(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 109314721 A
(43)申请公布日 2019.02.05

(21)申请号 201780036707.X

(22)申请日 2017.08.07

(30)优先权数据
62/422,751 2016.11.16 US
15/585,815 2017.05.03 US

(85)PCT国际申请进入国家阶段日
2018.12.13

(86)PCT国际申请的申请数据
PCT/CN2017/096233 2017.08.07

(87)PCT国际申请的公布数据
WO2018/090674 EN 2018.05.24

(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 郭雷 陈瑾 陈冲 柯晓棣 陈晨

(51)Int.Cl.
H04L 29/08(2006.01)

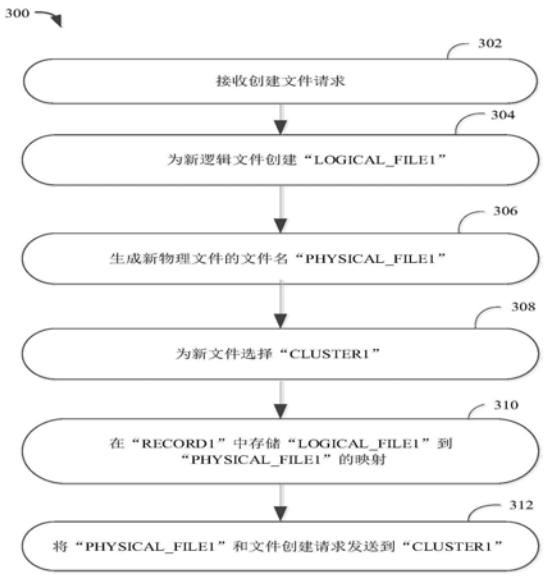
权利要求书4页 说明书11页 附图15页

(54)发明名称

分布式文件系统的多个集群的管理

(57)摘要

本发明涉及用于管理其上存储有集群文件的分布式文件系统的集群的方法和系统。其上运行有应用的用户设备与分布式文件系统的集群之间设有中间层,用于使用关于集群文件的元数据管理和协调多个集群之间的操作。



CN 109314721 A

1.一种用于管理分布式文件系统的多个集群的系统,其特征在于,所述多个集群具有集群文件,所述系统包括:

至少一个处理单元;

非瞬时性存储器,其通信地耦合至所述至少一个处理单元,并且包括所述至少一个处理单元可执行的以用于如下操作的计算机可读程序指令:

从用户设备上的应用接收创建新集群文件的请求;

创建与所述新集群文件对应的集群管理文件;

分配逻辑文件名给所述集群管理文件,并分配物理文件名给所述新集群文件;

在所述多个集群中为所述新集群文件分配物理文件位置;

将元数据存储在所述集群管理文件中,将所述集群管理文件映射到所述新集群文件,其中所述元数据包括所述物理文件名和所述物理文件位置;

使用所述物理文件名将创建所述新集群文件的所述请求传输到所述物理文件位置对应的所述集群中的一个集群;

使用所述逻辑文件名向所述应用确认所述集群文件的创建。

2.根据权利要求1所述的系统,其特征在于,所述分布式文件系统是Hadoop分布式文件系统或Hadoop兼容文件系统。

3.根据权利要求1所述的系统,其特征在于,可执行所述程序指令以用于实现至少一个客户端组件与所述应用和所述集群通信,以及用于实现至少一个管理器组件生成和存储所述元数据。

4.根据权利要求3所述的系统,其特征在于,所述至少一个客户端组件包括多个客户端组件,每个客户端组件用于与不同用户应用通信。

5.根据权利要求3所述的系统,其特征在于,所述至少一个管理器组件包括多个管理器组件,每个管理器组件用于与所述多个集群的不同分组通信。

6.根据权利要求1所述的系统,其特征在于,可执行所述程序指令以用于实现将所述系统实现为虚拟机。

建议增加关于多个集群管理系统之间的共识、分发策略和目录用例的权项。

7.一种用于管理分布式文件系统的多个集群的方法,其特征在于,所述多个集群具有集群文件,所述方法包括:

从用户设备上的应用接收创建新集群文件的请求;

创建与所述新集群文件对应的集群管理文件;

分配逻辑文件名给所述集群管理文件,并分配物理文件名给所述新集群文件;

在所述多个集群中为所述新集群文件分配物理文件位置;

将元数据存储在所述集群管理文件中,将所述集群管理文件映射到所述新集群文件,其中所述元数据包括所述物理文件名和所述物理文件位置;

使用所述物理文件名将创建所述新集群文件的所述请求传输到所述物理文件位置对应的所述集群中的一个集群;

使用所述逻辑文件名向所述应用确认所述集群文件的创建。

8.根据权利要求7所述的方法,其特征在于,还包括将创建所述新集群文件的所述请求从第一格式转换为第二格式,其中所述应用支持所述第一格式,所述集群支持所述第二格

式。

9.根据权利要求7所述的方法，其特征在于，为所述新集群文件分配物理文件位置包括：选择所述集群中离请求所述新集群文件的所述应用最近的集群。

10.根据权利要求7所述的方法，其特征在于，为所述新集群文件分配物理文件位置包括：从所述集群中选择与其它集群相比可用空间最大的集群。

11.根据权利要求7所述的方法，其特征在于，还包括：

接收访问所述新集群文件的请求，其中所述请求包括所述逻辑文件名；

使用所述逻辑文件名检索所述新集群文件对应的所述元数据；

根据所述元数据确定所述物理文件位置；

使用所述物理文件名将访问所述新集群文件的所述请求发送到所述集群中的至少一个集群。

12.根据权利要求11所述的方法，其特征在于，发送访问所述新集群文件的所述请求包括：通过考虑系统性能、系统一致性、本地数据可用性以及所述集群之间的负载均衡中的至少一个来选择所述至少一个集群。

13.一种计算机可读介质，其特征在于，所述计算机可读介质存储有可由处理器执行的程序指令，以用于管理分布式文件系统的多个集群，其中，所述多个集群具有集群文件，所述程序指令用于：

从用户设备上的应用接收创建新集群文件的请求；

创建与所述新集群文件对应的集群管理文件；

分配逻辑文件名给所述集群管理文件，并分配物理文件名给所述新集群文件；

在所述多个集群中为所述新集群文件分配物理文件位置；

将元数据存储在所述集群管理文件中，将所述集群管理文件映射到所述新集群文件，其中所述元数据包括所述物理文件名和所述物理文件位置；

使用所述物理文件名将创建所述新集群文件的所述请求传输到所述物理文件位置对应的所述集群中的一个集群；

使用所述逻辑文件名向所述应用确认所述集群文件的创建。

14.一种用于管理分布式文件系统的多个集群的系统，其特征在于，所述多个集群具有集群文件，所述系统包括：

至少一个处理单元；

非瞬时性存储器，其通信地耦合至所述至少一个处理单元，并且包括所述至少一个处理单元可执行的以用于如下操作的计算机可读程序指令：

接收访问所述集群中的至少一个集群内的集群文件的请求，其中所述请求包括逻辑文件名，所述请求从用户设备上的应用接收；

使用所述逻辑文件名检索元数据，其中所述元数据将逻辑文件映射到所述集群文件对应的物理文件；

根据所述元数据确定所述物理文件的位置；

使用物理文件名将访问所述集群文件的所述请求发送到所述物理文件的所述位置对应的所述集群中的一个集群。

15.根据权利要求14所述的系统，其特征在于，所述分布式文件系统是Hadoop分布式文

件系统或Hadoop兼容文件系统。

16.根据权利要求14所述的系统，其特征在于，可执行所述程序指令以用于实现至少一个客户端组件与所述应用和所述集群通信，以及用于实现至少一个管理器组件生成和存储所述元数据。

17.根据权利要求16所述的系统，其特征在于，所述至少一个客户端组件包括多个客户端组件，每个客户端组件用于与不同用户应用通信。

18.根据权利要求16所述的系统，其特征在于，所述至少一个管理器组件包括多个管理器组件，每个管理器组件用于与所述多个集群的不同分组通信。

19.根据权利要求14所述的系统，其特征在于，可执行所述程序指令以用于实现将所述系统实现为虚拟机。

20.一种用于管理分布式文件系统的多个集群的方法，其特征在于，所述多个集群具有集群文件，所述方法包括：

接收访问所述集群中的至少一个集群内的集群文件的请求，其中所述请求包括逻辑文件名，所述请求从用户设备上的应用接收；

使用所述逻辑文件名检索元数据，其中所述元数据将逻辑文件映射到所述集群文件对应的物理文件；

根据所述元数据确定所述物理文件的位置；

使用物理文件名将访问所述集群文件的所述请求发送到所述集群中的至少一个集群。

21.根据权利要求20所述的方法，其特征在于，还包括将访问所述新集群文件的所述请求从第一格式转换为第二格式，其中所述应用支持所述第一格式，所述集群支持所述第二格式。

22.根据权利要求20所述的方法，其特征在于，发送访问所述新集群文件的所述请求包括：将所述请求发送到所述物理文件的所述位置对应的所述集群中的所述至少一个集群，其中所述位置是根据所述元数据确定的。

23.根据权利要求20所述的方法，其特征在于，发送访问所述新集群文件的所述请求包括：通过考虑系统性能、系统一致性、本地数据可用性以及所述集群之间的负载均衡中的至少一个来选择所述集群中的所述至少一个集群。

24.根据权利要求20所述的方法，其特征在于，还包括：

接收修改所述新集群文件的请求，其中所述请求包括所述逻辑文件名；

使用所述逻辑文件名检索所述新集群文件对应的所述元数据；

根据修改所述新集群文件的所述请求生成新的元数据；

根据所述元数据确定所述物理文件的位置；

使用所述物理文件名将修改所述新集群文件的所述请求发送到所述集群中的至少一个集群；

存储与所述新集群文件相关联的所述新的元数据。

25.根据权利要求24所述的方法，其特征在于，发送修改所述新集群文件的所述请求包括：将所述请求发送到所述物理文件的所述位置对应的所述集群中的所述至少一个集群，其中所述位置是根据所述元数据确定的。

26.一种计算机可读介质，其特征在于，所述计算机可读介质存储有可由处理器执行的

程序指令，以用于管理分布式文件系统的多个集群，其中，所述多个集群具有集群文件，所述程序指令用于：

接收访问所述集群中的至少一个集群内的集群文件的请求，其中所述请求包括逻辑文件名，所述请求由用户设备上的应用发送；

使用所述逻辑文件名检索元数据，其中所述元数据将逻辑文件映射到所述集群文件对应的物理文件；

根据所述元数据确定所述物理文件的位置；

使用物理文件名将访问所述集群文件的所述请求发送到所述集群中的至少一个集群。

分布式文件系统的多个集群的管理

[0001] 相关申请案交叉申请

[0002] 本发明要求2016年11月16日递交的第62/422,751号美国临时专利申请案的在先申请优先权,该在先申请的全部内容以引入的方式并入本文本中。本发明还要求2017年5月3日递交的第15/585,815号美国临时专利申请案的在先申请优先权,该在先申请的全部内容也以引入的方式并入本文本中。

技术领域

[0003] 本发明涉及集中管理分布式文件系统的多个集群。

背景技术

[0004] 分布式文件系统的集群为基于客户端/服务器的应用,该应用允许用户(通过客户端)访问和处理来自多个主机的数据,其中这些主机通过计算机网络共享。

[0005] 由于文件系统的大小增加以及组织内有不同的需求,导致相互独立地创建和管理分布式文件系统的多个集群。这会带来一些挑战,例如与以下内容相关的挑战:在一集群中生成不同集群需要的数据,多个集群之间的应用负载均衡,以及用于容灾目的的数据复制需求。

[0006] 存在用于解决这些问题的某些工具,但这些工具很复杂并且针对的是个体需求,例如数据复制或同步命名空间。因此需要一种整体方法来管理分布式文件系统的多个集群。

发明内容

[0007] 本发明涉及用于管理其上存储有集群文件的分布式文件系统的集群的方法和系统。其上运行有应用的用户设备与分布式文件系统的集群之间设有中间层,用于使用关于集群文件的元数据管理和协调多个集群之间的操作。

[0008] 根据第一广义方面,提供了一种用于管理分布式文件系统的多个集群的系统,所述多个集群具有集群文件。所述系统包括至少一个处理单元和通信地耦合至所述至少一个处理单元并且包括计算机可读程序指令的非瞬时性存储器。所述程序指令可由所述至少一个处理单元执行以用于如下操作:从用户设备上的应用接收创建新集群文件的请求;创建与所述新集群文件对应的集群管理文件;分配逻辑文件名和物理文件名给所述新集群文件;在所述多个集群中为所述新集群文件分配物理文件位置;将元数据存储在所述集群管理文件中,将所述集群管理文件映射到所述新集群文件,其中所述元数据包括所述物理文件名和所述物理文件位置;使用所述物理文件名将创建所述新集群文件的所述请求传输到所述物理文件位置对应的所述集群中的一个集群;使用所述逻辑文件名向所述应用确认所述集群文件的创建。

[0009] 在任一前述实施例中,所述分布式文件系统是Hadoop分布式文件系统或Hadoop兼容文件系统。

[0010] 在任一前述实施例中，可执行所述程序指令以用于实现至少一个客户端组件与所述应用和所述集群通信，以及用于实现至少一个管理器组件生成和存储所述元数据。

[0011] 在任一前述实施例中，所述至少一个客户端组件包括多个客户端组件，每个客户端组件用于与不同用户应用通信。

[0012] 在任一前述实施例中，所述至少一个管理器组件包括多个管理器组件，每个管理器组件用于与所述多个集群的不同分组通信。

[0013] 在任一前述实施例中，可执行所述程序指令以用于实现将所述系统实现为虚拟机。

[0014] 根据另一广义方面，提供了一种用于管理分布式文件系统的多个集群的方法，所述多个集群具有集群文件。从用户设备上的应用接收创建新集群文件的请求。创建与所述新集群文件对应的集群管理文件。分配逻辑文件名和物理文件名给所述新集群文件。在所述多个集群中为所述新集群文件分配物理文件位置。将元数据存储在所述集群管理文件中，从而将所述集群管理文件映射到所述新集群文件，其中所述元数据包括所述物理文件名和所述物理文件位置。使用所述物理文件名将创建所述新集群文件的所述请求传输到所述物理文件位置对应的所述集群中的一个集群。使用所述逻辑文件名向所述应用确认所述集群文件的创建。

[0015] 在任一前述实施例中，所述方法还包括将创建所述新集群文件的所述请求从第一格式转换为第二格式，其中所述应用支持所述第一格式，所述集群支持所述第二格式。

[0016] 在任一前述实施例中，为所述新集群文件分配物理文件位置包括：选择所述集群中离请求所述新集群文件的所述应用最近的集群。

[0017] 在任一前述实施例中，为所述新集群文件分配物理文件位置包括：从所述集群中选择与其它集群相比可用空间最大的集群。

[0018] 在任一前述实施例中，所述方法还包括：接收访问所述新集群文件的请求，其中所述请求包括所述逻辑文件名；使用所述逻辑文件名检索所述新集群文件对应的所述元数据；根据所述元数据确定所述物理文件位置；使用所述物理文件名将访问所述新集群文件的所述请求发送到所述集群中的至少一个集群。

[0019] 在任一前述实施例中，发送访问所述新集群文件的所述请求包括：通过考虑系统性能、系统一致性、本地数据可用性以及所述集群之间的负载均衡中的至少一个来选择所述至少一个集群。

[0020] 根据另一广义方面，提供了一种计算机可读介质，所述计算机可读介质存储有可由处理器执行的程序指令，以用于管理分布式文件系统的多个集群，其中，所述多个集群具有集群文件。所述程序指令用于执行本文所述的任一方法。

[0021] 根据又一广义方面，提供了一种用于管理分布式文件系统的多个集群的系统，所述多个集群具有集群文件。所述系统包括至少一个处理单元和通信地耦合至所述至少一个处理单元并且包括计算机可读程序指令的非瞬时性存储器。所述程序指令可由所述至少一个处理单元执行以用于如下操作：接收访问所述集群中的至少一个集群内的集群文件的请求，其中所述请求包括逻辑文件名，所述请求从用户设备上的应用接收；使用所述逻辑文件名检索元数据，其中所述元数据将逻辑文件映射到所述集群文件对应的物理文件；根据所述元数据确定所述物理文件的位置；使用物理文件名将访问所述集群文件的所述请求发送

到所述物理文件的所述位置对应的所述集群中的一个集群。

[0022] 在任一前述实施例中，所述分布式文件系统是Hadoop分布式文件系统或Hadoop兼容文件系统。

[0023] 在任一前述实施例中，可执行所述程序指令以用于实现至少一个客户端组件与所述应用和所述集群通信，以及用于实现至少一个管理器组件生成和存储所述元数据。

[0024] 在任一前述实施例中，所述至少一个客户端组件包括多个客户端组件，每个客户端组件用于与不同用户应用通信。

[0025] 在任一前述实施例中，所述至少一个管理器组件包括多个管理器组件，每个管理器组件用于与所述多个集群的不同分组通信。

[0026] 在任一前述实施例中，可执行所述程序指令以用于将所述系统实现为虚拟机。

[0027] 根据另一广义方面，提供了一种用于管理分布式文件系统的多个集群的方法，所述多个集群具有集群文件。接收访问所述集群中的至少一个集群内的集群文件的请求，其中所述请求包括逻辑文件名，所述请求从用户设备上的应用接收。使用所述逻辑文件名检索元数据，其中所述元数据将逻辑文件映射到所述集群文件对应的物理文件。根据所述元数据确定所述物理文件的位置；使用物理文件名将访问所述集群文件的所述请求发送到所述集群中的至少一个集群。

[0028] 在任一前述实施例中，所述方法还包括将访问所述新集群文件的所述请求从第一格式转换为第二格式，其中所述应用支持所述第一格式，所述集群支持所述第二格式。

[0029] 在任一前述实施例中，发送访问所述新集群文件的所述请求包括：将所述请求发送到所述物理文件的所述位置对应的所述集群中的所述至少一个集群，其中所述位置是根据所述元数据确定的。

[0030] 在任一前述实施例中，发送访问所述新集群文件的所述请求包括：通过考虑系统性能、系统一致性、本地数据可用性以及所述集群之间的负载均衡中的至少一个来选择所述集群中的所述至少一个集群。

[0031] 在任一前述实施例中，所述方法还包括：接收修改所述新集群文件的请求，其中所述请求包括所述逻辑文件名；使用所述逻辑文件名检索所述新集群文件对应的所述元数据；根据修改所述新集群文件的所述请求生成新的元数据；根据所述元数据确定所述物理文件的位置；使用所述物理文件名将修改所述新集群文件的所述请求发送到所述集群中的至少一个集群；存储与所述新集群文件相关联的所述新的元数据。

[0032] 在任一前述实施例中，发送修改所述新集群文件的所述请求包括：将所述请求发送到所述物理文件的所述位置对应的所述集群中的所述至少一个集群，其中所述位置是根据所述元数据确定的。

附图说明

[0033] 进一步地，通过阅读以下结合附图所作的详细描述将容易了解本发明的特征和优势，附图包括：

[0034] 图1为示例计算环境的框图；

[0035] 图2为示例集群管理系统的框图；

[0036] 图3为示出了文件创建请求的示例实施例的流程图；

[0037] 图4为示出了文件打开请求的示例实施例的流程图；

[0038] 图5为具有多个客户端组件的示例集群管理系统的框图；

[0039] 图6为具有多个管理器组件的示例集群管理系统的框图；

[0040] 图7为具有多个连接在一起的子单元的示例集群管理系统的框图；

[0041] 图8为在集群管理系统内有多个子单元的示例计算环境的框图；

[0042] 图9A为用于实现集群管理系统的示例计算设备的框图；

[0043] 图9B为由图9B的计算设备实现的示例虚拟机的框图；

[0044] 图10A示出了文件创建的各种场景；

[0045] 图10B示出了文件访问的各种场景；

[0046] 图10C示出了文件复制的各种场景；

[0047] 图11A示出了以第一模式操作的集群管理系统的示例；

[0048] 图11B示出了以第二模式操作的集群管理系统的示例。

[0049] 需注意，在所有附图中，相同特征由相同附图标号标识。

具体实施方式

[0050] 根据本实施例，用户设备上的应用与一个或多个分布式文件系统的集群之间设有中间层。中间层在本文称为集群管理系统。中间层从用户设备上的应用接收对存储在分布式文件系统中的文件的请求。集群管理系统创建集群管理文件，集群管理文件是存储在中间层的逻辑文件，用于管理集群文件，集群文件是存储在分布式文件系统中的物理文件。元数据存储在集群管理文件中，以便将逻辑文件映射到物理文件。因此，元数据包括映射信息以及其它信息，例如分布式文件系统中的物理文件的名称和位置。

[0051] 参考图1，示出了计算环境100。至少一个用户设备102_1、102_2(统称为用户设备102)上运行有至少一个应用114_1、114_2(统称为应用114)。计算环境100包括多个集群104_1、104_2、104_3(统称为集群104)。每个集群104包括一个或多个分布式文件系统108_1、108_2、108_3、108_4、108_5、108_6、108_7(统称为DFS108)。每个DFS108存储可由用户设备102上的应用114访问的一个或多个文件110_1、110_2、110_3、110_4、110_5(统称为集群文件110)。集群文件110对应于以各种格式存储在DFS108中的数据和/或目录，并且可以被应用114访问和处理，犹如集群文件就在用户设备102中。集群文件110也可以称为物理文件，因为集群文件是在集群104的真实底层文件系统中创建的。

[0052] 在一些实施例中，DFS108为Hadoop分布式文件系统(Hadoop Distributed File System，HDFS)和/或Hadoop兼容文件系统(Hadoop Compatible File System，HCFS)，例如Amazon S3、Azure Blob存储器、Google云存储连接器等。每个集群104均可以包括HDFS等单一类型的DFS108，或者一种或多种类型的兼容分布式文件系统108，例如一个HDFS和两个HCFS等。也可以使用其它类型的分布式文件系统108。

[0053] 集群104_1等给定集群的DFS108位于相同或不同位置。例如，$DFS108_1$位于一组织的室内，$DFS108_2$位于云中。在另一示例中，$DFS108_1$位于一组织的第一分支中，$DFS108_2$位于同一组织的第二分支中，第一分支和第二分支位于不同地理位置，例如不同城市、不同国家，不同洲等。在又一示例中，$DFS108_1$和$DFS108_2$二者位于相同地理位置，但是对应不同部门或位于组织的不同楼层。集群104也可以设置在相同或不同位置。例如，集群104_1位于中国的

多个城市，集群104_2遍布欧洲，集群104_3位于佛罗里达州的迈阿密。

[0054] 集群管理系统106设为集群104和用户设备102之间的中间层。集群管理系统106是管理和协调集群104的操作的实体。集群管理系统106与来自用户设备102的应用104通信，以接收针对集群文件110的请求。针对集群文件110的请求可包括对集群文件110的各种操作，例如创建集群文件、修改集群文件、访问集群文件以及替换集群文件等。当接收到的请求要求创建或修改文件时，例如创建文件请求、修改文件名请求、替换文件请求等，集群管理系统106进行生成或更新，然后为集群文件110存储生成的或更新后的元数据。最终将在集群104中的一个集群中创建文件，更具体地，在DCS108中创建文件作为集群文件110。集群文件的元数据包括文件名以及集群文件在集群104中的位置。当接收到的请求要求访问但不修改文件时，集群管理系统106使用集群文件的元数据来定位文件从而相应地提供访问。

[0055] 图2示出了集群管理系统106的示例实施例。作为示例实施例，集群管理系统106包括客户端组件200和管理器组件202，二者协作以在全局命名空间内管理集群104。因此，集群文件虽然存在于集群104的多个物理位置中，但却是根据全局命名空间的统一结构进行管理。客户端组件200从应用114接收针对集群文件110的请求。客户端组件200向管理器组件202发送针对该请求的指令。

[0056] 当请求涉及创建或更新文件时，管理器组件202基于请求创建和/或更新集群文件110的元数据。

[0057] 图3示出了根据方法300创建新集群文件110的示例实施例。在步骤302处，接收创建新文件请求。该请求由集群管理系统106的客户端组件200从任一用户设备102上的任一应用114接收。在一些实施例中，文件创建请求中提供的唯一信息是请求本身，即一个创建文件的命令行。在一些实施例中，请求还包括文件名和/或文件目的地。文件目的地是指任何集群104中的给定位置，或集群104的给定集群中任何DFS108中的给定位置。如果新文件与任一集群104中已存在的现有文件有任何关系，也可以在请求中提供该信息。

[0058] 在步骤304处，为新集群文件创建集群管理文件。集群管理文件也可以称为在全局文件名空间中创建和管理的逻辑文件。每个逻辑文件可以具有一个或多个对应的物理文件(即集群文件110)。在本示例中，集群管理文件的文件名为“LOGICAL_FILE1”。

[0059] 在步骤306处，生成物理文件的文件名。在本示例中，物理文件的文件名称为“PHYSICAL_FILE1”。物理文件的文件名为待创建新文件的元数据。

[0060] 在步骤308处，在由集群管理系统106管理的各种集群104中选择物理文件“PHYSICAL_FILE1”的位置。当该位置不是请求中提供的信息的一部分时，可以根据各种因素选择该位置，这将在下文详述。物理文件的位置也是集群文件的元数据的一部分。

[0061] 集群管理系统106使用逻辑文件名，该示例中即“LOGICAL_FILE1”，与应用114进行通信。例如，从应用114接收的打开该文件的请求将采用“打开LOGICAL_FILE1”的形式。集群管理系统106使用物理文件名，该示例中即“PHYSICAL_FILE1”，与集群104进行通信。例如，发送到集群104中的合适集群的打开“LOGICAL_FILE1”请求将采用“打开PHYSICAL_FILE1”的形式。因此，按照步骤310，集群管理系统106在“RECORD1”中存储“LOGICAL_FILE1”到“PHYSICAL_FILE1”的映射。该映射包括由管理器组件202先前响应创建新文件请求而生成的元数据，并被存储在集群管理文件内。因此，元数据包括物理文件的名称和物理文件在集群104中的位置。

[0062] 在步骤312处，将文件创建请求和物理文件名“PHYSICAL_FILE1”传输到合适的集群，即“CLUSTER1”。然后，集群将相应地创建集群文件110。

[0063] 通过图2中集群管理系统106的示例架构，客户端组件200从应用接收请求，管理器组件202在集群管理文件中生成并存储元数据，客户端组件200将请求发送到合适的集群。也可以使用集群管理系统106的其它架构来实现方法300。

[0064] 元数据存储在一个或多个存储设备例如存储设备204内的集群管理文件中，该存储设备可以在集群管理系统106的本地或远端。

[0065] 图4是示出处理打开已创建文件的请求的示例实施例的方法400。应用114知道逻辑文件而非物理文件的存在。因此，按照步骤402，从应用114接收的任何打开文件请求将包括逻辑文件名。该请求由集群管理系统106的客户端组件200从任一用户设备102上的任一应用114接收。该请求包括要打开文件的名称，即“LOGICAL_FILE1”，并从客户端组件200发送到管理器组件202。

[0066] 在步骤404处，管理器组件202检索将“LOGICAL_FILE1”映射到“PHYSICAL_FILE1”的元数据，以便在步骤406处确定“PHYSICAL_FILE1”在集群104中的位置。元数据存储在存储设备204内的集群管理文件中。在步骤408处，将打开物理文件的请求发送到合适的集群。请求可能采用“打开PHYSICAL_FILE1”的形式并发送到“CLUSTER1”。

[0067] 在一些实施例中，客户端组件200向管理器组件202发送打开“LOGICAL_FILE1”的请求。管理器组件202从“RECORD1”检索出“LOGICAL_FILE1”到“PHYSICAL_FILE1”的映射并且检索出“CLUSTER1”为“PHYSICAL_FILE1”的位置。管理器组件202然后将“CLUSTER1”返回给客户端组件200，客户端组件200向“CLUSTER1”发送打开“PHYSICAL_FILE1”请求。

[0068] 再参照图2，管理器组件202中设有复制器206以便在集群104之间共享信息并且确保集群104之间的一致性。必要时，复制器206可以将来自一个集群，例如集群104_1，的数据复制到另一集群，例如集群104_2。这允许$DFS108_4$等将先前仅能在$DFS108_2$中可用的数据进行本地存储，从而允许集群104_2执行之前仅可以由集群104_1执行的操作。例如，这可以用于集群104之间的负载均衡或用于使用来自集群104_1和集群104_2两者的数据进行的连接运算。在一些实施例中，基于从应用114接收到的请求选择性地或根据需要执行数据复制。在一些实施例中，根据定义的时间表来执行数据复制，以确保数据始终可用于所有集群104。在一些实施例中，选择性地且周期性地执行数据复制。数据复制对应用114和用户设备102都是透明的。

[0069] 数据管理系统106通过负载均衡来改善集群104之间的工作负载分布。负载均衡旨在优化资源使用、最大化吞吐量、最小化响应时间以及避免任何单个资源过载。在一些实施例中，集群管理系统106，例如管理器组件202，用于通过如下操作来优化集群104的性能：根据接收到的请求选择集群104中的集群来执行特定任务。因此，任务可以更均匀地分布在集群104中，和/或因为各种原因集中于集群104中的特定几个集群。在接收到新请求时用于选择集群的一些选择标准是给定集群中数据的可用性、容量、速度、集群的可用性以及请求类型。因为数据可以从一个集群复制到另一集群，所以给定集群中数据的可用性仅仅是优化计算环境100性能的一个标准，会和其它标准一起权衡。

[0070] 在一些实施例中，集群管理系统106，例如客户端组件200，包括转换器208。转换器208用于从应用接收请求，其中该应用基于的DFS类型与集群104的DFS类型中的一个或多个

DFS类型不同。例如，如果应用114_1是基于HCFS，并且集群管理系统106选择将请求发送到集群104_3，集群104_3中DFS108_5、108_6、108_7是HDFS，则转换器208将请求从HCFS转换为HDFS。从应用114_1接收到的请求是HCFS格式，由集群管理系统106传输到集群104_3的请求是HDFS格式。转换器208可以用于执行除HDFS-HCFS和HCFS-HDFS之外的转换。

[0071] 如图1所示，可以在任一集群104与集群管理系统106之间设置代理112。在一些实施例中，代理112嵌入在集群管理系统106内。在一些实施例中，代理112设在集群管理系统106外部。该代理用于向应用114提供对集群104的访问。尽管仅示出了一个代理112，但是在计算环境100中也可以存在一个或多个代理112。

[0072] 根据图5，在一些实施例中，集群管理系统106包括多个客户端组件500_1、……、500_n。客户端组件500_1、……、500_n中的每一个可操作地连接到管理器组件202，并用于与用户设备102的一个或多个应用114通信。在一些实施例中，针对每个可以发送请求的应用114，集群管理系统106包括一个客户端组件500_n。每个客户端组件500_n包括用于将请求从第一格式转换为第二格式的转换器308_1、……、308_n。或者，一个或多个转换器208由集群管理系统106内的客户端组件500_1、……、500_n共享。

[0073] 图6示出了包括n个客户端组件500_1、……、500_n和m个管理器组件600_1、……、600_m的实施例。管理器组件600_1、……、600_m都包括用于存储一组集群文件的元数据的存储介质604_1、……、604_m以及用于跨集群104复制数据的复制器606_1、……、606_m。或者，一个或多个存储介质604_i(其中i＝1至m)和/或一个或多个复制器606_i由集群管理系统106内的管理器组件600_1、……、600_m共享。管理器组件600_1、……、600_m可操作地连接在一起，并能通过各自的共识引擎608_1、……、608_m协调信息交换和/或数据操作。

[0074] 共识引擎608_1、……、608_m用于确保管理器组件600_1、……、600_m在如何处理涉及由不同管理器组件600_1、……、600_m管理的集群104的操作上作达成共识。需要达成共识的操作的示例是数据复制、数据共享以及集群104间负载的重分配。共识也可以用于其它操作。在一些实施例中，定义一个或多个共识协议来协调对集群文件110的操作。共识协议的一些示例是Paxos、Chubby、Phase King、工作量证明、锁步、MSR类型和散列图。也可以使用其它共识协议。

[0075] 在一些实施例中，共识引擎通过共识协议来执行一个或多个集群管理文件中元数据的创建、更新和/或删除。例如，应用114_1发出删除集群文件110_5的请求。该请求由客户端组件500_1接收并传输到管理器组件600_1。共识引擎608_1向管理器组件600_2至600_m的共识引擎608_2至608_m发送用于修改的共识请求(即删除与集群文件110_5相关的元数据)。每个共识引擎608_2至608_m基于其当前状态独立于其它共识引擎针对修改进行投票。如果大多数共识引擎同意该修改请求，则共识引擎608_1向共识引擎608_2至608_m发送修改确认。然后，每个管理器组件600_1至600_m对其本地存储设备604_1至604_m中的本地集群管理文件进行修改。如果大多数共识引擎不同意该修改请求，则拒绝修改，不对管理器组件600_1至600_m中的任何一个进行修改。

[0076] 每个管理器组件600_1、……、600_m与一个或多个客户端组件500_1、……、500_n相关联以形成子单元，所有子单元连接在一起以形成集群管理系统106。在图7的一示例中，示出了三个子单元700_1、700_2、700_3。对于m＝3，在集群管理系统106中有三个管理器组件600_1、600_2、600_3和三个子单元700_1、700_2、700_3。每个管理器组件600_1、600_2、600_3与一个或多个客户端组

件500_1、……、500n形成子单元700_1、700_2、700_3。

[0077] 如图8所示，每个子单元700_1、700_2、700_3与单独的一组集群800_1、800_2、800_3以及单独的一组用户设备802_1、802_2、802_3通信。在一些实施例中，用户设备102在子单元700_1、700_2、700_3之间共享和/或集群104在子单元700_1、700_2、700_3之间共享。

[0078] 计算环境100中的通信、子单元700_1、700_2、700_3之间的通信、集群管理系统106与用户设备102之间的通信和/或集群管理系统106与集群104之间的通信以各种方式进行，包括通过一个或多个网络直接地和间接地通信。网络可以涉及有线连接、无线连接或其组合。网络可涉及不同网络通信技术、标准和协议，例如全球移动通信系统(Global System for Mobile Communications，GSM)、码分多址(Code division multiple access，CDMA)、无线本地环路、WiMAX、Wi-Fi、蓝牙、长期演进(Long Term Evolution，LTE)等等。网络可能涉及不同物理介质，例如同轴电缆、光纤、收发信台等。示例网络类型包括因特网、以太网、传统电话业务(plain old telephone service，POTS)线路、公共交换电话网络(public switched telephone network，PSTN)、综合业务数字网络(integrated services digital network，ISDN)、数字用户线(digital subscriber line，DSL)等等，包括它们的任何组合。网络可以包括局域网和/或广域网。

[0079] 图9A是用于实现集群管理系统106的计算设备910的示例实施例。计算设备910包括处理单元912和其中存储有计算机可执行指令916的存储器914。处理单元912可以包括能用于执行一系列步骤的任何合适设备，从而计算设备910或其它可编程装置执行指令916时，本文所述方法中的指定功能/动作/步骤得以执行。例如，处理单元912可以包括任何类型的通用微处理器或微控制器、数字信号处理(digital signal processing，DSP)处理器、中央处理器(central processing unit，CPU)、集成电路、现场可编程门阵列(field programmable gate array，FPGA)、可重构处理器、其它适合编程或可编程逻辑电路或其任何组合。

[0080] 存储器914可以包括任何合适的已知或其它机器可读存储介质。例如，存储器914可以包括但不限于电子、磁性、光学、电磁、红外或半导体系统、装置或设备或前述的任何适当组合的非瞬时性计算机可读存储介质。存储器914可以包括位于设备内部或外部的任何类型的计算机存储器的适当组合，例如随机存取存储器(random-access memory，RAM)、只读存储器(read-only memory，ROM)、光盘只读存储器(compact disc read-only memory，CDROM)、电光存储器、磁光存储器、可擦除可编程只读存储器(erasable programmable read-only memory，EPROM)、电可擦除可编程只读存储器(electrically-erasable programmable read-only memory，EEPROM)、铁电RAM(Ferroelectric RAM，FRAM)等。存储器914可以包括适合于可检索地存储处理单元912可执行的机器可读指令916的任何存储装置(例如设备)。

[0081] 在一些实施例中，计算设备910是其上实现一个或多个虚拟机的物理服务器，其示例在

[0082] 图9B中示出。虚拟机950是计算机系统的仿真，包括使用一组虚拟硬件956在操作系统954上运行的应用952。虚拟硬件956包括CPU、存储器、网络接口、磁盘等等。每个虚拟机950在外界看来都是一个真实机器，且是独立的并且不会受到同一物理服务器上其它虚拟机的干扰。在一些实施例中，使用一个或多个虚拟机950来实现集群管理系统106。

[0083] 为单个文件和/或目录生成元数据在响应来自应用114的请求时具有灵活性。请求可以由集群管理系统106发送到存储了原始数据的给定集群，或者可以被转发到另一集群，例如在地理位置上更接近用户设备102的集群，其中发送该请求的应用114在用户设备102中运行。转发请求可改善具有较大网络延迟的广域网(wide area network，WAN)环境的系统性能。注意，转发请求可用于某些类型的请求，例如不修改数据的读取操作。将涉及修改数据的请求，例如写入操作，发送至具有原始文件的集群。

[0084] 实际上，集群管理系统106支持可用于跨集群创建、访问和复制数据的灵活策略。图10A示出了各种数据创建场景的示例。集群管理系统106可以用于在最近的集群上创建文件以便优化性能。该示例由路径1002示出，在路径1002中，应用114_2在集群104_1上创建文件。系统106可以用于在可用空间最多的集群上创建文件，以便集群中的数据分布更加均匀。该示例由路径1004示出，在路径1004中，应用114_3在集群104_2上创建文件。系统106可以用于在指定的集群上创建文件，如路径1006所示，在路径1006中，应用114_6在集群104_3上创建文件。也可以使用其它配置和组合。

[0085] 图10B示出了各种数据访问场景。集群管理系统106可以用于从最近的集群访问文件以获得最佳性能。例如，当应用114_1请求某文件时，通过路径1008在集群104_1中访问该文件。当应用114_6请求上述同一文件时，通过路径1010在集群104_3中访问该文件。在一些实施例中，系统106用于从具有最新更新数据的集群访问文件，以确保强一致性。例如，当应用114_3和114_4中的任何一个请求文件时，即使应用114_3本可以在集群104_1中本地访问该文件，也仍然分别通过路径1012和1014在集群104_2中访问该文件。在一些实施例中，系统106用于在本地集群没有可用副本的情况下从远程集群访问文件。例如，应用114_5请求某文件，通过路径1016在集群104_1中访问该文件。在一些实施例中，为了负载均衡的目的，系统106用于从工作负载最少的集群访问所请求的文件。例如，当应用114_2请求访问某文件时，即使集群104_1中也有该文件，也仍然由集群104_3通过路径1018提供文件。也可以使用其它配置和组合。

[0086] 图10C示出了各种数据复制场景。系统可以用于复制选定文件或所有文件。可以将这些文件仅复制到选定集群或所有集群。例如，文件1024仅在两个集群中复制，而文件1026在三个集群中复制。可以使用管道路径来复制文件，如路径1020A、1020B所示。文件也可以从中心集群复制，例如使用路径1022A、1022B复制。在一些实施例中，集群管理系统106用于通过流模式复制数据以达到最佳可用性，或定期地批量复制数据以获得更好的性能，或通过以上方式的组合复制数据。也可以使用其它配置和组合。

[0087] 可以使用各种机制来进行集群104之间的数据复制，无论数据是否是单独的文件和/或目录。在一些实施例中，基于应用114接收的请求触发复制。在一些实施例中，根据常规时间表来规划复制。在一些实施例中，根据一个或多个策略触发复制。也可以使用这些实施例的任何组合。

[0088] 因此，集群管理系统106向构成不同集群的一部分的所有DFS108提供单个管理系统。集群管理系统106的添加并未改变集群文件和集群104的一般功能。

[0089] 在一些实施例中，集群104是HDFS/HCFS集群。客户端组件200是与应用114通信的HCFS兼容库。应用114可以基于HDFS/HCFS协议动态地加载客户端组件200。换言之，应用114可以根据需要加载客户端组件200。另外，客户端组件200可以为不同集群104加载不同驱动

程序。例如，客户端组件200为集群104_1加载第一驱动程序，客户端组件为集群104_2加载第二驱动程序。第一驱动程序用于HDFS版本1，而第二驱动程序用于HCFS版本x。可以为集群104_3加载第三驱动程序，适用于HDFS版本2。

[0090] 应用114可以使用统一资源标识符(Uniform Resource Identifier，URI)来表示要访问的数据文件。示例URI格式为“scheme://authority/path”。对于基于HDFS的应用，使用HDFS方案。对于基于HCFS的应用，可以使用各种方案类型和文件系统插件，例如对于基于Amazon S3的HCFS系统使用“s3”。客户端组件200用于向应用114提供基于HDFS的URI和/或基于HCFS的URI。使用上述实施例的示例为用于HCFS方案的“Pylon://user/LOGICAL_FILE1”，或者用于HDFS方案的“hdfs://temp/LOGICAL_FILE1”。

[0091] 在一些实施例中，客户端组件用于在两种或更多种模式下运行，每种模式根据使用的对应方案设置应用114的特定行为。表1示出了具有两种模式和两种方案的示例。

[0092]

	HDFS 方案	HCFS 方案
模式 1	应用未加载客户端组件	应用加载客户端组件以处理请求，客户端组件与管理器组件协调以进行数据访问
模式 2	应用加载客户端组件以处理请求，客户端组件与管理器组件协调以进行数据访问	

[0093] 表1

[0094] 图11A示出了模式1的示例。如图所示，将基于HDFS方案的应用114_1发送的请求直接发送到集群104_1。基于HCFS方案的应用114_1发送的请求在被转发到集群104_1前，先被发送到客户端组件200。图11B示出了模式2的示例。如图所示，基于HDFS和HCFS的请求在被转发到集群104_1之前，先被发送到客户端组件200。如果客户端组件200基于的是HCFS，转换器208将基于HDFS的请求转换成基于HCFS方案。

[0095] 通过使用存储介质204中存储的元数据，管理器组件202为基于HDFS/HCFS的多个集群114提供全局命名空间。管理器组件监管多个集群114，以及通过复制器206调度集群114之间的数据流。复制器206用于跟踪对文件的更改并规划跨集群114的复制任务。

[0096] 在存储介质204或别处创建并存储元数据以便管理集群文件110。在一些实施例中，每个集群文件110具有对应的集群管理文件。集群管理文件包含逻辑文件到物理文件的映射以及管理集群文件110所需的任何其它信息。逻辑文件到物理文件的映射以及管理集群文件110所需的任何其它信息为元数据。当存在多个管理器组件400时，通过共识协议协调集群文件上的操作。

[0097] 在一些实施例中，集群文件110根据目录编目结构来组织，集群管理文件用于存储关于集群目录和目录间关系的信息(即元数据)。每个集群目录可以包括一个或多个集群文件110，并且在一些情况下，引用子目录。还可以在集群之间复制包括集群文件的目录和目录间关系，而非仅仅将集群文件从一个集群复制到另一集群。

[0098] 在一些实施例中，目录在元数据管理层被当作逻辑概念使用，因此不需要物理集群层中的一对一映射。此时，一些与目录相关的操作不需要访问底层集群。例如，应用114发

送创建目录请求。集群管理系统106的客户端组件200接收该请求。客户端组件200将该请求传送给管理器组件202，管理器组件202创建关于该请求的元数据。元数据可以存储在新集群管理文件或现有集群管理文件中。在一些实施例中，当创建新目录时，为该目录创建新集群管理文件。存储在新集群管理文件中的元数据包括目录名称，例如“DIRECTORY1”，在“RECORD2”中创建和存储将新集群管理文件映射到“DIRECTORY1”的元数据。元数据还可以包括构成目录一部分的任何文件。客户端组件200向应用114返回目录名称。由于创建目录不会对集群文件的结构造成任何影响，因此该操作无需访问集群。

[0099] 应用114可以使用目录名称与集群管理系统106交互，例如通过请求列出“DIRECTORY1”的内容这种方式进行交互。该请求由客户端组件200接收并传送到管理器组件202。管理器组件202访问集群管理文件并检索关于“DIRECTORY1”的元数据。客户端组件200向应用114返回该信息。同样，由于列举目录文件不会对集群文件的结构造成任何影响，因此该操作无需访问集群。

[0100] 类似地，重命名目录只涉及元数据管理并不需要访问底层物理集群文件。当从应用114接收到重命名目录的请求时，集群管理系统106更新目录的集群管理文件以及目录中每个文件的所有集群管理文件。如果目录中有大量文件，这可能是一个耗时的操作，但重命名可以与其它正在进行的写入操作同时进行。

[0101] 删除目录等的其它请求将涉及访问集群文件，因为改变了集群文件的底层结构。目录删除操作的第一部分与目录创建操作相同。删除目录请求由客户端组件200接收并发送到管理器组件202。所有集群管理文件通过删除目录本身的条目(名称、内容、状态)以及目录下文件的条目(名称、状态)相应更新。客户端组件200向应用114发送确认删除。另外，集群管理系统106还向集群104发送通知以删除目录。

[0102] 这里描述的每个计算机程序可以用高级程序、或面向对象的编程、或脚本语言或其组合来实现，以与计算机系统进行通信。或者，程序可以用汇编或机器语言来实现。该语言可以是编译或解释的语言。每个这样的计算机程序可以存储在存储介质或者设备上，例如ROM、磁盘、光盘、闪存驱动器或者任何其它合适的存储介质或者设备。

[0103] 集群管理系统106的实施例也可以认为是通过其上存储有计算机程序的非瞬时性计算机可读存储介质来实现。

[0104] 计算机可执行指令可以为许多形式，包括程序模块，由一个或多个计算机或其它设备执行。一般而言，程序模块包括执行特定任务或实现特定抽象数据类型的例程、程序、对象、组件以及数据结构等。通常，该程序模块的功能可根据各实施例需要进行组合或分配。

[0105] 本计算环境100的各个方面可以单独使用、组合使用或者在前面描述的实施例中未具体讨论的各种情况中使用，因此，其应用不限于在前面描述中阐述的或者在附图中示出的部件的细节和布置。例如，一个实施例中描述的方面可以与其它实施例中描述的方面任意组合。虽然已经示出和描述了特定实施例，但是对于本领域技术人员来说显而易见的是，可以在本发明更广的方面进行改变和修改而不偏离本发明。所附权利要求在其范围内涵盖所有这些改变和修改。

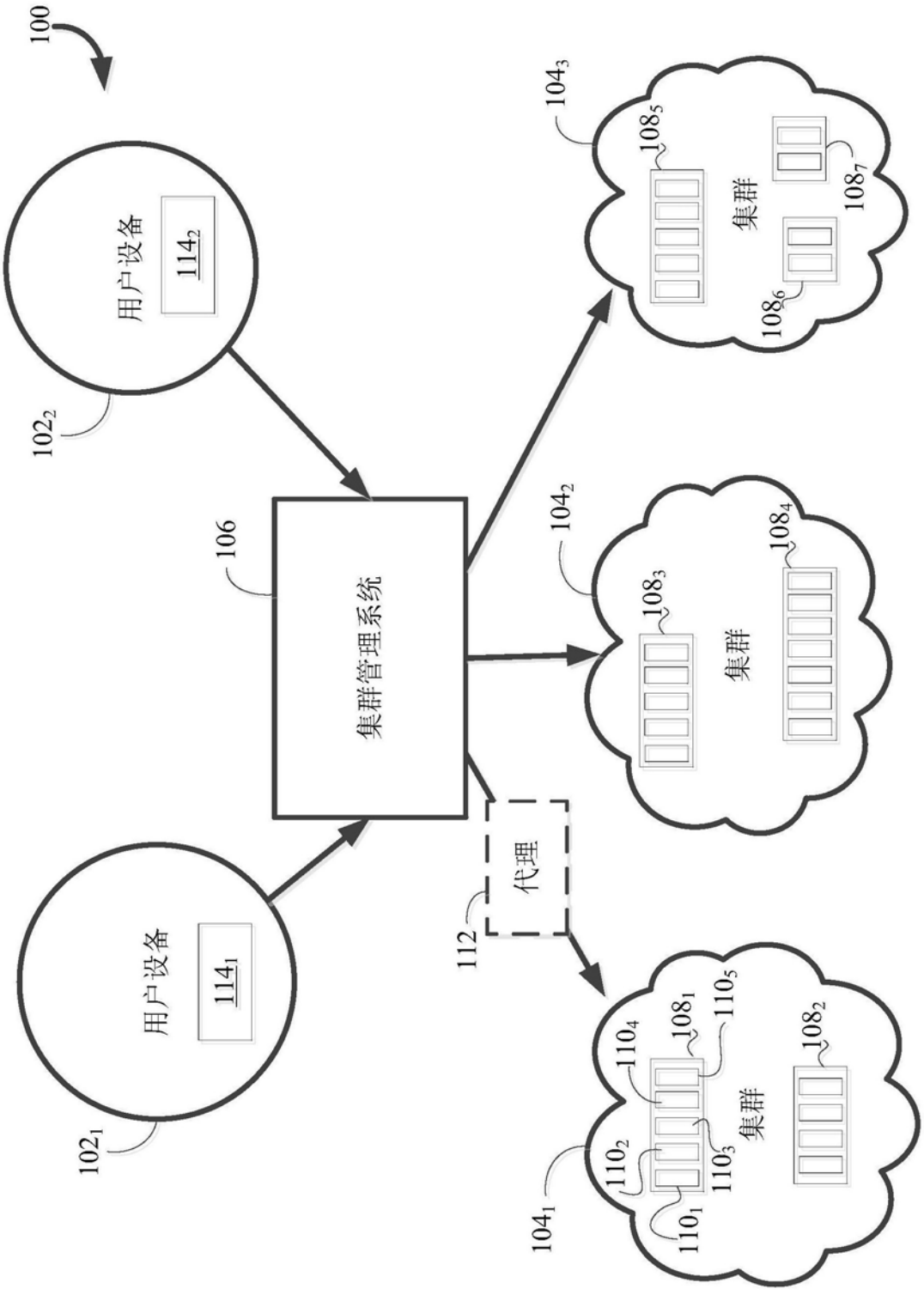

100
用户设备
114_2
102_2
106
集群管理系统
104_3
108_5
集群
108_7
108_6
104_2
108_3
集群
108_4
112
代理
用户设备
114_1
102_1
104_1
110_4
108_1
110_5
110_2
110_3
集群
108_2
110_1

图1

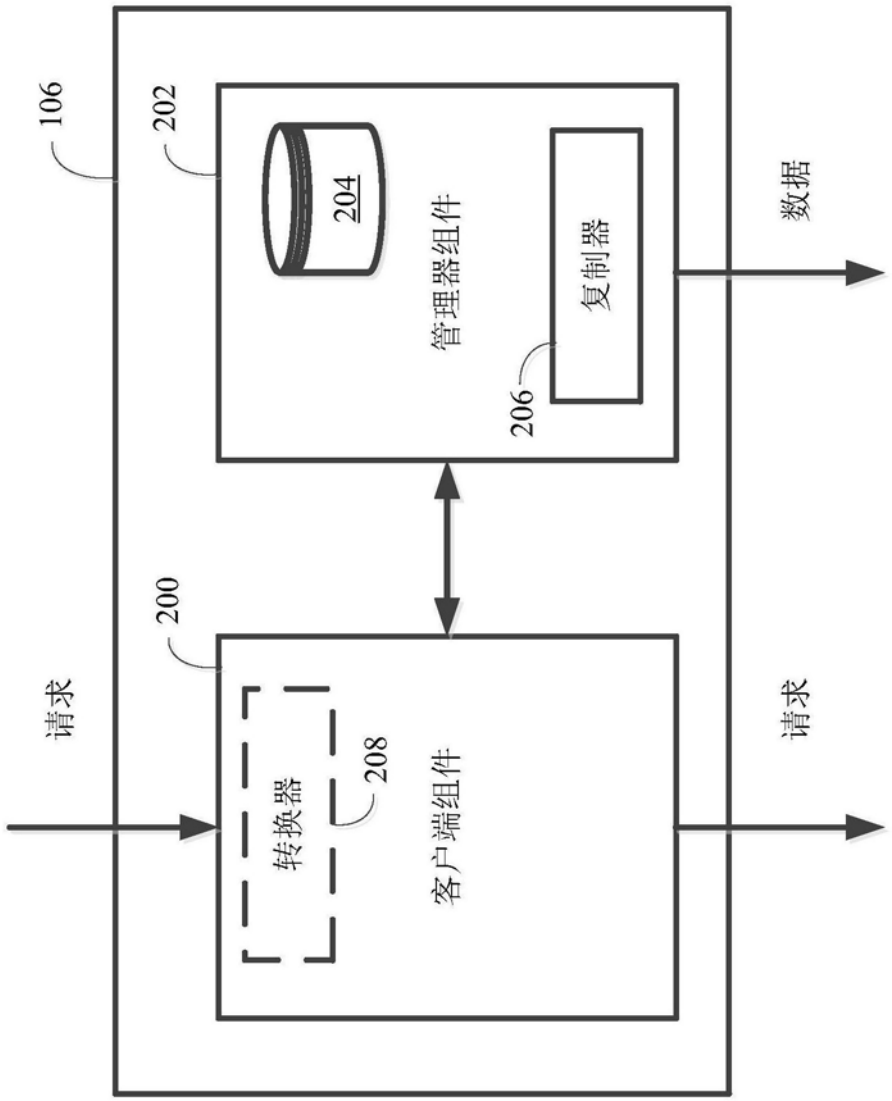


图2

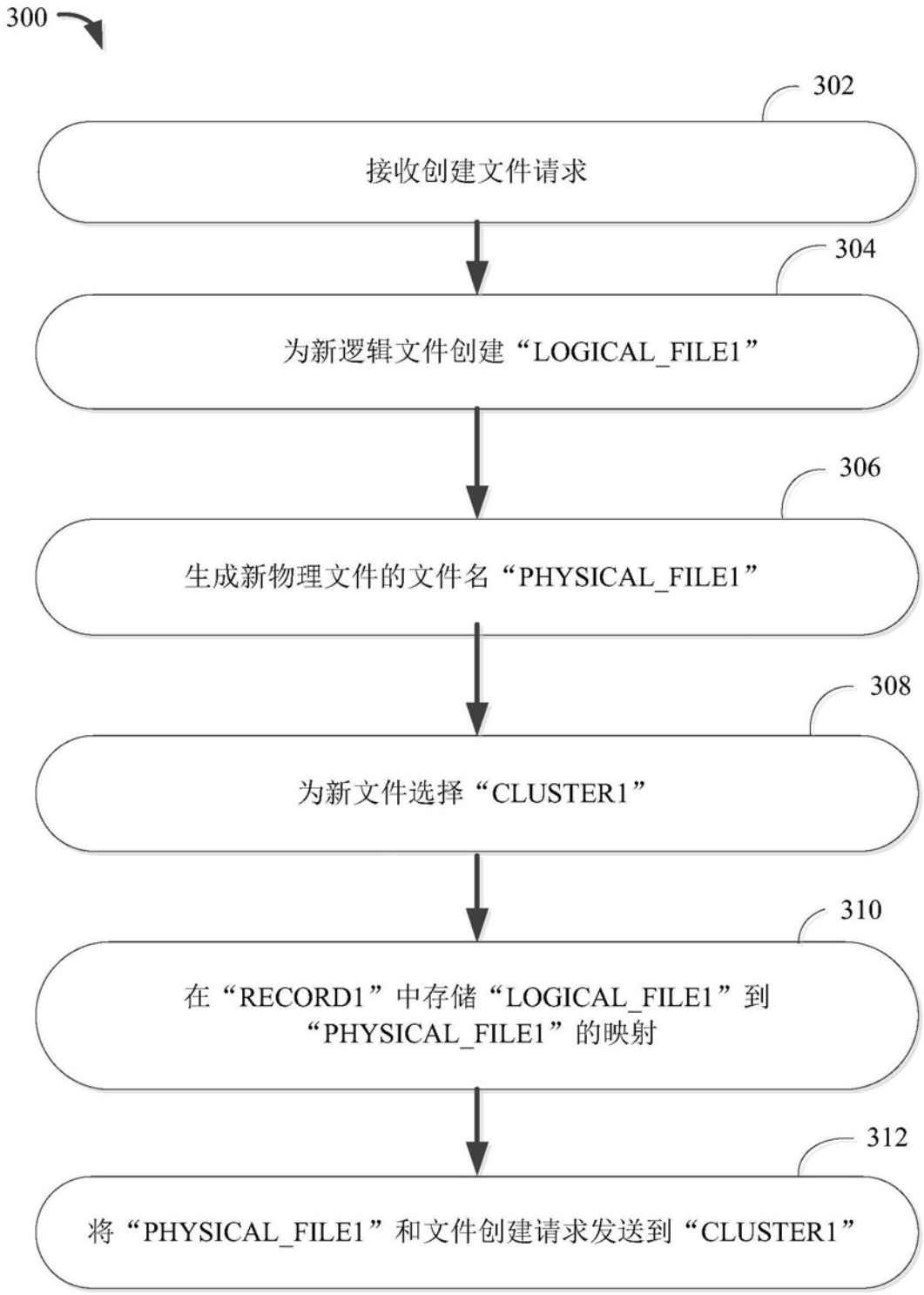

300
302
接收创建文件请求
304
为新逻辑文件创建“LOGICAL_FILE1”
306
生成新物理文件的文件名“PHYSICAL_FILE1”
308
为新文件选择“CLUSTER1”
310
在“RECORD1”中存储“LOGICAL_FILE1”到“PHYSICAL_FILE1”的映射
312
将“PHYSICAL_FILE1”和文件创建请求发送到“CLUSTER1”

图3

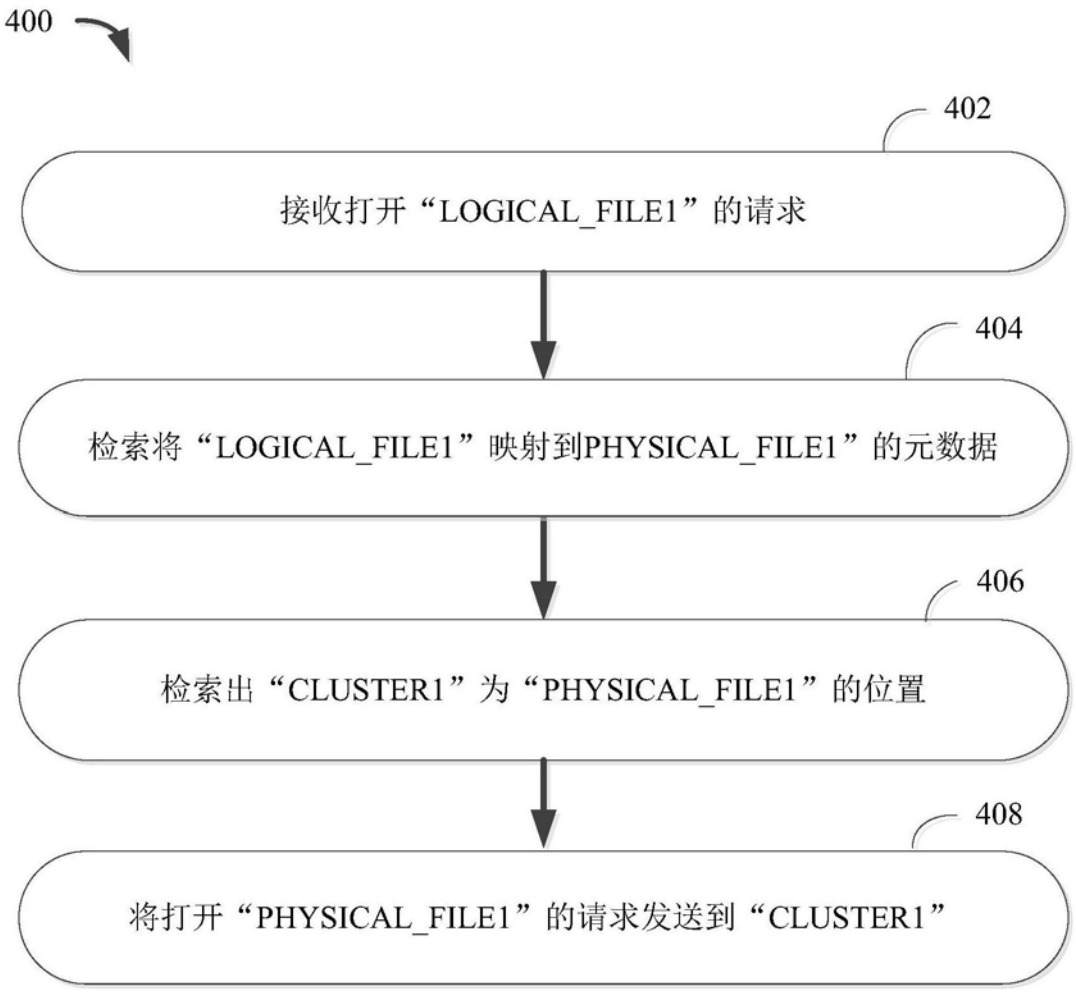

400
402
接收打开“LOGICAL_FILE1”的请求
404
检索将“LOGICAL_FILE1”映射到PHYSICAL_FILE1”的元数据
406
检索出“CLUSTER1”为“PHYSICAL_FILE1”的位置
408
将打开“PHYSICAL_FILE1”的请求发送到“CLUSTER1”

图4

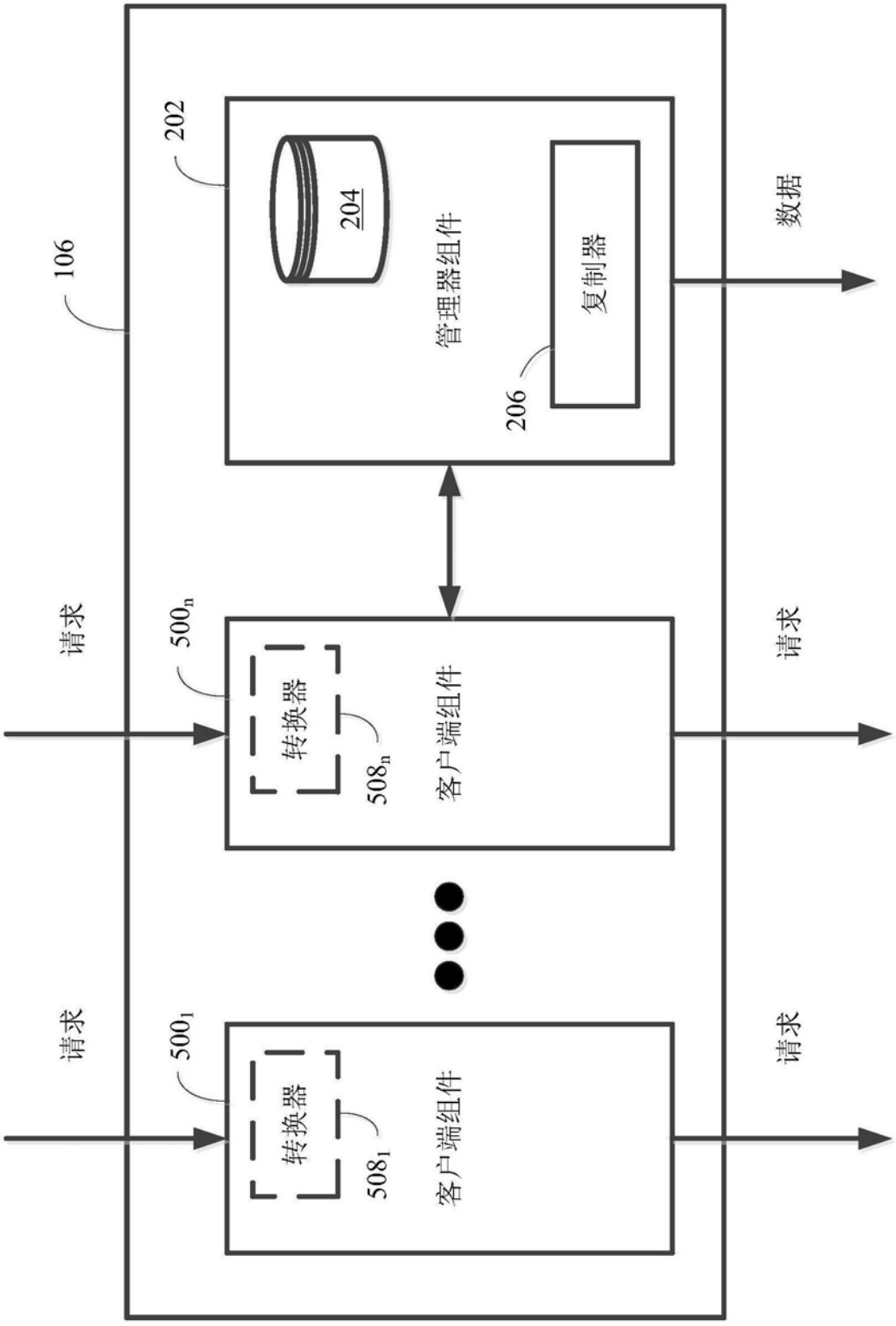

106
202
204
管理器组件
复制器
206
数据
请求
500_n
转换器
508_n
客户端组件
请求
请求
500_1
转换器
508_1
客户端组件
请求

图5

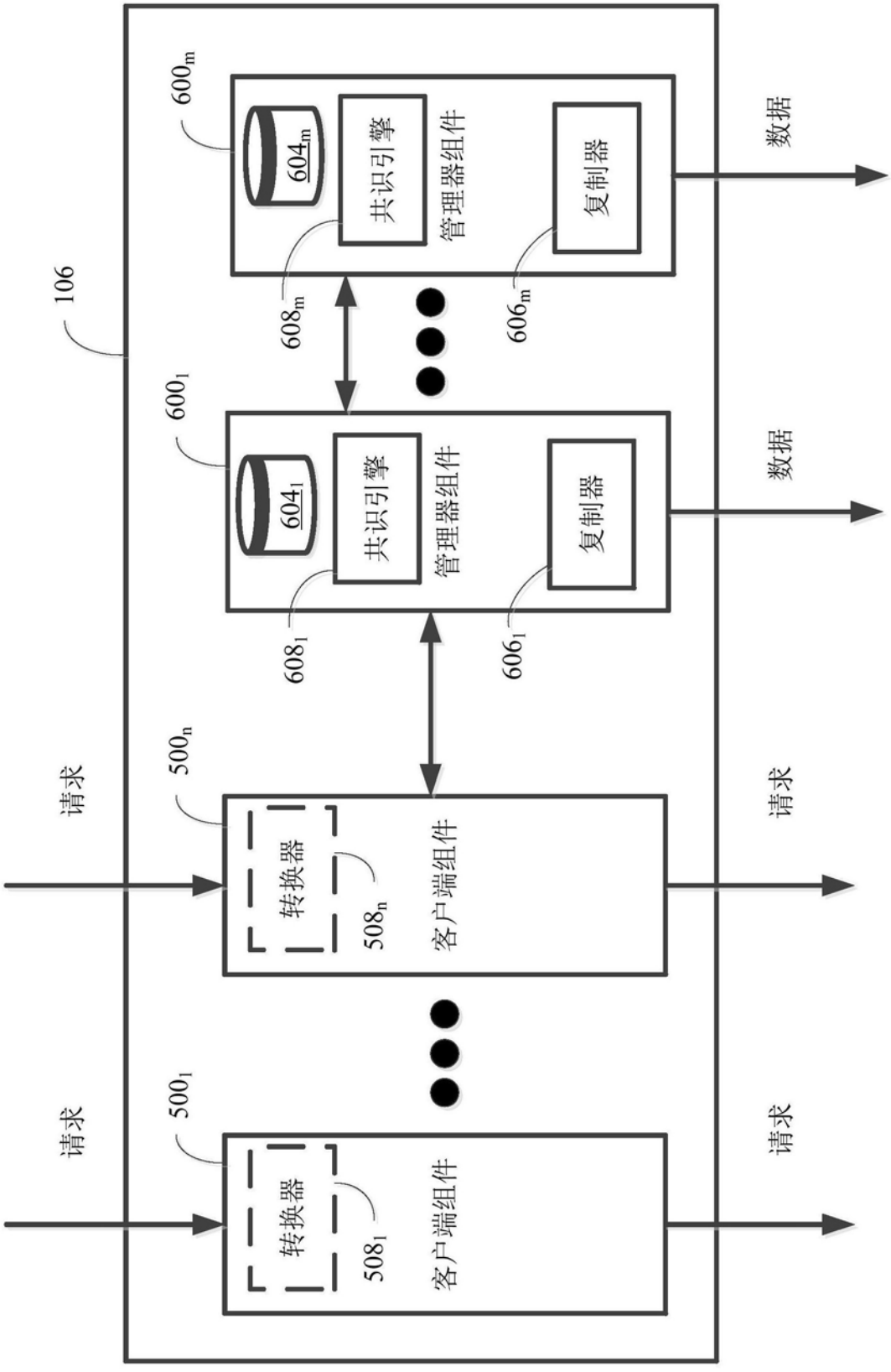

106
600m
604m
共识引擎
管理器组件
复制器
608m
606m
数据
600l
604l
共识引擎
管理器组件
复制器
608l
606l
数据
500n
转换器
508n
客户端组件
请求
请求
500l
转换器
508l
客户端组件
请求
请求

图6

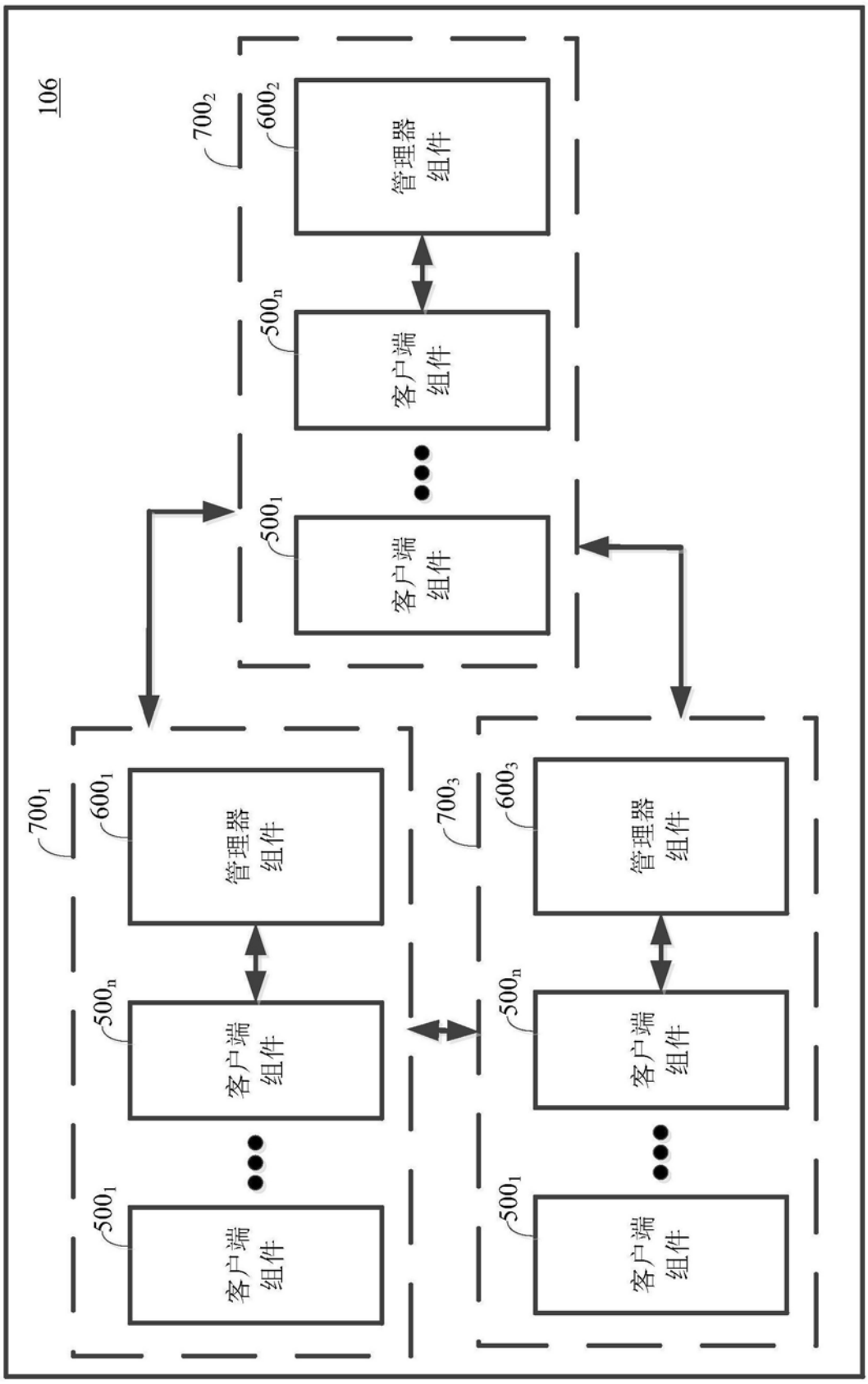

106
700₂
600₂
管理器组件
500ₙ
客户端组件
500₁
客户端组件
700₁
600₁
管理器组件
500ₙ
客户端组件
500₁
客户端组件
700₃
600₃
管理器组件
500ₙ
客户端组件
500₁
客户端组件

图7

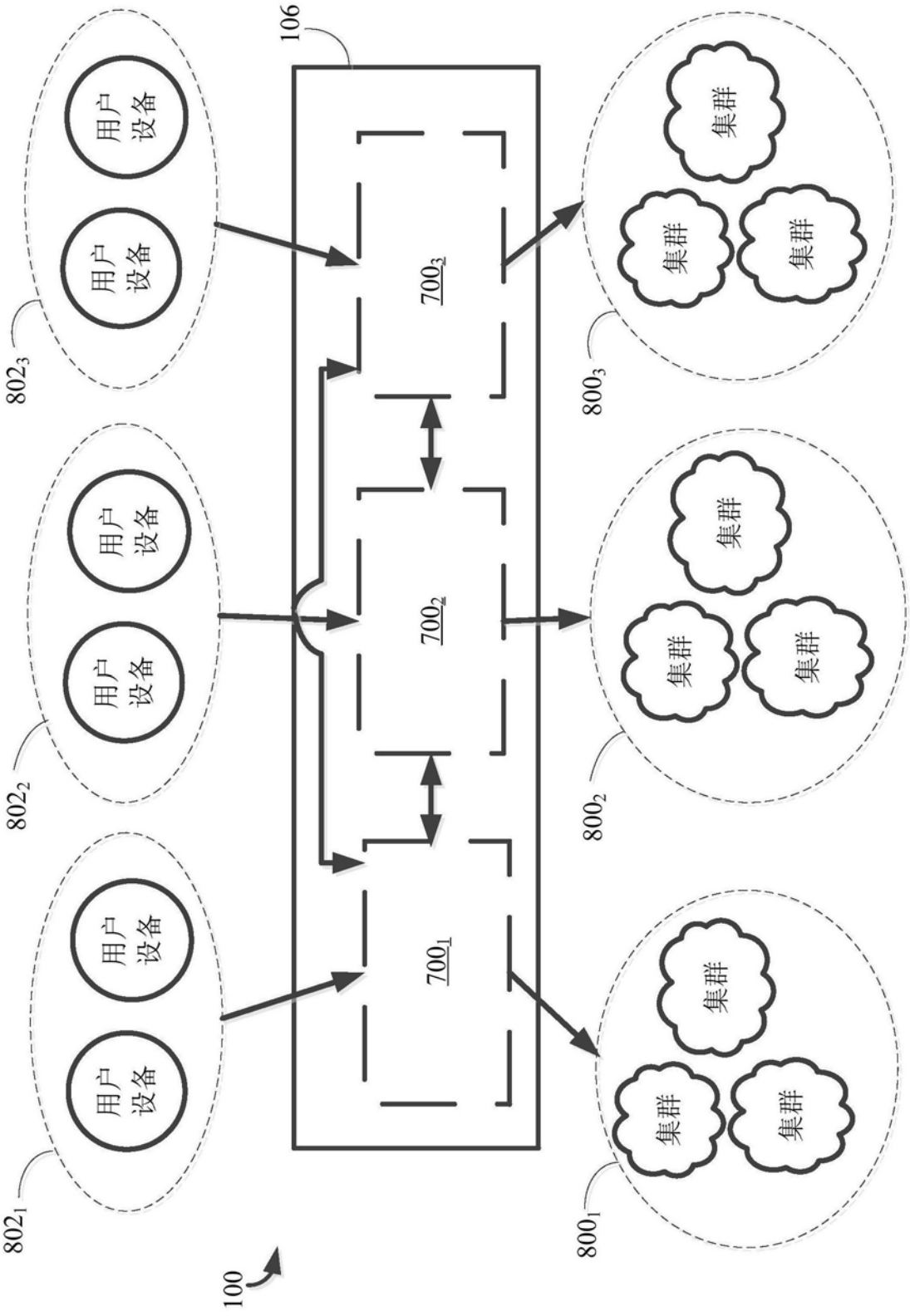


图8

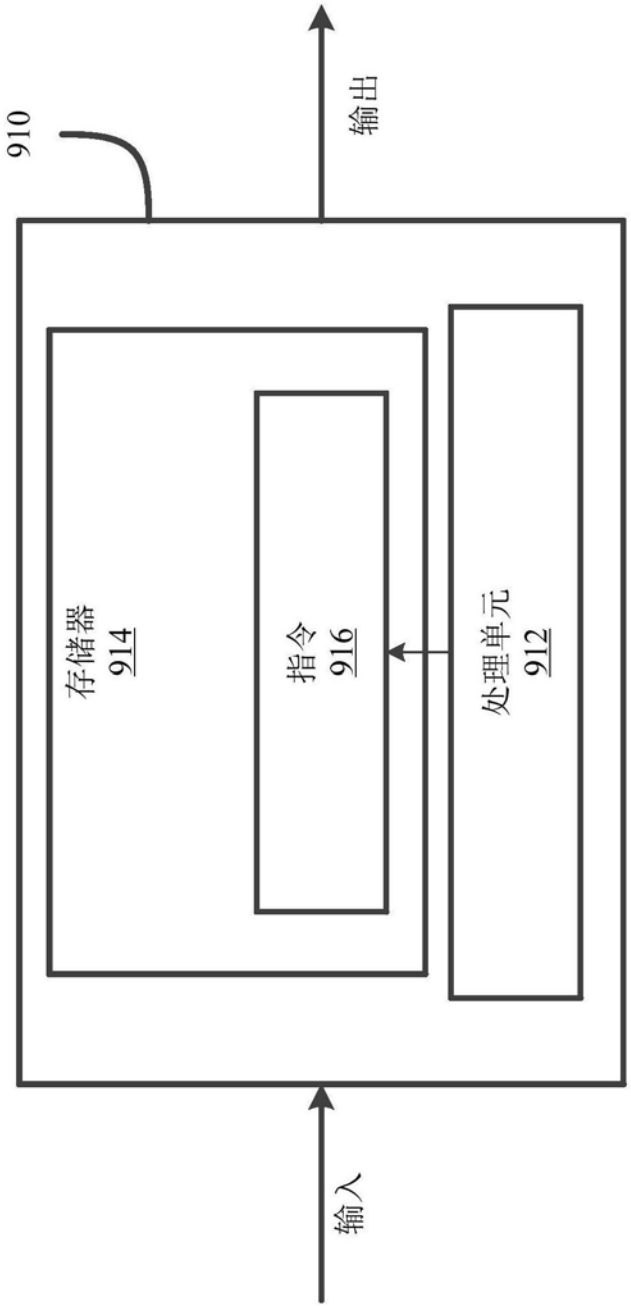


图9A

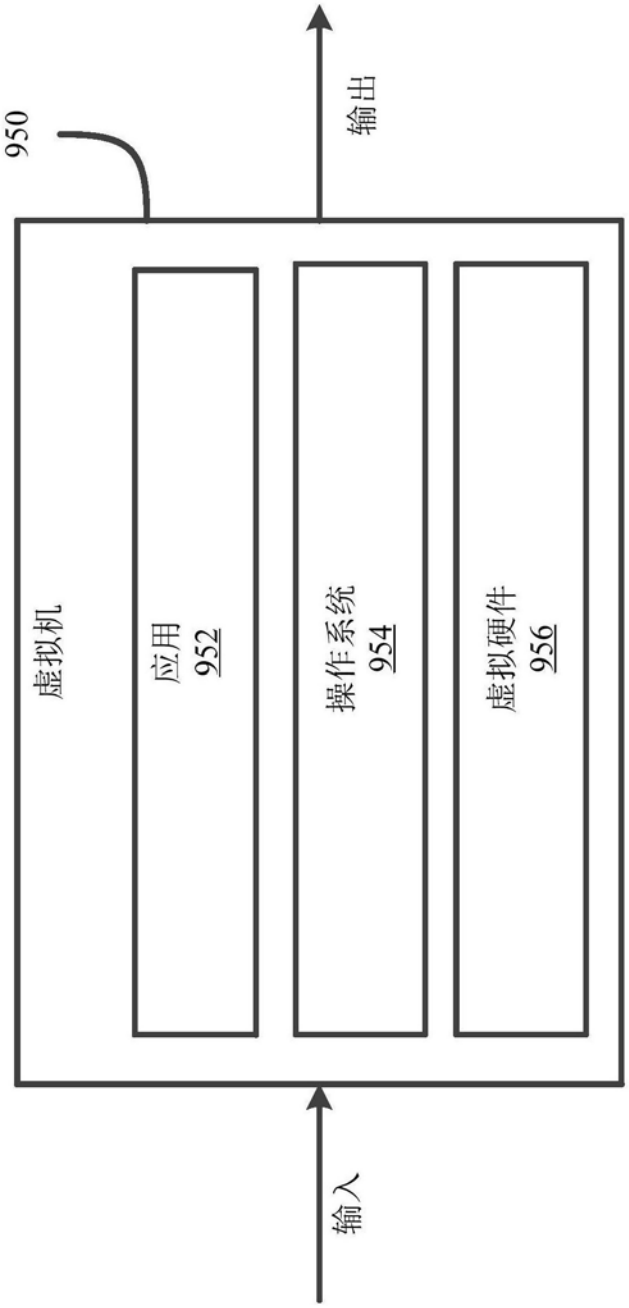


图9B

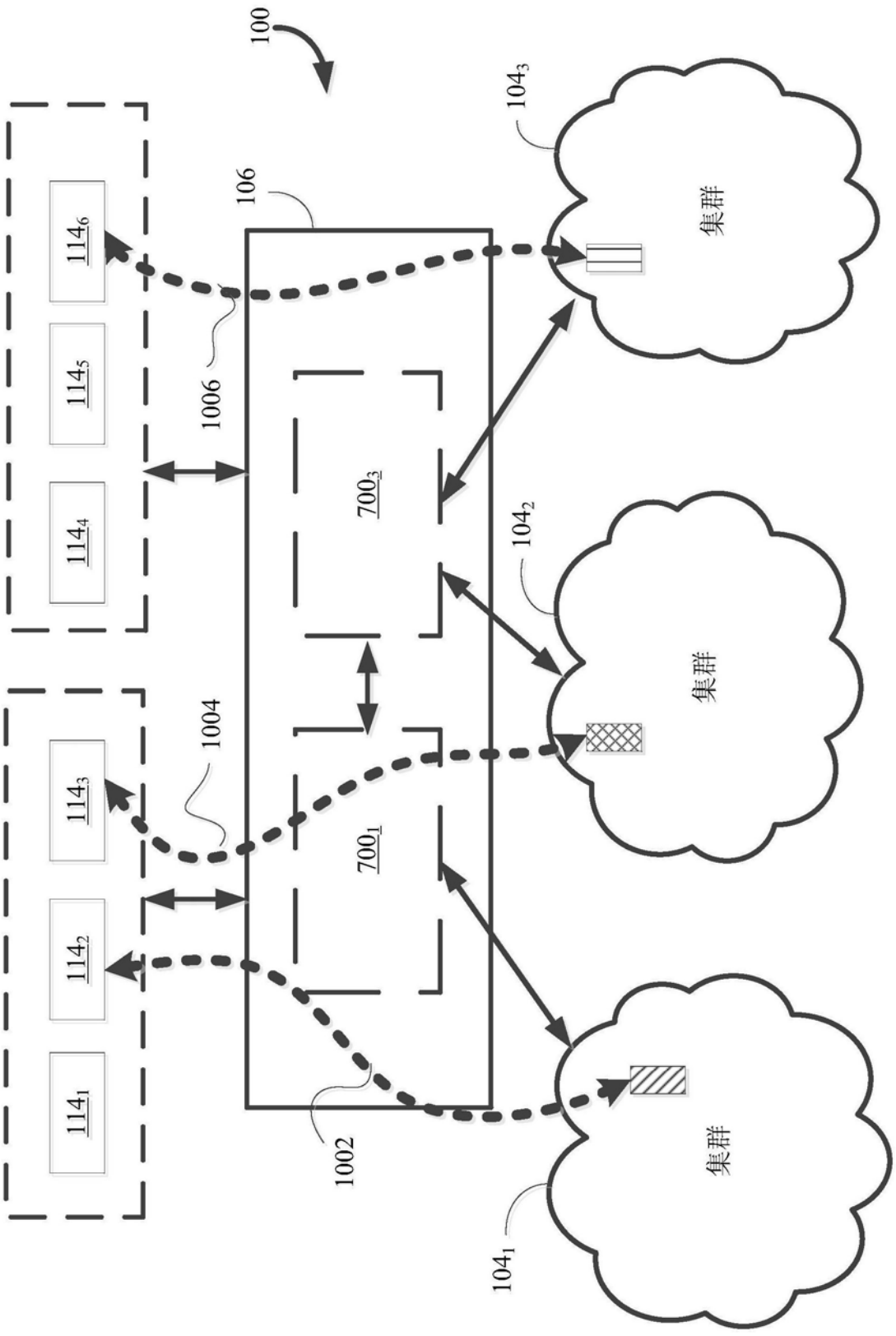

100
106
114_1
114_2
114_3
114_4
114_5
114_6
700_1
700_3
1002
1004
1006
104_1
104_2
104_3
集群
集群
集群

图10A

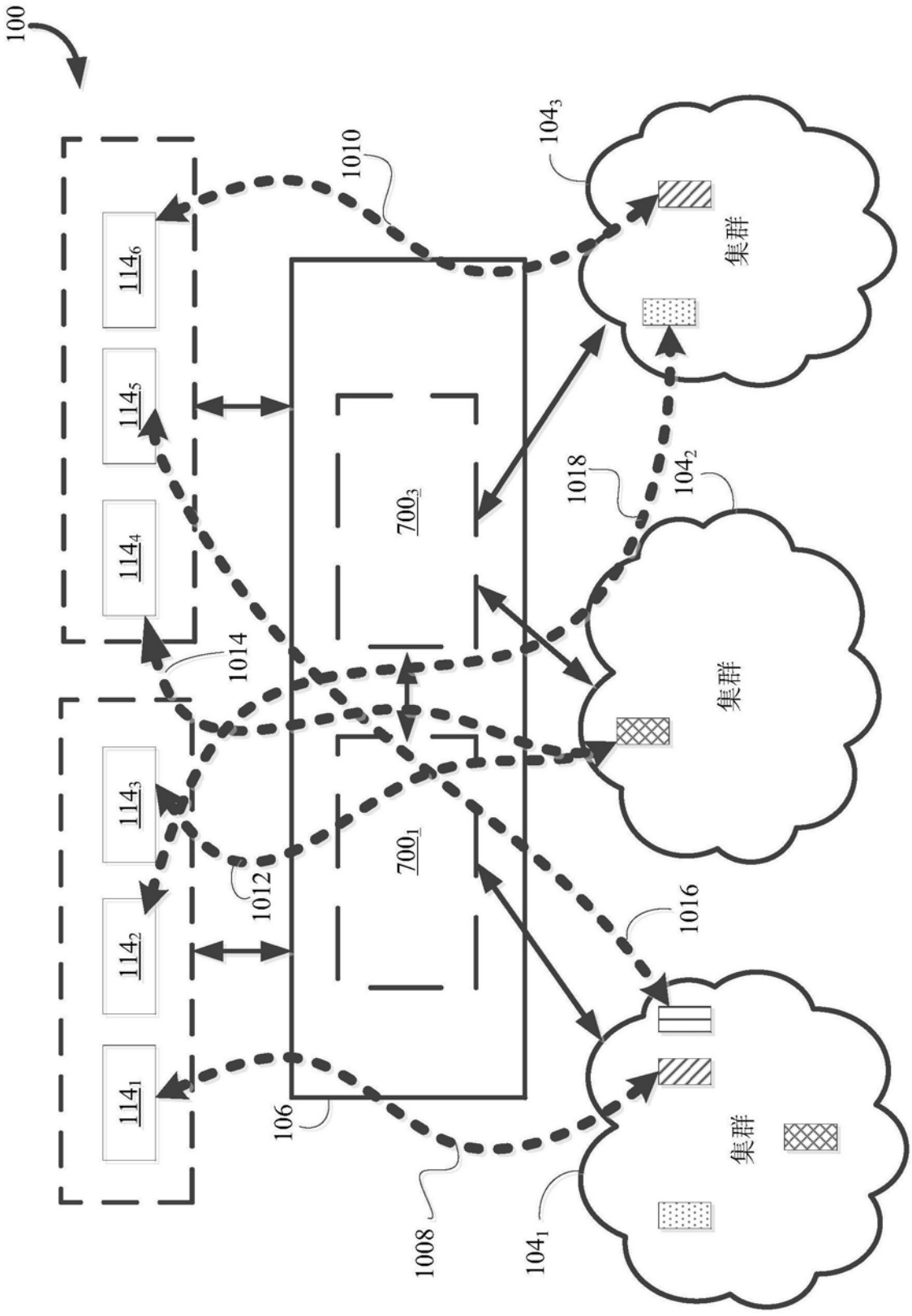

100
1010
104_3
集群
114_6
114_5
114_4
700_3
1018
104_2
1014
集群
114_3
700_1
1012
1016
114_2
114_1
106
集群
1008
104_1

图10B

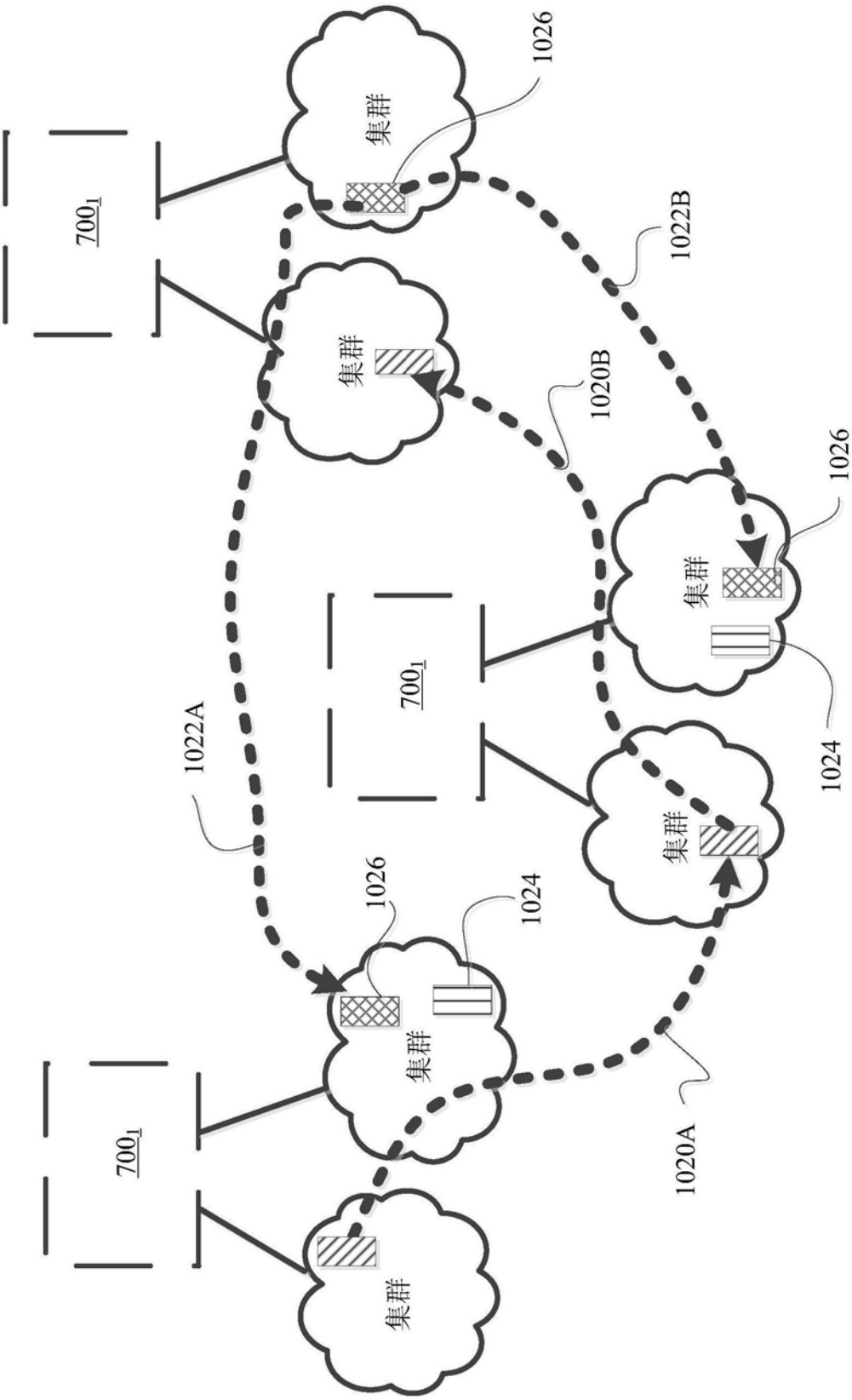


图10C

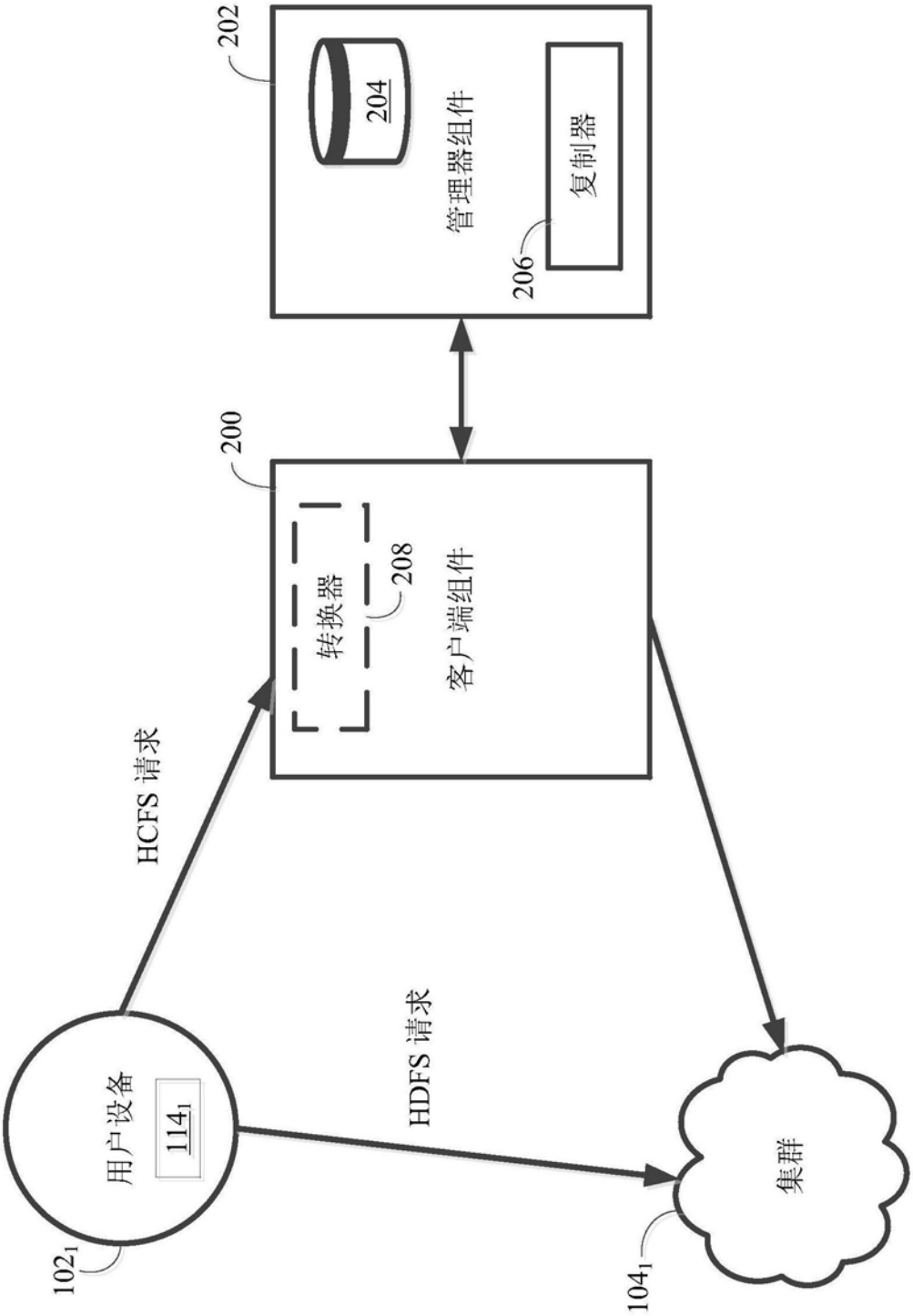

202
204
管理器组件
复制器
206
200
转换器
208
客户端组件
HCFS 请求
HDFS 请求
用户设备
114_1
102_1
集群
104_1

图11A

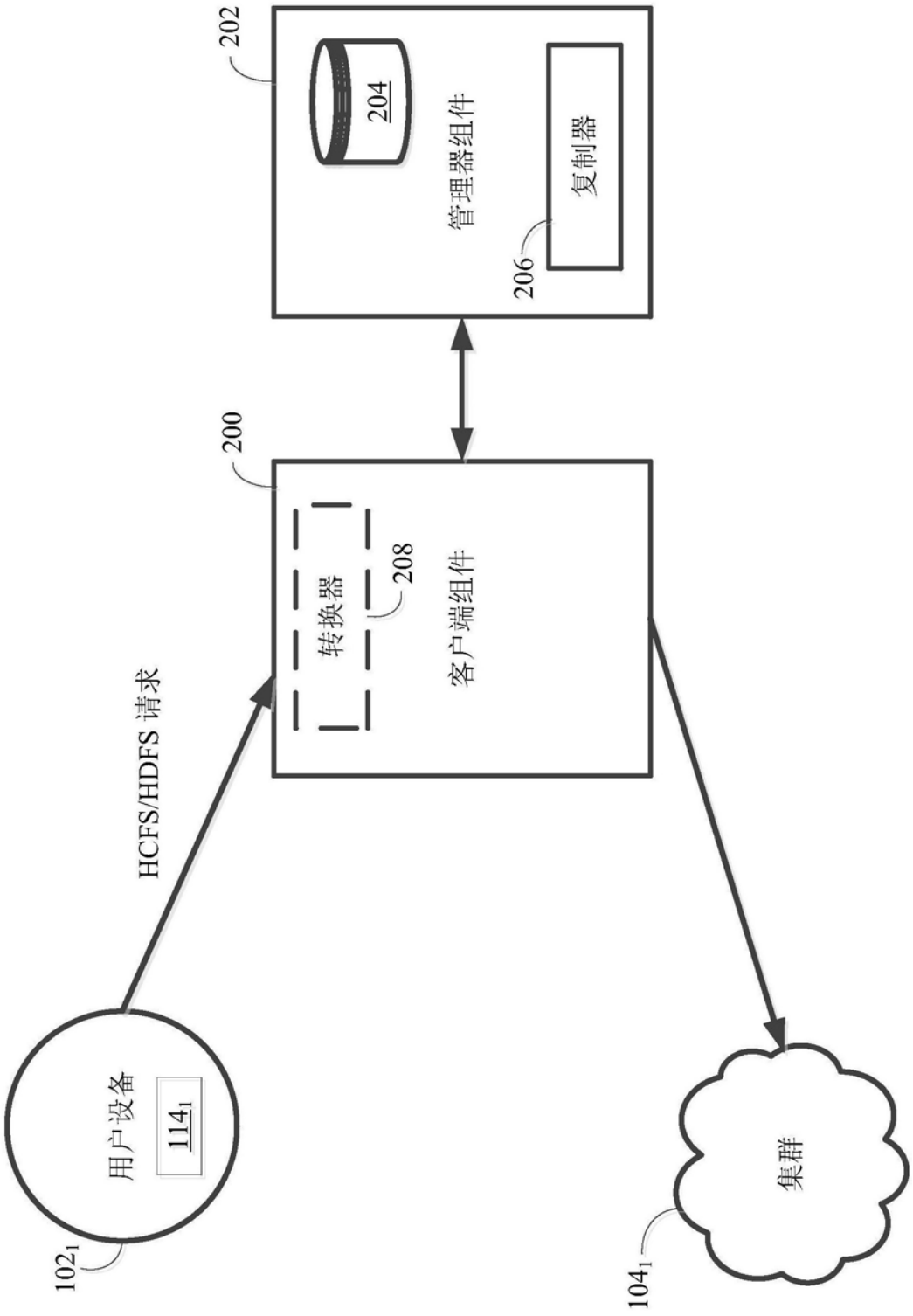

202
204
管理器组件
复制器
206
200
转换器
208
客户端组件
HCFS/HDFS 请求
用户设备
114_1
102_1
集群
104_1

图11B