(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 103186348 B
(45)授权公告日 2016.04.13

(21)申请号 201110444604.1

(22)申请日 2011.12.27

(73)专利权人 杭州信核数据科技股份有限公司
地址 311202 浙江省杭州市萧山区金城路1038号国际创业中心12楼

(72)发明人 施苗峰 任永坚 汪海 芮琨

(74)专利代理机构 北京银龙知识产权代理有限公司 11243
代理人 曾贤伟 杨继平

(51)Int.Cl.
G06F 3/06(2006.01)
H04L 29/08(2006.01)

(56)对比文件
CN 102193842 A,2011.09.21,
US 2008/0162605 A1,2008.07.03,
CN 101216772 A,2008.07.09,
刘杰．基于SAN的网络存储系统研究与实现．《中国优秀硕士学位论文全文数据库信息科技辑》.2009,(第9期),第5-36页及第68页．
刘杰．基于SAN的网络存储系统研究与实现．《中国优秀硕士学位论文全文数据库信息科技辑》.2009,(第9期),第5-36页及第68页．

审查员 徐菲

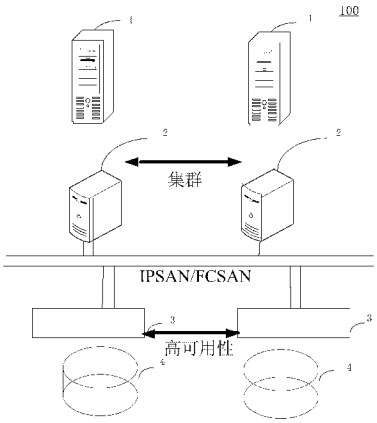
权利要求书1页 说明书5页 附图3页

(54)发明名称

存储系统及其数据读写方法

(57)摘要

本发明公开了一种存储系统及其数据读写方法。该存储系统在存储级和应用级上具有高可用性并包括：至少一台应用服务器，建立集群并向用户提供相同的应用服务；两台存储管理服务器，分别与所述应用服务器连接，利用映射卷技术实现所述存储系统的虚拟化存储；以及多台存储设备，分别与所述存储管理服务器连接，并在物理上存储用户的数据。通过所述至少两台存储管理服务器实现用户数据的读取/写入，当发生故障时，通过存储系统实现所述两台存储管理服务器的自动切换，实现应用透明，从而保证应用系统的持续运行。



CN 103186348 B

1.一种存储系统，所述存储系统在存储级和应用级上具有高可用性，其特征在于，所述存储系统包括：

多台客户端；

至少一台应用服务器，建立集群并分别与所述多台客户端连接以便向用户提供相同的应用服务；

至少两台存储管理服务器，分别通过IPSAN/FCSAN与所述至少一台应用服务器连接，利用映射卷技术实现所述存储系统的虚拟化存储；以及

多台存储设备，分别与所述至少两台存储管理服务器连接，并在物理上存储用户的数据，

其中，所述至少一台应用服务器和所述至少两台存储管理服务器被设置在IPSAN/FCSAN的不同侧；

其中，通过所述至少两台存储管理服务器实现用户数据的读取/写入并且所述至少两台存储管理服务器实时地保持有两份相同的在线数据，当发生故障时，通过所述存储系统实现所述至少两台存储管理服务器的自动切换，实现应用透明，从而保证应用系统的持续运行。

2.根据权利要求1所述的存储系统，其特征在于，当发生故障后维修修复了故障之后，通过自动反向同步技术，再次实现用户数据的同步状态。

3.根据权利要求1或2所述的存储系统，其特征在于，所述应用服务器支持Windows、Linux、Unix、Solaris、以及VMware操作系统。

4.根据权利要求1或2所述的存储系统，其特征在于，所述应用服务器支持Oracle、DB2、MS SQL、以及Sybase数据库。

5.根据权利要求1或2所述的存储系统，其特征在于，所述两台存储管理服务器通过交换网络与所述多台存储设备相连接。

6.根据权利要求1或2所述的存储系统，其特征在于，所述多台存储设备是由不同厂商制造的不同品牌的存储设备。

7.一种由根据权利要求1-6中任一项所述的存储系统实现的数据读写方法，其特征在于，所述方法包括如下步骤：

当接收到读取/写入命令时，判断要读取/写入的卷是否为镜像卷组中的一个镜像卷；

当要读取/写入的卷是镜像卷组中的一个镜像卷时，从镜像卷组中的一个镜像卷读取数据或者向镜像卷组中的一个镜像卷写入数据。

8.根据权利要求7所述的方法，其特征在于，当判断要读取/写入的卷不是镜像卷组中的一个镜像卷时，将读取/写入命令发送到下一层。

9.根据权利要求7或8所述的方法，其特征在于，所述方法还包括判断是本地卷还是镜像卷宕掉的步骤，

其中，如果是本地卷宕掉，则执行向镜像卷的读取/写入操作；并且

如果是镜像卷宕掉，则执行向本地卷的读取/写入操作。

存储系统及其数据读写方法

技术领域

[0001] 本发明涉及存储系统及其数据读写方法，尤其涉及在存储级和应用级上具有高可靠性的存储系统及其使用方法。

背景技术

[0002] 目前，对于现代化企业来说，利用计算机系统来提供及时可靠的信息和服务是必不可少的。对于计算机应用来说，最重要的是可持续的、具有一致性和完整性的数据访问。但是，计算机硬件与软件都不可避免地会发生故障，这些故障有可能给企业带来极大的损失，甚至造成整个服务的终止和网络的瘫痪。对于那些任何停工都将产生严重的财产损失、名誉损失、甚至生命损失的关键性应用的企业或公司，必须有适当的措施来确保计算机系统提供不间断的服务，以维护系统的可用性。因此，系统的高可用性显得尤为重要。

[0003] 高可用性(HA，High Availability)指的是通过尽量缩短因日常维护操作(计划)和突发的系统崩溃(非计划)所导致的停机时间，以提高系统和应用的可用性。高可用性方案利用冗余部件、由软件检测故障，一旦故障发生立即隔离损坏部件，通过提供故障恢复实现最大化系统和应用的可用性。HA的目标是尽量实现不停机操作。故障被掩饰掉，而且尽量不修改最终的应用程序。

[0004] 从客户端看来，集群(cluster)中的所有服务器是一个系统，就像一台大型的计算机系统，其上运行着客户端需要的应用服务。由于集群系统能够保证用户的业务是连续的并且具有持续可用的特性，即具有7×24小时的可用性。

[0005] 传统的HA结构

[0006] 传统的HA是应用服务器级的HA。如图1所示，两台应用服务器1、2之间做服务器集群，采用互备模式(Active/Active)或热备模式(Active/Standby)。

[0007] 互备模式：在正常情况下，两台服务器均为前端客户提供各自的应用服务，并互相监视对方的运行情况。当一台服务器出现故障情况，不能对客户端提供正常服务时，另一台服务器将接管对方的应用.

[0008] 热备模式：在正常情况下，一台服务器是工作机，另一台服务器为备份机。工作机在为信息系统提供服务时，备份机在监视工作机的工作。当工作机出现故障，不能对前端客户提供服务时，备份机接管工作机的应用，继续为客户端提供正常服务，从而保证信息系统的业务不间断。当工作机修复后，可重新接入系统要回自己的应用。

[0009] 继续为客户端提供正常服务，从而保证信息系统的业务不间断。服务器同时连接到同一个存储设备。在这种结构下，如果一台服务器宕机另一台就接管全部的应用处理服务，防止应用服务器级的单点故障。但是如果存储级的设备出现问题。整个架构都将不能使用。所以传统的HA架构不能真正完整地保护企业数据业务365×24×60的不间断性。企业一旦无法访问到任务关键数据，就会造成生产和供应链的延迟，这将给企业带来难以估量的损失。

[0010] 因此，需要提出一种改进的具有高可靠性的存储系统及其使用方法。

发明内容

[0011] 本发明的目的在于，提供一种在存储级和应用级上具有高可靠性的存储系统及其使用方法。

[0012] 根据本发明的一个方面，提供一种存储系统，所述存储系统在存储级和应用级上具有高可用性。所述存储系统包括：至少一台应用服务器，建立集群并向用户提供相同的应用服务；两台存储管理服务器，分别与所述应用服务器连接，利用映射卷技术实现所述存储系统的虚拟化存储；以及多台存储设备，分别与所述存储管理服务器连接，并在物理上存储用户的数据。通过所述至少两台存储管理服务器实现用户数据的读取/写入，当发生故障时，通过存储系统实现所述两台存储管理服务器的自动切换，实现应用透明，从而保证应用系统的持续运行。

[0013] 优选的，当发生故障后维修修复了故障之后，通过自动反向同步技术，再次实现用户数据的同步状态。

[0014] 优选的，所述应用服务器支持Windows、Linux、Unix、Solaris、以及VMware等操作系统并且支持Oracle、DB2、MS SQL、以及Sybase等数据库。

[0015] 优选的，所述两台存储管理服务器通过光交换网络(OSN)或普通交换网络与所述多台存储设备相连接。

[0016] 优选的，所述多台存储设备是由不同厂商制造的不同品牌的存储设备。

[0017] 根据本发明的另一方面，提供了一种由上述存储系统实现的数据读写方法，所述方法包括如下步骤：当接收到读取/写入命令时，判断要读取/写入的卷是否为镜像卷组中的一个镜像卷；当要读取/写入的卷是镜像卷组中的一个镜像卷时，从镜像卷组中的一个镜像卷读取数据或者向镜像卷组中的一个镜像卷写入数据。

[0018] 优选的，当判断要读取/写入的卷不是镜像卷组中的一个镜像卷时，将读取/写入命令发送到下一层。

[0019] 优选的，所述方法还包括判断是本地卷还是镜像卷宕掉的步骤，如果是本地卷宕掉，则执行向镜像卷的读取/写入操作；并且如果是镜像卷宕掉，则执行向本地卷的读取/写入操作。

[0020] 优选的，所述方法还包括当所述镜像卷组中的任一个卷的数据遭到破坏时，从所述镜像卷组中的其他卷将遭到破坏的数据恢复。

[0021] 相应地，本发明所取得的有益效果包括：

[0022] 保障业务连续性

[0023] 两个存储服务器实时地保持两份相同的在线数据，当其中之一发生故障时，存储服务会自动透明地切换到另一台存储上，从而保证了客户业务连续性。待维修完毕，通过自动反向同步技术，又可实现两台存储数据同步状态。高可用存储使数据实现双保险，业务连续性有了切实保障。

[0024] 广泛的兼容性

[0025] 存储相对独立，用户可自由选择主机与数据库类型。支持Windows、Linux、Unix、Solaris、VMware等主流操作系统，支持Oracle、DB2、MS SQL、Sybase等主流数据库。

[0026] 异构存储管理

[0027] 支持主流品牌的存储设备，可实现不同品牌存储设备之间的HA；支持IP、FC、SAS(Serial Attached SCSI)三种连接方式，支持不同连接方式的存储设备之间的HA。

[0028] 简化管理

[0029] 可快速地完成相关配置，提供易用的中文图形化操作界面和自动化监控系统。

[0030] 广泛的可扩展性

[0031] 可平滑扩展持续数据保护与容灾功能。

附图说明

[0032] 本发明的特征、实施例和优点，将参照附图在以下详细描述。

[0033] 图1是传统的应用服务器级的高可靠性(HA)存储系统的系统结构的示意图；

[0034] 图2是描绘了根据本发明实施例的存储级HA的存储系统的结构的示意图；以及

[0035] 图3是描绘了根据本发明实施例的存储级HA所实现的基本功能的示意图。

具体实施方式

[0036] 接下来，将结合附图进行详细描述本发明的实施例。只要可能，在整个附图中，相同的附图标记将指示相同的部件。

[0037] 硬件配置

[0038] 下面，参照图2来说明根据本发明实施例的存储级HA的存储系统的硬件配置。图2是描绘了根据本发明实施例的存储级HA的存储系统的结构的示意图。

[0039] 如图2所示，存储系统100包括多台客户端(例如，PC)1、建立集群并向用户提供相同的应用服务的多台应用服务器2、分别与应用服务器连接的两台存储管理服务器3、以及多台由不同厂商提供的性能不同的物理存储设备4。在存储管理服务器3上通过映射卷技术实现存储虚拟化以提供对数据的保护。

[0040] 通过两台存储管理服务器实现用户数据的读取/写入，当发生故障时，通过软件实现多台应用服务器的自动切换，从而保证存储系统的持续运行。

[0041] 如图2所示，应用服务器、存储管理服务器和存储设备的数目都是两台。然而，本领域技术人员可以理解的是，根据需要，这些组成部件的数目还可以是多于两台，本发明并不局限于此。

[0042] 另外，虽然图2中示出了存储管理服务器3和物理存储设备4之间通过光交换网络(OSN，Optical Switch Network)相连接，但是本领域技术人员可以理解的是，还可以在存储系统中采用其他类型的连接方式来实现相同或相似的功能，而本发明并不局限于此。

[0043] 与传统的普通HA相比较，通过根据本发明的存储级HA，能够实现数据的同步读取/写入。

[0044] 此外，如图2所示，通过应用级的集群(Cluster)，实现在出现故障时业务系统的自动切换，从而保证业务系统的持续运行。

[0045] 因此，根据本发明的存储级HA的特点包括：

[0046] 第一，独立于主机(应用服务器)与在其上运行的应用，能够在不影响现有应用的情况下，透明地实现存储集群。

[0047] 第二，独立于存储系统，可构建存储HA框架，为业务将来发展选择由不同厂商生产

的更多存储硬件品牌。

[0048] 第三，支持异构存储管理，可充分发挥现有IT的作用，简化管理。

[0049] 第四，在单点故障情况下全自动切换、恢复，从而实现99.99％以上的安全系数，最大程度的保障业务连续性。

[0050] 第五，基于虚拟存储，自动精简配置，可提高存储利用率。

[0051] 第六，提供现有存储的自动迁移服务，最大限度减少业务宕机时间。

[0052] 下面将会参考图3，来详细地说明根据本发明的存储级HA的I/O处理过程。

[0053] 首先，如图3所示，描绘了根据本发明的存储级HA中的由OSN对存储设备执行的三种基本任务，即读操作、写操作和恢复操作。图3是描绘了根据本发明的存储级HA所实现的基本功能的示意图。

[0054] 下面，将会通过三个实施例来分别详细地描述这三种操作。

[0055] 第一实施例(读操作)

[0056] 接下来，描述根据本发明的存储级HA的读操作。

[0057] 当卷接收到读命令时，首先判断这个卷是否是镜像卷组中的一个。如果不是，说明这个卷不是镜像卷，把读命令发送到下一层，程序结束。如果这个卷是镜像组中的一个卷，则继续判断这个卷是否拒绝I/O操作。这是因为有时候为了保护卷中的数据，会设置这个卷不可读写。然后，需要再判断这个卷和它的镜像卷是否都宕掉了。只要其中一个没宕掉，就继续判断这个读请求是否来自镜像卷。如果是来自镜像卷，执行从本地卷读取数据。如果读数据请求不是来自镜像卷，则还需要判断本地卷是否宕掉。如果本地卷宕掉，则执行从镜像卷读取数据。如果本地卷没宕掉，则执行从本地卷读取数据的操作。

[0058] 执行从本地卷读数据的回调函数判断读操作是否成功，是则设置读写操作状态为成功，程序结束。否则判断这个卷是否有镜像卷。如果没有镜像卷，则设置读写操作不成功，程序结束。如果这个卷有镜像卷，则判断是否从镜像卷读数据，是则返回不成功，否则判断镜像卷是否是正常连接(UP)状态，是则从镜像卷读取数据，执行回调函数判断读数据是否成功。

[0059] 第二实施例(写操作)

[0060] 接下来，将会描述写操作，其中不再赘述与第一实施例中的读操作相同的步骤，而仅描述两者不同之处。

[0061] 当写操作开始执行时，首先判断写的这个卷是不是镜像组中的卷，不是的话写命令传到下一层。如果这个卷不拒绝I/O操作而且这个卷所在的镜像组的其中至少一个卷能够正常工作，则判断这个I/O是来自镜像卷的I/O还是来自应用层的I/O。如果是来自镜像卷的I/O，则说明本地卷要进行恢复操作。如果来自应用层的I/O，则不仅要把这个I/O写入本地卷，还要写入这个卷的镜像卷中。

[0062] 第三实施例(恢复操作)

[0063] 接下来，将会描述恢复操作，其中不再赘述与第一、第二实施例中的读/写操作相同的步骤，而仅描述与这两者的不同之处。

[0064] 在镜像卷组中，如果其中一个镜像卷的数据遭到破坏，可以从其它镜像卷中恢复回来。当开始恢复时，首先检查设备是否准备好，例如磁盘是否已经正常扫描到等。当设备准备好后，判断两台服务器是否都要求做恢复，因为这样的情况下说明镜像组同时损坏，恢

复失败。如果这种情况没发生，则需判断数据是从本地卷恢复到镜像卷，还是从镜像卷恢复到本地卷。恢复的时候数据按程序设定的值(在本示例中为1M)为单位分成几次恢复。先把数据从正常的卷读出来再把数据写到需要恢复的卷，从而完成恢复过程。

[0065] 尽管给出一些实施例，但本发明并不限于此。本领域技术人员基于本发明实施例的任何变形、修改，都不会背离本发明所限定的权利要求的范围。

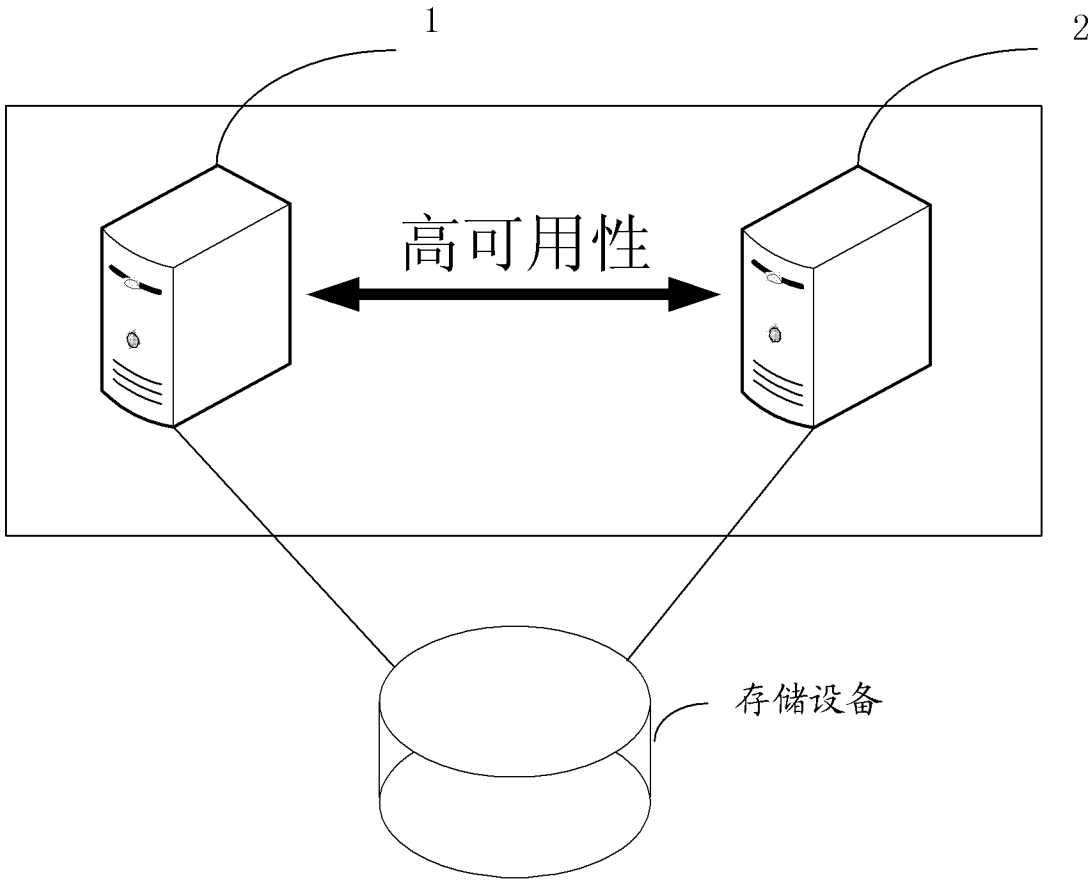


图1

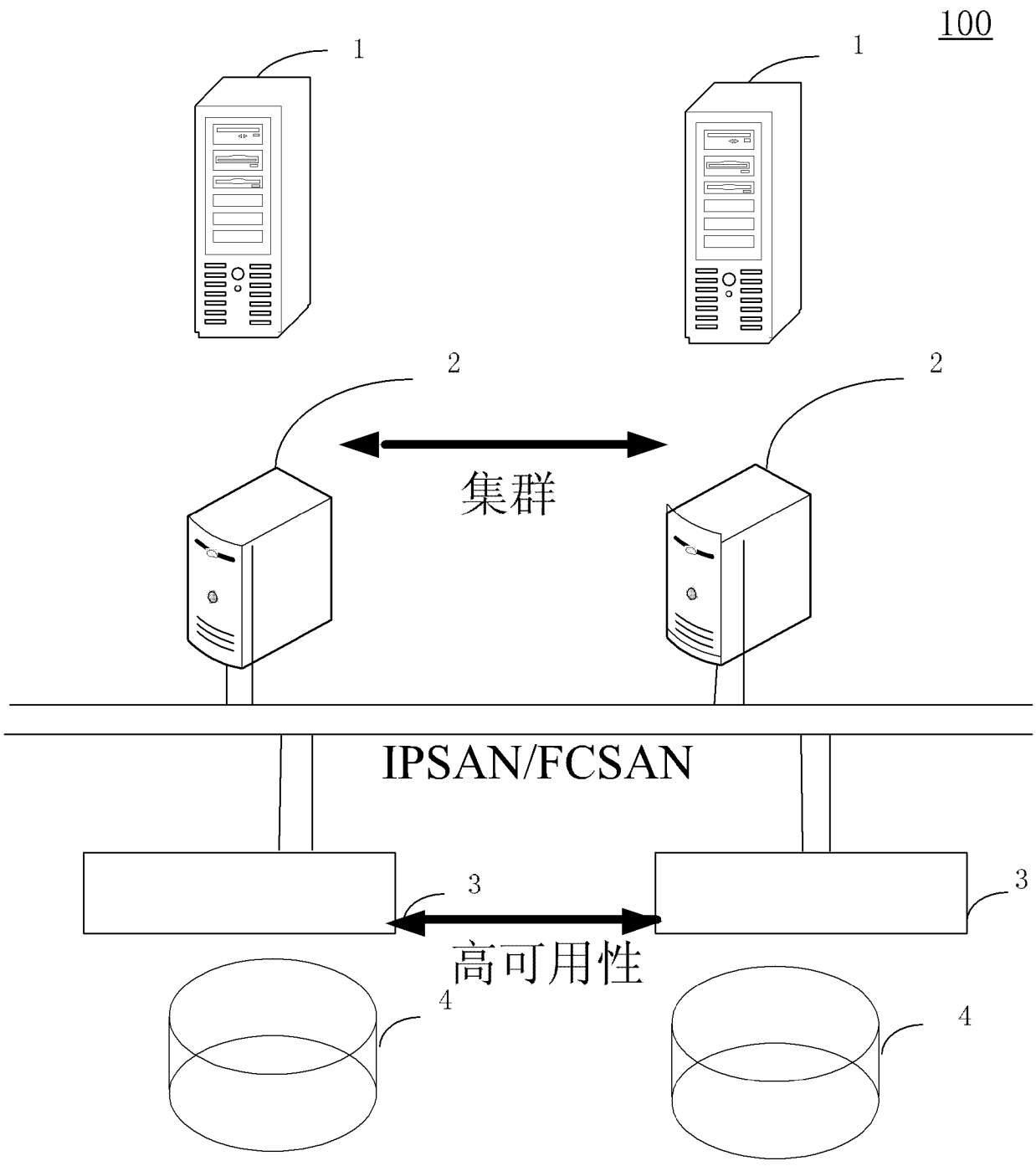


图2

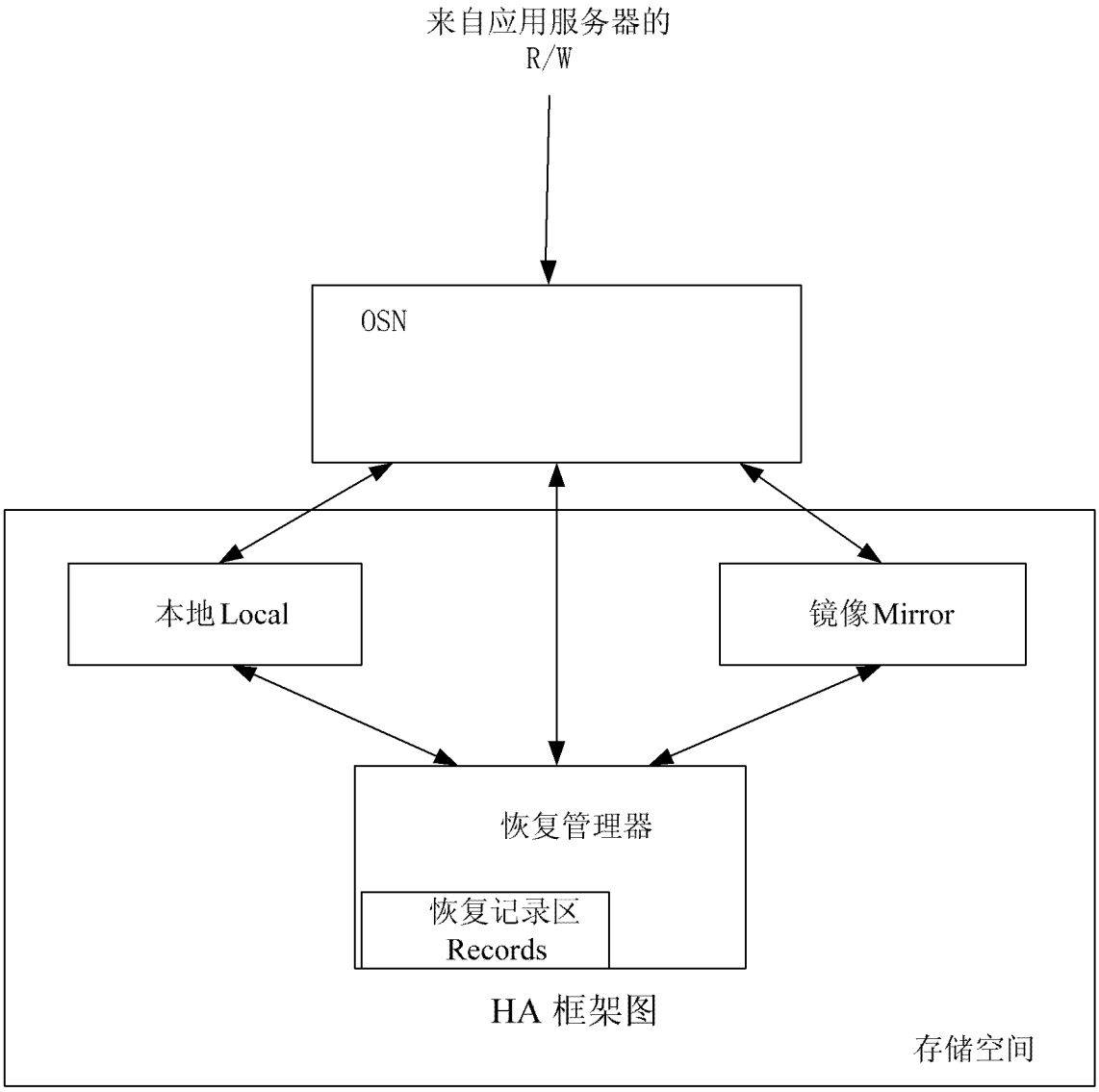

来自应用服务器的
R/W
OSN
本地 Local
镜像 Mirror
恢复管理器
恢复记录区
Records
HA 框架图
存储空间

图3