行文件系统的存储节点间建立逻辑镜像环来增强系统可靠性和可用性的机制. 同时, 建立了该机制的可靠性模型, 并应用不同的节点数目  $n$  和镜像深度  $m$  对该机制进行分析和求解. 分析结果表明, 该机制可大大提高并行文件系统的可靠性和可用性. 今后还需研究的内容包括: 在对系统的单个节点失效概率和恢复概率的假设中, 要考虑故障分布和故障切换对失效概率和恢复概率的影响; 存储服务器的状态监测和控制等.

(下转第 1096 页)