第33卷 第16期
Vol.33 No.16
计 算 机 工 程
Computer Engineering
2007年8月
August 2007
·开发研究与设计技术· 文章编号：1000—3428(2007)16—0261—03 文献标识码：A 中图分类号：N945

基于LVM的异步实时远程复制系统

王 锋，刘晓光，王 刚，刘 璟
（南开大学计算机科学与技术系，天津 300071）

摘 要： 在分析和讨论现有的远程复制方案特点的基础上，设计并实现了一种基于LVM的异步实时远程复制系统。它是在现有IP网络的基础上通过纯软件方式实现的异步的在线远程复制系统，无须任何昂贵的硬件支持，为远程数据容灾技术提供了一个简单经济的解决方案。

关键词： 远程复制；生产中心；容灾中心；数据一致性；逻辑卷管理器

Asynchronous and Real-time Remote Replication System Based on LVM

WANG Feng, LIU Xiao-guang, WANG Gang, LIU Jing
(Dept. of Computer Science and Technology, Nankai University, Tianjin 300071)

【Abstract】 This paper presents a design of real-time remote replication system based on LVM after discussing the characteristics of all kinds of approaches of the remote replication, which is an asynchronous and on-line remote replication system purely implemented by software in the existed IP network. And the system provides an easy and economic approach for the remote disaster tolerance technology without any expensive hardware.

【Key words】 remote replication; production site; disaster tolenance site; data consistency; logical volume manager

随着计算机系统的广泛应用和Internet技术的飞速发展，各类数据已经成为现代企业正常运作的重要技术基础。关键数据的丢失和损坏很可能会对企业造成致命的打击。远程容灾系统的核心思想[1]就是通过在异地建立和维护一个备份系统，利用地理上的分离来保证系统和数据对灾难性事件的抵御能力。远程复制技术就是其中最基本的技术之一。目前绝大多数的远程复制系统都是以昂贵的网络和存储硬件设备为基础实现的，不仅造价高，而且在距离上也有很大的局限性。

本文针对当前远程容灾系统普遍造价高昂的缺点，提出了一种简单实用的远程复制方案——基于LVM的异步实时远程复制。它将远程复制模块嵌入到LVM中，在LVM层之上实现远程复制功能，以存储设备的逻辑数据块为单位进行复制，对上层的文件系统和应用程序透明，无需特殊的物理环境，而是利用已有的IP网络使远程复制的范围提升到了Internet的广度。同时，它完全通过软件的方法实现，克服了目前远程复制系统普遍存在的造价昂贵的局限性，具有经济实用的优点。并且，该系统采用了主从方式实现异步的在线复制机制，在本地写数据的同时，可将数据更新到一个或多个远端备份系统而不影响本地主机对数据的响应速度。

1 远程数据复制技术

远程数据复制技术[2]是为尽量避免因数据丢失或损坏导致的数据灾难，而在本地建立生产中心，在远端建立一个或多个容灾中心，将数据从生产中心复制到各个容灾中心，形成一个或多个生产中心的远程镜像，使得对数据的访问操作在源数据损坏的情况下仍能继续进行。

1.1 远程复制的方式

数据复制技术可分为定时复制和实时复制2大类[3]。定时复制技术首先比较源设备和目标设备数据上的差异，然后将二者不一致的数据从源设备更新到目标设备。定时复制不能完全满足用户对数据持续复制的要求，而且可能会破坏目标和源设备间数据的一致性。实时复制指源设备上数据发生任何改变时，都会立即更新到目标设备上去，从而更好地维护了目标设备上数据的一致性。实时复制技术包括同步复制、半同步复制和异步复制3种方式：(1)同步方式。应用程序每发出一个写请求后，必须等待数据完全写入主从存储系统后才能继续执行。该模式适用于网络延迟较小的远程容灾系统（通常局限于几十公里以内）。(2)半同步方式。应用程序每发出一个写请求时，只须等待数据写入主存储系统之后即可继续执行，此后再由主从存储控制系统进行写请求的数据同步，但同步完成之前主存储控制系统暂停响应主机的下一个写请求。(3)异步方式。当上层发出多个写请求时，只须等待数据写入主存储系统之后即可继续执行，此后再由主存储控制系统与从存储控制系统完成写数据同步。显然，异步方式更适用于实际的远程容灾。

在同步和半同步方式下，数据一致性得到了保证，但也降低了源设备主机的I/O性能。异步方式对源设备主机的性能影响较小，但可能影响数据的一致性。因此，保证生产中心和容灾中心数据的一致性是保证容灾系统可靠性与可用性的关键。

1.2 异步远程复制方式下的数据一致性

如图1所示，在异步复制模式下，主机由写命令序列依次写逻辑块B1，B2，B3，B2。在t时刻主机的WB2命令出现了覆盖。对于生产中心，主存储系统可以把t时刻前的写命令序列按原顺序提交给本地磁盘，因此，不会出现写顺序颠倒的

基金项目： 国家自然科学基金资助项目(90612001)；天津市科技发展计划基金资助项目(043800311，043185111-14)；南开大学科技创新基金资助项目

作者简介： 王 锋(1982－)，男，硕士研究生，主研方向：并行与分布式系统；刘晓光、王 刚，副教授；刘 璟，教授、博士生导师

收稿日期： 2006-10-10 **E-mail：** wangfeng@mail.nankai.edu.cn

情况。但对于容灾中心，这个写命令序列的数据在被传输给容灾中心之前，由于 B2 先发生了覆盖，因此被传输的数据序列就变成了 B1，B3，B2。若序列 B1，B3，B2 能全部到达容灾中心，则容灾中心的数据状态就能成为生产中心的一个快照。但是，如果这个序列中只有 B1，B3 到达容灾中心，而 B2 却因为网络故障等原因没有到达容灾中心，则容灾中心最终所呈现的数据状态将是生产中心在任何时刻都不可能出现的，这就产生了写顺序不一致性问题。

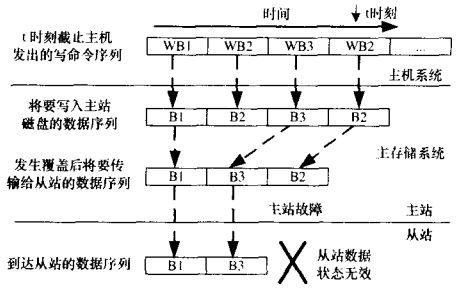


图 1　写顺序不一致性问题的产生

如果具有逻辑依赖关系的写操作在容灾中心出现错序，则容灾中心的数据可能处于无效状态而变为不可用，因此，异步复制技术中最大的困难在于如何保证容灾中心和生产中心的写顺序的一致，以及如何提高远程复制的效率。在实际的容灾系统设计中，根据系统的容灾级别通常选用严格的写顺序一致性和松散的写顺序一致性 2 种不同的方案来解决异步复制中的数据一致性问题。

严格的写顺序一致性，即严格按照生产中心写操作的顺序将其提交到容灾中心。容灾中心依照时间顺序保存生产中心某一时刻的快照，当生产中心出现故障时，就把其中的数据恢复到故障前的某个有效状态。为了实现严格一致性，只须按照生产中心记录的写操作顺序向容灾中心提交写操作即可。设计简单，数据丢失量小，数据安全性较高，但由于没有进行写覆盖等优化操作，因此效率相对较低。

松散的写顺序一致性，即在写操作的序列中，只要保证有逻辑依赖关系的写命令之间的提交顺序一致即可。但是它的实现相对复杂，潜在的数据丢失量较大，但效率相对较高。

在基于 LVM 实时异步远程复制系统中，考虑到系统实现的简洁，采取了严格的写顺序一致的方案，并对其进行了改进优化，以确保数据的可靠性和可用性。

2　设计实现与关键技术

2.1　逻辑卷管理器

逻辑卷管理器(logical volume manager, LVM)本质上是一个虚拟设备驱动，是在内核中块设备和物理设备之间添加的一个新的抽象层次[4]。它可以将几块磁盘即物理卷(physical volume)组合起来形成一个存储池或者卷组(volume group)。LVM 可以每次从卷组中划分出不同大小的逻辑卷(logical volume)创建新的逻辑设备。底层的原始磁盘不再由内核直接控制，而由 LVM 层控制。对于上层应用，卷组替代了磁盘块成为数据存储的基本单元。LVM 逻辑设备向上层应用提供了和物理磁盘相同的功能，如文件系统的创建和数据的访问等。但 LVM 逻辑设备不受物理约束的限制，逻辑卷不必是连续的空间，它可以跨越许多物理卷，并且可以在任何时候任意调整大小，相比物理磁盘，更易于磁盘空间的管理。

从用户态应用来看，LVM 逻辑卷相当于一个普通的块设备，对其的读写操作和普通的块设备完全相同。而从物理设备层来看，LVM 相对独立于底层的物理设备，并且屏蔽了不同物理设备之间的差异。因此，在 LVM 层上考虑数据的远程复制问题，无须单独考虑某种具体的物理设备，避免了远程复制中因物理设备之间的差异而产生的问题。

2.2　NBD

NBD(network block device)是一种 Linux 内核的设备驱动扩展，它可以通过和远端主机建立 TCP/IP 连接将远端资源映射成本地块设备，从而建立一个廉价安全实时的镜像，对本地映射设备的读写将通过 TCP/IP 连接转换为对远端主机的读写操作（图 2）。

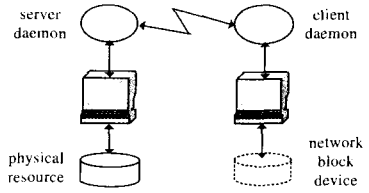


图 2　NBD 使远端资源映射成本地块设备

本地 NBD 设备在设备级跟远端主机实现了同步数据交互，对本地映射设备的读写请求的完成与否取决于远端主机是否完成读写请求，以及本地设备是否收到远端主机的应答。

由于 NBD 的工作模式采取了对远端设备实现本地的同步实时数据镜像，本文基于 NBD 协议并通过进一步改进其效率和连接可靠性，实现了具有高可靠性和高可用性的异步在线远程复制。当生产中心发生一系列写数据操作时，可以在 LVM 层将写操作数据块写入本地设备，再将该数据块的副本交给 NBD 处理，然后返回继续处理下一个 I/O 请求，所有的数据块副本向远端设备（容灾中心）的同步由 NBD 独立完成，极大地降低了系统实现的复杂性。

2.3　工作原理

将远程复制机制嵌入到 Linux 系统的 LVM 组件中，即可利用 LVM 优越的磁盘空间管理性能，屏蔽因物理存储设备的差异而导致的系统设计的复杂性，降低了实现的难度。系统基本工作原理是：生产中心在 LVM 层截获用户的写操作请求，通过 LVM 层的逻辑映射将请求发送到本地设备的 I/O 操作队列，即本地设备写操作完成；同时将写操作数据副本和 I/O 请求按照本地设备（即生产中心）的写操作顺序记录到远程复制 I/O 请求队列中。NBD 从远程复制请求队列中提取写操作请求，同样按照本地数据的写操作顺序，通过 NBD 协议将写请求及相应的数据副本发送到容灾中心，完成生产中心向容灾中心的数据更新，NBD 的整个工作周期是以同步方式进行的，从而保证了本地和远端的严格的数据一致性，即保证了容灾中心作为本地生产中心的实时镜像的数据的连续可用性。工作原理如图 3 所示。

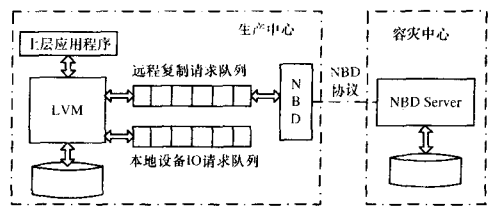


图 3　基于 LVM 的异步实时远程复制系统原理图

由于生产中心设备的写操作速度远高于通过 IP 网络对容灾中心数据的更新速度，在生产中心设备写操作完成后，即可立即返回继续处理下一 I/O 请求，无须等待确认容灾中心数据

更新操作的完成。而针对生产中心和容灾中心的数据一致性问题，由于在本地和远端之间采用了严格的数据一致性的方案，即NBD协议完全采用同步的方式和远程容灾中心进行数据交互，对于生产中心在一段时间内所有的写操作，严格按照生产中心写操作的顺序将这些写操作和数据副本提交到容灾中心，保证了生产中心和容灾中心严格的写顺序一致性，从而保证了容灾中心和生产中心数据的一致。

2.4 双链接模式下数据的更新

在远程复制系统实现的过程中，为了避免因为网络异常问题导致远程复制系统的瘫痪，提高系统可靠性，本文在原有的NBD协议的基础上，对其进行优化，通过建立2条可靠的网络链接来实现生产中心本地写操作向远程容灾中心的顺序更新。在双链接模式下，数据的更新可采用主动式和被动式2种方式：

(1)主动式：NBD创建2个内核线程[5]，每个线程分别主动地去远程复制队列中获取写请求和数据，然后选择一条可用链接将数据更新到远程容灾中心，容灾中心的确认更新数据收到后，该线程继续去队列中获取下一请求进行处理。这种方式两条链接同时工作，当一条链接失效另外一条链接仍可继续工作，效率比较高；但是由于线程并行工作和网络时延的不确定性，可能会导致写请求到达远程容灾中心的顺序的颠倒，从而导致容灾中心和生产中心数据的不一致。这就要求运行在容灾中心上的系统必须具有重新整合写请求的顺序的能力，无形中增加了系统实现的复杂性。

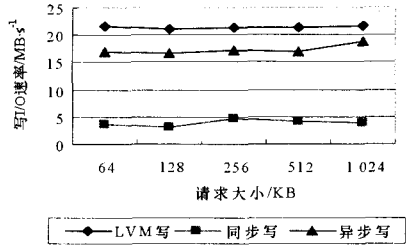
(2)被动式：本地系统在完成对本地设备的写I/O操作的同时，将写请求发送到远程复制请求队列，NBD被动地接受请求，指定其中一条链接将数据按请求顺序同步到远端。当检测到此链接失效则立即更换到另一链接继续工作。这种方式某一时刻只有一条TCP/IP链接在工作，因而必须显式地去检测链接的状态，从而确定是否需要进行更换。数据更新传播效率相对主动方式而言较低；但是这种方式具有简单易实现的优点。基于LVM的异步实时远程复制系统就是采用了这种被动的接收写请求进行处理的方式来实现的。

2.5 远程复制队列溢出问题

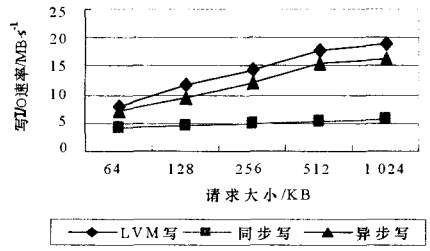
远程复制队列的空间有限，如果网络传输速度较慢或出现网络故障导致2条连接均不能正常工作，请求队列就会很快溢出。如果检测出至少还有一条网络保持连接，可以通过延迟响应上层写请求，使其等待，直到完成部分请求使队列中出现一定空闲的空间。如果检测出网络连接全部断开导致队列溢出，就立即对NBD尚未完成的更新请求和数据副本作上标记，写入本地空闲设备进行保存，直至远程复制机制恢复后，再将其同步到远程容灾中心。在同步的过程中，上层应用一直处在等待状态，直到同步操作完成才能继续响应上层应用的写请求，进入正常的远程复制状态由于完成的更新请求并不严格按照原来生产中心的写操作请求的顺序进行传播，因此在同步操作完成前，容灾中心和生产中心并不严格保持数据的一致性。

3 试验结果与性能分析

在实验室局域网内测试了该系统写操作处理性能，在多台Linux主机下利用IOZone软件分别测试了LVM本身的写操作性能、同步模式下数据的远程复制的性能，以及本文的异步实时远程负责的写操作性能。硬件环境为多台安装有Redhat企业版4.0的操作系统，CPU为P4 2.0GHz，内存256MB的PC主机系统，网络环境为基于IPv4的百兆IP网络。测试结果如图4所示。



(a) 连续I/O



(b) 随机I/O

图4 系统性能比较

由图4可以看出：当请求大小从64KB到1MB逐渐递增时，无论是连续的还是随机的用户写请求，逻辑卷管理器本身的写性能都优于同步复制和异步实时复制的性能。异步实时复制性能也远高于同步复制性能。这一结果完全符合理论分析。考虑到实际应用中远程复制的广度远大于实验过程中的测试距离，实际网络延迟会比测试环境中的延迟大得多，同步远程复制性能会随着网络延迟的增大而迅速降低，即在实际应用中，异步实时的远程复制在相比同步方式性能方面更加优越。

4 总结与展望

本文提出了一种基于LVM的实时异步的远程复制系统的设计方案。该系统采用实时异步的在线数据复制方式，大大减少传输到容灾中心的数据量，有效地利用网络带宽，提高了数据远程复制的效率，对本地主机的I/O性能影响较小。而且整个系统在原有系统结构的基础上以纯软件方式实现，不需要任何特殊的网络或存储设备，系统实现简单低廉。今后将在提高系统的复制效率（例如数据压缩传输机制的实现）、安全性和健壮性（例如采用工业标准的ISCSI协议来实现）等方面作继续深入的研究。此外，由于LVM本身还支持快照技术，因此还可以将远程复制技术和快照技术结合使用，更好地提高远程数据容灾系统的可靠性和可用性。

目前该系统主要被集成到存储系统逻辑卷管理器中，应用在基于Linux平台的中小型服务器系统中以数据的远程实时备份为中心的数据容灾等领域。事实证明，相比采用同步复制技术的容灾系统，该系统在通用性、易用性和性能方面具有很大的优势。

参考文献

1 邓玉洁，张忠能. IP网络存储技术研究[J]. 计算机工程与应用, 2004, 40(23): 148-151.

2 董欢庆，李战怀. Linux平台远程逻辑卷复制系统的设计[J]. 计算机工程与应用, 2004, 40(18): 109-112.

3 林 伟. 远程卷复制系统的研究与开发[D]. 西安：西北工业大学, 2005-03.

4 Hasenstein M. LVM Whitepaper[Z]. (2001-09).Http://www.sistina.com/lvm_whitepaper.pdf.

5 Beck M. Linux Kernel Programming[M]. [S. l.]: Pearson Education Deutschland GmbH, 2001.