



(12) 发明专利

(10) 授权公告号 CN 101031889 B

(45) 授权公告日 2011.06.08

(21) 申请号 200480044118.9

(51) Int. Cl.

(22) 申请日 2004.08.12

G06F 11/20 (2006.01)

(85) PCT申请进入国家阶段日

2007.03.29

(56) 对比文件

(86) PCT申请的申请数据

PCT/EP2004/009047 2004.08.12

US 6148383 A, 2000.11.14, 全文.

(87) PCT申请的公布数据

WO2006/015612 EN 2006.02.16

US 20040034808 A1, 2004.02.19, 说明书第2页第[0018]-[0026]段、附图1,2.

(73) 专利权人 意大利电信股份公司

CN 1497458 A, 2004.05.19, 全文.

地址 意大利米兰

EP 0670551 A1, 1995.09.06, 全文.

(72) 发明人 安德烈·迪吉里奥

CN 1264476 A, 2000.08.23, 全文.

拉法埃莱·吉拉尔迪

审查员 刘渊

欧金尼奥·M·马菲奥内

(74) 专利代理机构 中国国际贸易促进委员会专

权利要求书 5 页 说明书 14 页 附图 12 页

利商标事务所 11038

代理人 康建忠

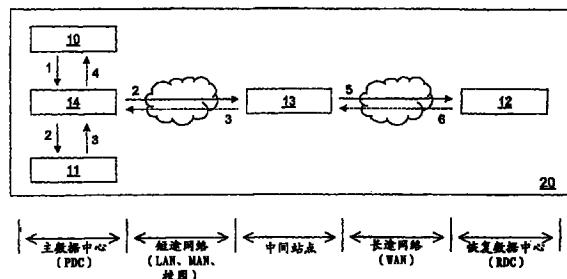
(54) 发明名称

用于更新数据集的系统、方法、设备和通信网

络

(57) 摘要

一种通过地理通信网络更新数据集的系统，该数据集存储在恢复数据中心(RDC)的恢复存储单元(12)中且需用由在主数据中心(PDC)的处理系统(10)生成的数据块更新，恢复数据中心(RDC)通过地理通信网络与配有临时存储区域(27)的设备(13)关联，临时存储区域(27)用于在对应存储位置临时存储由处理系统生成并用于更新数据集的到来数据块，其中一旦在临时存储区域(27)中写入数据块，就将用于确认在临时存储区域(27)中成功写入数据块的第一同步确认信号(3)提供给主数据中心(PDC)，以不阻塞或减缓运行处理系统(10)的正常方式，当接收到用于确认成功更新数据集的第二确认信号(6)时，使由数据块占用的临时存储区域(27)的位置对新的到来数据块可用。



1. 一种用于通过通信网络更新数据集的系统,所述数据集存储在位于恢复数据中心(RDC)中的恢复存储单元(12)中,并且必须用由与主数据中心(PDC)关联的至少一个处理系统(10)生成的到来数据块来更新,所述主数据中心(PDC)请求将所述数据块写入所述恢复数据中心(RDC)中的请求的地址中,所述用于通过通信网络更新数据集的系统包括:

- 通过所述通信网络置于所述主数据中心(PDC)和所述恢复数据中心(RDC)之间的设备(13),其包括临时存储区域(27),所述临时存储区域(27)用于在对应的存储位置临时写入由所述处理系统(10)生成的到来数据块,并将所述数据块转发到所述恢复数据中心(RDC),其中,所述设备(13)包括适用于以下功能的模块:

- 一旦在所述临时存储区域(27)中写入所述数据块,就将用于确认在所述临时存储区域(27)中成功写入所述数据块的第一同步确认信号(3)发送到所述主数据中心(PDC),

- 接收并管理由所述恢复数据中心(RDC)发送的第二确认信号(6),其用于在转发所述数据块之后确认成功更新所述数据集,以及

- 在接收所述第二确认信号(6)之后,使所述临时存储区域(27)的位置对于新的到来数据块可用,

其特征在于,

所述设备还包括适于按照以下方式将所述数据块写入所述临时存储区域(27)中的模块:

- 如果在所述临时存储区域中没有写入将被写入到请求的地址处的数据块,则在所述临时存储区域(27)的位置中写入所述数据块;

- 如果已经在所述临时存储区域的位置中写入将被写入到请求的地址处的数据块,但还未转发所述数据块,则通过替换先前的数据块而在所述临时存储区域(27)的所述位置中写入所述数据块;

- 如果已经在所述临时存储区域(27)的位置中写入将被写入到请求的地址处的数据块,已经转发所述数据块,还未接收到来自所述恢复数据中心(RDC)的确认,并且具有相同地址的第二数据块没有出现在所述临时存储区域(27)中,则在所述临时存储区域(27)中的新的位置中写入所述数据块。

2. 如权利要求1所述的系统,其中,所述处理系统(10)与存储所述数据集的拷贝的主存储单元(11)关联,所述主存储单元(11)位于所述主数据中心(PDC)内。

3. 如权利要求2所述的系统,其中,所述临时存储区域(27)具有小于主存储单元(11)和恢复存储单元(12)的存储容量的存储容量。

4. 如权利要求1所述的系统,其中,所述设备(13)和所述主数据中心(PDC)之间的距离小于所述设备(13)和恢复数据中心(RDC)之间的距离。

5. 如权利要求1所述的系统,其中,所述通信网络包括从包含iSCSI类型协议和光纤信道类型协议的组中选择的至少一个协议。

6. 如权利要求1所述的系统,其中,所述设备(13)包括:

- 接口模块(20,26),将所述设备(13)分别接口到所述主数据中心(PDC)和所述恢复数据中心(RDC);

- 过滤器模块(22),截取由所述主数据中心(PDC)发送的I/O命令,将所述命令有选择地发送到用于临时存储所述数据块的所述临时存储区域(27)或所述恢复数据中心(RDC);

- 控制模块 (24, 25), 从所述主数据中心 (PDC) 接收具有所述数据块的相对地址参数的所述数据块, 将所述数据块封装到为管理所述恢复存储单元 (12) 预留的命令内, 并将它们转发到所述恢复数据中心 (RDC)。

7. 如权利要求 1 所述的系统, 其中, 所述设备 (13) 包括 :

- 网络接口卡 (NIC) ;
- iSCSI 目标设备 (21), 布置在所述网络接口卡 (NIC) 的下游 ;
- 修改的 SCSI 目标设备 (22), 布置在所述 iSCSI 目标设备 (21) 的下游 ;
- 缓存系统 (27), 与所述修改的 SCSI 目标设备 (22) 关联 ;
- 逻辑卷管理器设备 (23), 布置在所述修改的 SCSI 目标设备 (22) 的下游 ;
- 本地 SCSI 子系统 (24), 布置在所述逻辑卷管理器设备 (23) 的下游 ;
- iSCSI 发起器设备 (25), 布置在所述 SCSI 子系统 (24) 的下游 ;
- 其它网络接口卡 (NIC), 被布置在所述 iSCSI 发起器设备 (25) 的下游并朝向所述恢复数据中心 (12)。

8. 如权利要求 6 所述的系统, 其特征在于, 所述设备 (13) 包括 :

- 主机总线适配器 (HBA), 朝向所述主数据中心 (PDC) ;
- 修改的 SCSI 目标设备 (22), 布置在所述主机总线适配器 (HBA) 的下游 ;
- 缓存系统 (27), 与所述修改的 SCSI 目标设备 (22) 关联 ;
- 逻辑卷管理器设备 (23), 布置在所述修改的 SCSI 目标设备 (22) 的下游 ;
- SCSI 子系统 (24), 布置在所述逻辑卷管理器设备 (23) 的下游 ; 以及
- 其它主机总线适配器 (HBA), 被布置在所述 SCSI 子系统 (24) 的下游并朝向所述恢复存储单元 (12)。

9. 如权利要求 7 或 8 所述的系统, 其中, 所述逻辑卷管理器设备由软件虚拟设备来替换。

10. 如前述权利要求中的任意一项所述的系统, 其中, 所述用于临时存储所述数据块的设备 (13) 包括与所述通信网络关联的网络设备。

11. 如权利要求 2 所述的系统, 其中, 所述恢复存储单元和所述主存储单元 (11, 12) 是盘存储单元。

12. 如权利要求 1 所述的系统, 其中, 所述临时存储区域 (27) 包括多个存储区域, 所述多个存储区域分别用于多个处理系统 (10) 和 / 或多个主数据中心 (PDC)。

13. 一种用于通过通信网络更新数据集的方法, 所述数据集存储在位于恢复数据中心 (RDC) 中的恢复存储单元 (12) 中, 并且必须用由与主数据中心 (PDC) 关联的至少一个处理系统 (10) 生成的到来数据块来更新, 所述主数据中心 (PDC) 请求将所述数据块写入所述恢复数据中心 (RDC) 中的请求的地址中, 所述方法包括以下步骤 :

- 在置于所述主数据中心 (PDC) 和所述恢复数据中心 (RDC) 之间的设备 (13) 所包括的临时存储区域 (27) 的对应的存储位置中, 通过所述通信网络临时写入由所述处理系统 (10) 生成的到来数据块 ;

- 一旦在所述存储区域 (27) 中写入所述数据块, 就由所述设备 (13) 并通过所述通信网络将用于确认在所述临时存储区域 (27) 中成功写入所述数据块的第一同步确认信号 (3) 提供给所述主数据中心 (PDC) ;

- 由所述设备并通过所述通信网络将所述数据块转发到所述恢复数据中心 (RDC)；
- 由所述设备并通过通信网络接收和管理第二确认信号 (6)，所述第二确认信号 (6) 确认将所述数据集成功更新到所述恢复数据中心 (RDC)；以及
- 在接收所述第二确认信号 (6) 之后，使所述临时存储区域 (27) 的位置对于新的到来数据块可用，

其特征在于，

在所述设备 (13) 中临时写入到来数据块的步骤包括如下步骤：

- 如果在所述存储区域中没有写入将被写入到请求的地址处的数据块，则在所述存储区域 (27) 中写入所述数据块；
- 如果已经在所述存储区域中写入将被写入到请求的地址处的数据块，但还未转发所述数据块，则通过替换先前的数据块而在所述存储区域 (27) 中写入所述数据块；
- 如果已经在所述存储区域 (27) 中写入将被写入到请求的地址处的数据块，已经转发所述数据块，还未接收到来自所述恢复数据中心 (RDC) 的确认，并且具有相同地址的第二数据块没有出现在所述临时存储区域 (27) 中，则在所述存储区域 (27) 中的新位置中写入所述数据块。

14. 如权利要求 13 所述的方法，包括以下步骤：

- 将所述数据集存储到与所述处理系统 (10) 相关联的主存储单元 (11) 内，所述主存储单元 (11) 位于所述主数据中心 (PDC) 内。

15. 如权利要求 13 所述的方法，其中，在所述存储区域 (27) 中临时写入到来数据块的步骤进一步包括以下步骤：

- 如果在所述存储区域 (27) 中两次存储将被写入到请求的地址处的数据块，则通过替换尚未转发的数据块来进行写入。

16. 如权利要求 13 所述的方法，其中，所述将所述数据块从所述设备 (13) 转发到所述恢复数据中心 (RDC) 的步骤包括以下步骤：

- 轮询所述设备 (13) 中所包括的所述存储区域 (27)，以检查是否存在将要转发到所述恢复数据中心 (RDC) 的任意数据块；

把将要转发的数据块发送到所述恢复数据中心 (RDC)，并同时启动具有预定超时延迟的定时器；以及

其中，接收和管理所述第二确认信号的步骤包括以下步骤：

- 检查所述超时延迟是否已经过去；以及

在尚未接收到第二确认信号 (6) 的情况下，

- 如果所述超时已经过去，则增加表示丢失的数据块的计数器；以及

如果丢失的数据块的计数器已经达到预定值，则将告警发送到主数据中心 (PDC)，以便对于数据块到所述设备 (13) 的任何其它发送进行锁定。

17. 如权利要求 16 所述的方法，其中，所述接收和管理所述第二确认信号的步骤还包括以下步骤：管理否定确认信号 (NACK)，所述否定确认信号 (NACK) 表示所述数据块已经到达恢复数据中心 (RDC) 但是受损的。

18. 如权利要求 16 所述的方法，其中，将所述超时延迟设置为大于从所述设备 (13) 发送数据块并由所述恢复存储单元 (12) 接收所述数据块的往返时间的值。

19. 如权利要求 13 所述的方法,其中,提供用于朝向主数据中心 (PDC) 恢复数据块的处理,其特征在于,进一步包括以下步骤:

- 如果在所述临时存储区域 (27) 中没有存储将被写入到请求的地址处的数据块,则由所述设备 (13) 向所述恢复数据中心 (RDC) 请求恢复的数据块;

- 如果在所述存储区域中存储有将被写入到请求的地址处的数据块,并且所述数据块尚未经历转发到恢复数据中心 (RDC) 的转发步骤,则从临时存储区域 (27) 读取恢复的数据块;

- 如果在所述临时存储区域中存储有将被写入到请求的地址处的数据块,并且已经将所述数据块转发到恢复数据中心 (RDC),并且在所述临时存储区域 (27) 中没有将被写入到相同的请求的地址的其它数据块,则从临时存储区域 (27) 读取恢复的数据块;

- 在所述主数据中心 (PDC) 的请求下,通过先前步骤中的至少一个步骤由所述设备 (13) 朝向所述主数据中心 (PDC) 恢复所述恢复的数据块。

20. 一种用于通过通信网络更新数据集的设备,所述设备包括临时存储区域 (27),所述存储区域 (27) 用于在对应的存储位置临时写入由与主数据中心 (PDC) 关联的处理系统 (10) 生成的到来数据块,并将所述数据块转发到位于恢复数据中心 (RDC) 的恢复存储单元 (12),所述主数据中心 (PDC) 请求将所述数据块写入所述恢复数据中心 (RDC) 中的请求的地址中,所述设备包括适用于以下功能的模块:

- 一旦在所述临时存储区域 (27) 中写入所述数据块,就通过所述通信网络将用于确认在所述临时存储区域 (27) 中成功写入所述数据块的第一同步确认信号 (3) 发送到所述主数据中心 (PDC),

- 通过所述通信网络接收并管理由所述恢复数据中心 (RDC) 发送的第二确认信号 (6),其用于在转发所述数据块之后确认成功更新所述数据集,以及

- 在接收所述第二确认信号 (6) 之后,使所述临时存储区域 (27) 的位置对于新的到来数据块可用,

其特征在于,

所述设备还包括适于按照以下方式将所述数据块写入所述临时存储区域 (27) 中的模块:

- 如果在所述临时存储区域中没有写入将被写入到请求的地址处的数据块,则在所述临时存储区域 (27) 的位置中写入所述数据块;

- 如果已经在所述临时存储区域的位置中写入将被写入到请求的地址处的数据块,但还未转发所述数据块,则通过替换先前的数据块而在所述临时存储区域 (27) 的所述位置中写入所述数据块;

- 如果已经在所述临时存储区域 (27) 的位置中写入将被写入到请求的地址处的数据块,已经转发所述数据块,还未接收到来自所述恢复数据中心 (RDC) 的确认,并且具有相同地址的第二数据块没有出现在所述临时存储区域 (27) 中,则在所述临时存储区域 (27) 中的新位置中写入所述数据块。

21. 如权利要求 20 所述的设备,其中,所述处理系统 (10) 与存储所述数据集的拷贝的、位于主数据中心 (PDC) 内的主存储单元 (11) 关联。

22. 如权利要求 20 所述的设备,所述设备 (13) 和所述主数据中心 (PDC) 之间的距离小

于所述设备 (13) 和恢复数据中心 (RDC) 之间的距离。

23. 如权利要求所述 20 的设备,其特征在于,包括适于和从包含 SCSI 类型协议、iSCSI 类型协议、光纤信道类型协议的组中选择的至少一个协议进行接口的模块。

24. 如权利要求 20 所述的设备,其中,所述设备包括:

- 接口模块 (20, 26),将所述设备 (13) 分别接口到所述主数据中心 (PDC) 和所述恢复数据中心 (RDC) ;

- 过滤器模块 (22),截取数据所关联的 I/O 命令,并判断是将所述命令发送到用于临时存储所述块的存储区域 (27) 还是所述恢复存储单元 (12) ;

- 控制模块 (24, 25),接收具有所述数据块的相对地址参数的所述数据块,将所述数据块封装到为管理所述恢复存储单元 (12) 预留的命令内,并将它们发送到所述恢复数据中心 (RDC) 。

25. 如权利要求 20 所述的设备,其中,所述设备包括:

- 网络接口卡 (NIC),朝向所述主数据中心 (PDC) ;

- iSCSI 目标设备 (21),布置在所述网络接口卡 (NIC) 的下游;

- 修改的 SCSI 目标设备 (22),布置在所述 iSCSI 目标设备 (21) 的下游;

- 缓存系统 (27),与所述修改的 SCSI 目标设备 (22) 关联;

- 逻辑卷管理器设备 (23),布置在所述修改的 SCSI 目标设备 (22) 的下游;

- 本地 SCSI 子系统 (24),布置在所述逻辑卷管理器设备 (23) 的下游;

- iSCSI 发起器设备 (25),布置在 SCSI 子系统 (24) 的下游,以及

- 其它网络接口卡 (NIC),被布置在所述 iSCSI 发起器设备 (25) 的下游并朝向所述恢复数据中心 (RDC) 。

26. 如权利要求 20 所述的设备,其中,所述设备包括:

- 主机总线适配器 (HBA),朝向所述主数据中心 (PDC) ;

- 修改的 SCSI 目标设备 (22),布置在所述主机总线适配器 (HBA) (20) 的下游;

- 缓存系统 (27),与所述修改的 SCSI 目标设备 (22) 关联;

- 逻辑卷管理器设备 (23),布置在所述修改的 SCSI 目标设备 (22) 的下游;

- SCSI 子系统 (24),布置在所述逻辑卷管理器设备 (23) 的下游;以及

- 其它主机总线适配器 (HBA),被布置在所述 SCSI 子系统 (24) 的下游并朝向所述恢复数据中心 (RDC) 。

27. 如权利要求 25 或 26 所述的设备,其中,所述逻辑卷管理器设备由软件虚拟设备来替换。

28. 如权利要求 20 所述的设备,其中,所述设备 (13) 包括与所述通信网络关联的网络设备。

29. 一种包括根据权利要求 1 至 12 之一所述的用于更新数据集的系统的通信网络。

用于更新数据集的系统、方法、设备和通信网络

技术领域

- [0001] 本发明涉及通过通信网络更新数据集的系统、方法和设备。
- [0002] 具体地说，本发明涉及通过地理通信网络并以同步模式将来自于位于主数据中心中的第一存储单元的数据集更新到位于地理上远离主数据中心的恢复数据中心的第二存储单元的系统、方法和设备，其中，可将存储在恢复数据中心中的数据用于涉及主数据中心的灾难情形的情况。
- [0003] 更具体地说，本发明涉及保护数据集的同步镜像技术。
- [0004] 背景技术
- [0005] 在高可靠性计算机架构中，通过镜像技术来提出在灾难的情况下保护重要数据不受损或丢失的扩展方式。所述技术提供维护存储在布置在两个不同站点中的至少两个不同存储单元中的重要信息的至少两个拷贝：第一本地拷贝，通常表示“工作拷贝”，由运行在位于主数据中心（PDC）的主计算系统（计算机）中的软件应用直接使用，而第二拷贝位于远程恢复数据中心（RDC），在主计算系统的故障的情况下，在灾难恢复过程的范围内使用。
- [0006] 本领域已知用于进行数据集复制的至少两种方法：第一技术或方法称为同步镜像，其中，在 RDC 处的远程拷贝的更新协同于在 PDC 处的本地拷贝中的数据项的修改；
- [0007] 第二技术称为异步镜像，根据批处理策略产生远程拷贝的更新。
- [0008] 作为本发明的有关技术的同步镜像技术通常提供以下步骤：
- [0009] a- 在本地存储单元上写入数据项；
- [0010] b- 在远程存储单元上写入数据项；
- [0011] c- 通过再次写入新的数据项在重复步骤 a) 和 b) 之前写入来自远程盘的确认信号 ACK。
- [0012] 因为同步镜像允许使至少两个存储单元在每一时刻完全地校准，所以在故障、数据丢失或灾难的情况下，其给予恢复主计算系统的状态更多的保证。
- [0013] 在灾难的情况下，通过使用同步镜像，能够将所谓的恢复时间目标或 RT0 保持为低，所述 RT0 是恢复与在灾难之前运行在 PDC 上的应用相同的软件应用的正常工作所需的时间间隔。
- [0014] 实际上，通过采用同步镜像技术，保证理想地以一个事务处理减少在主数据集以及用于重置软件应用的主数据集的拷贝之间的偏离。
- [0015] 在本领域，在主数据集及其拷贝之间按时间测量的偏离通常称为恢复点目标或 RPO。
- [0016] 需要指出，与同步镜像有关的步骤的序列要求把在 PDC 处的软件应用锁定下述时间间隔，即，从由软件应用自身产生数据项到在 PDC 处接收由在 RDC 处的远程存储单元写入数据项（写入操作）的确认所经过的时间间隔。可将该时间间隔估计为下述项的和：
- [0017] - 数据项的串行化时间；
- [0018] - 往返时间，这是传播延迟与存在于主计算系统和远程拷贝之间的连接的装置中的处理和队列时间之和；

- [0019] - 将在盘上的数据项写入到 RDC 的写入时间；
- [0020] - 确认信号的产生和串行化时间，所述时间关于数据项的串行化时间和往返时间可以忽略。
- [0021] 通常，最小往返时间不会低于与在使用的介质中的物理传播（传播延迟）有关的时间间隔，所述时间间隔是直接与主数据中心 (PDC) 中的计算机和恢复数据中心 (RDC) 中的远程存储单元之间的物理连接的距离成比例。
- [0022] 在存储产业中公知的是，以及存储软件和硬件供应商的出版物中详尽记载的那样，同步镜像降低了产生涉及同步副本（镜像）的数据的软件应用的性能。随着 PDC 和 RDC 之间的距离增加，应用性能成比例地降低。作为示例，假定限制效应仅为传播延迟，PDC 和 RDC 之间的距离从 10km 增加到 100km 提供了写入响应时间（往返时间）增加 10 倍；结果，取决于写入操作的速率的量，应用吞吐量可能减少达到 90%。
- [0023] 无论如何，申请人认为，不能以简单和单一的方式来定义所述距离限制，超过所述距离限制，这些性能使得维持 PDC 处的计算系统的正常功能变得不可接受，因为其严格取决于商务类型和有关的软件应用（写入操作的大小和频率）以及 PDC 和 RDC 之间的通信网络的物理特性（带宽、技术和拓扑）。
- [0024] 一些著作源指出关于同步镜像的一些距离限制：
- [0025] - Nortel Networks 白皮书“Storage Distance extension :increasing the reach and utility of networked storage applications”，指出甚至当使用高带宽链路时将 400km 作为距离限制；
- [0026] - Hitachi 白皮书“Business Continuity Solution Blueprint-Synchronous data replication”指出 50km，声明距离限制具体取决于应用响应时间；
- [0027] - 同步镜像的优先级解决方案指出，由于管理数据副本的特定软件导致不同的距离限制；更具体地说，IBM PPRC(白皮书“IBM and Cisco :Metro Optical Solution for Business Continuity and StorageNetworking”，2003 年 8 月) 指出 40-100km 作为距离限制。EMCSRDF(白皮书“EMC and Cisco Metro Optical Storage NetworkingSolution”，2001 年 6 月 27 日) 指出 80km 作为最大距离限制。
- [0028] 申请人注意到，甚至在存在高带宽情况下，由此串行化时间是可忽略的，以及在存在专用电路连接的情况下，由此往返时间减少到最小，在所述情况下，同步镜像技术通常不能应用于具有 PDC 和 RDC 之间的任意距离的连接。
- [0029] 这样的限制与高可用性计算架构的典型需求形成对照，根据所述高可用性计算架构，需要在位于较大距离（例如离开工作拷贝几百公里）处的站点中存储数据的副本，从而允许在大灾难的情况下的高的保护级别。
- [0030] 为了避免对于同步镜像技术固有的上述问题，已经提出了称为多跳盘镜像的技术。例如在“Asynchronous Cascading Implementation, TIPS0310”，IBM 红皮书-Hints & Tips, 2003 年 10 月 15 日（在网站 <http://publib-b.boulder.ibm.com/Redbooks.nsf>, 2004 年 6 月 14 日的互联网上可以找到）并且在“Reomote Mirroring of Business CriticalInformation”，EMC, 2002 年 6 月 10 日（在网站 <http://italy.emc.com/local/it/IT/download/pdf/giugno2002/Burns.pdf>, 2004 年 6 月 14 日的互联网上可以找到）已经公布了这样的技术。

[0031] 多跳镜像技术提出：在位于与由在 PDC 处的软件应用提出的限制兼容的距离的中间站点处进行同步镜像，并朝向 RDC 站点异步地复制数据。

[0032] 申请人注意到，多跳镜像在端到端链中具有引入复杂单元的缺点。

[0033] 根据现有技术，多跳镜像需要在中间站点引入存储单元，这样，必须至少具有 PDC 中的存储单元的相同的存储容量。

[0034] 这样的解决方案减少了架构的整个可靠性。

[0035] 此外，由于从中间站点到恢复站点（RDC）的更新通常以相对低的频率在批处理模式下产生，因此可能在主数据集和恢复的数据集之间产生有关的差别。

[0036] 因此，在具有包括主站点（PDC）和中间站点两者活动范围的灾难的情况下，可能不能获得很低的恢复点目标或 RPO。

[0037] 论文“Heterogeneous Midrange Storage with Local Mirroring and Remote IP Replication”，Falconstore，2002 年 10 月 10 日（在网站 <http://www.falconstor.com/Whitepapers/MidrangeSSFSolutionWhitePaper.pdf>，2004 年 6 月 14 日互联网上可以找到）和 PCT 专利申请 No. WO02/069159 公开了一种镜像技术，其中，提供一种位于 PDC 并且在运行软件应用的主计算系统和本地存储单元之间插入的设备。远程存储单元上的拷贝总是通过位于 PDC 处的设备而异步产生。

[0038] 总之，申请人认为，现有技术中的解决方案不能用于实现独立于 PDC 和 RDC 之间的距离并具有很低的 RPO 的同步镜像。

发明内容

[0039] 因此，本发明的目的在于提供一种用于通过通信网络更新数据集，从而允许实现与距离无关的同步镜像技术的系统、方法和设备。

[0040] 本发明的另一目的在于，在任意情况下在向客户确保可与同步镜像技术比较的 RPO 和 RTO 的同时，即使所述客户不处于集总式恢复数据中心的相同城域中，也允许具有例如集总式恢复数据中心的服务提供商向各自具有相应主数据中心的大量客户或客户机提供灾难恢复服务。

[0041] 本发明的另一目的在于提供能够管理为了实现根据本发明的方法所执行的操作的计算机程序或一组计算机程序产品。

[0042] 通过如所附权利要求所要求权利的系统、方法、计算机程序产品以及网络来实现本发明的上述目的。

[0043] 根据本发明，提供一种用于实现从主数据中心 PDC 到与恢复数据中心 RDC 关联的远程盘或存储器支持的数据集的同步更新的系统和方法，其中所述恢复数据中心 RDC 在地理上位于例如离开 PDC 很大距离的地方，其中，所述距离例如大于能够保证 RPO 在合理限制之下的距离。

[0044] 根据本发明，提供一种临时存储设备，具有用于存储由 PDC 生成的数据块的临时存储区域，其位于例如 PDC 和 RDC 之间的中间站点。

[0045] 所述临时存储设备配置有预定存储容量，并且包括智能缓存软件程序，其能够向客户机处理系统或 PDC 同步地提供成功写入的确认信号，即，在十分短的时间间隔内并且可与同步镜像技术的时间间隔（例如可表示为 1ms 内）相比较的确认信号。

[0046] 根据本发明，多个客户机处理系统（客户机）和 / 或主数据中心可共享临时存储设备的存储区域。

[0047] 根据本发明的另一特点，独立于将要更新的数据卷的大小来确定 在临时存储设备的存储区域中分配给客户机的缓存大小。

[0048] 具体地说，根据本发明优选实施例，取决于以下条件来确定缓存大小：

[0049] - 在客户机处理系统和临时存储设备之间使用的带宽；

[0050] - 在临时存储设备和恢复站点 RDC 之间使用的带宽；以及

[0051] - 在 PDC 处的客户机软件应用产生的数据速率，其中，例如，可将所述数据速率量化为在单位时间内产生或修改的数据的数量。

[0052] 总之，本发明相对于现有技术提供以下优点：

[0053] - 对于在小于主数据中心 PDC 和临时存储区域之间的距离的感兴趣的范围内的故障或损坏，也就是当未损坏存储设备时，对呈现出等于单跳同步镜像技术的 RPO（在本地盘和远程盘之间没有错误校准）的 RPO 进行优化；

[0054] - 对于还影响中间站点的事件，RPO 十分接近于零（受限于很少的应用事务处理的错误校准）；在此情况下，存在于中间站点中并且尚未被拷贝到远程盘的数据丢失，但根据本发明中公布的传送数据的方式，丢失的数据实际上是非常有限的量；

[0055] - RTO 很低，并且可与可通过在主数据中心和恢复数据中心之间应用的同步盘镜像技术获得的 RTO 相比较；

[0056] - 在就 RPO 和 RTO 而言提供同等可靠性的同时，独立于距离或关于本地盘和远程盘之间的距离没有限制，而在单跳同步镜像技术中情况恰好相反；

[0057] - 由于本发明提出的方法不需要中间站点包含主数据中心的本地盘的整个数据集，因此有限量的数据存储在中间站点中。

附图说明

[0058] 以下将参考本发明优选的但非限定性实施例的附图来公开本发明，其中：

[0059] 图 1 是示出根据本发明的盘镜像系统的框图；

[0060] 图 2 是根据本发明的设备的框图；

[0061] 图 3 是图 2 的设备的第一实施例的框图；

[0062] 图 4 是图 2 的设备的第二实施例的框图；

[0063] 图 5A、5B 和 5C 分别是由图 2 的设备进行的离台（destaging）、写入和读取过程的流程图；

[0064] 图 6 示出是图 1 的盘镜像系统的可能的故障状态的流程图；

[0065] 图 7 示例性示出根据本发明的存储服务提供商可实现镜像服务的情况；

[0066] 图 8 是示出客户机系统的可能的故障状态的流程图；

[0067] 图 9 和图 10 示出其中指出根据本发明的镜像服务的特征参数的网络；

[0068] 图 11 示出以批发形式提供镜像服务的网络。

[0069] 在所有附图中，相同标号已用于表示相同或基本实现等同功能的组件。

具体实施方式

[0070] 虽然现将参照盘镜像技术描述本发明,但要注意,可在进行无镜像应用的不同环境中,例如在本地盘的故障的情况下或不提供数据集的本地拷贝的那些应用中,成功实现相同的发明原理。

[0071] 图 1 示出根据本发明的盘镜像系统,其中,提供在一侧的处理系统或客户机系统 10(其中,软件应用正在运行)及其相关主数据中心 PDC 的本地盘单元 11 和在另一侧的恢复数据中心 RDC 的远程盘单元 12 之间放置的临时数据储存设备 13。设备 13 放置于例如本地盘 11 和远程盘 13 之间的地理上中间的站点,其中,例如中间站点和 RDC 之间的距离优选地大于 PDC 和中间站点之间的距离。

[0072] 根据本发明,术语“距离”表示例如用于将 PDC 连接到中间站点和用于将中间站点连接到 RDC 的光纤或通信线缆的长度。

[0073] 参照图 2,以模块 20-27 详细解释临时储存设备 13 的架构。

[0074] 根据本发明优选实施例,以软件实现模块 20-27,但本领域技术人员可理解,可通过包括集成电路或可编程逻辑的硬件模块实现这些 模块。

[0075] 模块 20 和 26 构成分别朝向主数据中心 PDC 和恢复数据中心 RDC 的前端。模块 20 和 26 可以是比如网络接口卡或 NIC 的两个等同的对象、例如单个 NIC 的相同对象、或是不同对象,其中比如模块 20 可以是 NIC 而模块 26 可以是主机总线适配器 HBA(反之亦然)。模块 20 和 26 位于传输协议的物质终接器设备中,这是管理通过网络的信息交换的设备。

[0076] 布置在模块 20 的下游的模块 21 包括协议的目标设备,用于远程镜像数据。目标设备 21 从主数据中心 PDC 接收 I/O 命令,并与模块 22 协同运行。目标设备执行由发起器设备(也就是请求 I/O 处理的设备)请求的操作。

[0077] 布置在模块 21 的下游的模块 22 实现软件过滤器,用于截取 I/O 命令并判断是否通过其余模块 23-26 将它们转发到缓存系统模块 27 或 RDC。

[0078] 布置在模块 22 的下游的模块 23 是虚拟模块,具有呈递给 PDC 逻辑存储区域的任务,所述 PDC 逻辑存储区域不同于为存在于 RDC 中的盘卷预留的物理存储区域。

[0079] 布置在模块 23 下游的模块 24 具有下述任务,即,用于接收具有数据块相对地址参数的数据块以及将数据块封装在为管理 RDC 中的存储盘预留的命令中。

[0080] 布置在模块 24 下游的模块 25 是发起器设备,与模块 26 协同运行。

[0081] 参照图 3,现将描述当使用用于远程传输数据块的 iSCSI/SCSI 协议(互联网小型计算机系统接口 / 小型计算机系统接口)时根据本发明的临时数据储存设备 13。

[0082] 可由传输协议终接器(例如 NIC)和用于远程存储数据的协议的目标设备(例如 iSCSI 目标设备的修改版本)分别表示模块 20 和 21。在此情况下,模块 22 是 SCSI 目标设备的修改的商用或免费版本。所述修改基本上包括:构建软件过滤器,其用来截取用于确定的逻辑单 元号 LUN 的 SCSI 命令并将所述命令寻址到缓存系统 27 或设备 13 的其余模块 23-26,以将命令转发到 RDC。

[0083] 在基于 Linux 操作系统的实现的情况下,这种互连模块 23-26 的链包括逻辑卷管理器 LVM 23。其它的模块 24-26 通常是可用的标准模块,在 Linux 中,它们可包括本地 SCSI 子系统或控制单元 24,与 iSCSI 标准发起器 25 和 NIC 26 协同运行。

[0084] 当在 Linux 环境中使用逻辑卷管理器 23 时,需要修改的 SCSI 目标设备 22 提供与

LVM 23 兼容的 I/O 接口。

[0085] 根据本发明的另一实施例，可避免逻辑卷管理器 23。

[0086] 然而，根据优选实施例，提供逻辑卷管理器 23 是有用的，因为它允许创建寻址下述存储区域的逻辑分区，所述存储区域属于例如可以是地理上分离的多个恢复数据中心。

[0087] 本领域技术人员可理解，LVM 23 可由软件虚拟设备来代替。

[0088] 在图 3 中提出的设备 13 仅仅是可用于设备 13 的可能的协议的一个示例，其中，SCSI 是存储管理协议，iSCSI 是用于在部件 PDC- 设备 13 和部件设备 13-RDC 中远程存储数据的协议。

[0089] 图 4 示出设备 13 的第二可能架构。在该版本中，SCSI 仍旧是存储管理协议，但光纤信道协议用于远程传输数据。

[0090] 在该实施例中，由光纤信道主机总线适配器 (HBA) 表示网络接口，从功能性的观点来看，所述 HBA 替代与图 3 中描述的实施例有关的 NIC 和 iSCSI 目标 / 发起器的集合。

[0091] 然而，还可能有混和解决方案，所述解决方案提供用于在由设备 13 分离的两个网络部分中远程存储数据的不同协议，其或者不使用 SCSI 作为存储设备的管理协议，或者仍旧提供像在 RDC 多于一个并且以不同技术被管理的情况下那样的较多通信技术。混和解决方案也影响由图 2 的模块 24 管理的协议类型。

[0092] 设备 13 可实现为专用器件或网络设备（例如路由器、光纤 / 数字交叉连接 (ODXC)、FC 交换机等）中的一组功能或模块。

[0093] 再次参照图 1，其中，箭头表示进行的操作，现将描述盘镜像系统 200 如何工作。

[0094] 在下面，术语“数据块”表示以下情况的逻辑关联：

[0095] • 将要在 RDC 中存储的信息（数据）；

[0096] • I/O 操作属性，例如目的地存储系统上的存储器位置的地址或地址集；

[0097] • 将连同所述信息一起存储的附加数据属性。

[0098] 下面，由在处理系统 10 上运行的软件应用来发出 I/O 请求 1，安装在主数据中心 PDC 中的镜像软件 14 将 I/O 请求 1 的副本 2 发送到本地盘 11 和设备 13。

[0099] 本地盘 11 和临时储存设备 13 通过将合适的确认信号 3 发送到镜像软件 14 来应答 I/O 请求 2，镜像软件 14 接着将对应的确认信号 4 发送到所述处理系统 10。其后，临时数据储存设备 13 负责将数据块 5 传送到恢复数据中心的远程盘 12，并管理有关的确认信号 6。

[0100] 因此，临时数据储存设备 13 具有以下任务：

[0101] – 通过盘镜像软件 14 将用于确认在设备 13 中接收到并写入数据项的信号 3 返回给软件应用，从而不阻塞或减缓运行所述软件应用的正常方式；

[0102] – 临时存储将要传送到远程盘 12 的数据块；

[0103] – 将数据块 5 异步传送到重新构建本地盘 11 的整个数据集的远程盘 12。

[0104] 主数据中心 PDC 和恢复数据中心 RDC 分别具有下述相同功能，当使用传统单跳盘镜像技术和引入临时数据储存设备 13 时，所述功能不改变它们运行的正常方式从而设备 13 实际上对软件应用是透明的。

[0105] 设备 13 可应用于网络中，其中，由相同管理者（例如公司）来管理主数据中心 PDC 和恢复数据中心 RDC，或在由存储服务提供商 SSP 管理 RDC 并且由购买由服务提供商提供的数据副本服务的用户公司拥有主数据中心 PDC 的服务情况下应用设备 13。

[0106] 此外,连接到相同主数据中心 PDC 的多个客户机可共享相同临时储存设备 13。

[0107] 至于有关架构(图 1)和临时数据储存设备 13(图 2)的操作,可识别三种主要模式(或阶段):

[0108] 1、设置阶段:这是服务于新的客户机 / 应用的设备 13 的通用架构;

[0109] 2、正常操作阶段:包括离台处理和过滤 I/O 命令;

[0110] 3、从故障事件阶段恢复。

[0111] 现将描述设置阶段,包括启用新的客户机 / 应用以使用根据本发明的设备所需的所有操作。提供至少两个不同步骤:

[0112] a) 通过临时数据储存设备 13 为 PDC 和 RDC 以透明方式启用存在于 RDC 中的物理卷盘的使用;

[0113] b) 如果在 PDC 和 RDC 之间实现同步镜像技术,则启用本发明的功能的使用。

[0114] 至于设置阶段的步骤 a),本领域技术人员可理解,新的客户机知道它通过合适地配置临时数据储存设备 13 的模块 23 所访问的逻辑卷的映像。

[0115] 对于作为存在于 RDC 中的物理分区的逻辑映像的所述逻辑卷,它被关联到由用于访问所述逻辑卷的客户机应用所使用的标识符。具体地说,设备 13 内的卷的逻辑映像对于发送用于错误地址的错误消息而不等待来自最终目的地的否定响应可能是有用的。

[0116] 至于步骤 b),实现所述步骤 b),从而客户机应用可通过设备 13 使用远程物理盘卷作为在其驻地存在的盘卷的同步映像(本地拷贝),在优选实施例中,提供以下功能:

[0117] - 将卷的一致拷贝传送到远程站点(第一同步);

[0118] - 为在设备 13 的缓存系统模块 27 中的所述客户机预留存储区域;

[0119] - 所述存储区域参考用于描述由客户机寻址的卷的特征的配置数据(地址集、卷大小、块大小等),所述配置数据存储在缓存系统模块 27 的内部数据库 27b 中;

[0120] - 在模块 22 内启用由客户机镜像软件发出的命令的过滤过程,从而通过模块链 23-26 朝向 RDC 来转发这些命令中的一些命令(例如清除远程盘单元的头、格式化盘等),而截取到一些其它命令(例如写入数据或读取数据)并将其转发到缓存系统模块 27。

[0121] - 通知缓存控制软件 27a 已经启用新的客户机并且已经将缓存系统模块 27 内的特定存储区域分配给所述客户机。

[0122] 在设置阶段之后进入正常运行阶段。该阶段至少包括两个主要过程:

[0123] 1、离台过程,用于将存储在设备 13 的缓存中的数据移动到 RDC;

[0124] 2、I/O 命令过滤过程,截取读取和写入操作并相关地管理设备 13 的缓存。

[0125] 参照图 5A、5B 和 5C,将详细描述控制涉及缓存系统模块 27 的离台 / 写入 / 读取操作。

[0126] 图 5A 表示示出在图 2 的缓存控制软件 27a 内部实现的离台过程如何运行的流程图。

[0127] 缓存软件 27a 首先执行与不同客户机对应的所有缓存存储区域的连续轮询(步骤 101),并检查缓存 27 中是否存在尚未转发到 RDC 的一些数据块。

[0128] 在由轮询器在缓存中识别数据块之后,在将所述数据块发送到 RDC 之前,所述过程可等待最大可配置时间(等待时间)。轮询周期、等待时间和在设备 13 和 RDC 之间的网络连接上发送数据块所需的时间(串行化时间)可以是可由管理器通过下述限制配置的参

数,所述限制即它们的和优选地必须小于用于要处于稳定状态的缓存的两个数据块之间的平均内部到达时间。

[0129] 此外,数据块到 RDC 的转发可遵循用户定义的策略,以修改数据块离台的优先级,例如让特定客户机或应用比其它客户机或应用处于有利位置(优先化)。总之,优先化策略必须不与上述关于离台延迟的考虑相冲突。

[0130] 必须转发到 RDC 的所有块被发送到 RDC(步骤 103)。

[0131] 一旦已经发送数据块,就不从缓存 27 按先后顺序擦除数据块,而是仅当由 RDC 发出的确认信号确认所述数据块实际上未受损地到达那里时将其擦除(步骤 107)。

[0132] 在此状态下,数据块被标记为“正在离台”,用于选择将要转发到 RDC 的数据块的过程负责避免将具有与已经处于“正在离台”的数据块相同的地址的另一数据块放置为“正在离台”。完成该过程以避免与以下描述的写入过程干扰。

[0133] 实际上,为了防止数据丢失,所述数据块仍旧存储在缓存 27 中,直到 RDC 实际已经存储从缓存发送的数据块。换句话说,当将数据块从设备 13 传送到 RDC 时存在一种用于防止数据丢失的传递控制。所述控制在图 5A 中由布置在发送块 103 之下的块来表示。

[0134] 在任何情况下,如果设备 23 和 / 或 24 采用传递已知类型的控制的某些机制,则这些机制应当与由所述缓存采用以避免干扰的机制相协调。

[0135] 当分析可能的状态时,数据块可能:

[0136] a) 无差错地到达 RDC;

[0137] b) 到达 RDC 但受损;

[0138] c) 没有到达 RDC。

[0139] 为了识别上述状态 a)、b) 和 c),在缓存控制软件 27a 中实现的过程提供:

[0140] - 在 a) 的情况下,提供肯定确认信号 ACK(步骤 111,向下箭头);

[0141] - 在 b) 的情况下,如果可由用于远程传送数据的协议正确地解释否定确认信号 NACK,则提供 NACK(步骤 111,向右箭头);

[0142] - 为了识别状态 c),设置超时延迟,所述超时延迟大于设备 13 和 RDC 之间的往返时间的估计;一旦所述超时延迟已经过去(步骤 109),并且肯定 ACK 或否定 NACK 确认信号尚未到达(步骤 107,向左箭头),则认为由缓存 27 发送的数据块丢失。当发送所述数据块时开始所述超时延迟(步骤 105)。

[0143] 在肯定确认信号之后的控制块 111、115 和 119(步骤 111,向下箭头)管理上述情况 a),以及控制块 111、113 和 117 管理情况 b),而由控制块 109 执行超时延迟,一旦预定超时延迟已经过去,控制块 109 假定 RDC 尚未接收到数据块。

[0144] 在优选实施例中,提供设置在与远程站点连接中的受损 / 丢失的连续块的最大数量 N_{MAX} ;如果达到这个最大数量 N_{MAX} (步骤 117),则缓存 27 通知客户机停止镜像服务(步骤 121),从而阻止把将要传送的其它数据从 PDC 发送到缓存。

[0145] 事实上,如果 RDC 不可到达,则缓存 27 通过客户机将连续接收数据,而不可能将它们转发到 RDC,并且因此缓存 27 将很快过载。

[0146] 在其它实施例中,在用于远程传递数据的协议不提供否定确认信号 NACK 的情况下,还对受损的数据块应用相同的超时延迟过程,也就是说,在已经到达连续受损的块的预定数量之后,停止镜像服务。

[0147] 图 2 的模块 22 实现的 I/O 命令过滤器过程：

[0148] • 解释从 PDC 到 RDC 的命令；

[0149] • 根据特定使用的协议考虑一个或多个 PDU（协议数据单元）而识别操作类型（写入、读取，其它操作）。

[0150] 在 I/O 命令不同于写入和读取的情况下，设备 13 将它们转发到 RDC。在写入和读取命令的情况下，第一步骤包括提取 I/O 操作的参数：

[0151] • 在写入命令的情况下，所述参数是例如整个数据块（也就是将要存储的数据）、操作属性（例如存储器位置的地址）以及附加数据属性；

[0152] • 在读取命令的情况下，由于将要获得数据，因此所述参数仅是例如操作属性。

[0153] 按以下描述的那样处理所述操作。

[0154] 参照图 5B，现将描述如何由缓存软件 27a 来管理由图 1 的镜像软件 14 发出并被转发到图 1 的设备 13 的写入命令。

[0155] 在设备 13 内部，由图 2 的块 22 截取由镜像软件 14 发出的写入命令，并根据图 5B 中描述的过程分析所述写入命令。

[0156] 在已经由缓存截取写入命令（步骤 151）之后，识别 I/O 操作的有关的参数（数据块）（步骤 152）。

[0157] 根据优选实施例，对于关注的将在缓存 27 中写入的数据块以及所述缓存内部的状况，提供四种可能的状态。本领域技术人员应理解，状态的数量可以小于四种或大于四种。

[0158] 具体地说，根据优选实施例的状态是：

[0159] - 情况 1)：如果具有请求的地址的数据块没有出现在缓存 27 中（步骤 153），则使用对所述客户机可用的新的缓存位置，并在所述新的位置写入所述数据块（步骤 161）；

[0160] - 情况 2)：如果具有请求的地址的数据块已经出现在缓存 27 中（步骤 153）并且其不是“正在离台”，则新的数据块替代现有旧的数据块（步骤 157）；

[0161] - 情况 3)：如果具有请求的地址的数据块已经出现在缓存 27 中（步骤 153），但处于“正在离台”（步骤 155）（也就是说，系统仍旧等待根据恢复数据中心 RDC 的确认信号）并且具有相同地址的第二数据块没有出现在缓存 27 中（步骤 159），则将所述数据块写入新的缓存位置（步骤 161）；

[0162] - 情况 4)：如果具有请求的地址的两个数据块已经出现在缓存中（所述数据块中的一个“正在离台”，而另一个不是“正在离台”），则新的块替代不是“正在离台”的现有的块（步骤 157）。

[0163] 在任意情况下，在已经完成数据块的记录之后，写入过程以信号将成功操作通知给 I/O 命令过滤器（图 2 的模块 22），所述 I/O 命令过滤器接着将当前使用的 I/O 协议所需的 ACK 信号返回到 PDC 中的镜像软件。

[0164] 如果出于任意原因（例如缓存没有空闲容量）而不能在缓存中记录数据块，则不返回 ACK，由 PDC 阻止发送进一步的写入命令，从而避免了将缓存填充超过其最大容量，并且避免了丢失数据块。

[0165] 图 5C 示出在图 2 的缓存控制软件 27a 中实现的过程如何管理由镜像软件发出给 RDC 中的远程盘并且因此到达图 1 的设备 13 的读取命令。

[0166] 虽然能够从缓存 27 读取数据,但应指出,缓存优先地没有替代 RDC 的任务。

[0167] 在设备 13 内,由图 2 的块 22 截取由镜像软件 14 发出的读取命令,并由图 5C 中描述的过程来分析该命令。

[0168] 在更新的数据存在于缓存 27 中的情况下(“正在离台”或不是“正在离台”),或在数据仅存在于 RDC 的情况下,读取处理的管理关注于确保在每一时刻运行在图 1 的处理系统 10 中的软件应用接收更新的数据。

[0169] 在已经由缓存截取读取命令(步骤 201)之后,识别有关的 I/O 操作的参数(例如存储器位置的地址)(步骤 202)。

[0170] 根据优选实施例,对于关注的将要读取的数据块,提供四种可能的状态。本领域技术人员应理解,状态的数量可以小于四种或大于四种。

[0171] 具体地说:

[0172] -情况 1):如果具有请求的地址的数据块没有出现在缓存 27 中(步骤 203),则通过图 2 的模块 23-26 将读取操作转发到 RDC;

[0173] -情况 2):如果具有请求的地址的数据块已经出现在缓存 27 中(步骤 203)并且其不是“正在离台”,则从所述缓存读取数据块,并将数据块返回 PDC(步骤 207);

[0174] -情况 3):如果具有请求的地址的数据块已经出现在缓存 27 中(步骤 203),但处于“正在离台”(步骤 205)(也就是说,系统仍旧等待根据 RDC 的确认信号)并且具有相同地址的第二数据块没有出现在缓存 27 中(步骤 209),则所述读取操作返回“正在离台”的数据块的数据内容(步骤 211);

[0175] -情况 4):如果具有请求的地址的两个数据块出现在缓存中(所述数据块中的一个“正在离台”,而另一个不是“正在离台”),则返回不是“正在离台”的数据块的数据内容(步骤 207)。

[0176] 现将参照图 6 描述从故障事件状态的恢复。

[0177] 图 6 示出可能的故障状态以及可以相应地采用的一些行为。具体地说,从正常操作的初始状态中(块 251),可设想至少两个异常的分类:

[0178] a) 在客户机和临时数据储存设备 13 之间的连接故障或设备 13 自身的故障(块 253),由此不能在中间站点的图 2 的缓存 27 中写入;

[0179] b) 在设备 13 和恢复数据中心 RDC 之间的连接故障或 RDC 自身的故障,由此不能访问 RDC。

[0180] 图 6 的流程图的左边分支详细解释了情况 a),可替换的是:

[0181] -在绝对要求根据本发明的同步传送的情况下,中断盘镜像(步骤 257);

[0182] -一旦已经恢复连接(步骤 267),就临时将数据存储在主数据中心的本地盘中(步骤 265)并完成将数据块传送到缓存 27,从而传送到 RDC;

[0183] -旁路不可用的缓存,并执行直接与 RDC 的异步镜像。

[0184] 图 6 的流程图的右左边分支详细解释了情况 b),可能的是:

[0185] -如果缓存具有充足的存储容量,并且估计不可用性在时间上非常短暂,则不进行操作(步骤 261),并在等待将要恢复的连接的同时在缓存中继续存储事务处理,从而通过使用图 5A 的离台过程在 RDC 中校准盘(步骤 269);

[0186] -作为可替换的是,将出现在缓存中的数据拷贝到可替换的站点(步骤 263),从而

如果不可用性在时间上相当长，则不饱和所述缓存。

[0187] 再次返回图 3，其中，当客户机希望使用设备 13 时，SCSI 是存储管理协议，iSCSI 是用于在部分 PDC 设备 13 和部分设备 13-RDC 中远程存储数据的协议，在启动阶段期间，它接收 LVM 中的逻辑卷的映像，可通过 iSCSI 协议访问所述 LVM。

[0188] 对于作为基于存在于 RDC 中的物理分区的逻辑映像的所述卷，它将根据将由用于访问它的客户机应用使用的 SCSI 技术而与具体 LUN(逻辑单元号)关联。

[0189] 此时，图 1 所示的通过镜像软件 14 与设备 13 交互的客户机应用可经由 iSCSI/SCSI 协议访问 RDC 站点远程卷。

[0190] 到此，为了应用可通过设备 13 将远程卷使用作为 PDC 中的逻辑卷盘的同步映像，优选地：

[0191] - 将逻辑卷盘的一致拷贝传送到远程站点（第一次同步）；

[0192] - 为客户端应用的 LUN 在设备 13 的缓存 27 中保留存储区域；

[0193] - 所述存储区域参考描述所述 LUN 卷的特性的配置数据（地址集、卷大小、块大小等），并且被存储在缓存系统模块 27 的内部数据库 27b 中；

[0194] - 在模块 22 内能够过滤由客户机镜像软件发出的 SCSI 命令的过程，从而通过模块链 23-26 朝向恢复数据中心转发这些命令中的一些，而截取一些其它命令并将其朝向缓存系统模块 27 转发。

[0195] - 通知缓存控制软件 27a 已经启用新的客户机，并且已经将缓存系统模块 27 内的 LUN 描述器和特定存储区域分配给存储区域客户机。

[0196] 具体地说，可用四个步骤来连接通过使用 SCSI 协议的从 PDC 到 RDC 的写入操作：

[0197] a) 发起器设备将写入请求发送到目标数据，以写入数据块的数量 N；

[0198] b) 目标数据将确认信号发送到发起器设备；

[0199] c) 在接收到确认信号之后，发起器发送 N 个数据块；

[0200] d) 目标设备确认成功写入 N 个数据块。

[0201] 必须由修改的 SCSI 目标设备 22 来执行任务 b) 和 d)。更具体地说：

[0202] - 修改的 SCSI 目标木刻 22 从主数据中心中的镜像应用（以及结合至其的 SCSI 发起器）接收用于写入特定数量的 N 个数据块的写入请求；

[0203] - 模块 22 截取写入命令，作为写入操作，识别所述命令是涉及缓存系统 27 的这些命令中的一个：因此验证缓存系统是否能够存储这些数据块；在肯定的情况下，模块 22 根据 SCSI 标准通过将数据请求消息发送到镜像应用来应答；

[0204] - 镜像应用发送所述 N 个数据块；

[0205] - 模块 22 检索所述数据块的目的地址，并将它们转移到专用于发送数据的客户机的缓存 27 的存储区域：更具体地说，模块 22 将从 SCSI 提取块，并根据数据内容先前已经指示的内容而在缓存中存储 SCSI 协议的写入命令的头 (DPO(禁用页面输出)、FUA(强制单元访问) 等) 的有意义的信息和卷内部的地址；

[0206] - 缓存软件 27a 接管进行离台过程并将写入请求转发到已经描述的 RDC。

[0207] 此外，按照相同的方式，将由缓存 27 截取读取命令，因为在否定情况下，由于没有更新的数据项仍旧出现在缓存 27 中，所以能够从 RDC 读取没有更新的数据项。

[0208] 因此，当读取请求到达模块 22 时，模块 22 通过缓存管理软件 27a 进行控制，以根

据图 5C 中描述的程序来检查块是否出现在缓存中。

[0209] 如前所述,缓存大小无需太大,相反,本发明的一个有利之处在于,可将其保持为显著地较小。以下描述将解释当使用典型网络参数时如何估计所需缓存大小来分配客户机 / 应用。

[0210] 这种估计是基于这样的假设的:设备 13 运行在一个或多个新的数据块从 PDC 的到达与它们转发到 RDC(离台)之间的平衡状态中。这种平衡状态通过先前描述的写入和离台过程来保证。因此,这种估计对于测量设备 13 尺寸是有用的工具,但对于其稳定操作则不是必要条件。

[0211] 如果在缓存 27 和恢复数据中心 RDC 之间传输的数据块没有丢失或受损,则其保留在缓存 27 内部达到由串行化时间加上缓存和 RDC 之间的往返时间来确定的一段时间间隔,也就是从 RDC 接收确认信号所需的时间间隔。在已经接收到所述确认信号之后,由于已经成功将数据块存储在 RDC 中,因此可从缓存 27 中将其擦除。

[0212] 可确定用于单个客户机 / 应用的缓存中出现的数据块的平均数量 的估计 (作为 PDC 写入操作的后续),例如,

$$[0213] N_{\text{CACHED}} = [(T_{\text{RT2}} + T_s) / (1/R)] + [T_{\text{MAX}} / (1/R)] + [T_{\text{RT1}} / (1/R)]$$

[0214] 其中,方括号表示将结果向上取整为整数 (下面定义该公式中的不同参数的意义)。这种增加合适的裕量来考虑参数值的波动的块的平均数量 N_{CACHED} 的估计可用于测量已经定义了单个数据块的大小的设备 13 的缓存存储器的尺寸。上述公式对于 N_{CACHED} 包括 3 次相加:

[0215] - 第一相加, $[(T_{\text{RT2}} + T_s) / (1/R)]$, 表示到达缓存的块的数量,并可被确定为串行化时间 (T_s)、设备 13 和 RDC 之间的网络部分的往返时间 (T_{RT2}) 以及表示两个连续数据块之间的平均内部到达时间的 $1/R$ 的函数;

[0216] - 第二相加, $[T_{\text{MAX}} / (1/R)]$, 表示出于某些原因在指定时限 (超时) 内尚未到达 RDC 的“正在离台”的块的数量。可将其确定为 T_{MAX} 的函数, T_{MAX} 表示在将导致丢失的定义数据块的设备 13 的接收与由设备 13 管理超时事件的最后的块的超时之间逝去的时间间隔。可将 T_{MAX} 确定为例如:

$$[0217] T_{\text{MAX}} = T_{\text{OUT}} + T_s + (N_{\text{MAX}} - 1) \cdot \max(1/R, T_s)$$

[0218] 其中, T_{OUT} 是可设置的参数,表示从设备 13 发送到 RDC 的单个块的超时间隔, N_{MAX} 是可设置的参数,表示由设备 13 管理超时事件的块的最大数量;

[0219] - 第三相加, $[T_{\text{RT1}} / (1/R)]$, 表示在缓存已经将盘镜像服务临时可用通知给客户机之前由客户机发送到缓存的块的数量, T_{RT1} 是在 PDC 和设备 13 之间的网络部分的往返时间。

[0220] 申请人的实验已经证明,在前面报告的类型的公式的应用给出结果:对于能够每秒产生几十次事务处理的每一客户机 / 应用,缓存存储器的大小在从几百 kB 到几 MB 的范围内。

[0221] 因此,通过应用本发明的架构,无需在中间站点复制一个或多个在 PDC 中出现的整个数据集,与多跳架构相反,能够节省盘存储和对应的维护和人力成本。

[0222] 涉及模块 27、27a 和 27b 的可能的改进包括个性化位于中间站点 并与客户机应用关联的盘存储的哪些区域具有特别频繁的访问 (称为“热点区域”),从而关于较少访问的区域而不同地处理“热点区域”;例如,仅当所述“热点区域”处于稳定状态并且不连续时可

以将它们传送到 RDC, 从而在设备 13 和 RDC 之间的网络部分中允许可观的带宽节省。

[0223] 参照图 7, 现将解释服务提供商 (SP) 可以如何使用本发明。更具体地说, 值得一提的是, 本发明允许 SP 增加可提供同步镜像服务的区域的数量, 并且同时限制 RDC 的数量。

[0224] 根据上述情形, SP 提供在每一城市区域中包含设备 13 的站点以及远离所述城市区域的地理区域中的一个或多个 RDC 12。本领域技术人员应理解, 设备 13 无需是单机设备, 而可以集成在网络装置中。

[0225] 图 8 示出在灾难 / 恢复情形中 (具体地说, 当客户机盘经历故障 (步骤 301) 并且触发真实恢复过程的情况下时执行的操作) 本发明的使用。恢复过程可根据不同方式而产生, 但所有恢复过程必须早于具有仍旧存储在缓存 (离台) 中的数据的 RDC 盘的重新校准。更新可通过以下方式来产生: 需要恢复过程的客户机数据的“标记”过程 (步骤 305), 从而它们具有相对于其它客户机的数据的优先级, 或通过让所述中间站点忽略存在的故障情况, 从而其可继续至 RDC 的正常传送行为。

[0226] 用于恢复客户机数据和应用的主要替换为:

[0227] - 通过使用在离台到 RDC 之后可用的数据, 能够在 RDC (经典灾难 / 恢复方法) 重新开始客户机应用 (307);

[0228] - 通过以高比特率可能地提供替换连接的激活, 在 PDC 的新的盘上从 RDC 中的副本恢复客户机数据;

[0229] - 临时启动对 RDC 数据的访问作为主盘。

[0230] 在灾难 / 恢复状态之后, 修复客户机 PDC (309), 并以用于受所述故障影响的客户机的新的设置阶段 (311) 作为开始, 在此创建同步镜像副本。

[0231] 本发明还允许创建用于支持同步数据副本服务的原始网络服务。网络服务的可能的客户的简档基本上是下述主体的简档, 所述主体想要将信息从第一站点同步复制到位于离开所述第一站点相当长距离 (几百公里) 的第二站点, 所述主体不具有位于第一和第二站点之间的中间距离的可用性, 或不愿意管理中间存储装置。

[0232] 具有这些需求的客户可以是公司 (“零售客户”), 其期望通过使用同步镜像技术来保护其自身的数据, 或是服务提供商 SP, 其倾向于将灾难恢复和 / 或商业连续性服务提供给其最终客户。

[0233] 可将所述特定网络服务定义为 “长距离同步镜像加速”。实际上, 通过所述服务, 并且由于由设备 13 返回的确认信号, 从客户机应用的观点来看, 在数据项的本地拷贝及其远程拷贝之间的同步复制如同在两个站点之间的距离短于真实距离那样而产生。

[0234] 这样允许进行同步镜像, 表明阻塞数据产生, 直到接收到先前数据的成功写入的确认信号, 而两个站点之间的真实距离不会对阻塞应用来等待确认信号的时间产生影响并对其性能产生影响。

[0235] 所述服务对于客户机完全透明, 这说明关于同步镜像的经典配置, 客户机不修改其系统和其运行方式。

[0236] 图 9 示出服务架构, 其中, 所述数据的两个拷贝存储在 PDC 和 RDC 中。由最终客户和 / 或 SP 来管理这两个站点, 因此不是由提供加速服务的运营商来管理所述两个站点。图 9 指出在运营商管理 (也就是网络基础架构, 例如城域网或 MAN 以及广域网或 WAN) 下的单元和加速服务的边界, 以及位于中间站点 (例如运营商的城市交换机) 的设备 13。

[0237] 图 9 和图 10 还指出服务主要特点,具体地说:

[0238] -L1 和 L2 个性化受保护不受灾难的两个区域。对于具有低于 L1(灾难区域 A1) 的扩展范围的事件,也就是设备 13 不感兴趣的事件,所述服务确保 RPO 等于一个事务处理;对于具有低于 L2(灾难区域 A2) 并且高于 L1 的扩展范围的灾难,取决于缓存离台策略,所述服务提供:可能丢失的数据最多是当发生灾难时由设备 13 传递的那些数据;

[0239] -T1 和 T2 分别定义写入确认产生的最大时间间隔和在 RDC 中完成数据项的拷贝的最大时间间隔。

[0240] 基于所述四个参数 L1、L2、T1、T2,定义客户的服务的服务级别协定 SLA。具体地说,应注意的是,在灾难影响灾难区域 A2 的情况下,丢失事务处理的数量等于在最大时间间隔 T2 期间由客户机应用处理的那些事务处理的数量。

[0241] 总之,服务提供:

[0242] - 在低于传统同步镜像技术的特征时间的固定时间限制 T1 内远程拷贝命令的至少一个,例如 $T1 < 0.5$ 毫秒;

[0243] - 甚至大于在本地或最接近的同步镜像技术的情况下允许的最大距离的远程拷贝的距离 L2,例如 $L2 >> 50\text{km}$;

[0244] - 在最大时间间隔 T2 内在远程盘可用的数据项的拷贝,取决于城域网 WAN 的分配的带宽和离台策略,最大时间间隔 T2 可呈现出十分接近于覆盖距离 L2 的整个传播时间;

[0245] - 对于具有低于 L1 的范围的灾难,确保最多丢失一个事务处理,并且对于在范围 $L1 < R < L2$ 内的故障,确保最多丢失在时间间隔 T2 期间产生的事务处理。

[0246] 图 11 示出运营商以批发形式提供的加速服务。

[0247] 主要特点为:

[0248] - 在由运营商拥有的交换机中在城域网 MAN 中使设备 13 离位;

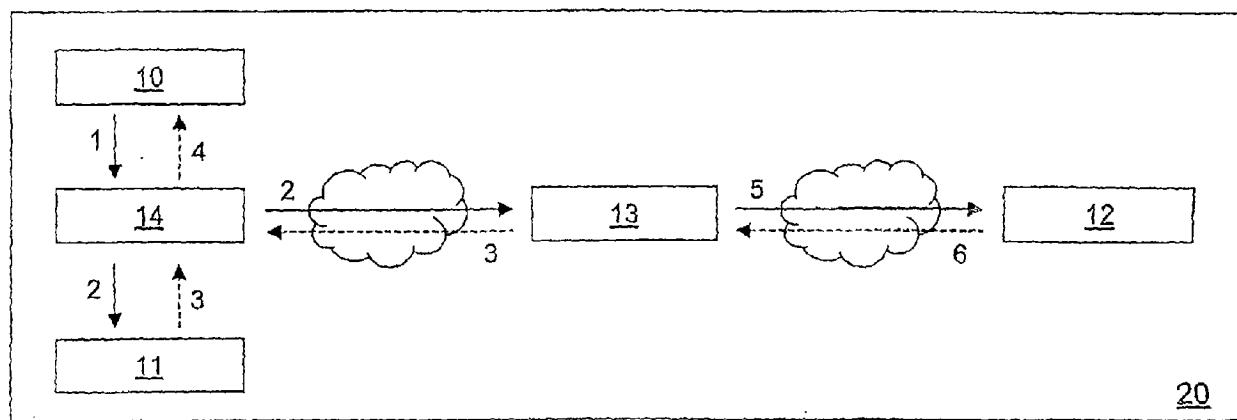
[0249] - 由运营商将服务提供给不同服务提供商 SP,SP 接着将灾难恢复的服务提供给其最终客户;在图 10 中,由框内部的相同阴影标识其各个客户;

[0250] - 由运营商提供的服务的边界(虚线)从提供商的站点限制延伸到最终客户的站点。

[0251] 本发明将通过以下方式解决 PDC 和 RDC 之间的同步镜像数据复制的距离限制的问题:在离开 PDC 的一定距离处插入存储的中间站点从而确保主要应用的同步性,并且在另一方面,在离开 PDC 很长距离处放置 RDC,从而最小化双重故障的风险(损坏 PDC 和 RDC 两者)的风险,并因此增加存储的信息的可用性。

[0252] 虽然已经参照实际的优选实施例示出了本发明,但对本领域技术人员明显的是,本发明通常会有落入本发明范围内的其它应用和修改。

[0253] 灾难恢复情形仅仅是本发明的可能的使用和结局的一个示例。应注意的是,本发明本身关于客户机应用以同步方式执行对出现在第一站点中的确定的数据卷到在地理上离开所述第一站点很远的第二站点的拷贝。不同于提供用于恢复和 / 或使用远程存储的数据的服务的其它模式可使用该功能。



20

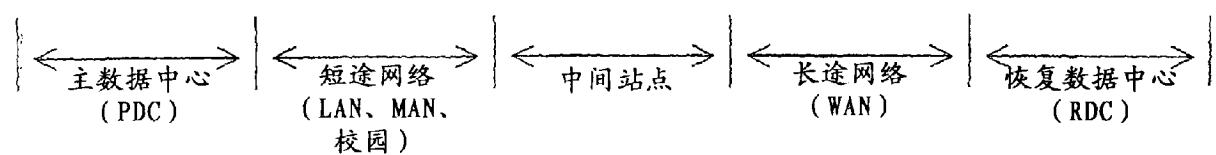


图 1

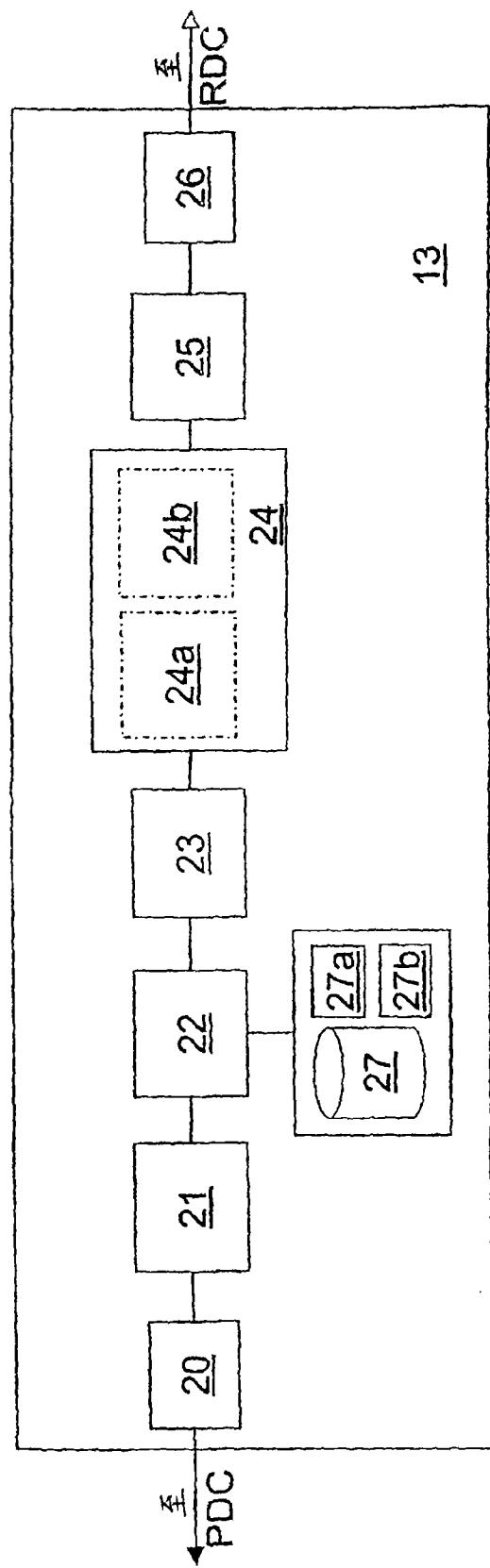


图 2

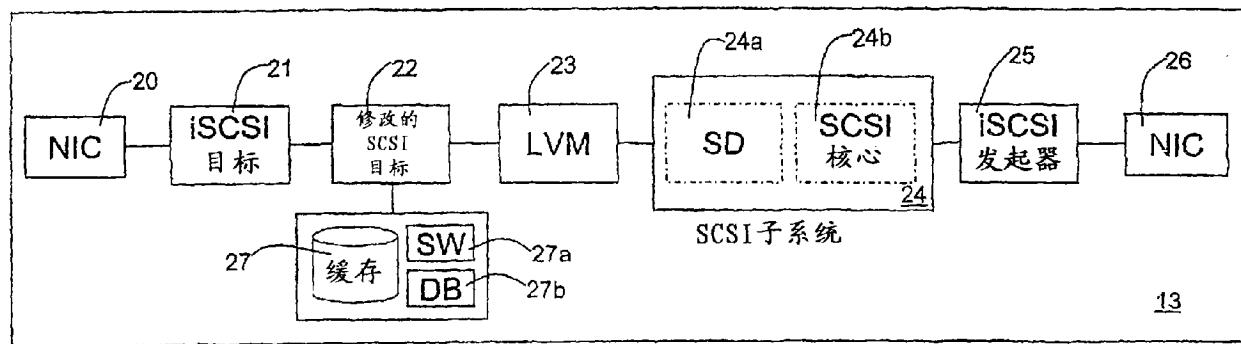


图 3

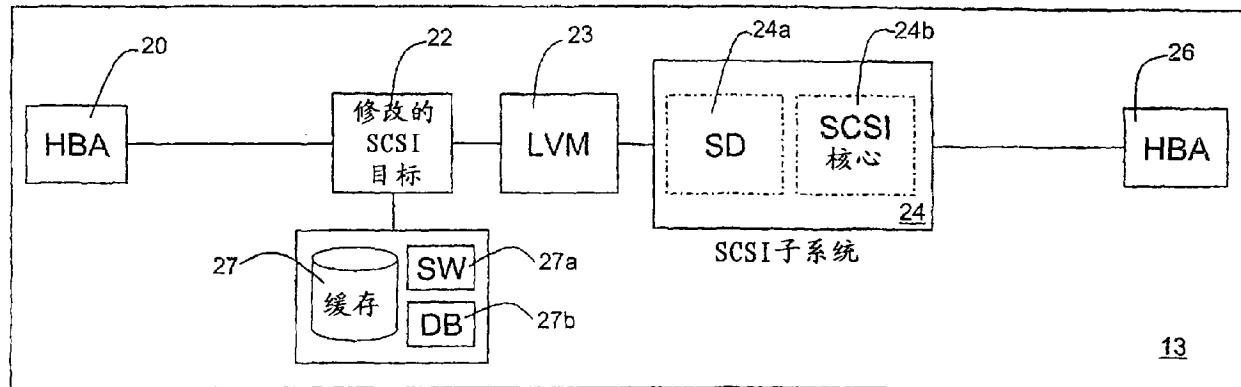


图 4

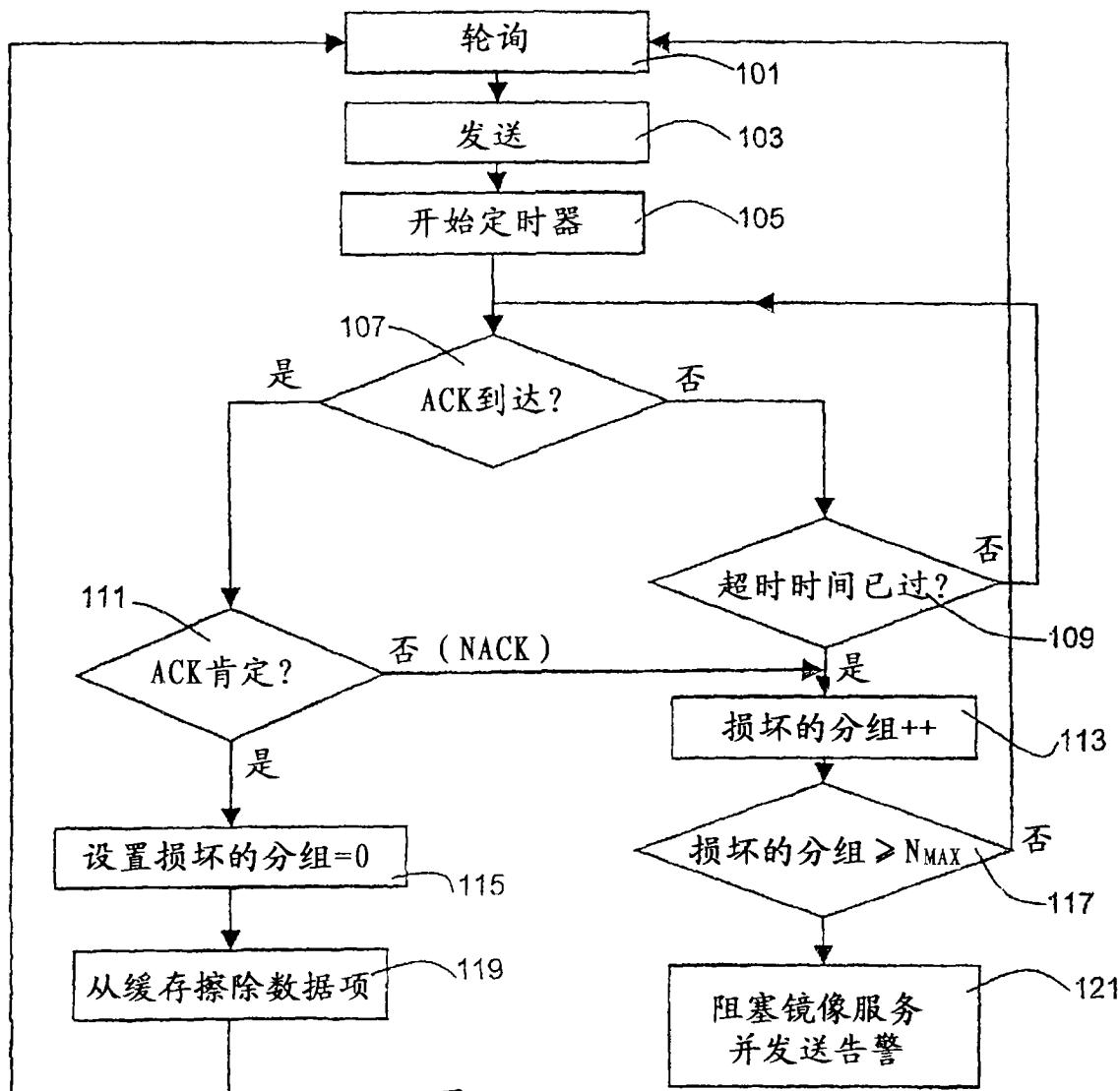


图 5A

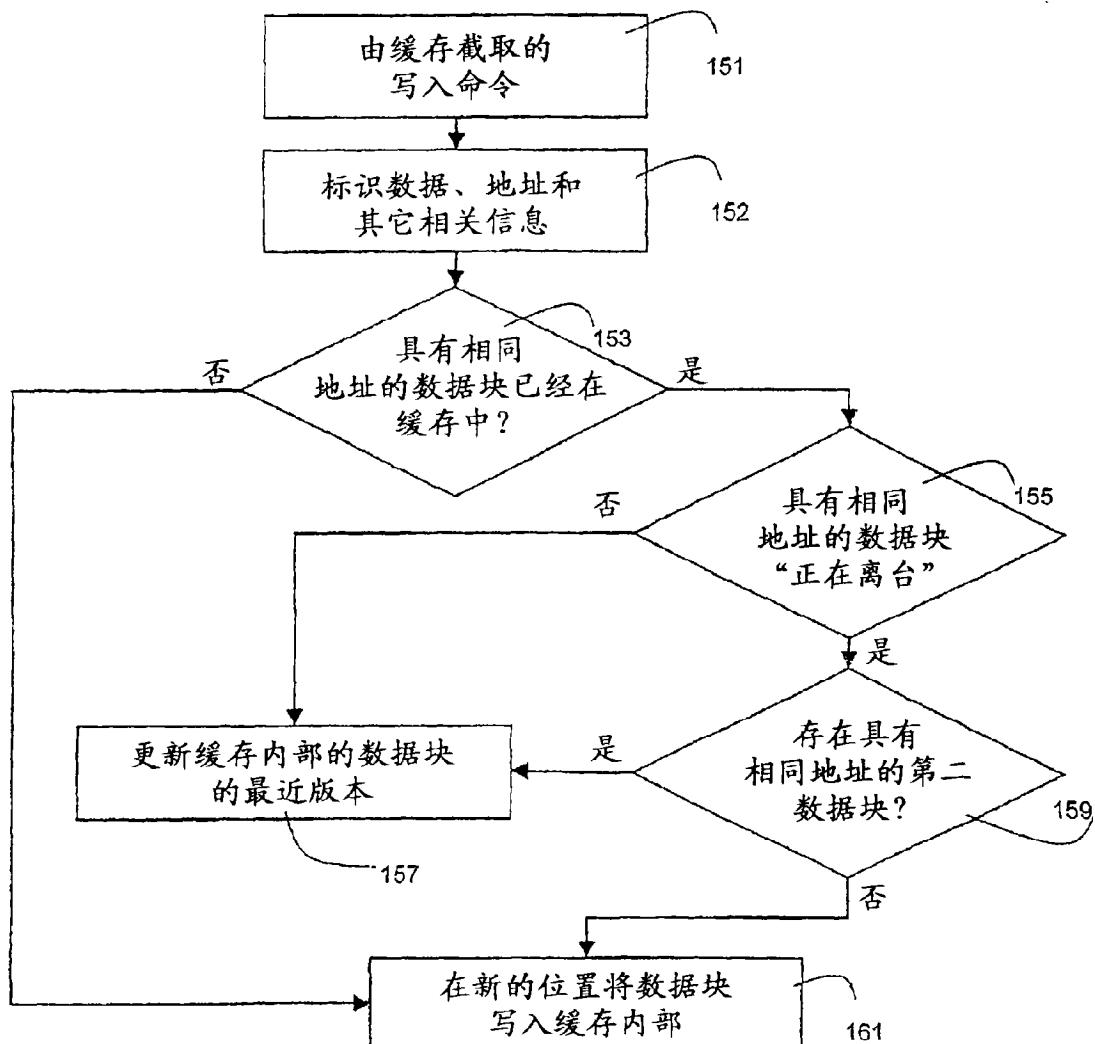
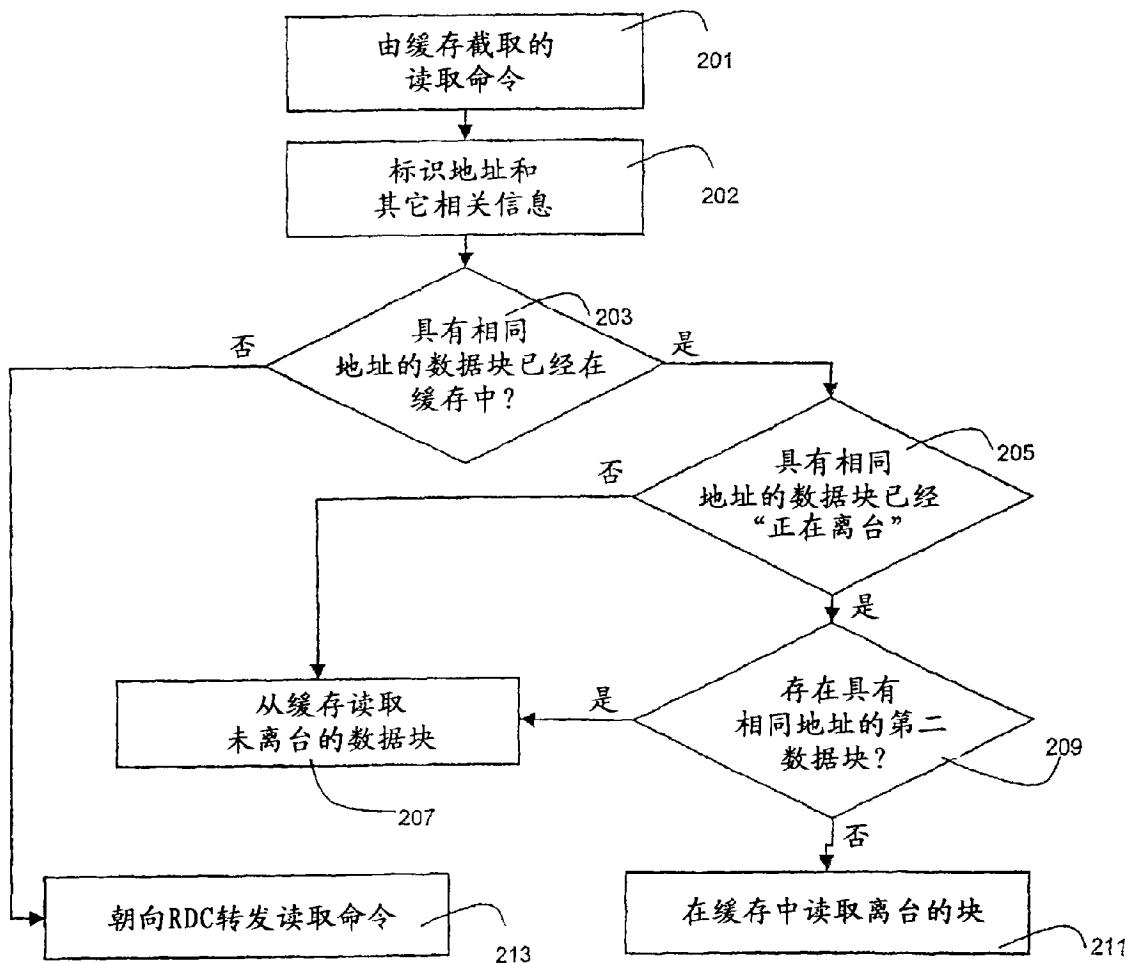


图 5B

图 5C



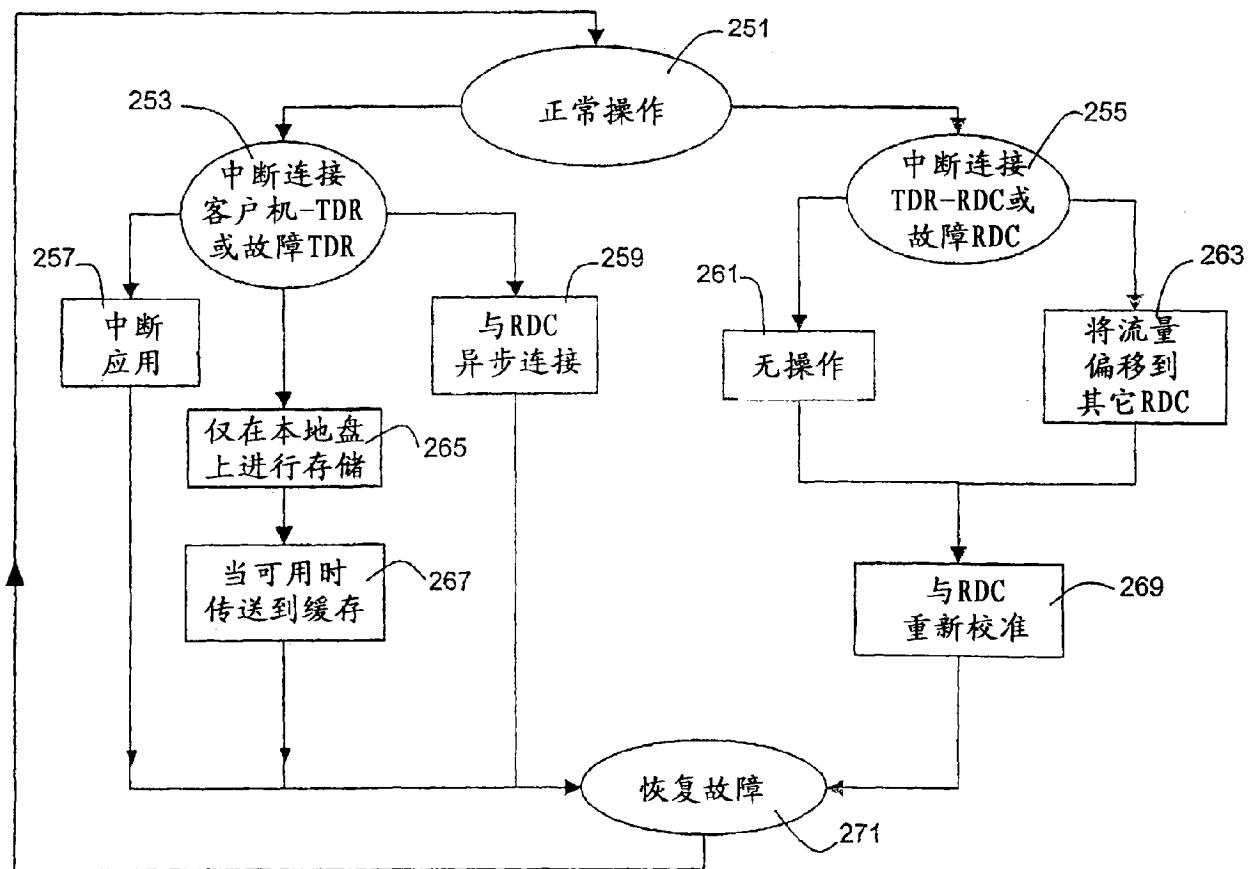


图 6

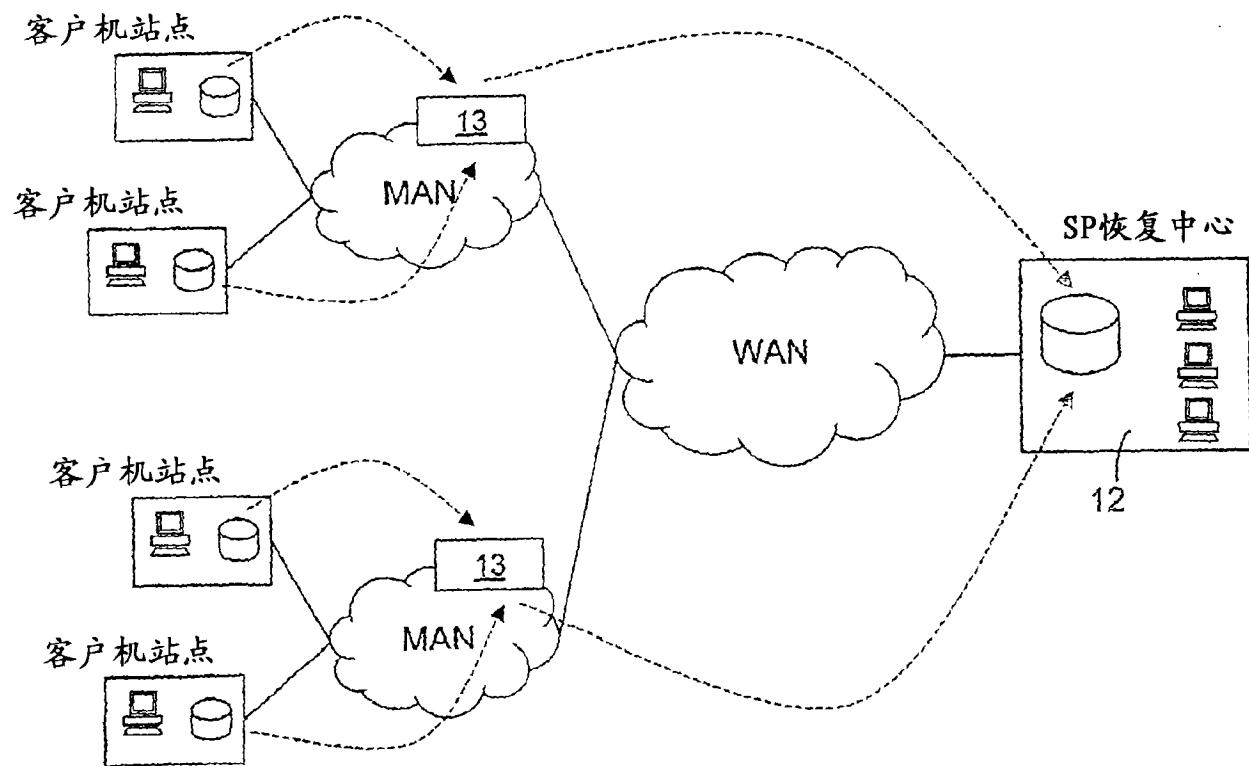


图 7

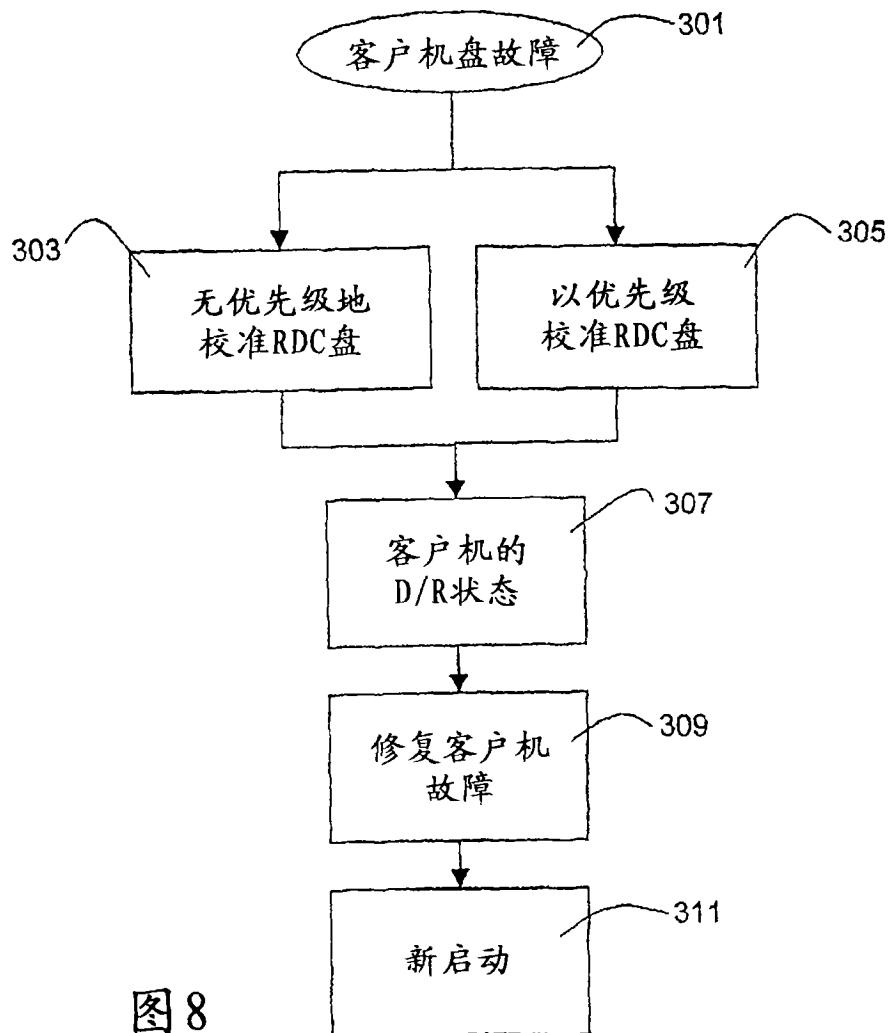


图 8

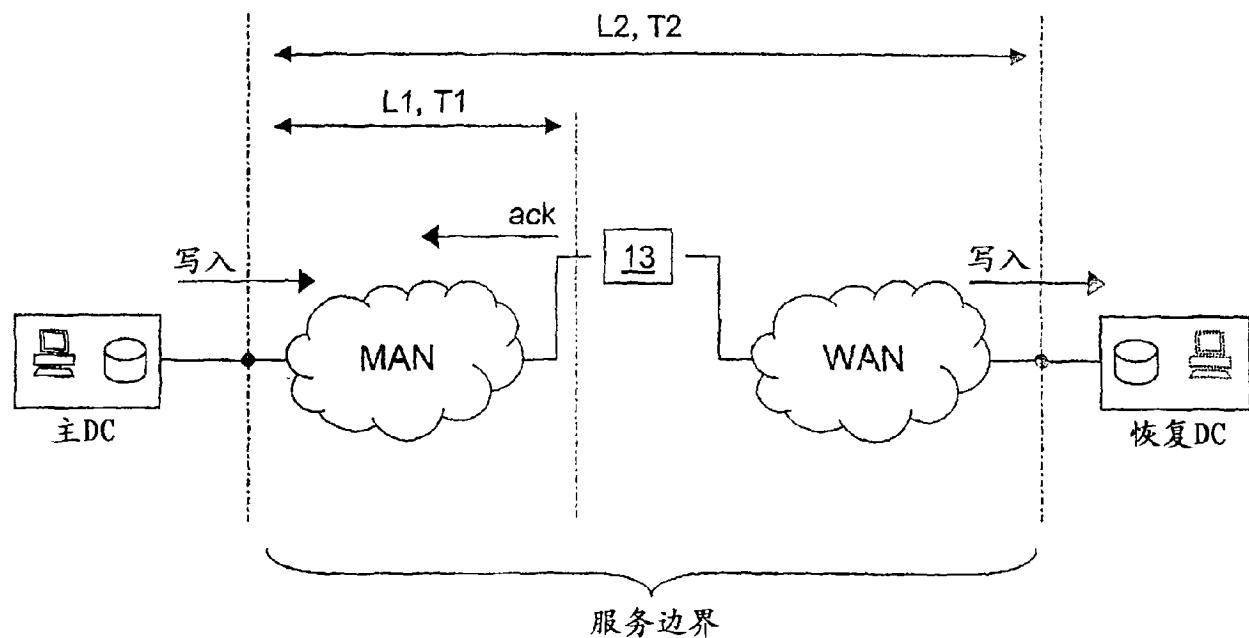


图 9

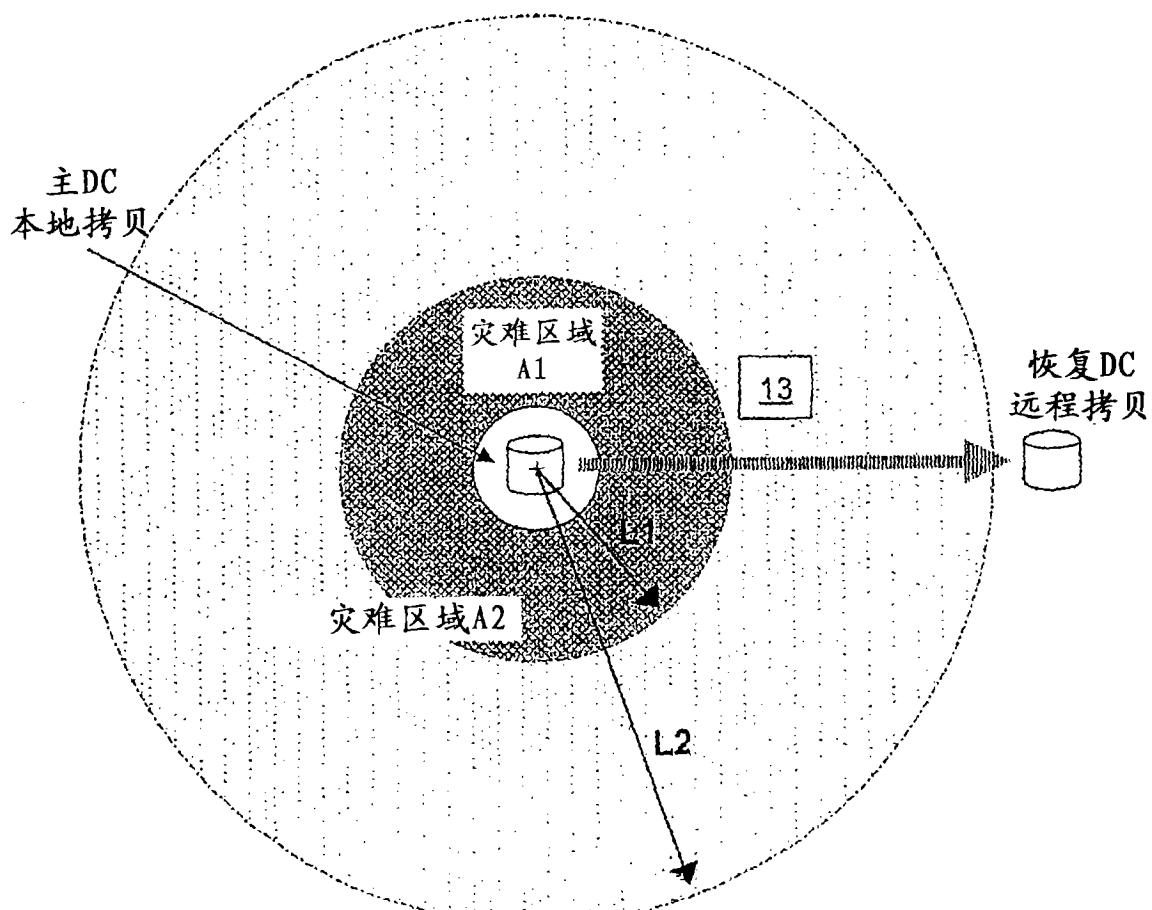


图 10

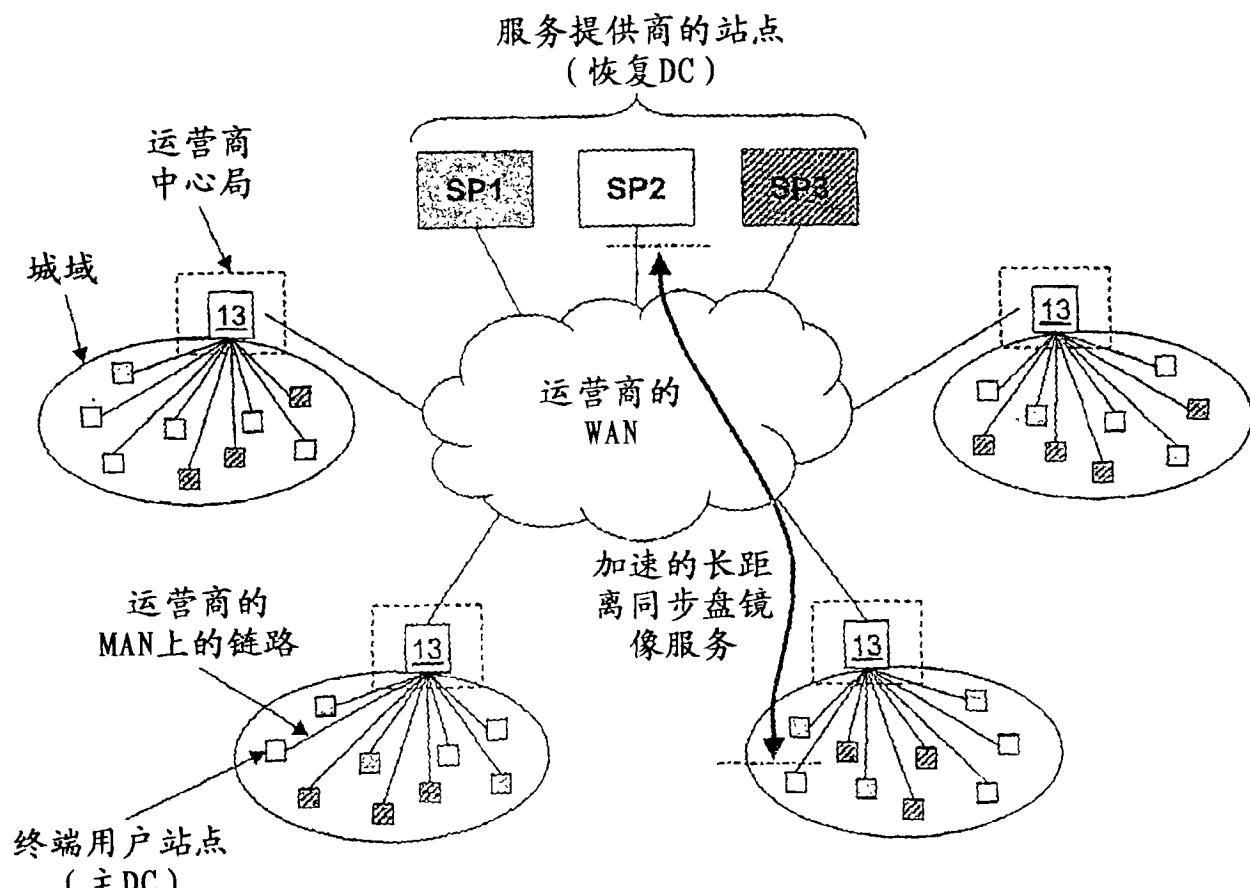


图 11