(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 102317939 A
(43)申请公布日 2012.01.11

(21)申请号 200980156985.4
(22)申请日 2009.12.22
(30)优先权数据
61/139,853 2008.12.22 US
(85)PCT申请进入国家阶段日
2011.08.17
(86)PCT申请的申请数据
PCT/US2009/069228 2009.12.22
(87)PCT申请的公布数据
WO2010/075401 EN 2010.07.01
(71)申请人 谷歌公司
地址 美国加利福尼亚州
(72)发明人 亚历山大·克塞尔曼
亚历山大·德罗贝切夫
(74)专利代理机构 中原信达知识产权代理有限责任公司 11219
代理人 周亚荣 安翔
(51)Int.Cl.
G06F 17/30(2006.01)

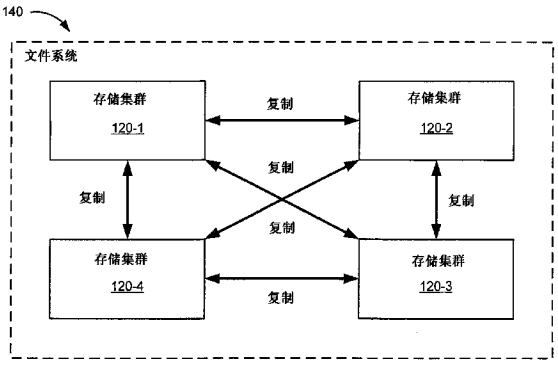
权利要求书 3 页 说明书 10 页 附图 8 页

(54)发明名称

用于复制的存储集群的异步分布式垃圾收集

(57)摘要

一种方法可以由分布式数据复制系统中的设备组中的设备执行。该方法可以包括：将对象存储在数据存储中，对象中的至少一个或多个是利用分布式数据复制系统来复制的；以及进行对数据存储中的对象的扫描。该方法可以进一步包括：将对象中的一个识别为没有指向对象的引用；将删除协商消息存储为与对象中的一个相关联的元数据；以及将带有删除协商消息的元数据复制到设备组中的一个或多个其它设备。



CN 102317939 A

1. 一种由分布式数据复制系统中的多个设备中的设备执行的方法，所述方法包括：

将对象存储在数据存储中，所述对象中的至少一个或多个是利用所述分布式数据复制系统来复制的；

进行对所述数据存储中的所述对象的扫描；

将所述对象中的一个识别为没有指向所述对象的引用；

将删除协商消息存储为与所述对象中的所述一个相关联的元数据；以及

将带有所述删除协商消息的所述元数据复制到所述多个设备中的一个或多个其它设备。

2. 根据权利要求 1 所述的方法，其中带有所述删除协商消息的所述元数据是使用所述分布式多主站数据复制系统的基础复制层来复制的。

3. 根据权利要求 1 所述的方法，进一步包括：

将所述对象中的另一个识别为具有在与所述对象中的所述另一个相关联的元数据中的另一个删除协商消息。

4. 根据权利要求 3 所述的方法，进一步包括：

将对所述另一个删除协商消息的应答、否定应答或确认中的一个存储为与所述对象中的所述另一个相关联的元数据；以及

将带有对所述删除协商消息的所述应答或所述否定应答的所述对象中的所述另一个的所述元数据复制到所述多个设备中的一个或多个其它设备。

5. 根据权利要求 3 所述的方法，进一步包括：

如果另一个删除协商消息指示失败的协商，则从所述对象元数据删除所述协商消息。

6. 根据权利要求 3 所述的方法，进一步包括：

如果另一个删除协商消息指示成功的协商，则删除所述对象。

7. 根据权利要求 1 所述的方法，其中存储所述删除协商消息包括：

写协商消息指示符、存储集群标识以及唯一协商请求标识。

8. 根据权利要求 1 所述的方法，其中新的引用不能被添加到带有所述删除协商消息的所述对象中的所述一个。

9. 一种分布式数据复制系统中的多个设备中的设备，所述设备包括：

用于将数据存储中的对象识别为具有与所述对象相关联的协商消息的装置；

用于将新的协商消息与所述对象进行关联的装置，所述新的协商消息基于所述对象的状态；

用于将所述新的协商消息复制到存储集群的装置；

用于接收与所述对象的复制品相关联的其它协商消息的装置；以及

用于如果所述其它协商消息指示成功的协商，则删除所述对象的装置。

10. 根据权利要求 9 所述的系统，进一步包括：

用于如果所述其它协商消息指示失败的协商，则删除所述新的协商消息和所述其它协商消息的装置。

11. 一种系统，包括：

存储器，所述存储器用来存储指令和数据存储；以及

处理器，所述处理器用来执行所述存储器中的所述指令以：

识别所述数据存储中的对象的状态，所述状态与所述对象是否具有引用以及删除协商消息是否与所述对象相关联有关；

基于所述对象的所述状态将新的协商消息写入所述对象元数据，

将带有所述新的协商消息的所述元数据复制到一个或多个设备，以及

从所述一个或多个设备中的一个接收与所述对象相关联的其它协商消息，

其中所述新的协商消息和所述其它协商消息为所述对象的删除协商提供共识。

12. 根据权利要求 11 所述的系统，其中所述新的协商消息和所述其它协商消息被嵌入在与所述对象相关联的元数据中，以及其中所述协商消息使用分布式多主站数据复制环境中的复制层来交换。

13. 根据权利要求 12 所述的系统，其中所述处理器被进一步配置为：

如果最终状态指示失败的删除协商，则从所述对象元数据删除所述新的协商消息。

14. 根据权利要求 12 所述的系统，其中所述处理器被进一步配置为：

如果最终状态指示成功的删除协商，则删除所述对象。

15. 根据权利要求 11 所述的系统，其中所述新的协商消息包括：

协商消息指示符；

存储集群标识；以及

唯一协商请求标识。

16. 根据权利要求 15 所述的系统，其中所述协商消息指示符包括以下中的一个：

用于发起删除协商的删除指示符，

用于提供对所述删除协商的肯定应答的应答指示符，

用于提供对所述删除协商的否定应答的否定应答指示符，或

用于提供已从其它存储集群接收应答的确认的同步指示符。

17. 一种方法，包括：

在分布式多主站数据复制系统内的存储集群之间交换一个或多个删除协商消息，其中所述删除协商消息中的每一个被包括在为所述删除协商消息的主题的对象的元数据中，以及其中所述删除协商消息是使用所述分布式多主站数据复制系统的复制层来在所述存储集群之间发送的；以及

如果基于所述一个或多个删除协商消息在所述存储集群之间达成分布式共识，则删除所述对象。

18. 根据权利要求 17 所述的方法，其中所述删除协商消息中的每一个包括：

协商消息指示符，

存储集群标识，以及

唯一协商请求标识。

19. 根据权利要求 17 所述的方法，其中所述协商消息指示符包括：

用于发起删除协商的删除指示符，

用于提供对所述删除协商的肯定应答的应答指示符，

用于提供对所述删除协商的否定应答的否定应答指示符，和

用于提供已从其它存储集群接收应答的确认的同步指示符。

20. 根据权利要求 19 所述的方法，其中所述删除指示符或所述应答指示符阻止发起所

述删除指示符或所述应答指示符的所述存储集群向所述对象添加新的引用。

21. 一种包括计算机可执行的指令的计算机可读存储器，所述计算机可读存储器包括：

用来识别数据存储中的对象的状态的一个或多个指令，所述状态与所述对象是否具有引用以及删除协商消息是否与所述对象相关联有关；

用来基于所述对象的所述状态来将与所述对象相关联的新的协商消息写入所述对象的元数据的一个或多个指令；

用来将带有所述新的协商消息的所述对象元数据复制到存储集群的一个或多个指令；

用来从一个或多个其它设备接收与所述对象相关联的其它协商消息的一个或多个指令；以及

用来基于与所述对象相关联的所述其它协商消息为所述对象的删除协商确定共识的一个或多个指令。

22. 根据权利要求 21 所述的计算机可读存储器，进一步包括：

用来如果所述共识指示失败的协商，则删除所述新的协商消息和所述其它协商消息的一个或多个指令；以及

用来如果所述共识指示成功的协商，则删除所述对象的一个或多个指令。

23. 根据权利要求 21 所述的计算机可读存储器，进一步包括：

用来维护所述新的协商消息和所述其它协商消息的写序保真度的一个或多个指令。

24. 根据权利要求 21 所述的计算机可读存储器，进一步包括：

用来在发起了删除协商消息之后阻止向所述对象添加新的引用的一个或多个指令。

用于复制的存储集群的异步分布式垃圾收集

背景技术

[0001] 企业计算局面经历了存储体系结构的根本性转变，因为中央服务体系结构已经让位于分布式存储集群。随着企业寻求增加存储效率的方式，商品计算机的这样的集群构建可以以与庞大盘阵列相比的成本的一小部分，为新的数据密集型应用提供高性能、可用性和可伸缩性。为了开启存储集群的全部潜力，跨多个地理位置复制数据，以增加可用性，并且减少距客户端的网络距离。

[0002] 垃圾收集((garbage collection)对于管理众多分布式对象的管理分权的存储系统会是一个问题。垃圾收集器负责通过删除不再被引用的对象来回收盘空间。机器和网络分区的常见故障使存储集群中的分布式垃圾收集更加复杂，这使得得到对象和其引用的全局同步视图即使并非不可能，也变得困难。

发明内容

[0003] 根据一个实施方式，一种方法可以由分布式数据复制系统中的设备组中的设备执行。该方法可以包括：将对象存储在数据存储中，对象中的至少一个或多个利用分布式数据复制系统来复制；进行对数据存储中的对象的扫描；将对象中的一个识别为没有指向对象的引用；将删除协商消息存储为与对象中的一个相关联的元数据；以及将带有删除协商消息的元数据复制到设备组中的一个或多个其它设备。

[0004] 根据另一个实施方式，分布式数据复制系统中的设备组中的设备可以包括：用于将数据存储中的对象识别为具有与对象相关联的协商消息的装置；用于将新的协商消息与对象进行关联的装置，新的协商消息基于对象的状态；用于将新的协商消息复制到存储集群的装置；用于接收与对象的复制品相关联的其它协商消息的装置；以及用于如果其它协商消息指示成功的协商，则删除对象的装置。

[0005] 根据又一个实施方式，一种系统可以包括用来存储指令和数据存储的存储器以及处理器。处理器可以执行存储器中的指令来：识别数据存储中的对象的状态，所述状态与对象是否具有引用以及删除协商消息是否与对象相关联有关；基于对象的状态来将新的协商消息写入对象元数据；将带有新的协商消息的元数据复制到一个或多个设备；以及从一个或多个设备中的一个接收与对象相关联的其它协商消息，其中新的协商消息和其它协商消息为对象的删除协商提供共识(consensus)。

[0006] 根据又一个实施方式，一种方法可以包括：在分布式多主站(multi-master)数据复制系统内的存储集群之间交换一个或多个删除协商消息，其中删除协商消息中的每一个被包括在为删除协商消息的主题的对象的元数据中，以及其中删除协商消息使用分布式多主站数据复制系统的复制层来在存储集群之间发送；以及如果基于一个或多个删除协商消息在存储集群之间达成分布式共识，则删除对象。

[0007] 根据进一步实施方式，一种计算机可读存储器可以包括计算机可执行的指令。计算机可读存储器可以包括：用来识别数据存储中的对象的状态的一个或多个指令，所述状态与对象是否具有引用以及删除协商消息是否与对象相关联有关；用来基于对象的状态来

将与对象相关联的新的协商消息写入对象的元数据的一个或多个指令；用来将带有新的协商消息的对象元数据复制到存储集群的一个或多个指令；用来从一个或多个其它设备接收与对象相关联的其它协商消息的一个或多个指令；以及用来基于与对象相关联的其它协商消息为对象的删除协商确定共识的一个或多个指令。

附图说明

[0008] 合并入并且构成本说明书的一部分的附图图示了在此描述的一个或多个实施例，并且与描述一起说明了这些实施例。在附图中：

[0009] 图 1 是在其中可以实现在此描述的系统和方法的示例性网络的图；

[0010] 图 2 是图 1 的文件系统的示例性配置的图；

[0011] 图 3 是图 1 的存储集群的示例性组件的图；

[0012] 图 4 是图 1 的示例性存储集群的功能框图；

[0013] 图 5 是根据与在此描述的系统和方法一致的一个实施方式的可以使用的消息结构的示例性图；

[0014] 图 6 是根据与在此描述的系统和方法一致的一个实施方式的用于在分布式多主站数据复制系统中执行垃圾收集的示例性过程的流程图；

[0015] 图 7 是根据与在此描述的系统和方法一致的一个实施方式的用于写协商消息的示例性过程的流程图；

[0016] 图 8 是根据与在此描述的系统和方法一致的一个实施方式的用于创建对对象的新的引用的示例性过程的流程图；以及

[0017] 图 9 是根据一个实施方式图示了示例性删除协商的一部分的图。

具体实施方式

[0018] 下面的详细描述参考附图。在不同附图中的相同参考数字可以识别相同或相似元素。并且，下面的详细描述不限制本发明。

[0019] 在此描述的系统和 / 或方法可以执行用于复制的存储集群的异步分布式垃圾收集。在此描述的实施方式可以使用分布式多主站数据复制系统的基础复制层来在分布式多主站数据复制系统的不同集群之间运送删除协商消息。当达成活引用或复制的引用均不存在于系统中的分布式共识时可以删除对象。

[0020] 示例性网络配置

[0021] 图 1 是在其中可以实现在此描述的系统和方法的示例性系统 100 的图。系统 100 可以包括经由网络 130 连接的客户端 110-1 至 110-N（统称为客户端 110）以及存储集群 120-1 至 120-M（统称为存储集群 120）。存储集群 120 可以形成文件系统 140（如图 1 中虚线所示）。

[0022] 网络 130 可以包括一个或多个网络，诸如局域网 (LAN)、广域网 (WAN)、诸如公共交换电话网络 (PSTN) 的电话网络、内联网、因特网、相似或相异网络、或网络的组合。客户端 110 和存储集群 120 可以经由有线和 / 或无线连接连接到网络 130。

[0023] 客户端 110 可以包括一种或多种类型的设备，诸如个人计算机、无线电话、个人数字助理 (PDA)、膝上型计算机或另一种类型的通信设备、运行于这些设备中的一个上的线程

或进程、和 / 或由这些设备可执行的对象。在一个实施方式中，客户端 110 包括或被联接到应用，在所述应用的代表下客户端 110 与存储集群 120 通信以读取或修改（例如，写）文件数据。

[0024] 存储集群 120 可以包括一个或多个服务器设备、或其它类型的计算或通信设备，其可以以在此描述的方式存储、处理、搜索和 / 或提供信息。在一个实施方式中，存储集群 120 可以包括能够为文件维护大型、随机读 / 写访问数据存储的一个或多个服务器（例如，计算机系统和 / 或应用）。如果发生改变，则存储集群 120 的数据存储可以允许索引系统快速更新索引的部分。存储集群 120 的数据存储可以包括一个或多个表（例如，可以包括每统一资源定位符 (URL) 一行的文档表、由 URL 之外的值作为键的辅助表等）。在一个示例中，存储集群 120 可以被包括在用于管理结构化数据（例如，文档的随机存取存储集群）的分布式存储系统（例如，如在第七届 OSDI 论文集 (2006 年 11 月)、第 205-218 页、Chang 等人的“Bigtable :A Distributed Storage System for Structured Data(Bigtable :用于结构化数据的分布式存储系统)”中阐述的 " Bigtable " ）中，所述结构化数据可以被设计为缩放到非常大的大小（例如，跨数千服务器的千兆兆字节的数据）。

[0025] 尽管在图 1 中未示出，但是系统 100 可以包括多种其它组件，诸如一个或多个专用客户服务器或集线器。如在此所使用的，组件可以包括硬件或软件和硬件的组合。客户服务器例如可以存储来自一个或多个存储集群 120 的数据存储的只读副本以供客户端 110 访问。集线器例如可以存储来自一个或多个存储集群 120 的数据存储的只读副本以分发到一个或多个客户服务器。

[0026] 示例性存储集群配置

[0027] 图 2 是文件系统 140 的示例性配置的图。如图 2 中所示，文件系统 140 可以包括存储集群 120-1、120-2、120-3 以及 120-4。在一个实施方式中，文件系统 140 可以是分布式多主站数据复制系统，其中存储集群 120-1、120-2、120-3 以及 120-4 中的每一个对于其它存储集群可以充任主服务器。在文件系统 140 中，可以跨存储集群 120-1、120-2、120-3 以及 120-4（例如，在多个地理位置）复制数据以增加数据可用性以及减少距客户端（例如，客户端 110) 的网络距离。通常，可以在不同的存储集群 120 中动态地创建、变异、克隆以及删除分布式对象和引用，以及基础数据复制层（未示出）维护写顺序保真度以确保所有存储集群 120 以数据的相同版本结束。因此，数据复制层重视对单个对象的相同复制品的写顺序。

[0028] 尽管图 2 示出了文件系统 140 的示例性功能组件，但是在其它实施方式中，文件系统 140 可以包含比在图 2 中所描绘的更少、另外、不同或不同布置的组件。在另外其它实施方式中，文件系统 140 的一个或多个组件可以执行被描述为由文件系统 140 的一个或多个其它组件执行的一个或多个其它任务。

[0029] 图 3 是存储集群 120 的示例性组件的图。存储集群 120 可以包括总线 310、处理器 320、主存储器 330、ROM 340、存储设备 350、输入设备 360、输出设备 370 以及通信接口 380。总线 310 可以包括允许在存储集群 120 的组件之间的通信的一个或多个导线。

[0030] 处理器 320 可以包括可以解释并且执行指令的任何类型的处理器或微处理器。主存储器 330 可以包括可以存储用于处理器 320 执行的信息和指令的 RAM 或另一种类型的动态存储设备。ROM 340 可以包括可以存储用于处理器 320 使用的静态信息和指令的 ROM 设

备或另一种类型的静态存储设备。存储设备 350 可以包括磁和 / 或光记录介质以及其对应的驱动。例如，存储设备 350 可以包括提供持久存储的一个或多个本地盘 355。在一个实施方式中，存储集群 120 可以在诸如主存储器 330 和 / 或存储设备 350 的一个或多个计算机可读介质内为存储在文件系统 140 中的对象维护元数据。例如，存储集群 120 可以为存储设备 350 内的对象存储版本号、时间戳、类别和 / 或引用指示符。

[0031] 输入设备 360 可以包括允许操作者向存储集群 120 输入信息的一个或多个机制，诸如键盘、键区、按钮、鼠标、笔等。输出设备 370 可以包括向操作者输出信息的一个或多个机制，包括显示、发光二极管 (LED) 等。通信接口 380 可以包括使存储集群 120 能够与其它设备和 / 或系统通信的任何类收发器机制。例如，通信接口 380 可以包括用于与其它存储集群 120 和 / 或客户端 110 通信的机制。

[0032] 图 4 图示了存储集群 120 的功能框图。如图 4 中所示，存储集群 120 可以包括数据存储 410 和垃圾收集器逻辑 420。在一个实施方式中，如图 4 中所图示的，数据存储 410 可以在存储集群 120 内提供。在其它实施方式中，数据存储 410 可以在与存储集群 120 通信的系统 100 的一个或多个其它设备内提供，诸如外部存储器设备或与索引系统（未示出）相关联的设备。

[0033] 数据存储 410 可以包括为搜索系统提供一个或多个索引的文档表和次级表。在一个示例中，文档表和次级表可以以 URL 的特性为键，以帮助访问和 / 或更新与 URL 相关联的信息。每一个数据存储 410 中的至少一部分可以在多个存储集群 120 上被复制。每一个数据存储 410 的复制品的数量可以是用户可配置的。

[0034] 垃圾收集器逻辑 420 可以包括移除无引用的内容，诸如先前删除的文件，的逻辑。垃圾收集器逻辑 420 可以从例如数据存储 410 移除无引用的内容。例如，垃圾收集器逻辑 420 经由遍历存储集群 120 并且移除无引用的对象的功能（例如，MapReduce 功能）可以确定来自数据存储 410 的对象（例如，文档）是否不再被引用（即，不包括任何指向对象的链接的对象），以及可以从存储集群 120 移除任何不再被引用的对象（例如，文档）。如果对象具有指向该对象的链接，则对象可以是“被引用的”或“活的”。因此，垃圾收集器逻辑 420 可以在维护活对象时从存储集群 120 移除不必要的信息。

[0035] 移除对象不如只是删除对象般简单，因为该对象可以存在于其它存储集群 120 中。因此，垃圾收集器逻辑 420 可以聚集可以在文件系统 140 的不同存储集群 120 之间发送的删除协商消息。当（例如，在文件系统 140 的包含该对象的复制品的所有存储集群 120 之间）达成活引用或复制的引用均不存在于系统中的分布式共识时，垃圾收集器逻辑 410 可以删除对象。垃圾收集器逻辑 420 可以将删除协商消息包括在为删除协商的主题的对象的元数据中。消息然后可以被异步复制到包含对象的复制品的所有其它存储集群 120。

[0036] 垃圾收集器逻辑 420 所生成的消息例如可以包括用于发起删除协商的“Delete”指示符、用于为删除协商提供肯定应答的应答（" ACK" ）指示符、用于为删除协商提供否定应答的否定应答（" NACK" ）指示符、以及用于提供已从其它存储集群 120 接收应答的确认的同步（" GotAll" ）指示符。在一个实施方式中，没有新的引用可以被添加到具有未解决的 (outstanding)Delete 或 ACK 消息的对象。在下面更加详细地描述了消息格式和用途。

[0037] 尽管图 3 示出了存储集群 120 的示例性功能组件，但是在其它实施方式中，存储集

群 120 可以包含比在图 3 中所描绘的更少、另外、不同或不同布置的功能组件。在另外其它实施方式中，存储集群 120 的一个或多个功能组件可以执行被描述为由一个或多个其它功能组件执行的一个或多个其它任务。

[0038] 示例性消息结构

[0039] 图 5 提供了在一个示例性实施方式中可以使用的协商消息的示例性消息结构 500 的图示。如图 5 中所示，消息结构 500 可以包括消息部分 510、存储集群标识部分 520 以及协商请求标识部分 530。消息部分 510 例如可以包括 " Delete" 指示符、" ACK" 指示符、" NACK" 指示符或 " GotAll" 指示符。存储集群标识部分 520 可以包括发起消息部分 510 中的消息的存储集群 120 的唯一标识（例如，集群 ID(Cluster ID)）。协商请求标识部分 530 可以包括原始删除协商的唯一标识（例如，ReqID)。

[0040] 消息结构 500 可以以 Message:Cluster ID:ReqID 的形式列出。例如，对象的删除协商可以由存储集群 120-1 以消息 " Delete:01:5555" 发起，其中 " 01" 是存储集群 120-1 的集群 ID 以及 " 5555" 是 ReqID。存储集群 120-2 对协商的应答可以是 " ACK:02:5555"，其中 " 02" 是存储集群 120-2 的集群 ID 以及对于该应答（以及与原始协商有关的任何未来消息）" 5555" 仍然是 ReqID。

[0041] 示例性过程流程

[0042] 图 6 是用于在分布式多主站数据复制系统（例如，文件系统 140) 中执行垃圾收集的示例性过程 600 的流程图。在一个实施方式中，过程 600 可以由存储集群 120 中的一个来执行。在另一个实施方式中，过程 600 的部分或全部可以由另一个设备或设备组—包括或排除存储集群 120- 来执行。过程 600 可以在每一个存储集群 120 中周期性实现，以及可以包括对存储集群 120 中的对象的全部或一部分的扫描。对于在下面描述的过程 600 的特定示例，可以参考文件系统 140 的存储集群 120-1，其中存储集群 120-1 包括集群 ID“01”。

[0043] 如图 6 中所图示，过程 600 可以以进行对对象的扫描（块 610) 以及识别无引用和协商的对象（块 620) 开始。例如，（使用例如垃圾收集器逻辑 420 的）存储集群 120-1 可以进行对存储在存储集群 120-1 中（例如，数据存储 410 中）的对象的全部或一部分的扫描。扫描可以通过读取与对象相关联的元数据来识别例如没有引用的对象以及带有删除协商消息的对象。

[0044] 可以确定是否为对象识别完成的删除协商（块 630)。完成的删除协商例如可以是对成功或失败的删除协商的指示。例如，存储集群 120-1 可以识别带有确认成功或失败的删除协商的元数据的对象。

[0045] 如果为对象识别完成的删除协商（块 630- 是），则发起的存储集群可以删除成功的删除协商的对象或失败的删除协商消息（块 640)。在一个示例性实施方式中，如果存储集群 120-1 识别对象中的指示下述的元数据 :存储集群 120-1 先前为对象发起了删除协商以及存储对象的复制品的所有其它存储集群（例如，通过将 ACK 消息和 / 或 GotAll 消息写入对象元数据）识别出对象可以被删除，则存储集群 120-1 可以删除对象和相关联的元数据。例如，如果在存储集群 120-1 中的扫描遇到带有 " Delete:01:ReqID" 和来自存储对象的复制品的所有其它存储集群（例如，存储集群 120-2、120-3 以及 120-4) 的 " GotAll:*:ReqID"（其中“*”指示每一个存储集群 120 的存储集群 ID) 的对象，则可以删除对象和元数据。因此，存储集群 120-1 可以是成功协商的发起者。

[0046] 仍然参考块 640，在另一个示例性实施方式中，如果存储集群识别对象中的指示下述的元数据：存储集群 120-1 先前为对象发起了删除协商，以及至少一个其它存储集群 120 通过写入 NACK 消息来指示对象不应当被删除，则存储集群 120-1 可以删除包含原始协商消息和来自其它存储集群 120 的所有对应的消息的元数据。例如，如果在存储集群 120-1 中的扫描遇到带有″ Delete:01:ReqID″、来自所有其它存储集群 120 的″ ACK:*:ReqID″以及″ NACK:*:ReqID″（其中“*”指示存储集群 ID）的对象，并且存在至少一个 NACK 消息，则可以从与对象相关联的元数据删除对应于 ReqID 的所有 Delete、ACK 和 NACK 消息。因此，存储集群 120-1 可以是失败协商的发起者。

[0047] 如果没有为对象识别完成的删除协商（块 630- 否），则可以基于对象状态将协商消息写入对象元数据（块 650）。如在此进一步描述的，基于对象的状态，消息（例如，″ Delete″、″ ACK″、″ NACK″、″ GotAll″）可以被写入一个集群中的对象元数据，以及被复制到包含对象复制品的所有其它集群。例如，取决于对象状态，存储集群 120-1 可以写删除对象的新的协商消息。替选地，存储集群 120-1 可以响应于正在进行的协商写 ACK 消息、NACK 消息或 GotAll 消息。参考图 7 进一步描述了对协商消息的使用。

[0048] 可以将对象元数据复制到其它存储集群（块 660）。例如，存储集群 120-1 可以使用分布式多主站数据复制系统 140 的基础复制层来将协商消息复制到存储集群 120-2、存储集群 120-3、存储集群 120-4 等。因此，协商消息可以与对象元数据复制品一起被分发到其它集群，而不是作为单独消息。

[0049] 可以重复过程 600，直到存储集群（例如，存储集群 120-1）中的所有对象均被扫描，以及可以周期性重复过程 600。过程 600 可以由分布式多主站数据复制系统（例如，文件系统 140）中的其它存储集群（例如，存储集群 120-2、120-3、…、120-M）中的每一个类似地执行。因此，响应于来自存储集群 120-1 的协商消息，从其它存储集群复制的对象元数据可以包含协商消息。存储集群中的每一个可以继续在文件系统的复制层中交换协商消息，以异步执行对由其它存储集群标记用于删除的对象的协商。

[0050] 图 7 是用于写在图 6 中引用的协商消息的示例性过程 650 的流程图。过程 650 可以由分布式多主站数据复制系统（例如，文件系统 140）中的存储集群（例如，存储集群 120 中的一个）来执行。对于过程 650 的特定示例，可以参考分布式多主站数据复制系统的存储集群 120-1（具有集群 ID“01”）和存储集群 120-2（具有集群 ID“02”）。

[0051] 可以确定是否存在正在进行的协商（块 710）。例如，（使用例如垃圾收集器逻辑 420 的）存储集群 120-1 可以确定对象的元数据是否包括删除协商消息。在一个实施方式中，对对象的删除协商可能先前已由存储集群 120-1 发起或可能已例如由另一个存储集群（例如，存储集群 120-2、120-3 或 120-4 中的一个）发起。

[0052] 如果确定没有正在进行的协商存在（块 710- 否），则可以确定任何引用是否指向对象（块 715）。例如，（使用例如垃圾收集器逻辑 420 的）存储集群 120-1 可以（例如，通过分析引用的有向图）确定特定对象是否具有任何引用。如果确定没有引用指向对象（块 715- 否），则可以写新的“delete”消息（块 720）。例如，如果在存储集群 120-1 中的扫描遇到没有引用的对象，并且没有正在进行的协商（例如，没有″ Delete″消息），则存储集群 120-1 可以生成唯一 ReqID，并且为对象写新的删除协商消息（例如，″ Delete:01:ReqID″）。如果确定引用指向对象（块 715- 是），则不需要消息（块

790)。例如,如果在存储集群 120-1 中的扫描遇到带有引用的对象,并且没有正在进行的删除协商,则在此时对象可以不需要另外的处理。

[0053] 如果确定存在正在进行的协商(块 710- 是),则可以确定任何引用是否指向对象(块 730)。例如,(使用例如垃圾收集器逻辑 420 的)存储集群 120-1 可以确定特定对象是否具有任何引用。如果确定引用指向对象(块 730- 是),则可以确定先前的否定应答是否已被存储在对象的元数据中(块 735)。例如,(使用例如垃圾收集器逻辑 420 的)存储集群 120-1 可以确定来自存储集群 120-1 的 NACK 消息(例如," NACK:01:ReqID")是否已被包括在对象的元数据中。

[0054] 如果确定先前的否定应答尚未被存储在对象的元数据中(块 735- 否),则可以写否定应答(" NACK")消息(块 740)。例如,如果在存储集群 120-1 中的扫描遇到带有引用的对象和来自另一个存储集群(例如,存储集群 120-2)的正在进行的协商(例如," Delete:02:ReqID"),则存储集群 120-1 可以将否定应答(例如," NACK:01:ReqID")写入对象的元数据。如果确定先前的否定应答已被存储在对象的元数据中(块 735- 是),则在此时不需要对对象的进一步处理(块 790)。

[0055] 如果确定没有引用指向对象(块 730- 否),则可以确定是否已接收所有的 ACK(块 750)。例如,(使用例如垃圾收集器逻辑 420 的)存储集群 120-1 可以确定来自系统 140 中的每一个存储集群 120 的应答(例如," ACK:*:ReqID",其中“*”指示存储集群 ID)是否已被包括在对象的元数据中。如果确定已接收所有的 ACK(块 750- 是),则可以写" GotAll" 消息(块 760)。例如,如果在存储集群 120-1 中的扫描遇到带有删除消息(例如," Delete:02:ReqID")的对象和来自系统 140 中的每一个存储集群 120 的应答(例如," ACK:*:ReqID",其中“*”指示存储集群 ID),则存储集群 120-1 可以写应答确认消息(例如," GotAll:01:ReqID)以供发起者存储集群 120-2 使用。如果确定尚未接收所有的 ACK(块 750- 否),则可以确定先前的应答是否已被存储在对象的元数据中(块 770)。例如,(使用例如垃圾收集器逻辑 420 的)存储集群 120-1 可以确定来自存储集群 120-1 的 ACK 消息(例如," ACK:01:ReqID")是否已被包括在对象的元数据中。

[0056] 如果确定先前的应答尚未被存储在对象的元数据中(块 770- 否),则可以写新的应答(" ACK")消息(块 780)。例如,如果在存储集群 120-1 中的扫描遇到没有引用的对象和来自另一个复制品(例如,存储集群 120-2)的正在进行的协商(例如," Delete:02:ReqID"),则存储集群 120-1 可以将应答(例如," ACK:01:ReqID")写入对象的元数据。如果确定先前的应答已被存储在对象的元数据中(块 770- 是),则在此时不需要对对象的进一步处理(块 790)。

[0057] 图 8 提供了根据与在此描述的系统和方法一致的一个实施方式的用于创建对对象的新的引用的示例性过程 800 的流程图。过程 800 可以由分布式多主站数据复制系统(例如,文件系统 140)中的存储集群(例如,存储集群 120 中的一个)来执行。对于过程 800 的特定示例,可以参考文件系统 140 的存储集群 120-1(具有集群 ID“01”)。

[0058] 可以接收对对象的引用指示(块 810)。例如,存储集群 120-1 可以接收向对象添加新的引用的请求。可以为在存储集群处发起的协商消息查看对象元数据(块 820)。例如,存储集群 120-1 可以查看对象的元数据以识别任何删除协商消息,尤其先前由存储集群 120-1 发起的任何 Delete 或 ACK 协商消息(例如," Delete:01:ReqID"

或" ACK:01:ReqID")。在此描述的实施方式中,存储集群 120-1 可以不写对对象的新的引用,所述对象在对象元数据中具有带有由存储集群 120-1 发起的 Delete 或 ACK 消息的正在进行的协商。

[0059] 可以确定是否存在 Delete 或 ACK 消息 (块 830)。如果存在 Delete 或 ACK 消息 (块 830- 是),则在另一个存储集群中的复制品可以用作失效备援 (块 840)。例如,如果存储集群 120-1 在对象元数据中识别" Delete:01:ReqID" 消息,则消息将有效地锁定存储集群 120-1 不能写对对象的新的引用。因此,将对对象的引用写入存储集群 120-1 的请求将被转送到另一个存储集群 (例如,存储集群 120-2)。

[0060] 如果没有 Delete 或 ACK 消息存在 (块 830- 否),则可以写对对象的新的引用 (块 850)。例如,存储集群 120-1 可以仅仅写所请求的对活对象的引用。

[0061] 示例

[0062] 图 9 提供了根据在此描述的实施方式实现了示例性删除协商的一部分的示例性网络。垃圾收集算法可以在存储集群 XX、YY 和 ZZ 的每一个中周期性地运行,以及可以扫描存储集群中的所有对象。消息 (例如,Delete、ACK、NACK、GotAll) 可以由垃圾收集器写入一个集群 (例如,存储集群 YY) 中的对象的元数据,以及被复制到包含对象复制品的所有其它集群 (例如,存储集群 XX 和 ZZ)。

[0063] 垃圾收集器所使用的垃圾收集算法可以使用与在此描述的原理一致的导则来操作。如果垃圾收集器的在存储集群 YY 中的扫描遇到没有引用的对象,并且不存在正在进行的协商 (例如,没有 Delete:YY:ReqID 消息),则存储集群 YY 中的垃圾收集器可以生成唯一 ReqID(例如,22222),并且将" Delete:YY:22222" 写入对象的元数据。如果垃圾收集器的在存储集群 XX 中的扫描首次遇到来自另一个复制品 (例如,来自存储集群 YY) 的删除协商 (例如, Delete:YY:22222),则在对象没有引用的情况下垃圾收集器写" ACK:XX:22222",否则写" NACK:XX:22222" 。存储集群 XX 不能向具有未解决的 Delete:XX:ReqID 或 ACK:XX:ReqID 消息的对象添加新的引用。如果垃圾收集器的在存储集群 XX 中的扫描首次遇到 Delete:YY:22222 和来自所有其它复制品的 ACK:*:22222,则垃圾收集器可以写 GotAll:XX:22222。在这种情况下,存储集群 XX 不是发起者。如果垃圾收集器的在存储集群 YY 中的扫描遇到 Delete:YY:22222 和来自所有其它复制品的 GotAll:*:22222,则删除对象和元数据。(存储集群 YY 是成功协商的发起者)。如果垃圾收集器的在存储集群 YY 中的扫描遇到 Delete:YY:22222、来自其它复制品的 ACK:XX:22222 和 NACK:ZZ:22222,以及由于存在至少一个 NACK 消息,则从对象的元数据删除与 ReqID 22222 相对应的所有 Delete、ACK 和 NACK 消息。在这种情况下,存储集群 YY 是失败协商的发起者。

[0064] 在图 9 的文件系统中,存储集群 XX、YY 和 ZZ 每一个可以被指定为存储对象元数据的复制品。图 9 示出了对象“对象 1”的元数据的复制品 (“元数据 1A”)。元数据 1A 包括存储集群 YY 发起的正发送给存储集群 XX 的删除协商。作为响应,存储集群 XX 可以将回复消息添加到对象元数据,并且将元数据复制品 (“元数据 1B”) 发送给存储集群 YY。在图 9 的文件系统中,元数据 1A 和元数据 1B 还将被复制到存储集群 ZZ(未示出)。在达成删除对象 1 或使对象 1 未受影响并且删除与由存储集群 YY 发起的协商有关的消息的分布式共识之前,在存储集群 XX、YY 和 ZZ 之间发送的随后元数据复制品 (未示出) 可以将另外协商消息包括在对象 1 元数据中。

[0065] 在此描述的系统和/或方法的应用可以提供协议可用性保证，使得带有活复制品的对象不能被删除，并且能够一直可用。例如，仍然参考图 9，如果在存储集群 XX 中存在对象 1 的活复制品，则存储集群 XX 不会肯定应答对象删除协商，因此对象 1 不会被删除。并且，当存储集群 YY 发起的删除协商正在进行时，在存储集群 YY 中的克隆请求（例如，创建新的对象引用的请求）通过到存储集群 XX 中的对象 1 的活复制品的失效备援（例如，自动转换）将成功。

[0066] 在此描述的系统和/或方法的应用可以提供协议活跃度保证。例如，对于任何发起的删除协商请求 Delete:XX:ReqID，垃圾收集器的在存储集群 YY 中的扫描过程将最终写 ACK:YY:ReqID 或 NACK:YY:ReqID，以及当所有这些 ACK 和/或 NACK 已被复制时，存储集群 XX 中的协商过程将以是/否判定结束。然后，如果判定是肯定的，则 GotAll:*:ReqID 可以由所有存储集群写入对象的元数据，其将最终触发将通过基础复制层传播到其它存储集群 YY 和 ZZ 的经由存储集群 XX 的实际删除。在例如存储集群 ZZ 中存在活复制品的情况下，判定是否定的，并且发起者（例如，存储集群 XX）可以通过删除协商消息来清除对象元数据。所清除的元数据最终将传播到写 ACK 的所有存储集群，以及对象在那里将变得可用。

[0067] 在此描述的系统和/或方法的应用还可以提供在对象被删除后，以后没有虚引用能够再现的保证。例如，假设在 XX 中首先删除对象 1。基于垃圾收集算法，情况必须是：在删除发生之前，来自其它存储集群的 GotAll:*:ReqID 已被复制到存储集群 XX。通过该推演，指定到存储集群 YY 的所有复制数据不受复制自又另一个存储集群 ZZ 的虚引用的污染。这应归于事实：当来自其它存储集群，特别地存储集群 ZZ 的所有 ACK 已被接收时，存储集群 YY 写 GotAll:YY:ReqID，而在 ACK:ZZ:ReqID 被写之后没有新的引用能够被添加在存储集群 ZZ 中，并且在此时在存储集群 ZZ 中没有活引用。注意到，在 ACK:YY:ReqID 被写之后并且在 ACK:ZZ:ReqID 被复制之前，仍然可以存在从存储集群 ZZ 复制到存储集群 YY 的引用，但是所有这样的引用在 ACK:ZZ:ReqID 被复制到存储集群 YY 时可以被删除，因为复制层重视对单个复制品的写顺序。

[0068] 在此描述的系统和/或方法的应用可以进一步提供非协议垃圾保证。例如，如果删除协商失败，则发起者将删除 Delete、ACK 和 NACK 协商消息，并且该删除将通过复制传播到其它对象复制品。由于算法被配置为使得所有的相关消息必须由发起者在经由发起者的删除发生之前接收，所以仍然没有垃圾。

[0069] 结论

[0070] 在此描述的系统和/或方法可以为复制的存储集群提供异步分布式垃圾收集算法，其提供可用性、活跃度和一致性保证。该算法使用基础复制层来在不同集群之间运送消息。每一个删除协商由集群中的一个中的垃圾收集器逻辑发起，并且具有唯一标识符。该算法支持多个并发的协商。当达成分布式共识时，发起者可以删除对象；否则，可以使协商无效。

[0071] 对实施方式的前面的描述提供了说明和描述，但是并不意在穷举或将本发明限制在所公开的精确形式。修改和变化根据上述教导是可能的或可以从本发明的实践获得。

[0072] 例如，在另一个实施方式中，可以使用垃圾收集算法的同步版本，其中在不同的存储集群中的垃圾收集器直接而不是使用复制层来通信。

[0073] 并且，虽然关于图 6 和 7 来描述了块系列，但是在其它实施方式中，可以修改块的

顺序。此外，可以并行执行非依赖性的块。

[0074] 将显而易见的是，在此描述的实施例可以以在附图中所图示的实施方式中的软件、固件以及硬件的许多不同形式实现。用于实现在此描述的实施例的实际软件代码或专用控制硬件并不是对本发明的限制。因此，没有引用特定软件代码描述了实施例的操作和行为—应该理解的是，软件和控制硬件可以基于在此的描述被设计来实现实施例。

[0075] 此外，在此描述的某些实施方式可以被实现为执行一个或多个功能的“逻辑”。该逻辑可以包括：硬件，诸如处理器、微处理器、专用集成电路或现场可编程门阵列；或硬件和软件（例如，由处理器执行的软件）的组合。

[0076] 应当强调的是，词语“包括”在本说明书中使用时被采用来明确说明所述特征、完整物、步骤或组件的存在，但是并不排除一个或多个其它特征、完整物、步骤、组件或其组群的存在或添加。

[0077] 尽管在权利要求书中记载和/或在说明书中公开了特征的特定组合，但是这些组合并不意在限制本发明的公开。实际上，可以以未在权利要求书中明确记载和/或未在说明书中明确公开的方式对这些特征中的许多特征进行组合。

[0078] 除非明确描述如此，在本申请的描述中使用的元素、动作或指令均不应当被解释为对本发明是关键的或至关重要的。同时，如在此所使用的，不加数量词的项意指包括一个或多个项。在意指仅仅一个项时，使用词语“一个”或类似语言。此外，除非另外明确说明，如在此所使用的短语“基于”意在表示“至少部分基于”。

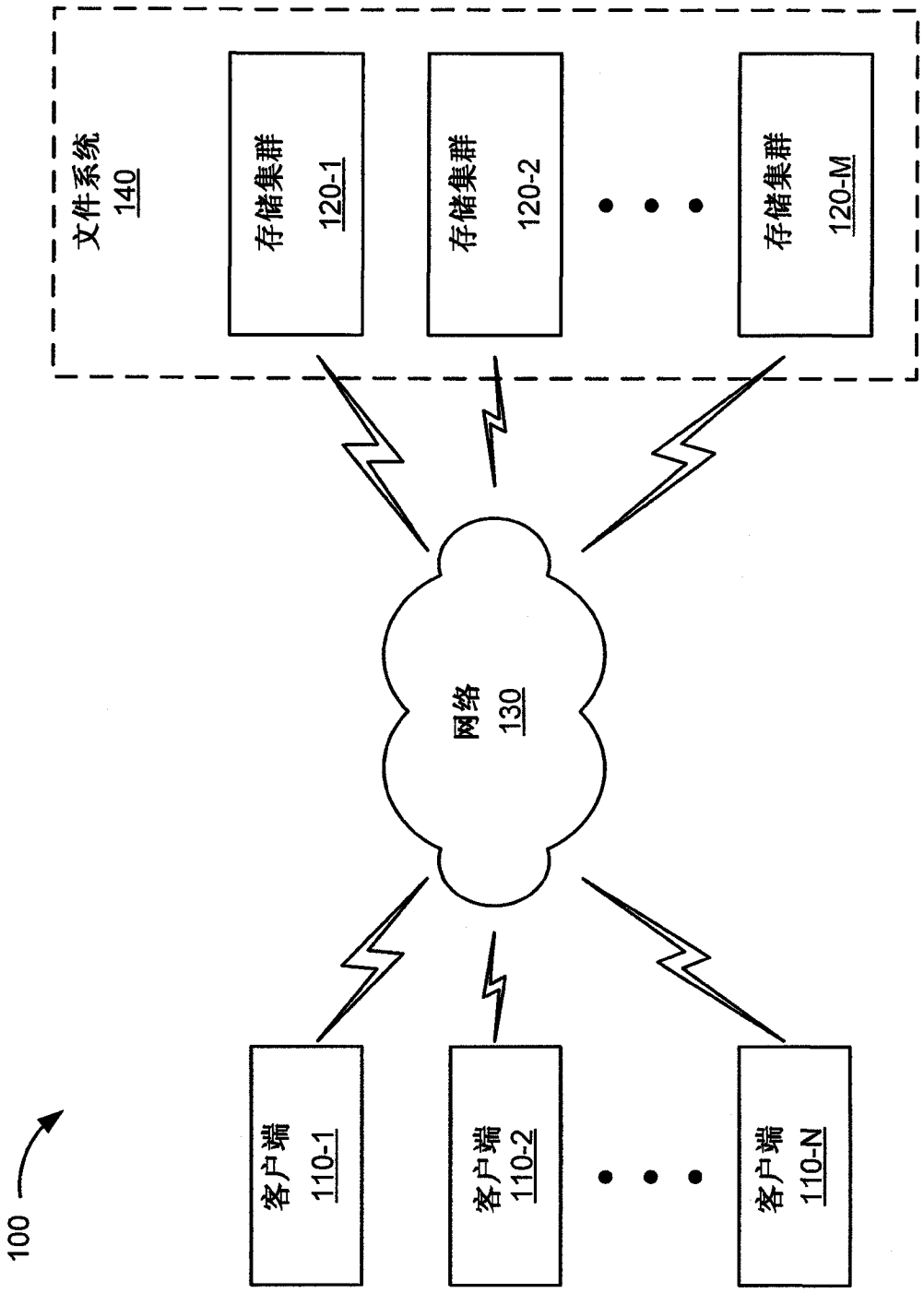

文件系统
140
存储集群
120-1
存储集群
120-2
存储集群
120-M
网络
130
客户端
110-1
客户端
110-2
客户端
110-N
100

图 1

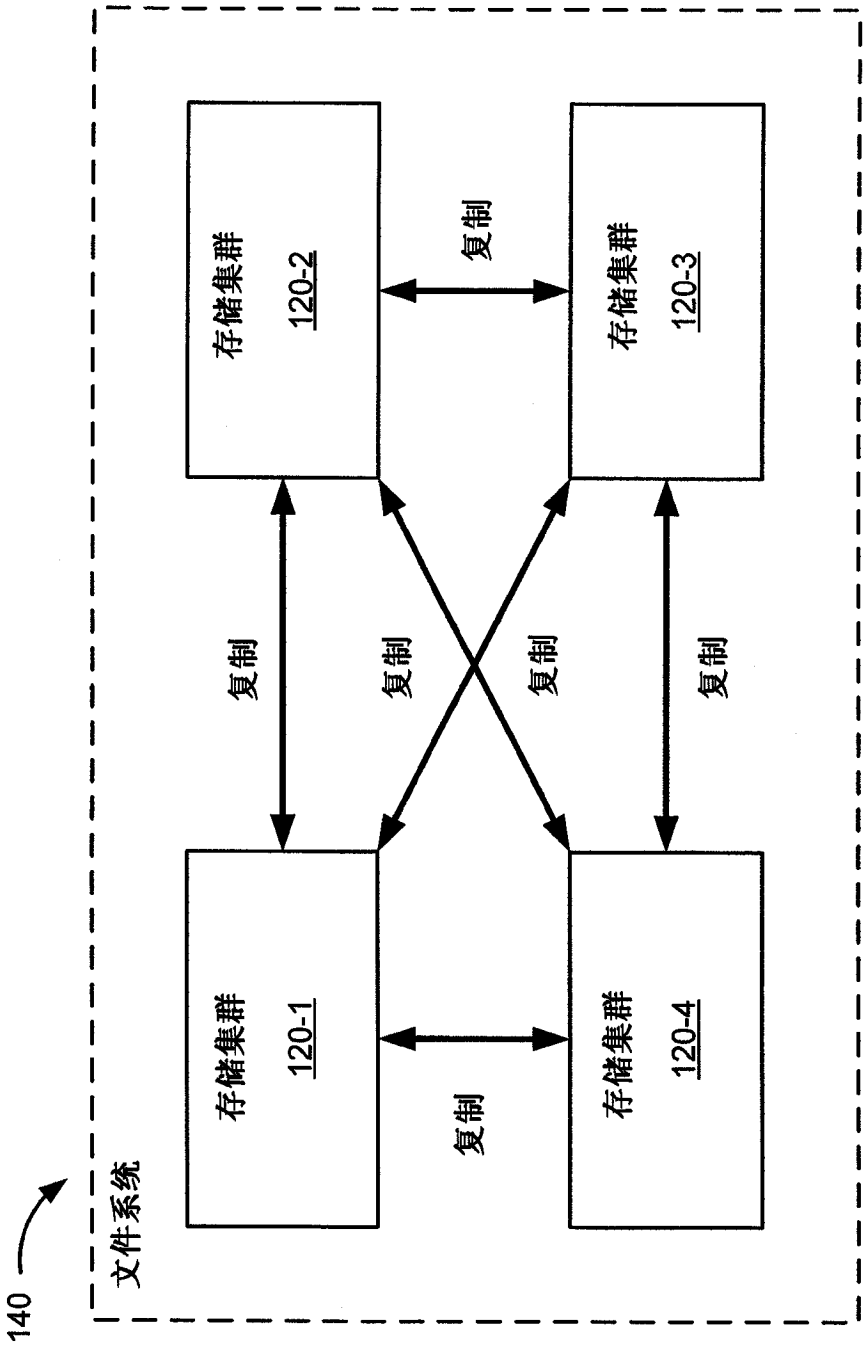

140
文件系统
存储集群 120-1
存储集群 120-2
存储集群 120-3
存储集群 120-4
复制
复制
复制
复制
复制
复制

图 2

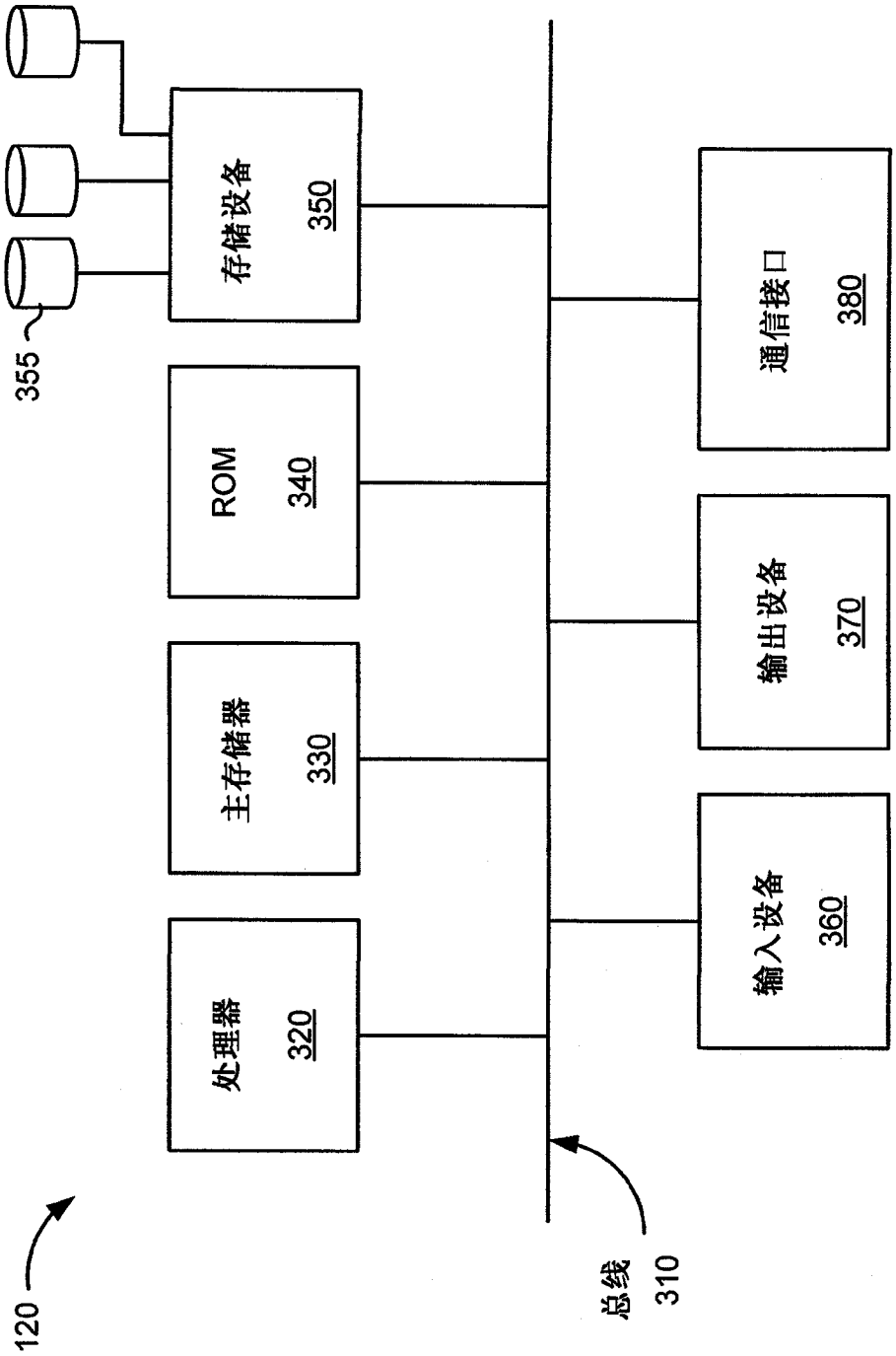

存储设备
350
355
ROM
340
主存储器
330
处理器
320
120
总线
310
通信接口
380
输出设备
370
输入设备
360

图 3

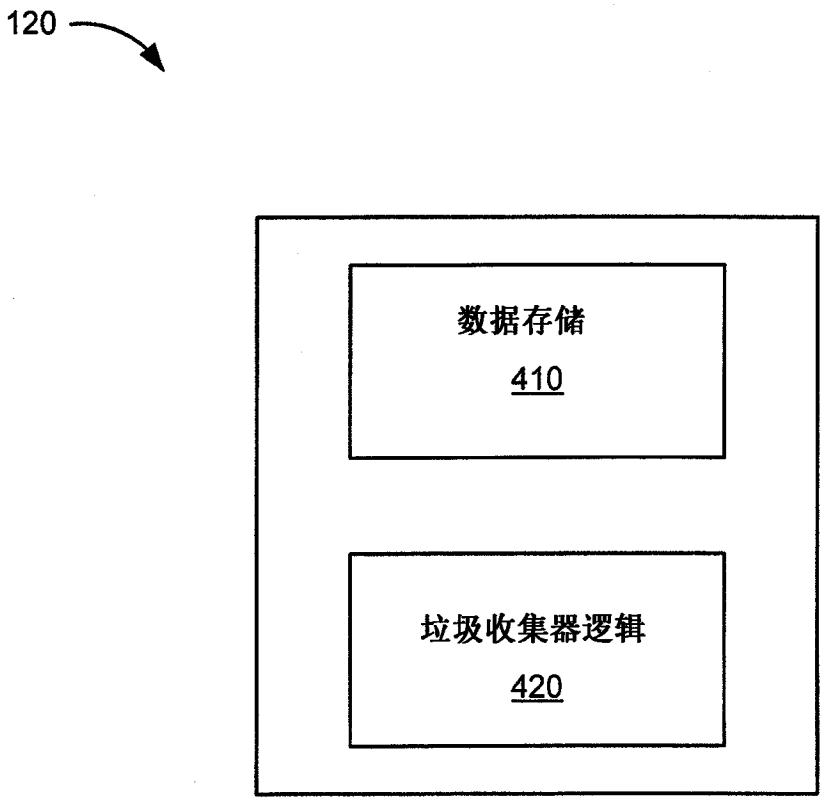

120
数据存储
410
垃圾收集器逻辑
420

图 4

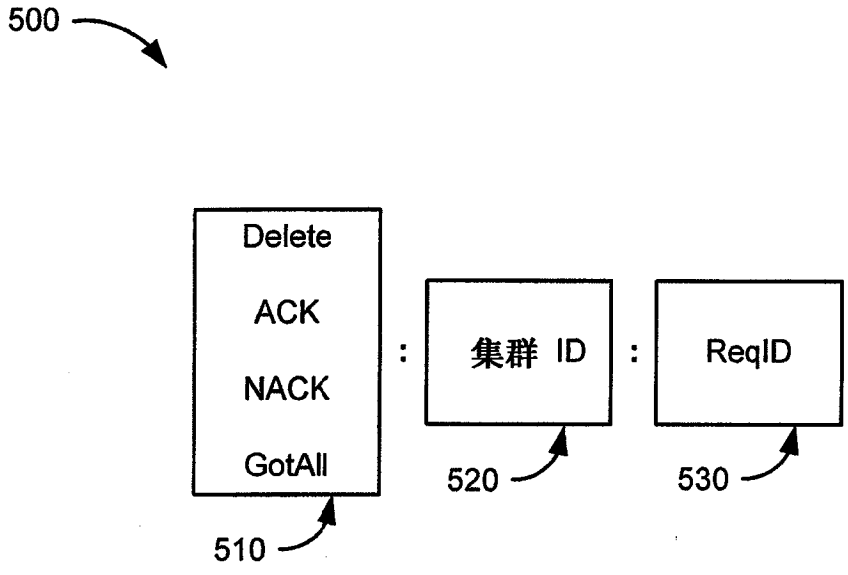

500
Delete
ACK
NACK
GotAll
:
集群 ID
:
ReqID
510
520
530

图 5

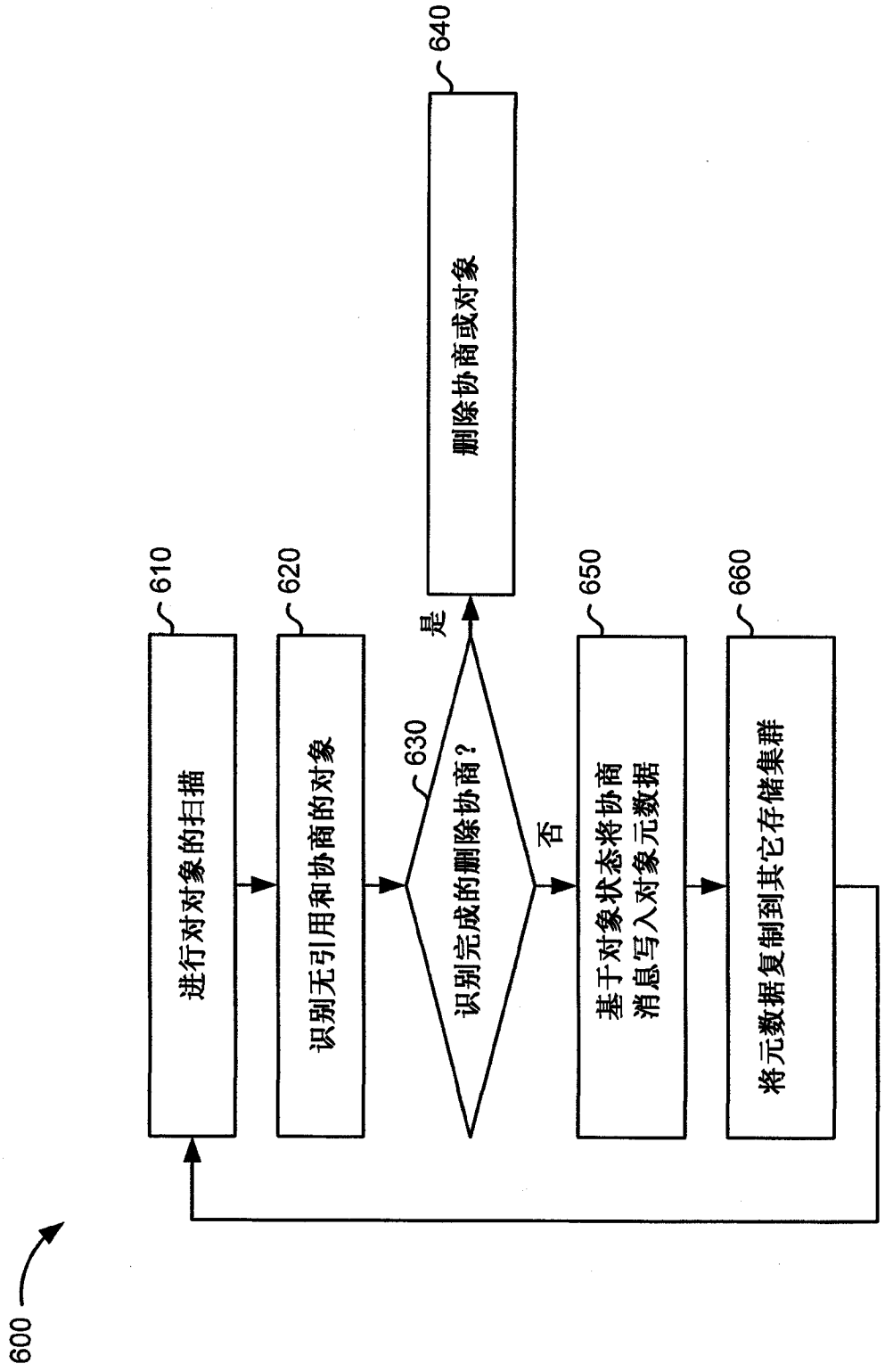

600
610
进行对对象的扫描
620
识别无引用和协商的对象
630
识别完成的删除协商？
是
否
640
删除协商或对象
650
基于对象状态将协商
消息写入对象元数据
660
将元数据复制到其它存储集群

图 6

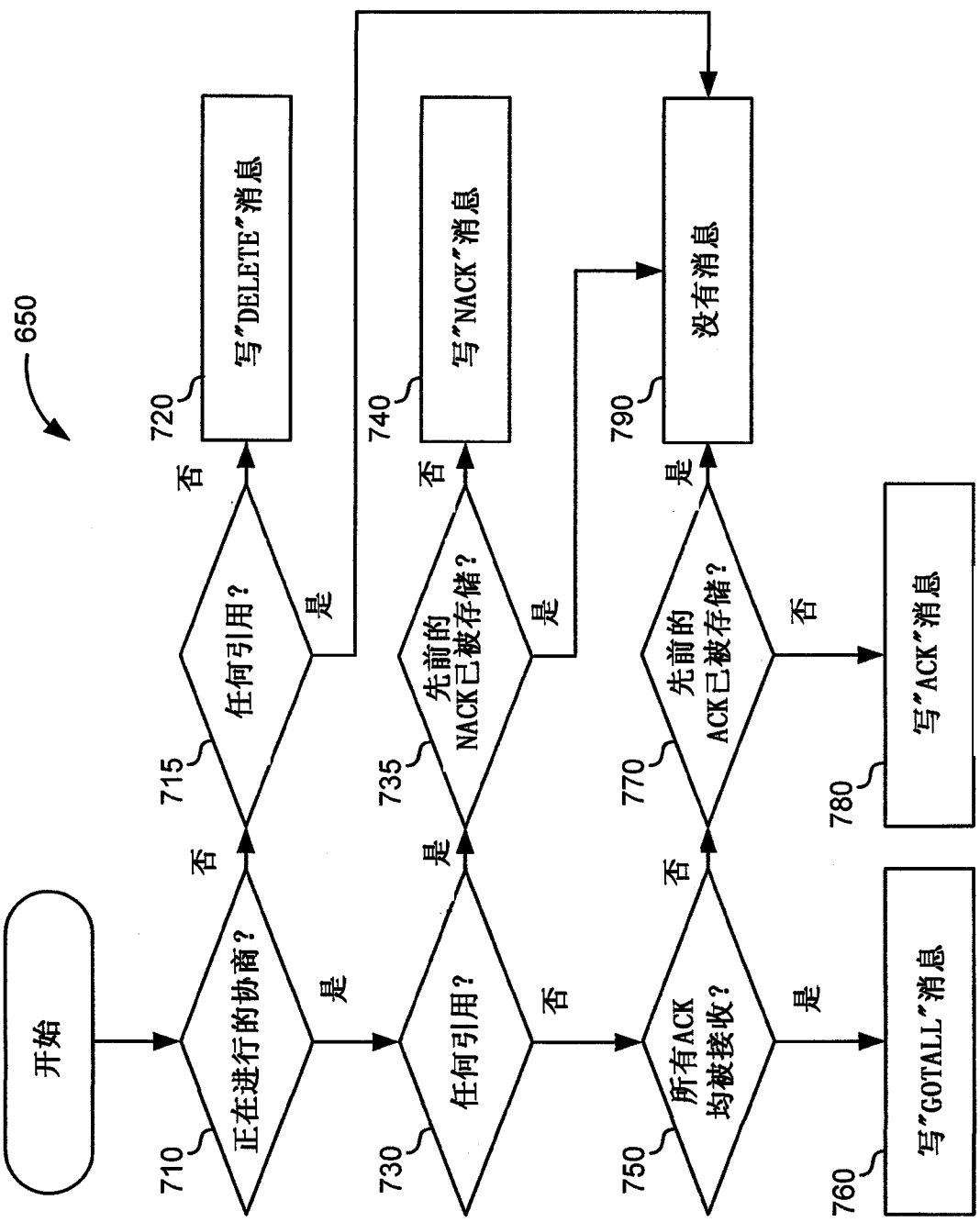

650
开始
正在进行的协商？
710
是
否
任何引用？
715
是
否
写"DELETE"消息
720
任何引用？
730
是
否
先前的NACK已被存储？
735
是
否
写"NACK"消息
740
所有ACK均被接收？
750
是
否
写"GOTALL"消息
760
先前的ACK已被存储？
770
是
否
写"ACK"消息
780
没有消息
790

图 7

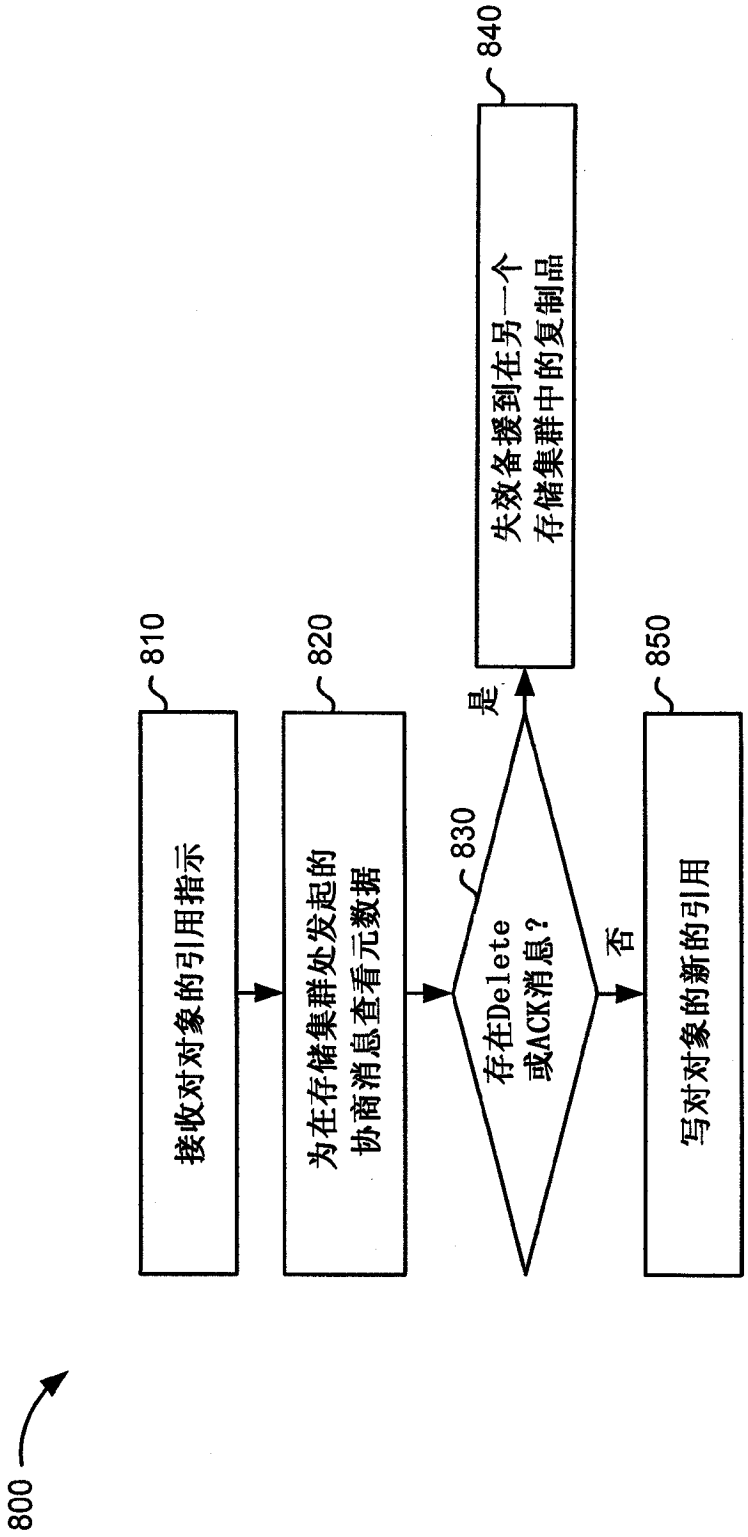

800
810
接收对对象的引用指示
820
为在存储集群处发起的
协商消息查看元数据
830
存在Delete
或ACK消息？
是
840
失效备援到在另一个
存储集群中的复制品
否
850
写对对象的新的引用

图 8

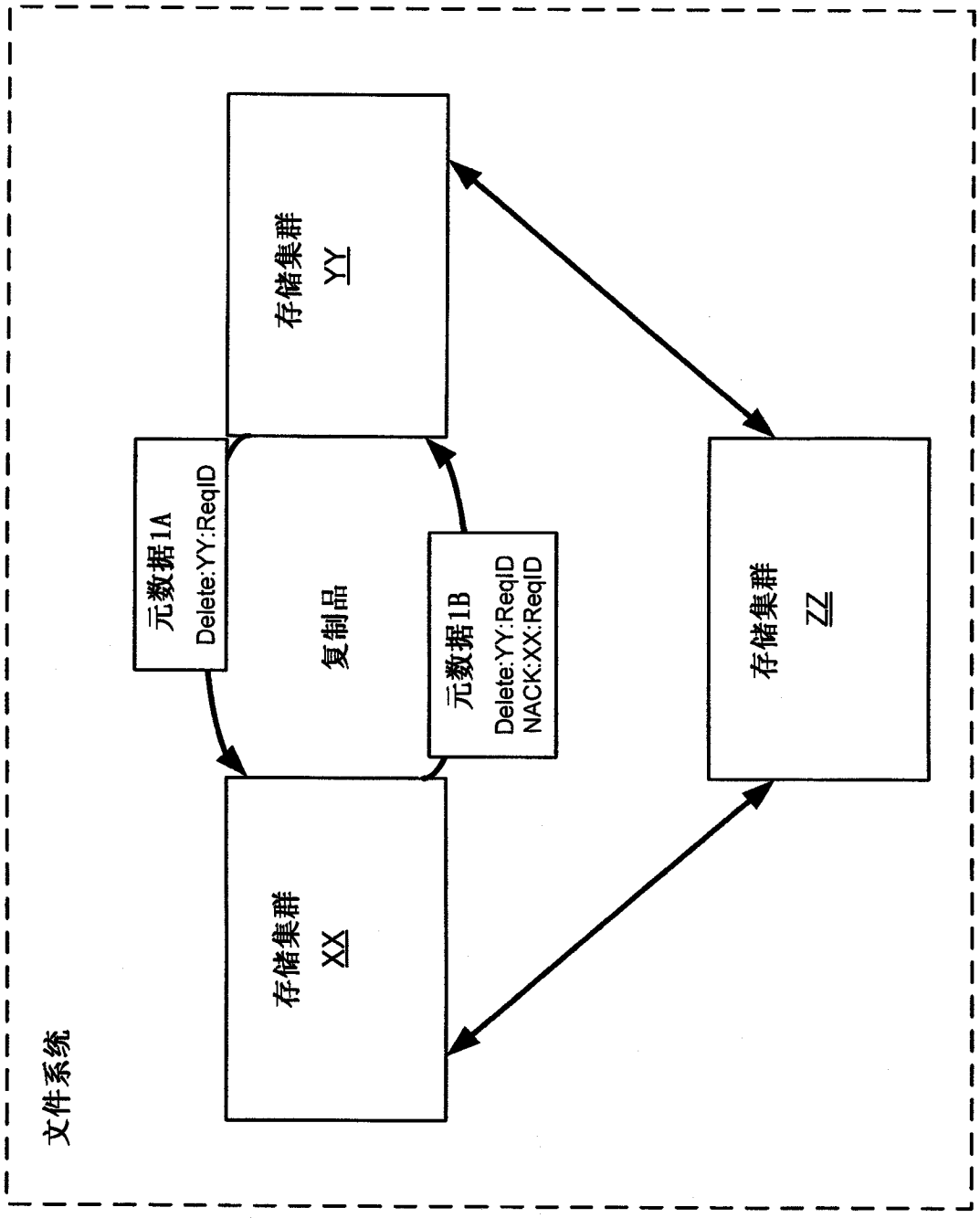

文件系统
存储集群
YY
存储集群
XX
存储集群
ZZ
元数据1A
Delete:YY:ReqID
复制品
元数据1B
Delete:YY:ReqID
NACK:XX:ReqID

图 9