(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 102024044 A
(43)申请公布日 2011.04.20

(21)申请号 201010587235.7

(22)申请日 2010.12.08

(71)申请人 华为技术有限公司
地址 518129 广东省深圳市龙岗区坂田华为总部办公楼

(72)发明人 张辉 范家星 姜南 吴波

(74)专利代理机构 北京同立钧成知识产权代理有限公司 11205
代理人 王庆龙

(51)Int.Cl.
G06F 17/30(2006.01)

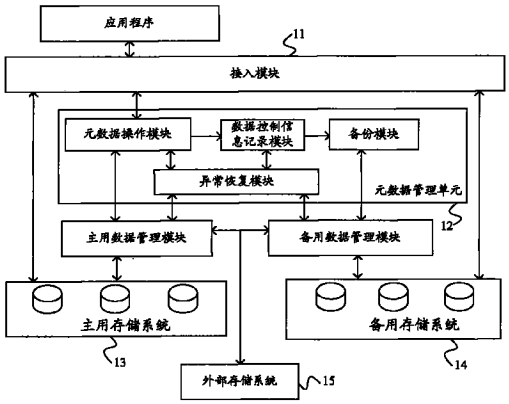
权利要求书 2 页 说明书 10 页 附图 7 页

(54)发明名称

分布式文件系统

(57)摘要

本发明实施例提供一种分布式文件系统。本包括分布式文件系统接入模块、元数据管理单元，主用存储系统、备用存储系统和外部存储系统；其中，所述接入模块、所述元数据管理单元、所述主用存储系统和所述备用存储系统之间通过系统总线连接；所述外部存储系统通过网络与所述主用存储系统和所述备用存储系统连接。本发明实施例提供的分布式文件系统中，通过将数据写入到主用存储系统中，采用异步备份机制备份到备用存储系统，不影响高速设备的读写性能；而且可以实现在线恢复数据，数据对外服务过程中自动实现数据恢复，加快恢复过程。



CN 102024044 A

1. 一种分布式文件系统，其特征在于，包括：接入模块、与所述接入模块连接的元数据管理单元，以及分别与所述元数据管理单元连接的主用存储系统和备用存储系统，所述接入模块还分别与所述主用存储系统和所述备用存储系统连接；所述接入模块、所述元数据管理单元、所述主用存储系统和所述备用存储系统之间通过系统总线连接；还包括外部存储系统，所述外部存储系统通过网络与所述主用存储系统和所述备用存储系统连接；其中：

所述接入模块，用于接收读 / 写数据请求，向所述元数据管理单元发送元数据请求以获取所请求的数据对应的元数据，并应用所述元数据向所述主用存储系统或所述备用存储系统读写数据；

所述元数据管理单元，用于在所述接入模块请求所述元数据时，查找所请求的数据在所述主用存储系统或所述备用存储系统上的位置，并构造所述元数据，返回给所述接入模块；所述主用存储系统，用于在所述分布式文件系统处于正常状态时为所述接入模块提供所请求的数据；所述备用存储系统，用于在所述分布式文件系统处于异常状态或恢复状态时，为所述主用存储系统提供数据备份；所述外部存储系统，用于为所述主用存储系统提供数据备份。

2. 根据权利要求 1 所述的分布式文件系统，其特征在于，所述元数据管理单元包括：元数据操作模块、备份模块和异常恢复模块；其中：

所述元数据操作模块，分别与所述接入模块、所述主用存储系统连接，用于接收所述元数据请求，向所述主用存储系统或所述异常恢复模块请求元数据并向所述接入模块返回所述元数据；还用于根据接收的所述主用存储系统上报的设备状态事件，更新所述系统状态；还用于记录所述读 / 写数据请求中的数据控制信息；

所述备份模块，与所述元数据操作模块和所述备用存储系统连接，用于读取所述元数据操作模块记录下来的所述数据控制信息，生成数据备份操作控制信息并发送给所述备用存储系统，以将所述主用存储系统上的数据备份到所述备用存储系统中；

所述异常恢复模块，分别与所述元数据操作模块、所述备份模块、所述主用存储系统和所述备用存储系统连接，用于在所述异常状态和所述恢复状态下获取所述元数据并返回给所述元数据操作模块。

3. 根据权利要求 2 所述的分布式文件系统，其特征在于，所述元数据管理单元还包括：

数据控制信息记录模块，连接在所述元数据操作模块和所述备份模块之间，用于存储所述数据控制信息。

4. 根据权利要求 2 所述的分布式文件系统，其特征在于，所述异常恢复模块包括：

第一处理子模块，用于在所述异常状态下，对于读数据请求，对从缓存或从所述主用存储系统获得的元数据进行缺失块检测，若检测到所述元数据有缺失块，则向所述备用存储系统发送请求信息包括缺失块信息和所述外部存储系统的地址；在所述备用存储系统返回所述元数据后，进行块地址重绑定操作，重组可用的元数据并发送；

第二处理子模块，用于在所述异常状态下，对于写数据请求，仅向所述备用存储系统请求元数据并返回给所述元数据操作模块；还用于记录写数据操作，以在数据恢复过程中根据记录的所述写数据操作将数据同步到所述主用存储系统中；

第三处理子模块，用于在所述恢复状态下，对于读数据请求，对从缓存或从所述主用存储系统获得的元数据进行缺失块检测，若检测到所述元数据有缺失块，则向所述备用存储系统发送请求信息包括缺失块信息，在所述备用存储系统返回元数据后，构造缺失块恢复操作信息，以供所述主存储系统进行数据恢复；并更新所述主存储系统的故障存储设备位图，返回所述元数据；

第四处理子模块，用于在所述恢复状态下，对于写数据请求，向所述主用存储系统请求元数据，获得所述元数据后，若检测到所述元数据有缺失块，则更新所述主存储系统的故障存储设备位图，返回所述元数据。

5. 根据权利要求 1 或 2 或 3 或 4 所述的分布式文件系统，其特征在于，还包括：

主用数据管理模块，连接在所述主用存储系统和所述元数据管理单元之间，用于管理所述主用存储系统所存储的数据，响应元数据请求和数据操作请求；以及

备用数据管理模块，连接在所述备用存储系统和所述元数据管理单元之间，用于管理所述备用存储系统所存储的数据，响应元数据请求和数据操作请求。

6. 根据权利要求 5 所述的分布式文件系统，其特征在于，所述主用存储系统包括数个高速存储设备，所述高速存储设备包括但不限于高数据传输率的 SCSI 硬盘、SATA 硬盘、SSD；

所述备用存储系统包括数个高速存储设备和 / 或低速存储设备，其中，所述高速存储设备包括 SCSI 硬盘、SATA 硬盘、SSD。

7. 根据权利要求 1 所述的分布式文件系统，其特征在于，

所述正常状态指主用存储系统没有出现故障；

所述异常状态指主用存储系统出现故障，主用存储系统和备用存储系统共同工作，协调完成数据存储，保存协调结果，其中，所述协调结果用于数据恢复过程；

所述恢复状态是指经过异常状态后，使用备用存储系统上的数据恢复主用存储系统中的数据。

分布式文件系统

技术领域

[0001] 本发明实施例涉及数据备份技术，尤其涉及一种分布式文件系统。

背景技术

[0002] 随着互联网向更宽更广方向发展，各行各业应用越来越多，特别是流媒体应用，内容发布网络（Content Delivery Network；以下简称：CDN）中的内容供应服务器性能及可靠性越来越重要，而承载这些应用数据的核心文件系统对性能和可靠性要求也越来越高。

[0003] 对于可靠性，现有技术主要采用独立冗余磁盘阵列（Redundant Arrayof Independent Disk；以下简称：RAID）技术使用冗余备份特性保证。简单的说，RAID是一种把多块独立的硬盘（物理硬盘）按不同的方式组合起来形成一个硬盘组（逻辑硬盘），从而提供比单个硬盘更高的存储性能和提供数据备份技术。而对于性能，现有技术通常采用分布式文件系统的存储条带化管理来叠加聚合硬盘带宽。所述的条带化为一种管理功能，其作用是将数据按一定步长分散在多个存储设备上，使得读取时并行从多个物理存储设备上获取，实现多物理存储设备性能叠加。RAID技术的数据冗余与数据条带化特性，保证了分布式文件系统的高可靠性与高性能。

[0004] 一般来说，分布式文件系统可以理解成建立在一个网络存储系统上的。在高性能需求不断增加形式下，硬盘逐渐被固态硬盘取代的趋势发展，由于固态盘极为昂贵，此时采用固态硬盘备份固态硬盘的RAID技术导致了服务器成本开销剧增。RAID技术常采用RAID1与RAID5实现。其中，RAID1是第1级RAID技术，采用一种完整镜像备份，需要两个同质的存储系统同步进行读写操作，互为镜像，即使有一个磁盘损坏，系统仍能正常工作。RAID5是一种存储性能、数据安全和存储成本兼顾的存储解决方案，不对存储的数据进行备份，而是把数据和相对应的奇偶校验信息存储到组成RAID5的各个磁盘上，并且奇偶校验信息和相对应的数据分别存储于不同的磁盘上。当RAID5的一个磁盘数据发生损坏后，利用余下的数据和相应的奇偶校验信息去恢复被损坏的数据。

[0005] 在实现本发明过程中，发明人发现现有技术中至少存在如下问题：在CDN网络中，使用RAID技术存在以下缺点：同等容量，组成RAID需要更多的磁盘，特别是RAID1需要双倍的磁盘，并且固态盘极为昂贵，因而存储成本导致系统成本过高；RAID5任何数据的修改需要重写校验，导致写入数据稍慢，数据恢复时间久，可能影响业务；而且，数据备份的恢复受到损坏磁盘的数量的限制。

发明内容

[0006] 本发明实施例提供一种分布式文件系统，包括：接入模块、与所述接入模块连接的元数据管理单元，以及分别与所述元数据管理单元连接的主用存储系统和备用存储系统，所述接入模块还分别与所述主用存储系统和所述备用存储系统连接；所述接入模块、所述元数据管理单元、所述主用存储系统和所述备用存储系统之间通过系统总线连

接；还包括外部存储系统，所述外部存储系统通过网络与所述主用存储系统和所述备用存储系统连接；其中：

[0007] 所述接入模块，用于接收读 / 写数据请求，向所述元数据管理单元发送元数据请求以获取所请求的数据对应的元数据，并应用所述元数据向所述主用存储系统或所述备用存储系统读写数据；

[0008] 所述元数据管理单元，用于在所述接入模块请求所述元数据时，查找所请求的数据在所述主用存储系统或所述备用存储系统上的位置，并构造所述元数据，返回给所述接入模块；所述主用存储系统，用于在所述分布式文件系统处于正常状态时为所述接入模块提供所请求的数据；所述备用存储系统，用于在所述分布式文件系统处于异常状态或恢复状态时，为所述主用存储系统提供数据备份；所述外部存储系统，用于为所述主用存储系统提供数据备份。

[0009] 本发明实施例提供的分布式文件系统中，通过将数据写入到主用存储系统中，采用异步备份机制备份到备用存储系统，不影响高速设备的读写性能；而且可以实现在线恢复数据，数据对外服务过程中自动实现数据恢复，加快恢复过程。

附图说明

[0010] 为了更清楚地说明本发明实施例或现有技术中的技术方案，下面将对实施例或现有技术描述中所需要使用的附图作一简单地介绍，显而易见地，下面描述中的附图是本发明的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其他的附图。

[0011] 图 1 为本发明一实施例分布式文件系统组成示意图；

[0012] 图 2 为本发明另一实施例分布式文件系统组成示意图；

[0013] 图 3 为本发明实施例分布式文件系统在正常状态下处理流程示意图；

[0014] 图 4 为本发明实施例分布式文件系统在异常状态下并且为读数据请求情况下的处理流程示意图；

[0015] 图 5 为本发明实施例分布式文件系统在异常状态下并且为写数据请求情况下的处理流程示意图；

[0016] 图 6 为本发明实施例分布式文件系统在恢复状态下响应读数据请求的处理流程示意图；

[0017] 图 7 本发明实施例分布式文件系统在恢复状态下响应写数据请求的处理流程示意图；

[0018] 图 8 为本发明实施例分布式文件系统在系统恢复状态下数据恢复过程的处理流程示意图。

具体实施方式

[0019] 为使本发明实施例的目的、技术方案和优点更加清楚，下面将结合本发明实施例中的附图，对本发明实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例是本发明一部分实施例，而不是全部的实施例。 基于本发明中的实施例，本领域普通技术人员在没有作出创造性劳动前提下所获得的所有其他实施例，都属于本发明保护

的范围。

[0020] 图 1 为本发明一实施例分布式文件系统组成示意图，如图 1 所示，该分布式文件系统 1 包括接入模块 11、与接入模块 11 连接的元数据管理单元 12，以及分别与元数据管理单元 12 连接的主用存储系统 13 和备用存储系统 14，接入模块 11 还分别与主用存储系统 13 和备用存储系统 14 连接；其中，接入模块 11、元数据管理单元 12、主用存储系统 13 和备用存储系统 14 均位于内部网络中，各功能模块之间通过系统总线连接；该分布式文件系统 1 还包括位于外部网络中的外部存储系统 15，外部存储系统 15 通过主用存储系统 13 和备用存储系统 14 网络与连接。

[0021] 分布式文件系统 1 中的接入模块 11，用于接收读 / 写数据请求，向元数据管理单元 12 发送元数据请求以获取所请求的数据对应的元数据，并应用元数据向主用存储系统 13 或备用存储系统 14 读写数据；

[0022] 接入模块 11 是分布式文件系统 1 访问数据的入口，接收应用程序的读写数据请求，向元数据管理单元 12 获取元数据，使用元数据信息向主用存储系统 13、备用存储系统 14 读写数据。作为一个功能模块可以单独部署在一个处理单元，如 PC 机，单板。一般接入模块有多个，以向系统外部提供高吞吐量，至少有一个。

[0023] 元数据管理单元 12，用于在接入模块 11 请求所述元数据时，查找所请求的数据在主用存储系统 13 或备用存储系统 14 上的位置，并构造所述元数据，返回给接入模块 11；还用于根据主用存储系统 13 的设备状态事件转换本分布式文件系统 1 的系统状态。分布式文件系统 1 具有三种状态：正常状态、异常状态和恢复状态：其中：

[0024] 正常状态，指主用存储系统 13 的存储设备没有出现故障，数据存储在主用存储系统上，有需要时会备份到备用存储系统 14 上。

[0025] 异常状态，指主用存储系统 13 有存储设备出现故障，此时需要主用存储系统 13、备用存储系统 14 共同工作，协调完成数据存储，并保存协调结果，协调结果用于数据恢复过程。

[0026] 恢复状态，指主用存储系统 13 恢复了故障存储设备后，触发系统进行系统数据恢复，主要使用备用存储系统 14 上的备份数据恢复到原来存储在故障存储设备上的数据。

[0027] 进一步地，元数据管理单元 12 负责在接入模块 11 请求元数据时，查找数据在存储系统上的位置并构造元数据，返回给接入模块 11。元数据管理单元 12 也负责管理系统状态，其接收主用存储系统 13 设备状态事件，根据事件信息，转换系统状态、选择存储系统，决定数据存储设备（位置）；还负责系统的可靠性，其根据数据访问信息，自动备份数据。在设备故障时，根据主用、备用、外部存储系统的数据分布情况，自动组成可用元数据，保证系统的可用性；以及负责系统数据自动在线恢复，确保数据的可用性与数据一致性。数据管理单元 12 可以单独部署在一个服务器上。

[0028] 主用存储系统 13，用于在所述的正常状态下为接入模块 11 提供所请求的数据。具体地，主用存储系统 13 是分布式文件系统 1 的主要存储，用来保存系统所有数据，以高读写性能为目的，系统正常状态下使用主用存储系统 13 读写数据并且全部数据都保存在主用存储系统 13 上，由高速存储设备组成，使用存储条带化技术支持存储带宽聚合提高读写性能，提供块访问方式，接入模块 11 可以以块方式直接读写其上存储的数据。

[0029] 备用存储系统 14，用于在所述的异常状态和所述的恢复状态下，为主用存储系统 13 提供数据备份。 具体地，备用存储系统 14 是分布式文件系统 1 的次要存储，用来备份数据以支持系统的可靠性、可用性、数据可恢复性，使用存储条带化技术支持存储带宽聚合提高读写性能，提供块访问方式，接入模块 11 可以以块方式直接读写其上存储的数据。

[0030] 外部存储系统 15，用于为主用存储系统 13 提供数据备份。 具体地，外部存储系统 15 指存储有更多数据的存储系统，可以是上一层或同层其它系统。 外部存储系统 15 通过网络与系统的其他模块连接起来，使用网络访问方式读写数据，用作外部数据备份补充，如果在主用存储系统 13、备用存储系统 14 找不到相应数据，就可以向外部存储系统 15 请求数据。

[0031] 上述的内部网络是连接接入模块 11、元数据管理单元 12、主用存储系统 13 和备用存储系统 14 直连的网络，可以是以太网也可以是内部总线（如 PCIe 总线）。 元数据主要是描述数据属性的信息，用来支持如指示存储位置、历史信息、资源信息以及文件记录等，例如可以是保存请求的数据（请求的长度和偏移）的存储地址（可以是包括存储设备号、存储块号、数据偏移）。 元数据也可以是数据的索引节点号。 数据的索引节点保存在存储设备中，并标识数据的存储地址。 根据数据的索引节点号，可以计算出索引节点的存储地址。

[0032] 图 2 为本发明另一实施例分布式文件系统组成示意图，基于上述实施例，如图 2 所示，该分布式文件系统 1 包括接入模块 11、元数据管理单元 12、主用存储系统 13 和备用存储系统 14 和外部存储系统 15。 进一步地，元数据管理单元 12 包括元数据操作模块、备份模块和异常恢复模块；其中：

[0033] 所述元数据操作模块，分别与接入模块 11、主用存储系统 13 连接，用于接收所述元数据请求，向主用存储系统 13 或异常恢复模块请求元数据并向接入模块 11 返回所述元数据；还用于根据接收的主用存储系统 13 上报的设备状态事件，更新所述系统状态；还负责系统的状态管理，记录所述读 / 写数据请求中的数据控制信息。 具体地，所述元数据操作模块接收接入模块 11 所有的元数据请求，首先检索其元数据缓存，如果没有查找到元数据，则需要向主用存储系统 13 或备用存储系统 14 请求元数据，缓存得到元数据，最后返回元数据给接入模块 11。

[0034] 在正常状态下，检索元数据缓存，如果没有，则向主用存储系统 13 请求元数据，同时记录下接收到的元数据请求中的数据控制信息，然后返回元数据给接入模块 11。 数据控制信息可以记录到内存，也可以持久化到数据库。 本实施例中，元数据管理单元 12 可以包括一个数据控制信息记录模块，连接在元数据操作模块和备份模块之间，用于永久性存储所述的数据控制信息。

[0035] 在异常状态下，元数据操作模块把接收模块 11 发来的元数据请求转发给异常恢复模块，由异常恢复模块负责向主用存储系统 13、备用存储系统 14 获取元数据信息。

[0036] 在恢复状态下，写数据过程跟正常状态下一样，元数据操作模块向主用存储系统 13 请求元数据；读数据过程，请求异常恢复模块负责处理。

[0037] 元数据操作模块负责状态管理，主用存储系统 13 负责上报设备状态事件，当接收到设备故障事件，记录故障设备 ID，系统状态由正常转为异常；当接收到设备恢复事

件时，系统状态由异常转为恢复；当恢复操作完成时，由异常恢复模块通报元数据操作模块后，系统状态由恢复转为正常。

[0038] 所述备份模块，与所述元数据操作模块和备用存储系统 13 连接，用于读取所述元数据操作模块记录下来的所述数据控制信息，生成数据备份操作控制信息并发送给备用存储系统 14，以将主用存储系统 13 上的数据备份到备用存储系统 14 中。 具体地，所述备份模块在正常状态下，以后台线程运行，读取前一段时间的元数据操作模块记录下来的数据控制信息，根据备份策略分析出数据使用状况并生成数据备份请求，然后向备用存储系统 14 上发出备份操作控制信息（操作类型，目标文件路径，源文件路径），要求将主用存储系统上的文件备份到备用存储服务器上指定位置。 所述备份模块结合备份策略，可以灵活实现各种数据备份方案，包括全备份方案，分析数据时只分析写数据操作，一旦是有写数据就需要生成备份请求；包括热点备份方案，分析数据时只分析读数据使用情况（读数据请求次数、读数据频率），根据策略中的数据热点条件（可以使用次数、读数据频率）；包括备份指定数据方案，可以由策略指定数据特征码，分析数据时分析出指定数据特征码并生成备份请求；包括分析出数据里备份请求信息，根据备份请求信息生成备份请求。

[0039] 所述异常恢复模块，分别与所述元数据操作模块、所述备份模块、主用存储系统 13 和备用存储系统 14 连接，用于在所述异常状态和所述恢复状态下获取所述元数据并返回给所述元数据操作模块。 所述异常恢复模块负责异常、恢复状态下获取元数据信息并维护元数据缓存的有效性，以及控制主用存储系统 13 数据恢复操作，以支持主用存储系统 13 发生故障下系统可用性。

[0040] 所述异常恢复模块可以包括如下子模块：

[0041] 第一处理子模块，用于在所述异常状态下，对于读数据请求，对从缓存或从主用存储系统 13 获得的元数据进行缺失块检测，若检测到所述元数据有缺失块，则向备用存储系统 14 发送请求信息包括缺失块信息和外部存储系统 14 的地址；在备用存储系统 14 返回所述元数据后，进行块地址重绑定操作，重组可用的元数据并发送。 具体地，在异常状态下，对于读数据请求，所述异常恢复模块通过第一处理子模块首先检索其元数据缓存，如果没有查找到元数据，向主用存储系统 13 请求元数据，接着对元数据进行缺失块（数据存储在故障存储设备上，检测其存储设备 ID）检测，如果发现有缺失块，再向备用存储系统 14 请求元数据，请求信息包括缺失块信息（块号，数据 ID，偏移）与外部存储系统 15 地址，在备用存储系统 14 返回元数据后，进行块地址重绑定操作。 块地址重绑定操作指把备用存储系统 14 上的缺失块对应的元数据信息替换掉主用存储系统 13 上的缺失块对应的元数据信息，重组成可用的元数据并缓存此元数据以加快元数据获取。

[0042] 第二处理子模块，用于在所述异常状态下，对于写数据请求，仅向备用存储系统 14 请求元数据并返回给所述元数据操作模块；还用于记录写数据操作，以在数据恢复过程中根据记录的所述写数据操作将数据同步到主用存储系统 13 中。 具体地，在异常状态下，对于写数据请求，所述异常恢复模块通过第二处理子模块自动选择备用存储系统 14 为存储目标，只向备用存储系统 14 请求元数据并返回请求结果到元数据操作模块，这样，接入模块 11 根据元数据信息会把数据都写到备用存储系统 14 上，异常恢复模块还会记录下写数据操作，数据恢复过程中，会根据这些记录把数据同步到主用存储系统 13。

[0043] 恢复状态下，异常恢复模块负责在线数据恢复，可以同时进行数据服务与数据恢复。 异常恢复模块通过故障存储设备 bitmap 位图表示缺失块恢复情况，恢复过程中，异常恢复模块负责维护 bitmap 的更新。 检测缺失块时，先根据存储设备 ID，再对比数据块所对应的 bitmap 是否已经恢复。

[0044] 异常恢复模块还可以包括第三处理子模块，用于在所述恢复状态下，对于读数据请求，对从缓存或从主用存储系统 13 获得的元数据进行缺失块检测，若检测到所述元数据有缺失块，则向备用存储系统 14 发送请求信息包括缺失块信息，在备用存储系统 14 返回元数据后，构造缺失块恢复操作信息，以供主存储系统 13 进行数据恢复；并更新主存储系统 13 的故障存储设备位图，返回所述元数据。

[0045] 具体地，在恢复状态下，对于读数据请求，异常恢复模块通过第三处理子模块首先检索其元数据缓存，如果没有查找到元数据，则向主用存储系统 13 请求元数据，接着对从缓存或主用存储系统 13 得到的元数据进行缺失块检测，如果发现有缺失块，再向备用存储系统 14 请求元数据，传递缺失块信息（块号，数据 ID，偏移），在备用存储系统 14 返回元数据后，构造主用存储系统 13 缺失块恢复操作信息，主用存储系统 13 根据此信息或者向备用存储系统 14 或者向外部存储系统 15 恢复数据。 恢复后，更新故障存储设备 bitmap 位图，返回元数据。

[0046] 异常恢复模块还可以包括第四处理子模块，用于在所述恢复状态下，对于写数据请求，向主用存储系统 13 请求元数据，获得所述元数据后，若检测到所述元数据有缺失块，则更新主存储系统 13 的故障存储设备位图，返回所述元数据。 具体地，在恢复状态下，对于写数据请求，异常恢复模块通过第四处理子模块向主用存储系统 13 请求元数据，得到元数据后，以故障存储设备 ID 检测缺失块，如果有缺失块则直接更新故障存储设备 bitmap 位图，然后返回请求结果到元数据操作模块。

[0047] 在恢复状态下，异常恢复模块启动一个后台恢复线程。 恢复线程根据恢复设备的类型，选择恢复过程。 如果是存储设备没有读写故障（可能是存储设备热插拔后又插回来），只需要把在异常期间写到备用存储系统的数据保存到主用存储系统，并删掉重绑定的元数据缓存。

[0048] 如果是存储设备读写故障，遍历主用存储系统的数据，其恢复过程如下：异常恢复模块首先检索其元数据缓存，如果查找到元数据缓存，就要检查是否有数据在备用存储系统，如果有数据在备用存储系统（也就是有重绑定过的元数据缓存），则构造主用存储缺失块恢复操作信息，主存储系统根据此信息向备份存储系统恢复数据。 恢复后，删掉重绑定过的元数据缓存并更新故障存储设备 bitmap 位图。

[0049] 如果在元数据缓存没有查找到元数据，则向主用存储系统请求元数据，接着对从主用存储系统得到的元数据进行缺失块检测，如果发现有缺失块，再向备用存储系统请求元数据，请求信息包括缺失块信息（块号，数据 ID，偏移），但不包括外部存储地址，在备用存储系统返回元数据后，构造主用存储系统缺失块恢复操作信息，主存储系统根据此信息或者向备份存储系统或者向外部存储系统恢复数据。 恢复后，更新故障存储设备 bitmap 位图。

[0050] 后台恢复线程在遍历主用存储系统的数据后，还遍历在异常期间异常恢复模块记录下写数据操作，负责把写在备用存储系统的数据保存到主用存储系统，然后更新故

障存储设备 bitmap 位图并删掉元数据缓存。

[0051] 如图 2 所示，该分布式文件系统中还可以包括主用数据管理模块，连接在主用存储系统 13 和元数据管理单元 12 之间，用于管理主用存储系统 13 所存储的数据，响应元数据请求和数据操作请求。 具体地，主用数据管理模块负责响应元数据请求，还负责接收恢复操作信息并实现数据恢复，恢复操作信息有两类，一种有备用存储系统地址信息，一种有外部存储系统地址信息的。 主用数据管理模块根据恢复操作信息，或者向备用存储系统或者外部存储系统恢复数据块。

[0052] 该分布式文件系统中还可以包括备用数据管理模块，连接在备用存储系统 14 和元数据管理单 12 元之间，用于管理备用存储系统 14 所存储的数据，响应元数据请求和数据操作请求。 具体地，备用数据管理模块负责响应元数据请求，还负责接收备份请求并实现数据备份。 元数据请求有两类，第一类有外部存储地址信息，第二类没有外部存储系统地址，备用数据管理模块对这两类元数据请求的区别是：如果在备用存储系统 14 检索不到元数据，则处理第一类时会向外部存储系统 15 请求所有数据并存储下来，返回存储后的元数据。 备用存储系统 14 处理备份请求时，根据备份请求，直接向主用存储系统 13 请求数据并保存下来。

[0053] 上述的分布式文件系统中，主用存储系统包括数个高速存储设备，所述的高速存储设备包括但不限于高数据传输率的 SCSI 硬盘、SATA 硬盘、SSD。 主用存储系统也负责监控自身存储设备状态，并上报设备事件（如设备故障，设备恢复）。上述设备故障包括存储设备读写故障、热插拔存储设备。 备用存储系统包括数个高速存储设备和 / 或低速存储设备，其中，高速存储设备包括但不限于高数据传输率的 SCSI 硬盘、 SATA 硬盘、SSD；低速存储设备包括但不限于低数据传输率的存储设备。

[0054] 上述实施例中所述的恢复操作信息有两类，一种有备用存储系统地址信息，一种有外部存储系统地址信息的。 主用数据管理模块根据恢复操作信息，或者向备用存储系统或者外部存储系统恢复数据块。 也就是说，有备用存储系统地址信息的恢复操作信息，主用数据管理模块从备用存储系统恢复数据块有外部存储系统地址信息的恢复操作信息，主用数据管理模块从外部存储系统恢复数据块在恢复状态下，读数据时，就会构造上面的恢复操作信息，这样，既提供读数据服务又马上恢复所访问的数据，根据数据的局部性原理，最近访问的数据也就是大多数用户关注的数据，优先恢复这些数据，有利于提高性能。

[0055] 关于外部存储地址：在异常状态下，读数据时，需要向备用存储系统请求元数据，这是包括外部存储地址的，这希望备用存储系统本身没有所请求的元数据时，备用存储系统先从外部存储系统请求数据（相当于把所有数据从外部存储系统备份到备用存储系统），再返回元数据。这样，既保证读数据服务又把数据从外部存储系统备份到备用存储系统中，而恢复时可以从备用存储系统中恢复数据，从而加速数据恢复。 在恢复状态下，读数据时或后台数据恢复线程时，可能会向备用存储系统请求元数据，这是不包括外部存储地址的，这样，如果备用存储系统本身没有所请求的元数据，备用存储系统不会向外部存储存储系统请求数据，备用存储系统返回空。 接着主用存储系统使用外部存储地址，向外部存储系统请求数据（只请求缺失块所包含的数据，数据恢复时只需要恢复缺失块所包含的数据）。

[0056] 本发明实施例提供的分布式文件系统，将数据写入到主用存储系统中，采用异步备份机制备份到备用存储系统，不影响高速设备的读写性能；可以在线恢复数据，数据对外服务过程中自动实现数据恢复，加快恢复过程；而且恢复数据时不需要计算，即使数据没有备份，仍可以通过外部存储恢复；在 CDN 环境下，使用策略备份数据机制与可用外部存储获取没有备份的数据，可实现部分数据备份而不影响可用性；另外，可以用廉价的存储设备组成备用存储系统，降低产品成本。

[0057] 图 3 为本发明实施例分布式文件系统在正常状态下处理流程示意图，如图 3 所示，该流程包括：

[0058] 步骤 1、备份管理模块读取数据控制信息；

[0059] 步骤 2、备份管理模块按照备份策略对数据控制信息进行读写请求情况分析，如果满足策略要求，组成备份控制信息，发到备用数据管理模块，要求将主用存储系统上的文件备份到备用存储服务器上；

[0060] 步骤 3、备用数据管理模块接收备份管理模块发过来的备份控制请求（操作类型，目标文件路径，源文件路径），根据备份控制请求信息，向主用数据管理模块发出备份请求；

[0061] 步骤 4、主用数据管理模块接收备份请求，向主用存储系统中的存储设备读取数据；

[0062] 步骤 5、主用数据管理模块返回数据给备用数据管理模块；

[0063] 步骤 6、备用数据管理模块把数据写到备用存储系统中的存储设备上；

[0064] 步骤 7、返回备份情况给备份管理模块。

[0065] 图 4 为本发明实施例分布式文件系统在异常状态下并且为读数据请求情况下的处理流程示意图，如图 4 所示，该流程包括：

[0066] 步骤 1、应用程序向接入模块发出读数据的请求；

[0067] 步骤 2、接入模块向元数据操作模块发出读取元数据请求；

[0068] 步骤 3、元数据操作模块把读取元数据请求转发给异常恢复模块；

[0069] 步骤 4、异常恢复模块先在元数据信息缓存查找元数据，如果找到转向第 6 步；

[0070] 步骤 5、异常恢复模块向主用数据管理模块发起元数据请求，主用数据管理模块接收备份请求，向主用存储系统中的存储设备读取元数据，然后返回元数据给异常恢复模块；

[0071] 步骤 6、异常恢复模块检查从元数据缓存或主用存储系统返回来的元数据信息是否有缺失块存在，有则向备用存储系统发起元数据请求并带上一个外部备份控制信息（包括外部存储系统位置，数据位置信息），没有则返回元数据信息到元数据操作模块，并转向第 9 步；

[0072] 步骤 7、如果此数据已经备份在备用存储系统，则备用数据管理模块返回元数据，否则备用数据管理模块根据外部备份控制信息向外部存储系统请求数据，把所得数据存储在备用存储系统，并返回元数据信息；

[0073] 步骤 8、异常恢复模块进行块地址重绑定操作，修改缺失块的元数据的块映射表，将在备用存储系统上应用相应的块地址替换掉缺失块地址，并将生绑定的元数据缓存起来，返回所有的元数据给元数据操作模块；

[0074] 步骤 9、元数据操作模块返回元数据给接入模块；

[0075] 步骤 10、接入模块根据返回的元数据信息，向相应的主用存储系统、备用存储系统发起 IO 数据请求；

[0076] 步骤 11、主用存储系统、备用存储系统返回数据给接入模块；

[0077] 步骤 12、接入模块响应数据返回给应用程序。

[0078] 图 5 为本发明实施例分布式文件系统在异常状态下并且为写数据请求情况下的处理流程示意图，如图 5 所示，该流程包括：

[0079] 步骤 1、应用程序向接入模块发出写数据的请求；

[0080] 步骤 2、接入模块向元数据操作模块发出读取元数据请求；

[0081] 步骤 3、元数据操作模块把读取元数据请求转发给异常恢复模块；

[0082] 步骤 4、异常恢复模块直接向备用存储系统发起元数据请求；

[0083] 步骤 5、备用数据管理模块接收元数据请求并构造元数据，返回元数据；

[0084] 步骤 6、异常恢复模块返回元数据信息给元数据操作模块；

[0085] 步骤 7、元数据操作模块返回元数据给接入模块；

[0086] 步骤 8、接入模块根据返回的元数据信息，向备用存储设备发起 IO 数据请求；

[0087] 步骤 9、接入模块写数据到备用存储系统；

[0088] 步骤 10、接入模块返回写数据结果给应用程序。

[0089] 图 6 为本发明实施例分布式文件系统在恢复状态下响应读数据请求的处理流程示意图，如图 6 所示，该流程包括：

[0090] 步骤 1、应用程序向接入模块发出写数据的请求；

[0091] 步骤 2、接入模块向元数据操作模块发出读取元数据请求；

[0092] 步骤 3、元数据操作模块把读取元数据请求转发给异常恢复模块；

[0093] 步骤 4、异常恢复模块检索元数据缓存，如果没有找到，则向主用存储系统请求元数据，如果找到跳到第 6 步；

[0094] 步骤 5、主用存储系统接收请求元数据，返回元数据；

[0095] 步骤 6、异常恢复模块检查有没有缺失块，如果有则向备用存储系统请求元数据，如果没有，则跳到第 11 步；

[0096] 步骤 7、异常恢复模块根据备用存储系统返回的元数据，构造恢复控制信息发给主用存储系统；

[0097] 步骤 8、主用存储系统接收到恢复控制信息，并执行数据恢复；

[0098] 步骤 9、返回数据恢复结果到异常恢复模块；

[0099] 步骤 10、异常恢复模块更新已经恢复的缺失块 bitmap 位图；

[0100] 步骤 11、返回从元数据缓存或主用存储系统得到的元数据给元数据操作模块；

[0101] 步骤 12、元数据操作模块返回元数据给接入模块；

[0102] 步骤 13、接入模块根据返回的元数据信息，向相应的主用存储系统、备用存储系统发起 IO 数据请求；

[0103] 步骤 14、主用存储系统、备用存储系统返回数据给接入模块；

[0104] 步骤 15、接入模块响应数据返回给应用程序。

[0105] 图 7 本发明实施例分布式文件系统在恢复状态下响应写数据请求的处理流程示

意图，如图 7 所示，该流程包括：

[0106] 步骤 1、应用程序向接入模块发出写数据的请求；

[0107] 步骤 2、接入模块向元数据操作模块发出读取元数据请求；

[0108] 步骤 3、元数据操作模块把读取元数据请求转发给异常恢复模块；

[0109] 步骤 4、异常恢复模块向主用存储系统请求元数据；

[0110] 步骤 5、主用存储系统接收请求元数据，返回元数据；

[0111] 步骤 6、异常恢复模块检查有没有缺失块，如果有，则更新已经恢复的缺失块 bitmap 位图，把缺失块所对应的位图设置为 1；

[0112] 步骤 7、返回从元数据给元数据操作模块。

[0113] 图 8 为本发明实施例分布式文件系统在系统恢复状态下数据恢复过程的处理流程示意图，如图 8 所示，该流程包括：

[0114] 步骤 1、后台数据恢复线程在元数据缓存中查找将要恢复的数据的元数据，如果没有找到则跳到第 6 步；

[0115] 步骤 2、如果找到元数据缓存，后台数据恢复线程检查是否有数据在备用存储系统上，如果没有数据在备用存储系统上，跳到第 11 步；

[0116] 步骤 3、如果有数据在备用存储系统上，后台数据恢复线程构造主用存储缺失块恢复操作信息并发给主用存储系统；

[0117] 步骤 4、主用存储系统接收到缺失块恢复操作信息并执行，将备用存储系统数据恢复到缺失块，返回缺失块恢复操作结果；

[0118] 步骤 5、后台数据恢复线程删掉重绑定过的元数据缓存并跳到第 11 步；

[0119] 步骤 6、如果在元数据缓存没有找到元数据，则后台数据恢复线程向主用存储系统请求元数据；

[0120] 步骤 7、对返回的元数据进行缺失块检查，如果没有缺失块，跳到第 11 步；

[0121] 步骤 8、如果有缺失块，后台数据恢复线程请求向备用存储系统恢复数据；

[0122] 步骤 9、如果第 8 步备份成功，则跳到第 11 步；

[0123] 步骤 10、如果第 8 步备份失败，则后台数据恢复线程请求向外部存储系统恢复数据；

[0124] 步骤 11、更新故障存储设备 bitmap 位图，则此数据恢复完成。

[0125] 本领域普通技术人员可以理解：实现上述方法实施例的全部或部分步骤可以通过程序指令相关的硬件来完成，前述的程序可以存储于一计算机可读取存储介质中，该程序在执行时，执行包括上述方法实施例的步骤；而前述的存储介质包括：ROM、RAM、磁碟或者光盘等各种可以存储程序代码的介质。

[0126] 最后应说明的是：以上实施例仅用以说明本发明的技术方案，而非对其限制；尽管参照前述实施例对本发明进行了详细的说明，本领域的普通技术人员应当理解：其依然可以对前述各实施例所记载的技术方案进行修改，或者对其中部分技术特征进行等同替换；而这些修改或者替换，并不使相应技术方案的本质脱离本发明各实施例技术方案的精神和范围。

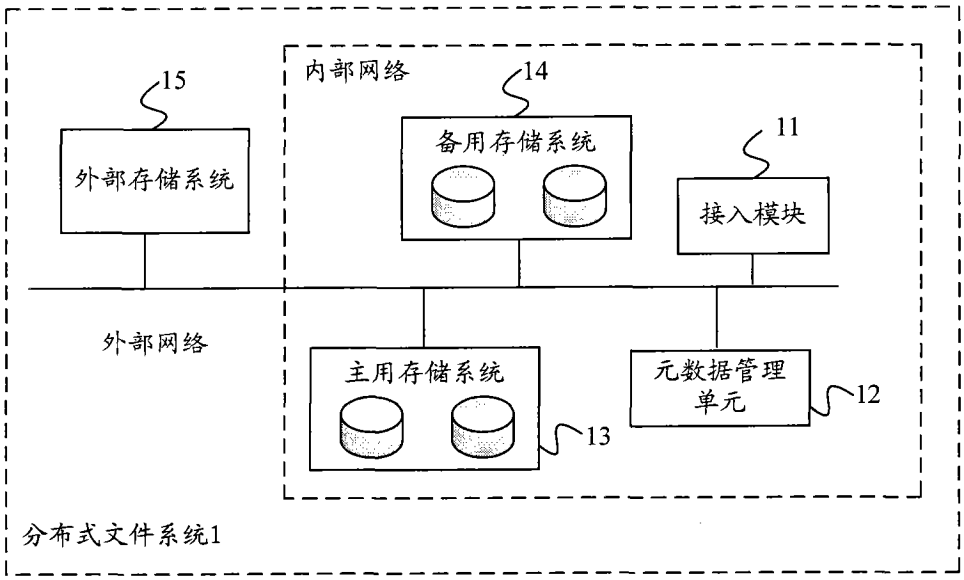

15
外部存储系统
内部网络
14
备用存储系统
11
接入模块
外部网络
主用存储系统
13
元数据管理单元
12
分布式文件系统1

图 1

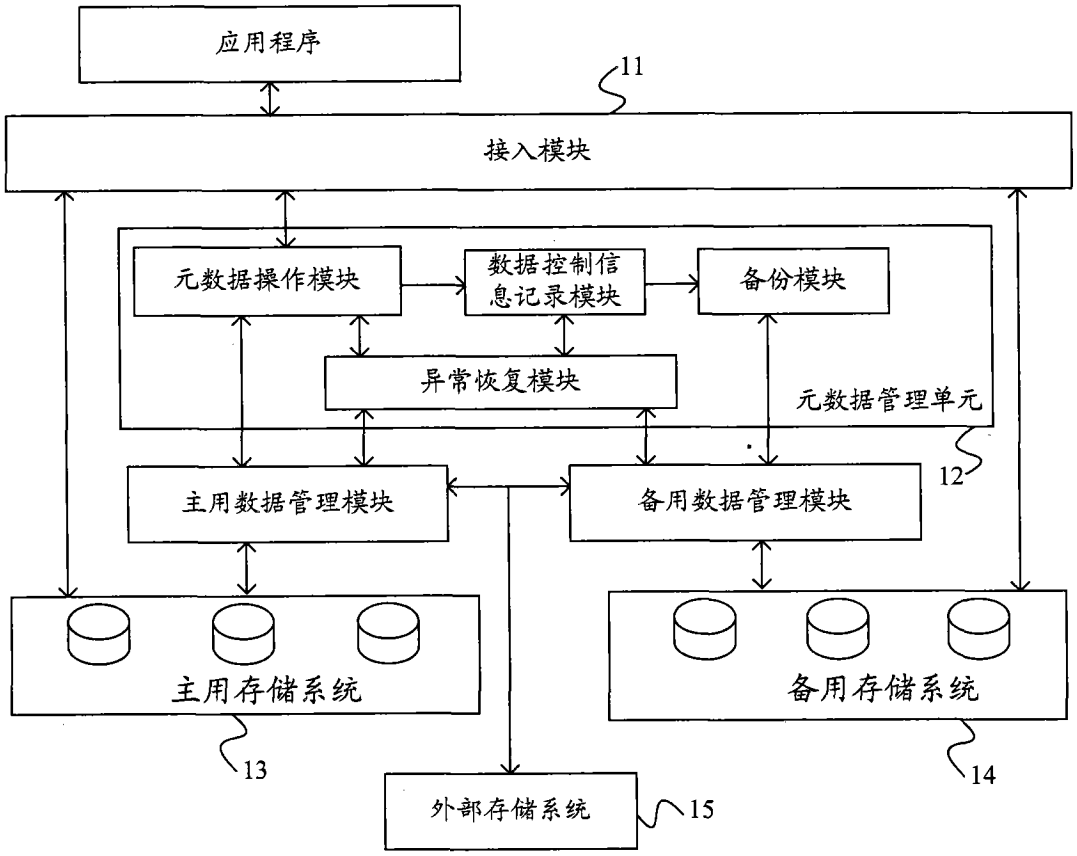

应用程序
11
接入模块
元数据操作模块
数据控制信息记录模块
备份模块
异常恢复模块
元数据管理单元
12
主用数据管理模块
备用数据管理模块
主用存储系统
13
备用存储系统
14
外部存储系统
15

图 2

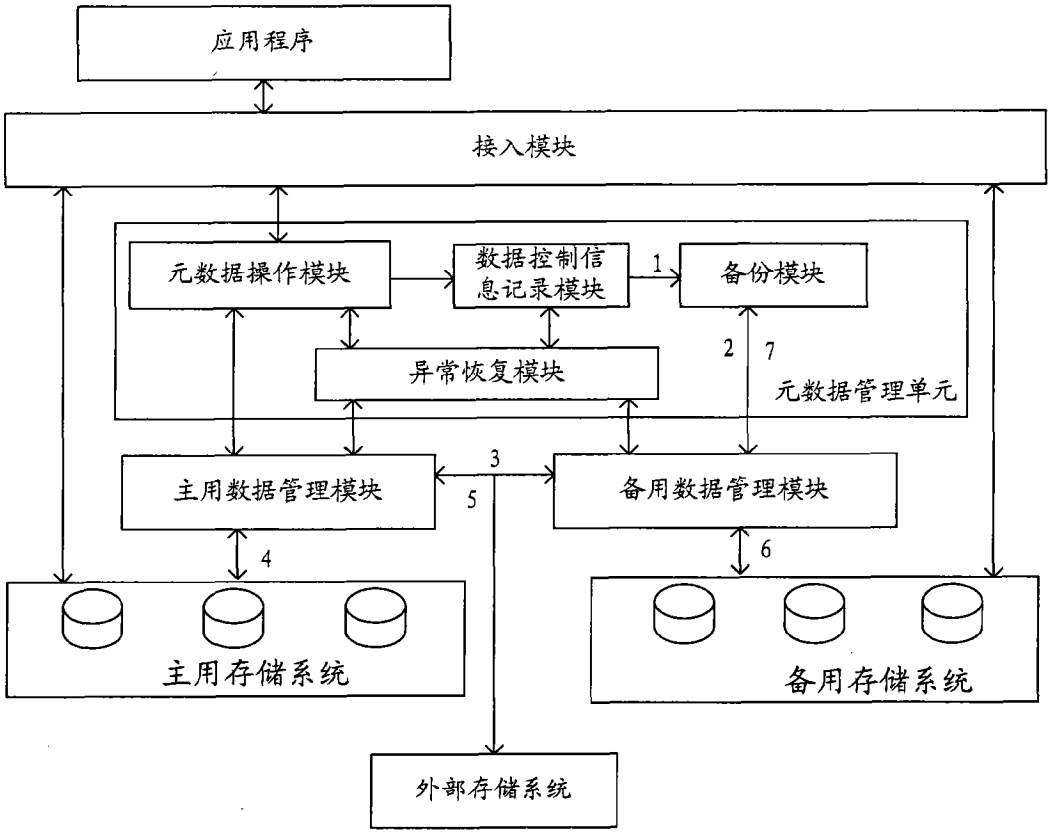

应用程序
接入模块
元数据操作模块
数据控制信息记录模块
1
备份模块
异常恢复模块
2
7
元数据管理单元
主用数据管理模块
3
5
备用数据管理模块
4
6
主用存储系统
备用存储系统
外部存储系统

图 3

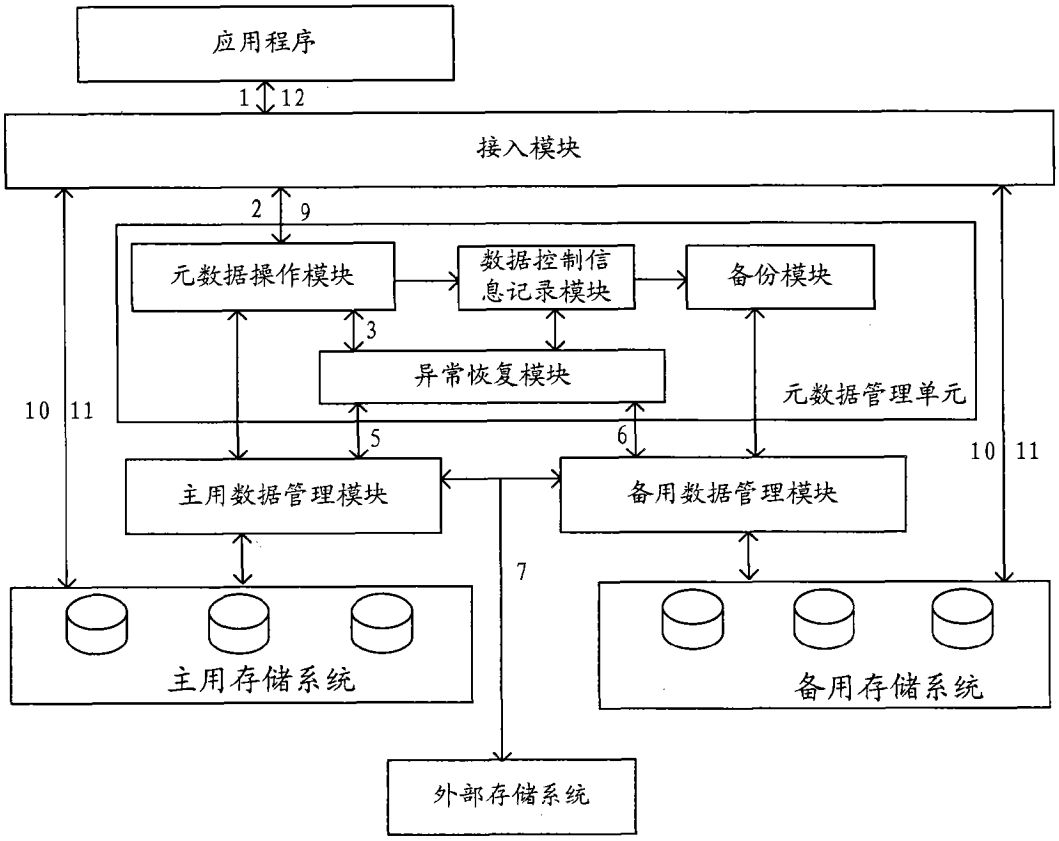

应用程序
1
12
接入模块
2
9
元数据操作模块
数据控制信息记录模块
备份模块
3
异常恢复模块
元数据管理单元
10
11
5
6
10
11
主用数据管理模块
备用数据管理模块
7
主用存储系统
备用存储系统
外部存储系统

图 4

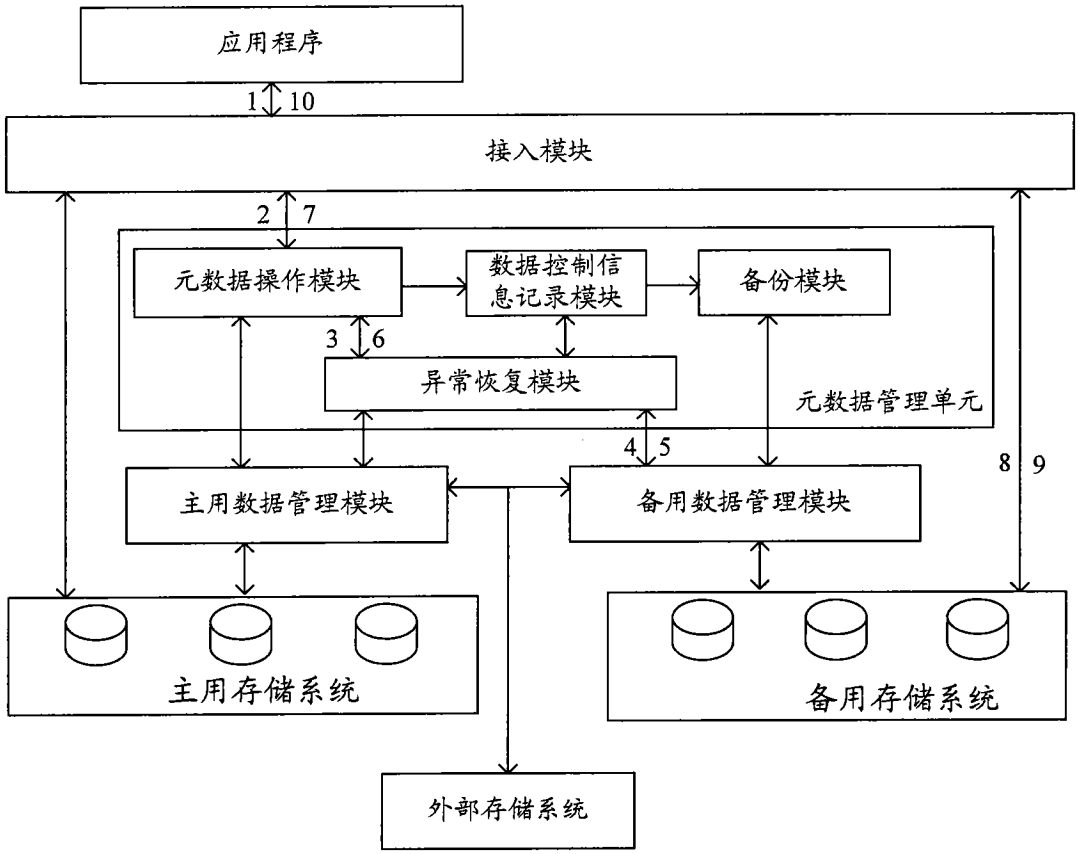

应用程序
1
10
接入模块
2
7
元数据操作模块
数据控制信息记录模块
备份模块
3
6
异常恢复模块
元数据管理单元
4
5
8
9
主用数据管理模块
备用数据管理模块
主用存储系统
备用存储系统
外部存储系统

图 5

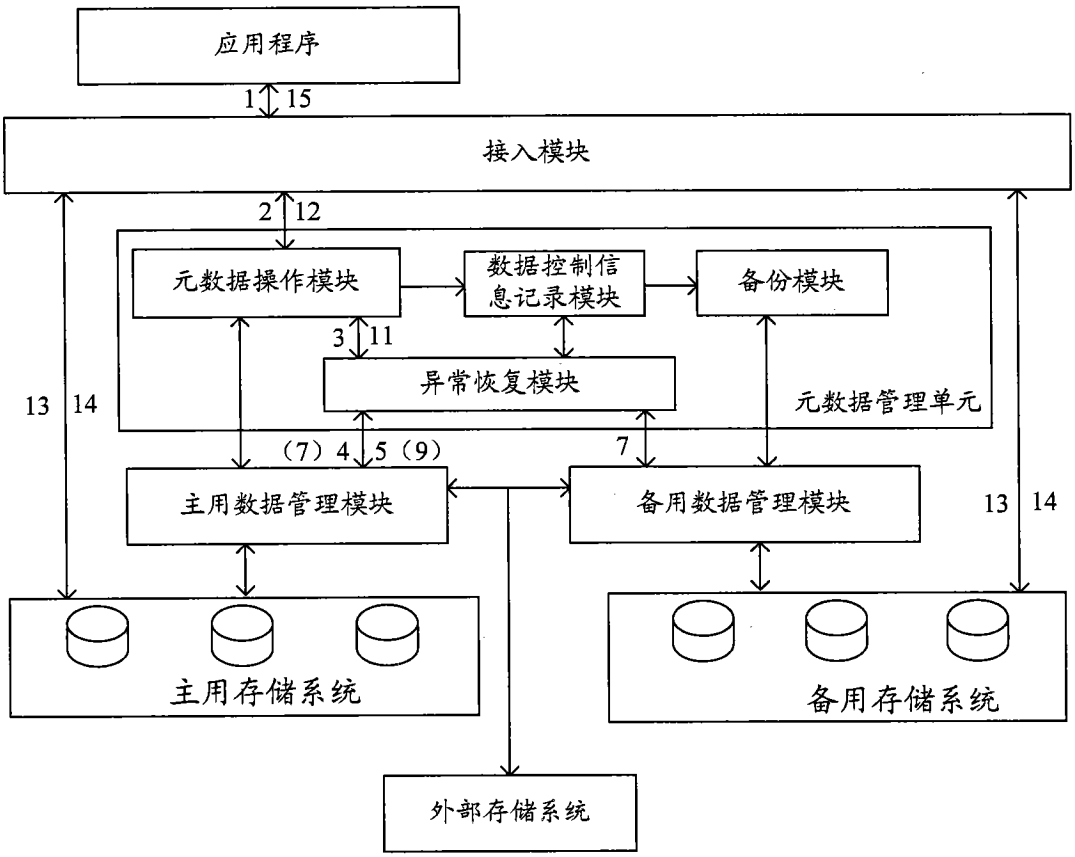

应用程序
1
15
接入模块
2
12
元数据操作模块
数据控制信息记录模块
备份模块
3
11
异常恢复模块
元数据管理单元
13
14
(7) 4
5 (9)
7
主用数据管理模块
备用数据管理模块
13
14
主用存储系统
备用存储系统
外部存储系统

图 6

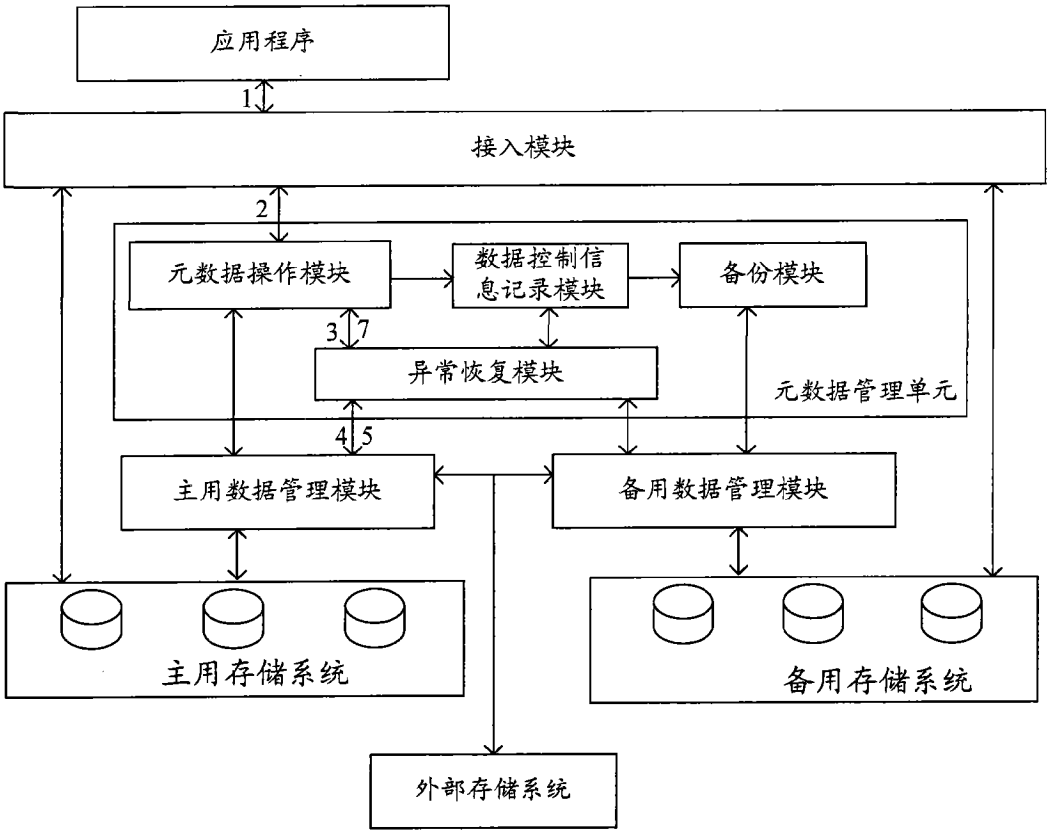

应用程序
1
接入模块
2
元数据操作模块
数据控制信息记录模块
备份模块
3
7
异常恢复模块
元数据管理单元
4
5
主用数据管理模块
备用数据管理模块
主用存储系统
备用存储系统
外部存储系统

图 7

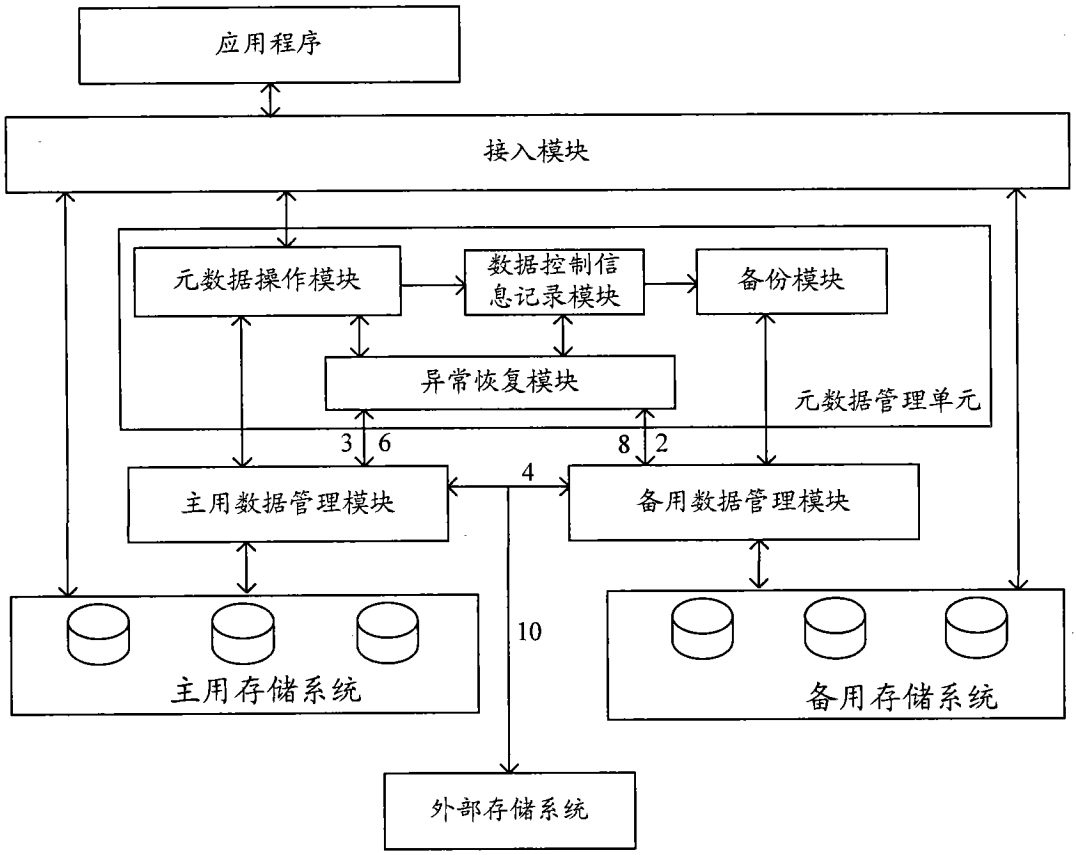

应用程序
接入模块
元数据操作模块
数据控制信息记录模块
备份模块
异常恢复模块
元数据管理单元
3
6
8
2
4
主用数据管理模块
备用数据管理模块
10
主用存储系统
备用存储系统
外部存储系统

图 8