第 34 卷 第 11 期 计 算 机 工 程 2008 年 6 月
Vol.34 No.11 Computer Engineering June 2008
·博士论文· 文章编号：1000—3428(2008)11—0007—03 文献标识码：A 中图分类号：TP319

蓝鲸集群文件系统中资源交互一致性协议

张军伟[1,2]，贾瑞勇[1]，贾亚军[3]，张建刚[1]，许 鲁[1]
(1. 中国科学院计算技术研究所，北京 100080；2. 中国科学院研究生院，北京 100039；
3. 新疆油田公司勘探开发研究院地球物理研究所，乌鲁木齐 830013)

摘 要：在蓝鲸集群文件系统中，分布式资源交互在系统异常的情况下会出现资源状态不一致的情况，为解决这一问题，该文提出分布式资源交互一致性协议 S2PC-RT。S2PC-RT 引入资源交互序号保证一致性，通过增加资源申请空间减少消息等待。证明协议的正确性，并在蓝鲸集群文件系统中实现了协议。测试结果表明，S2PC-RT 能够保证资源的一致性，有效提高分布式资源的交互性能。
关键词：蓝鲸集群文件系统；集群文件系统；资源交互；一致性

Consistency Protocol of Resource Transference in Blue Whale Cluster File System

ZHANG Jun-wei[1,2], JIA Rui-yong[1], JIA Ya-jun[3], ZHANG Jian-gang[1], XU Lu[1]
(1. Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080; 2. Graduate University of Chinese Academy of Sciences, Beijing 100039; 3. Geophysical Research Institute, Exploration and Development Institute of Xinjiang Oil, Urumqi 830013)

【Abstract】 In Blue Whale Cluster File System(BWFS), inconsistency of resource state may exist during distributed resource transference under system exception. To solve that problem, the consistency protocol S2PC-RT is presented. Resource transference sequence no. is introduced to protect consistency and space for resource applying added to reduce message wait. The correctness of S2PC-RT is proven. Implementation and test are done in BWFS, and test results show that S2PC-RT can guarantee the resource consistency and enhance the performance of distributed resource transference operations.
【Key words】 Blue Whale Cluster File System(BWFS); cluster file system; resource transference; consistency

1 概述

随着信息技术的发展，科学计算及信息处理等应用要求存储系统具有更高性能和更大容量。蓝鲸集群文件系统(Blue Whale File System, BWFS)[1]采用元数据与数据分离机制，元数据请求通过元数据控制协议由元数据服务器集群处理，数据请求则直接由多个数据服务器并行处理，可以很好地满足对高性能和大容量的要求。在 BWFS 中，元数据服务器(Metadata Server, MS)集群使用系统中的资源，包括 inode、元数据块以及数据块等资源，采用集中资源管理者(Central Resource Manager, CRM)管理系统中的所有资源。

MS 通过资源申请和回收与 CRM 完成资源交互操作。在资源申请过程中，CRM 把资源状态修改为已经占用，MS 把获取的资源状态修改为 MS 本地空闲可用；在资源回收过程中，CRM 把回收的资源状态修改为空闲，MS 把本地已回收资源清除，表示不能再使用这些资源。在 MS 与 CRM 资源交互过程中，存在如何在系统异常情况下保证资源一致性的问题。如果 MS 与 CRM 的资源状态修改不能保证原子性，则会出现资源丢失或重用，从而导致资源状态不一致。

当前，有多种机制保证分布式操作的一致性，包括 2 阶段提交协议 2PC (Two-Phase-Commit) [2]、DCFS2[3]文件系统中采用的 S2PC-MP[3] 协议等，但上述协议没有考虑分布式资源交互的特点。本文提出针对分布式资源交互操作的一致性协议 S2PC-RT(Simple 2PC for Resource Transference)，该协议引入资源交互序号保证一致性，增加资源申请空间以减少参与者的消息等待，提高了资源交互性能。

2 分布式资源交互一致性协议 S2PC-RT

分布式资源交互操作中，只要 MS 发出资源交互请求，就隐含着已经 vote yes，因此，CRM 可以根据本身的执行状态决定是 Commit 还是 Abort。MS 只有收到 CRM 的 Commit 消息后，才进行资源状态修改操作，否则 MS 端的状态不受影响。因为 MS 的资源状态修改操作简单，所以假设总是成功的，除非由于系统宕机而造成处理中断。

在资源交互操作中，交互的内容只有资源，因此，可以为每次资源的交互设置唯一的序号，通过序号来区分每次的资源交互，从而根据序号保证一致性。在申请操作中，把资源首先转移到暂存空间，以满足资源申请的重复请求；在回收操作中，对重复回收的资源只要丢弃即可。

2.1 协议数据结构

在 MS 中需要维护资源交互状态的数据结构如下：

```
struct C_state{
req_seq:    资源申请操作序号，同时表明前一个申请已经被确认；
```

基金项目：国家“973”计划基金资助项目(2004CB318205)；中科院计算所创新基金资助项目(20056420)
作者简介：张军伟(1977－)，男，博士研究生，主研方向：网络存储，集群文件系统；贾瑞勇，助理研究员、博士；贾亚军，高级工程师；张建刚，副研究员、博士；许 鲁，研究员、博士生导师
收稿日期：2007-06-10 **E-mail**：zhangjunwei@nrchpc.ac.cn

reclaim_seq：资源回收操作序号；

}

在 CRM 中需要维护的资源交互状态的数据结构如下：

struct P_state{

req_seq：　　期望接收的资源申请序号，同时表明期望接收的确认资源申请序号为 req_seq-1；

reclaim_seq：期望接收的资源回收序号；

apply_block：在资源申请过程中，把资源首先转移到该域中；

}

以上数据结构在系统格式化时都设置为 0，在每次节点重新启动时从磁盘中读取相应结构内容。

2.2　资源申请流程

资源申请具体流程如图 1 所示。

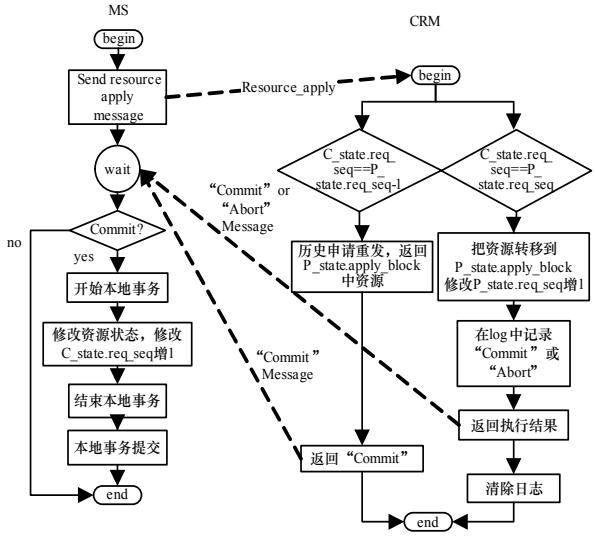

图 1　分布式资源交互一致性协议 S2PC-RT 资源申请流程

资源申请步骤如下：

(1)MS 向 CRM 发送资源申请请求 resource_apply，请求中包含 C_state.req_seq。

(2)CRM 收到 resource_apply 消息后，进行 C_state.req_seq 的检测。

1)之后的重复申请，CRM 把 P_state.apply_block 中的资源内容返回给 MS。

2)如果 C_state.req_seq 等于 P_state.req_seq，表示是新的资源申请。CRM 把资源首先转移到 P_state.apply_block 中，P_state.req_seq 增 1；如果以上操作成功，在 log 中记录“Commit”，否则记录“Abort”，并向 MS 返回相应消息，然后进行 log 记录的清除。

(3)MS 如果收到“Abort”回应信息，不做任何操作；如果收到“Commit”回应消息，首先启动本地的事务操作，根据申请到的资源内容修改本地资源状态，C_state.req_seq 增 1，事务提交后，就表示资源申请交互完毕。

2.3　资源回收流程

资源回收具体流程与资源申请流程类似，如图 2 所示，具体步骤如下：

(1)MS 发送资源回收请求 resource_reclaim，请求中包含回收序号 C_state.reclaim_seq 和回收的资源内容。

(2)CRM 收到 resource_reclaim 后，检测 C_state.reclaim_seq。

1)如果 C_state.reclaim_seq 等于 P_state.reclaim_seq-1，表示资源重新回收，CRM 只要向 MS 回应“Commit”消息，不必进行其他工作。

2)如果 C_state.reclaim_seq 等于 P_state.reclaim_seq，表示是新的资源回收操作。CRM 首先把资源进行回收，P_state.reclaim_seq 增 1，如果以上操作成功，记录“Commit”，否则记录“Abort”，并向 MS 返回相应消息，然后进行 log 记录的清除。

(3)MS 如果收到“Abort”消息，不做任何操作；如果收到“Commit”消息，则首先启动本地事务操作，修改本地的资源状态为已回收，C_state.reclaim_seq 增 1，事务提交后，资源交互完毕。

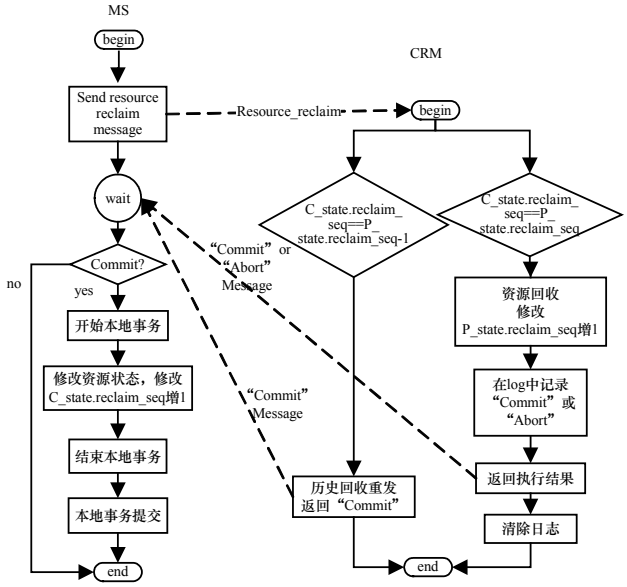

图 2　分布式资源交互一致性协议 S2PC-RT 资源回收流程

3　S2PC-RT 的恢复机制与证明

在恢复中假设故障只有节点失效和网络异常情况，并且可以在有限时间内恢复，失效节点的状态一致性由节点本身的事务机制保证，在恢复过程中允许系统再次异常。

断言 1　在 S2PC-RT 分布式资源交互一致性协议中，资源申请操作中 C_state.req_seq 只可能取值 P_state.req_seq 或者 P_state.req_seq-1。

证明　初始化时，设置为相等；在资源交互过程中，MS 只有在 CRM 增加 P_state.req_seq 并回应消息后，才进行 C_state.req_seq 的增加。如果 MS 在交互过程中失效，或者在 C_state.req_seq 修改前失效，那么在恢复后，C_state.req_seq 与 P_state.req_seq-1 相同；或者已经修改了 C_state.req_seq 失效，那么在恢复后，C_state.req_seq 与 P_state.req_seq 相同。

如果 CRM 在交互过程中失效，或者还没有修改 P_state.req_seq，那么在恢复后，C_state.req_seq 取值 P_state.req_seq；或者已经修改了 P_state.req_seq 增 1，那么在恢复后 C_state.req_seq 取值 P_state.req_seq-1。

断言 2　在 S2PC-RT 分布式资源交互一致性协议中，资源回收操作中 C_state.reclaim_seq 只能取值 P_state.reclaim_seq 或者 P_state.reclaim_seq-1。

证明过程同断言 1。

断言 3　在 S2PC-RT 分布式资源交互一致性协议中，P_state.apply_block 中的资源是对应 P_state.req_seq-1 申请序号的资源。

证明 不考虑初始化情况下，在 P_state.req_seq 资源申请过程中，把资源转移到 P_state.apply_block 操作与 P_state.req_seq++操作是一个原子操作序列，因此，在 CRM 失效恢复到一致状态后，P_state.apply_block 中的内容为 P_state.req_seq-1 申请序号所对应的申请资源。

断言 4 如果在资源交互操作中 MS 失效，那么根据操作的日志记录 MS 和 CRM 能够达到一致资源状态。

证明 如果在资源申请操作中 MS 失效，则根据断言 1，C_state.req_seq 只可能取值 P_state.req_seq 或者 P_state.req_seq-1。如果 C_state.req_seq == P_state.req_seq-1，则根据断言 3，把 P_state.apply_block 中的资源内容返回即可。如果 C_state.req_seq == P_state.req_seq 标志为新的申请，资源状态已经一致。

如果在资源回收操作中 MS 失效，根据断言 2，C_state.reclaim_seq 只可能取值 P_state.reclaim_seq 或者 P_state.reclaim_seq-1。如果 C_state.reclaim_seq == P_state.reclaim_seq-1，标志是回收操作的重发，丢弃该请求，避免了重新回收造成的资源不一致；如果 C_state.reclaim_seq == P_state.reclaim_seq，表示是新的资源申请请求，资源状态已经一致。

断言 5 如果在资源交互操作中 CRM 失效，那么根据操作的日志记录 MS 和 CRM 能够达到一致资源状态。

证明 过程同断言 4。

断言 6 如果在资源交互操作中网络异常，那么网络恢复后，系统可以达到一致状态。

证明

(1)在 MS 发送资源交互请求时网络异常，那么 CRM 没有接收到资源交互请求。当网络恢复后 MS 重新发送资源交互请求，资源处于一致状态。

(2)在 CRM 接收到资源交互请求后网络异常。CRM 按照正常的流程进行处理，但 MS 接收不到处理完毕后的结果。当网络恢复后，MS 重新发送资源交互请求。此时，如果是资源申请操作，则 C_state.req_seq == P_state. req_seq-1；如果是资源回收操作，则 C_state.reclaim_seq == P_state.reclaim_seq-1；根据断言 4 的证明，这 2 种情况下都能达到一致的状态。

(3)在 CRM 发送执行结果回应给 MS 过程中网络异常，这种情况同(2)。

因此，根据断言 4～断言 6 可以得到如下结论：采用 S2PC-RT 分布式资源交互一致性协议，在节点失效或网络异常的情况下，在节点或网络恢复后，能够恢复资源状态的一致性。

4 测试结果

S2PC-RT 协议已经在 BWFS 文件系统中得到了实现，MS 运行在内核态，本地一致性采用 journal 机制保证；CRM 运行在用户态，采用信息复制[4]的方式提供本地的一致性保证和状态恢复。在 BWFS 文件系统中，对该协议进行了正确性和性能的测试，并与 S2PC-MP 协议进行了对比。

4.1 正确性测试

对协议的正确性进行了测试，模拟各种网络异常和节点失效的情况，进行故障恢复后，测试结果表明资源处于一致状态。

4.2 性能测试

对比测试了 S2PC-MP 同 S2PC-RT 分布式资源交互操作的性能，并对消息传递次数进行统计；测试了 2 000 次连续资源交互操作的时间，得出每次交互的平均时间，测试性能结果如图 3 所示。

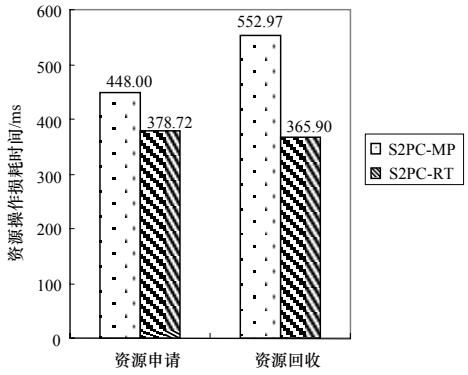

图 3 S2PC-RT 与 S2PC-MP 性能对比

消息传递次数统计结果为：S2PC-MP 为 6 000 次，S2PC-RT 为 4 000 次。

测试结果表明：S2PC-RT 降低了网络消息传递，提高了资源交互性能。

5 结束语

BWFS 采用元数据与数据分离机制满足应用日益增长的对性能和容量的要求，使用多个 MS 处理元数据请求，利用 CRM 为各个 MS 提供资源服务。为解决 MS 同 CRM 资源交互的一致性问题，引入资源交互序号，增加资源申请空间，提出了分布式资源交互一致性协议 S2PC-RT，在节点宕机或网络异常恢复后，能够保证资源的一致性，提高分布式资源交互性能。

BWFS 中采用的元数据与数据分离和集中的资源管理体系结构是新一代高性能海量集群文件系统的主要特征，目前 DCFS2, GPFS 中都采用了该结构，而资源交互一致性是这类系统需要解决的关键问题之一，因此，S2PC-RT 协议具有广泛的应用意义。

参考文献

[1] Özsu M T, Valduriez P. Principles of Distributed Database Systems[M]. 2nd ed. [S. l.]: Prentice Hall, Inc., 1999.

[2] 黄 华, 张建刚, 许 鲁. 蓝鲸分布式文件系统的分布式分层资源管理模型[J]. 计算机研究与发展, 2005, 42(6): 1034-1038.

[3] 熊 劲, 范志华, 马 捷, 等. DCFS2 的元数据一致性策略[J]. 计算机研究与发展, 2005, 42(6): 1019-1027.

[4] 黄 华. 蓝鲸分布式文件系统资源管理[D]. 北京: 中国科学院研究生院, 2005.