融合存储系统的设计与实现①

张宗平[1], 秦磊华[2], 关锦明[1]

[1](广东出入境检验检疫局信息中心, 广州 510623)
[2](华中科技大学 计算机科学与技术学院, 武汉 430074)

摘 要: 设计并实现了 IP SAN 和 FC SAN 的融合存储控制器, 支持存储协议交换和存储虚拟化功能, 实现异构应用服务器对异构存储资源的透明访问.设计并实现了融合式存储容灾原型系统. 通过对存储容灾系统原型的测试, 验证了基于融合存储控制器构建存储容灾系统的合理性与可行性.

关键词: 融合存储; 复制; 恢复

Design and Implementation of Integrated Storage System

ZHANG Zong-Ping[1], QIN Lei-Hua[2], GUAN Jin-Ming[1]

[1](Computer Information Center, Guangdong Entry/Exit Inspection and Quarantine Bureau of the P.R.China, Guangzhou 510623, China)
[2](Computer School of Huazhong University of Science and Technology, Wuhan 430074, China)

Abstract: The paper designed and implemented a storage controller which can integrate IP SAN and FC SAN and it can support the functions of storage protocol interchange and storage virtualization. We also implemented integrated storage disaster recovery prototype system based on our integrated controller, It validated the feasibility of building storage disaster recovery system based on integrated storage controller via the test for integrated storage disaster recovery prototype system.

Key words: integrated storage; replication; recovery

1 背景

FC SAN 与 IP SAN 凭借各自的技术特点, 在存储行业应用中相互取长补短, 并都得到了较为广泛的应用. 在可见的未来, 存储架构难以统一, FC SAN 和 IP SAN 还将并存[1]. 由于 IPSAN 和 FCSAN 在通信协议和存储协议等方面存在一定的差异性[2], 这些差异性会增加数据管理, 尤其是数据容灾和备份的复杂性[3]. 由此, 有必要研究融合存储系统, 以屏蔽 IP SAN 与 FC SAN 技术上的差异性, 简化数据复制与恢复操作流程, 提高数据的可用性, 确保应用系统可靠运行.

2 融合存储系统的设计与实现

2.1 融合存储体系架构

融合存储体系架构解决 IP-SAN 和 FC-SAN 两种不同类型存储设备互联互通的融合性问题, 通过 IP、FC 两种不同连接方式将本地或远程 FC 或 IP 存储资源进行映射, 并由存储服务器虚拟化后, 对应用服务器提供 IP 或 FC 的块级存储服务. 基于 ATCA 的融合存储平台由应用服务器层、融合存储服务控制层和融合存储设备层[4]. 融合存储体系架构如图 1 所示.

应用服务器为融合存储体系的最上层, 它们通过各自原有的连接方式连接到中间层的融合存储服务器, 对客户端提供对融合存储资源的透明访问.

中层为基于 ATCA 架构设计的融合存储服务控制层, 该层是实现融合存储访问的关键. 中层将底层的融合存储设备划分为逻辑单元, 并映射给上层的各类应用服务器使用, 实现对融合存储系统中不同类型存储设备的统一访问. 另外, 为适应应用服务器层种类各异的存储访问接口, 在 ATCA 存储服务器上配置 FC HBA、以太网卡等接口.

① 基金项目:国家质检总局科技计划项目(2012IK256)
收稿时间:2013-08-14;收到修改稿时间:2013-11-11

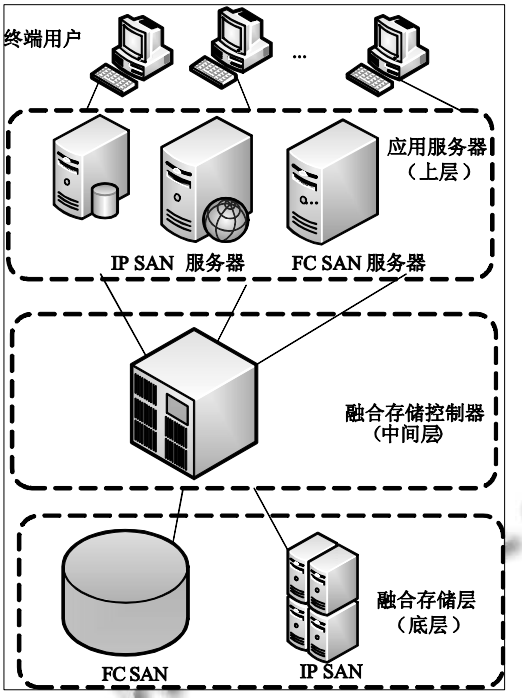

图 1 融合存储系统体系结构

底层为包含 FC SAN 和 IP SAN 的融合存储设备层，存储空间被映射到中层进行储虚拟化后，将被划分逻辑单元进行统一管理，并供应用服务器使用.

合存储系统中的各层都统一使用 SCSI 协议进行数据访问，相互间采用 FCP、iSCSI 等主流存储协议进行块级的数据处理和传输.

2.2 融合存储控制器的设计

2.2.1 融合存储控制器的硬件结构设计

融合存储控制器是融合存储体系的核心，它是应用存储服务器和融合存储系统的中间层. 为实现对不同存储服务提供统一的存储访问服务，必须要求融合存储控制器对下层的 FC SAN 和 IP SAN 表现为不同 initiator 发起端，对上层不同应用表现为不同的 target 目标端，并在融合存储控制器中实现不同存储访问发起端和目标端之间的数据交换. 基于 ATCA 平台设计的融合存储控制器的硬件结构如图 2 所示.

在具体实现上，融合存储控制器硬件开发平台基于开放标准先进电信运算架构(Advanced Telecommunications Computing Architecture，简称 AdvancedTCA，ATCA). ATCA 为 PCI Industrial Computers Manufacturing Group(PICMG)制定的规范，也称 PICMG 3.x 规范. 在具体选型上，存储控制单元采用 Radisys 公司的 ATCA-4300 计算模块，多协议交换单元选用 Radisys 公司的 ATCA-2100. 另外，通过 AMC 卡的形式为存储控制单元提供本地存储功能和额外的 FC 接口.

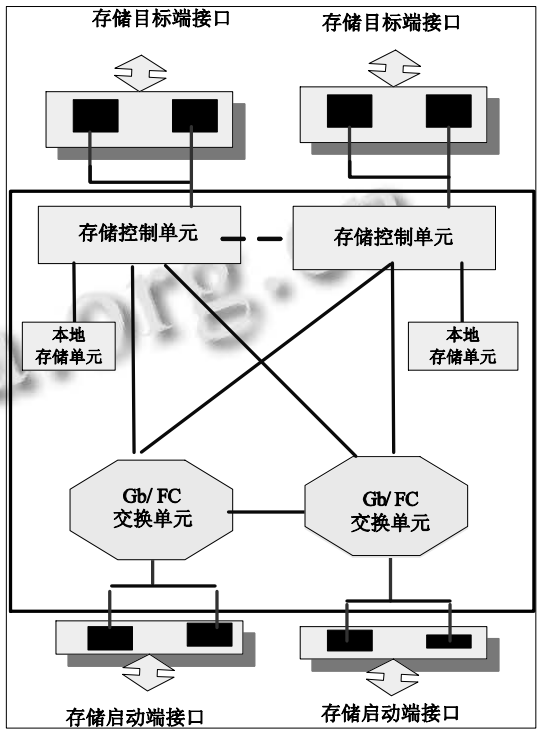

图 2 融合存储控制器的硬件结构

2.2.2 统一 SCSI 目标端模块设计

融合存储控制器通过虚拟化子系统将所有存储空间组成统一的存储卷组，并根据需求以逻辑卷的方式分配. 统一 SCSI 目标端通过逻辑卷与统一 SCSI 发起端交互，需实现与逻辑卷间的 I/O 操作. 本文利用 Linux 内核中块设备驱动框架来实现统一 SCSI 目标端与逻辑卷间的 I/O 功能[5].

Linux 中块设备 I/O 是通过统一的框架和接口实现的，基于 LVM 的逻辑卷是标准块设备，通过 Linux 内核中的块 I/O 操作是最直接自然的对基于 LVM 的逻辑卷进行操作的方式. 本文通过在统一 SCSI 目标端和基于 LVM 的逻辑卷间添加融合逻辑卷模块来实现，如图 3 所示.

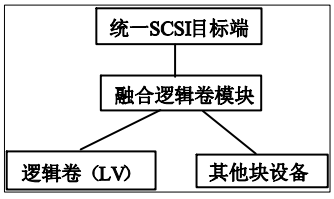

图 3 虚拟磁盘模块层次

融合逻辑卷模块封装了对基于 LVM 逻辑的 I/O 操作, 并实现基于逻辑卷的数据容灾功能(如数据复制、快照)等. 通过这样的方式, 统一 SCSI 目标端以用统一的方式处理底层存储逻辑卷, 底层逻辑卷也对上以统一的方式提供存储接口. 存储容灾功能仅在融合逻辑卷模块中实现, 从而为整个系统提供了良好的功能扩展性: 新功能的添加仅在融合逻辑卷模块中进行, 对统一 SCSI 目标端和底层逻辑卷透明, 保持了模块间接口的透明稳定性[6]. 图 4 和图 5 显示了存储服务层与存储虚拟化层之间通过存储 API 接口交互, 完成读/写操作的主要流程框架.

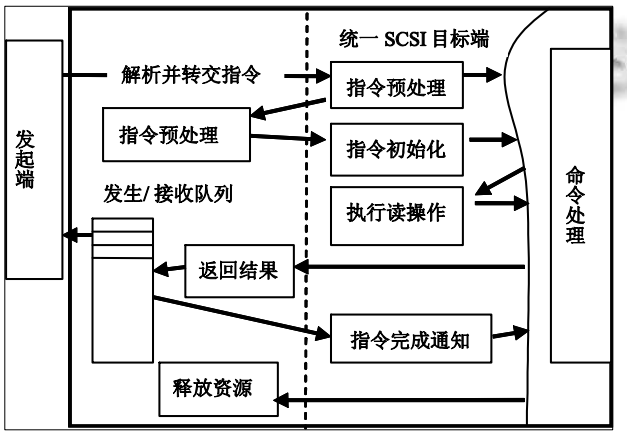

图 4　读命令流程

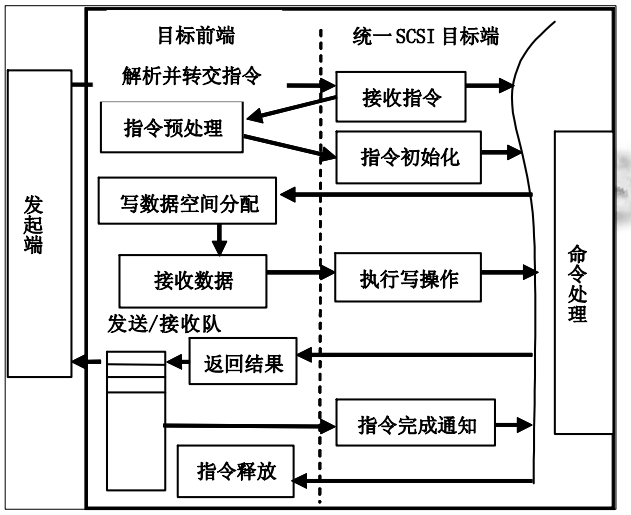

图 5　写命令流程

2.3 融合存储模式下同步复制的设计与实现

目标前端的 I/O 请求将被解析成 SCSI 指令描述块 CDB, 并根据 SCSI 指令的具体内容(具体设备、I/O 地址、数据长度、I/O 类型等)分别对本地设备和镜像设备执行 I/O 请求. 最终同步复制模块将检查本地设备的 I/O 请求是否正确执行完毕, 如果都正确执行, 则继续下一个 I/O 请求.

同步复制设计采用的先写源设备, 再写备份设备的思路. 考虑执行效率, 按串联的方式实现, 不需要单独针对备份设备把 SCSI 命令数据截取出来执行, 那样既不会消耗多余的存储空间也无需要考虑两个设备的响应的同步机制. 在源设备执行结束后, 直接把针对源设备的 SCSI 命令的逻辑设备改成目标设备, 然后发送给备份执行. 这样, 就实现了同步复制目的, 因基本没有额外的 I/O 开销, 所以 I/O 执行效率高, 所需要的工作量也是最小的. 同步复制流程如图 6.

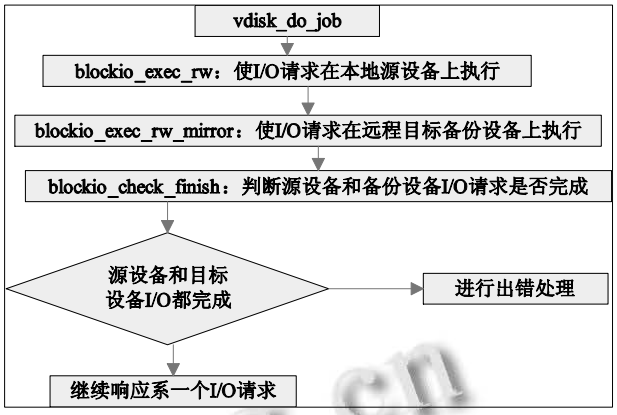

图 6　同步复制的实现流程图

首先对 I/O 请求按读写进行分类, 若是读请求, 则先对源设备和目标设备的状态进行查看, 查看是否至少有一个设备正常, 如果有, 则从正常设备上读取所需数据, 默认从源设备读取数据, 否则, 返回设备读错误.

如果是写请求, 同样先对源设备和目标设备的状态进行查看, 判断是否至少有一个正常, 如果没有, 则返回设备写错误; 如果至少有一个正常, 还要进一步判断是否两个设备都正常, 如果是, 则对两个设备进行写请求, 否则, 对正常设备进行 I/O 写请求, 并产生日志记录修改数据. 如果日志文件满, 则启动位图记录修改数据.

2.4 融合存储模式下异步复制的设计与实现

异步复制操作的主要设计思路是: 在执行本地写

操作前，首先判断专门为异步镜像操作服务的工作队列链表是否已满，如果未满，则把数据拷贝一份并添加进该链表中. 然后执行原有的本地操作，仅在第一次本地操作结束时激活异步镜像操作服务的工作队列线程，从而处理器会在某时刻调度这个线程，触发异步镜像操作；如果已满，则直接执行本地操作，等本次操作完成以后，把本次所做的操作记录下来并添加入另一个由专门线程控制的并为异步镜像操作服务的链表中. 在执行异步镜像操作的时候，首先从工作队列链表中取出数据项，然后从线程控制的链表中取出数据项进行操作，当本次异步操作完成时，需要把完成的数据项从链表中删除掉，并继续对剩余的数据项进行操作. 若两个链表中都没有数据项，则说明本次异步操作完成. 异步复制时工作队列的首次激活流程如图 7.

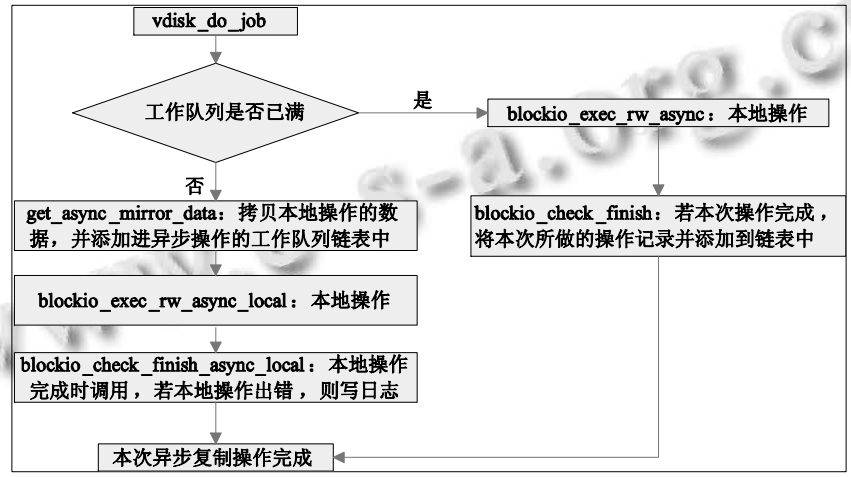

图 7 异步复制时工作队列的流程图

2.5 融合存储模式下数据恢复的设计与实现

数据恢复过程的流程图如图 8. 函数接口 scst_suspend_activity()和 scst_resume_activity()的作用分别是暂停和重启系统接受新的 I/O 请求. reset_bitmap()的作用是清空位图文件，所有的设备块标记二进制位全部为 0.

函数接口 recovery()的流程逻辑是按照发生故障时，I/O 请求记录的顺序，先从日志文件中直接恢复数据，日志文件恢复完成后从位图文件中按照设备块号从小到大的顺序构造日志来恢复数据.

如果采用镜像功能配置了多个设备，并且刚好在同一时间有多个设备出现故障，那样就会对应每一个设备生成一个位图文件；在各个设备都恢复正常时进行数据恢复，先使用日志文件恢复完后，就要恢复每个位图文件上记录的设备更新，把每个故障设备上的数据都恢复到与源设备相同的状态.

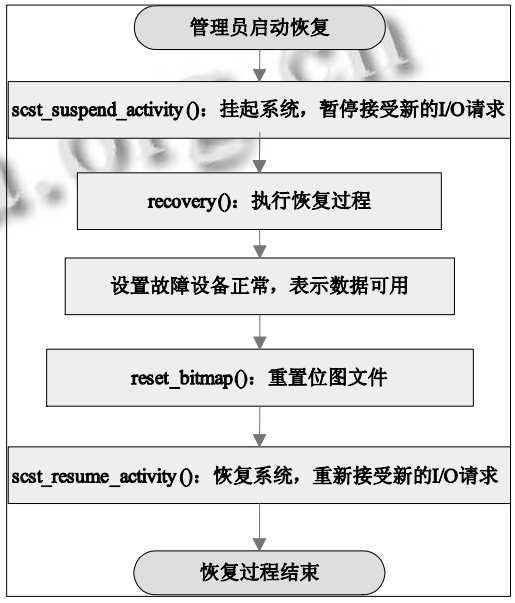

图 8 数据恢复过程流程图

3 实验效果

建立如表 1 所示的测试环境.

测试主要包括二部分，第一是测试系统融合对磁盘 I/O 性能的影响；其次，测试测试同步复制和异步复

制的性能. 测试是通过使用硬盘 IO 性能测试工具 bonnie++, 来对不同情况下的磁盘 I/O 进行测试的. 具体测试结果分别如表 2 所示.

表 1 原型系统设备软/硬件配置

设备名	硬件配置	软件配置
客户端	CPU: Intel Pentium Dual 1.8GHz 内存: 2GB 网卡: 1 个 100Mbps	OS: Windows XP 软件:ftp 客户端/Oracle 客户端
服务器 1 (FC/iSCSI)	CPU: Intel Xeon 3220 2.4GHz 内存: 8GB 网卡: 2 个 1000Mbps 光纤卡:QLogic 2342 4Gb/s	OS: RHEL5 软件: iSCSI/光纤卡 HBA 发起端, 心跳客户端软件
服务器 2, 3(iSCSI)	CPU: Intel Xeon 3220 2.4GHz 内存: 8GB 网卡: 2 个 1000Mbps	OS: RHEL5 / WinXP 软件: iSCSI 发起端
融合存储控制器(控制单元)	CPU: Intel Xeno 2.0GHz 内存:2GB 网卡: 2 个 1000Mbps 光纤卡: QLogic 2312 2Gb x 2	OS RHEL5 软件: 融合存储控制器、存储管理
交换机	D-Link 16 口千兆交换机	无
广域网模拟器	CPU: Intel Xeon 2.4GHz 内存: 1GB 网卡: 2 个 1000Mbps	OS: RHEL5 软件: NistNet[137]
FC 阵列	EMC CX500(2Gb)	EMC 存储管理软件
iSCSI 阵列 1, 2, 3	CPU: Intel Xeon 3220 2.4GHz 内存: 8GB 网卡: 2 个 1000Mbps	OS: RHEL5 软件: iSCSI Target 软件

表 2 融合存储系统性能测试与对比

	链接方式	写操作平均速度(KB/sec)	读操作平均速度(KB/sec)
参考值	ATCA--FC	186627	99027
	ATCA--iSCSI	163118	36282
融合前	FC--ATCA--FC	163118	46239
	iSCSI--ATCA--iSCSI	82090	55162
融合后	iSCSI--ATCA--FC	94485	100205
	FC--ATCA--iSCSI	84142	30488

融合确实对系统有所影响, 这主要是由协议之间的转换所影响的.

复制操作影响了系统本身的性能, 并且异步操作明显比同步操作的性能要好. 而且, 同步复制的性能受远程存储池链路带宽的影响, 而异步复制的性能受影响程度低.

表 3 复制对系统的影响

复制方式	链接远程存储池的带宽设置(Mbit/s)	写操作平均速度(KB/sec)	读操作平均速度(KB/sec)
同步复制	5	465	54676
	40	5341	53524
	100	13233	54695
异步复制	5	77192	57591
	40	55150	53079
	100	56156	50272

4 结语

基于融合存储控制器可构成模块化、可扩展、可动态存储分区的网络存储体系, 辅以全方位的远程数据镜像、数据快照与恢复、广域高可用服务诊断与接管等功能的存储容灾手段, 为检验检疫数据中心及各类信息系统提供统一的、满足服务质量要求的容灾网络存储服务. 目前, 我们研发的系统原型, 虽然理论上能够满足用户的指标要求, 但是, 距离真实上线应用还有一段距离, 系统的稳定性及技术风险还存在, 但对于指导研究新型的备份系统研究上有一定的应用价值和指导意义.

参考文献

1 FC SAN 与 IP SAN 的融合.http://www.ciotimes.com/infrastructure/cc/78920.html.
2 郑轲.存储系统中 iSCSI 和 FC 链路融合技术研究与实现[学位论文].武汉:华中科技大学,2008.
3 秦磊华,余胜生,周敬利等.粗波分复用技术在同城容灾系统中的应用.华中科技大学学报(自然科学版),2008,36(1):85–87.
4 汪琦晔.融合平台下 SAN 设备的访问控制设计与实现[学位论文]. 武汉:华中科技大学,2008.
5 李瑞. IP 与 FC 融合式存储系统的安全体系设计与实现[学位论文]. 武汉:华中科技大学,2011.
6 秦磊华,苏彦君,张宗平等.光纤通道存储区域网扩展研究.计算机工程,2007,33(24):109-111.