(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 110100240 B
(45)授权公告日 2023.12.12

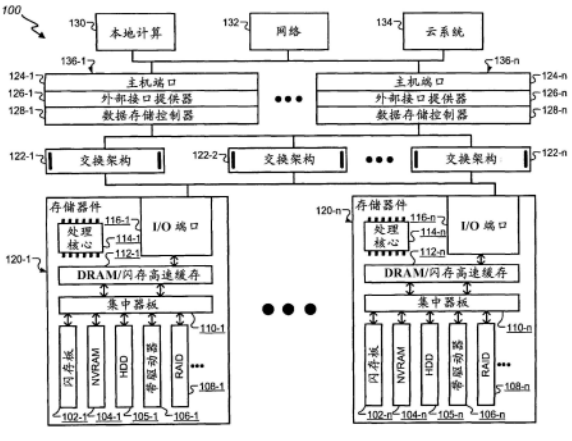
(21)申请号 201780080042.2
(22)申请日 2017.12.20
(65)同一申请的已公布的文献号
申请公布号 CN 110100240 A
(43)申请公布日 2019.08.06
(30)优先权数据
62/443,391 2017.01.06 US
15/610,370 2017.05.31 US
(85)PCT国际申请进入国家阶段日
2019.06.24
(86)PCT国际申请的申请数据
PCT/US2017/067708 2017.12.20
(87)PCT国际申请的公布数据
WO2018/128825 EN 2018.07.12
(73)专利权人 甲骨文国际公司
地址 美国加利福尼亚
(72)发明人 M·梅比 J·克雷默
V·拉图什金 G·吉布森
(74)专利代理机构 中国贸促会专利商标事务所有限公司 11038
专利代理师 周衡威
(51)Int.Cl.
G06F 16/11(2019.01)
G06F 16/182(2019.01)
G06F 3/06(2006.01)
G06F 9/455(2018.01)
G06F 11/14(2006.01)
G06F 12/0868(2016.01)
G06F 12/0897(2016.01)
G06F 12/128(2016.01)
G06F 21/60(2013.01)
H04L 9/06(2006.01)
H04L 12/66(2006.01)
(56)对比文件
EP 2615566 A2,2013.07.17
US 2013110779 A1,2013.05.02
CN 105637487 A,2016.06.01
CN 105260377 A,2016.01.20
CN 102016852 A,2011.04.13
审查员 许强
权利要求书6页 说明书43页 附图28页

(54)发明名称

用于ZFS快照生成和存储的云网关

(57)摘要

本文描述的技术涉及数据存储的系统和方法,更具体而言,涉及在对象接口上提供文件系统功能的分层。在某些实施例中,文件系统功能可以在云对象接口上分层以提供基于云的存储,同时允许从遗留应用预期的功能。例如,POSIX接口和语义可以在基于云的存储上分层,同时以与对名称层次结构中的数据组织的基于文件的访问一致的方式提供对数据的访问。各种实施例还可以提供数据的存储器映射,使得存储器映射改变被反映在持久存储装置中,同时确保存储器映射改变和写入之间的一致性。例如,通过将ZFS文件系统基于盘的存储变换成ZFS基于云的存储,ZFS文件系统获得了云存储的弹性。



CN 110100240 B

1.一种创建远离块存储系统的云对象存储库上的树层次结构的快照的方法,所述云对象存储库是基于对象的系统,所述块存储系统是基于文件的系统,所述方法包括:

在所述块存储系统的应用层处通过所述块存储系统的接口层的系统调用接口接收用于存储或修改文件的第一请求,所述第一请求包括文件数据;

在所述块存储系统的事务对象层处生成多个数据块,所述多个数据块中的每个数据块与所述文件数据的至少一部分对应;

在所述块存储系统的所述事务对象层处生成与所述多个数据块对应的多个元数据块,所述多个元数据块被配置为分层地指向在与所述文件相关联的树层次结构中位于比所述多个元数据块更低级别处的块,从而与用于所述文件的所述树层次结构的至少一部分对应,其中:

所述多个元数据块中的每个元数据块包括一个或多个地址指针,所述一个或多个地址指针中的每个地址指针指向所述多个数据块中的数据块或指向所述多个元数据块中的元数据块;

所述多个数据块中的每个数据块由所述多个元数据块中的至少一个元数据块指向;

所述多个元数据块包括位于用于所述文件的所述树层次结构的顶部处的根块和一个或多个非根元数据块;以及

所述多个元数据块中的每个非根元数据块由用于所述文件的所述树层次结构的所述多个元数据块中的至少一个元数据块指向;

通过将所述多个数据块和所述多个元数据块发送到混合云存储系统,使得一组云存储对象被存储在所述云对象存储库中,所述混合云存储系统管理所述云对象存储库中的数据存储,使得一组云存储对象被存储在所述云对象存储库中包括:

经由云接口设备,将数据块和相应的元数据转化为一组云存储对象;

通过所述云接口设备,将所述一组云存储对象传送到远程的所述云对象存储库;

通过多个虚拟设备中的所述云接口设备,创建第一逻辑树的所述树层次结构中的每个逻辑块到远程的所述云对象存储库中的相应云存储对象的映射;

创建第二逻辑树,所述第二逻辑树是所述树层次结构的基于云的实例化,并包括指向所述云存储对象的地址指针;

向所述混合云存储系统发送对于一组地址的一个或多个第二请求,所述一组地址中的每个地址与所述多个数据块中的数据块或所述多个元数据块中的元数据块对应;

由所述云接口设备将所述一个或多个第二请求转换为对象接口请求,用于执行针对所述树层次结构的存储在所述云对象存储库中的所述基于云的实例化的操作;

在所述块存储系统处,经由所述云接口设备从所述混合云存储系统接收对所述一个或多个第二请求的一个或多个响应,所述一个或多个响应中的每个响应识别与所述多个数据块中的数据块或所述多个元数据块中的元数据块对应的地址,所述地址识别所述云对象存储库中的存储位置;

在所述块存储系统的应用层处通过所述块存储系统的所述接口层的所述系统调用接口接收用于生成所述树层次结构的快照的第三请求;以及

通过以下操作在所述块存储系统的数据和快照层处生成所述树层次结构的所述快照:

存储对活的根块的快照引用,其中所述活的根块是在接收到所述第三请求时开始的时

间段内处于使用中的根块；

将能够从所述快照引用访问的所述多个数据块和所述多个元数据块标记为具有只读能力;以及

将所述树层次结构的所述快照存储在所述映射中。

2.如权利要求1所述的创建远离块存储系统的云对象存储库上的树层次结构的快照的方法,其中通过以下操作从所述快照创建克隆：

在所述块存储系统的应用层处通过所述块存储系统的所述接口层的所述系统调用接口接收用于生成所述克隆的第四请求；

通过以下操作在所述数据和快照层生成克隆：

保存对所述活的根块的克隆引用；

将所述克隆引用链接到所述快照引用,使得在删除所述克隆引用之前无法删除所述快照引用;以及

将能够从所述克隆引用访问的所述多个数据块和所述多个元数据块标记为具有读取和写入能力。

3.如权利要求1所述的创建远离块存储系统的云对象存储库上的树层次结构的快照的方法,其中所述快照引用被存储在自适应替换高速缓存、二级自适应替换高速缓存、或所述云对象存储库中。

4.如权利要求2所述的创建远离块存储系统的云对象存储库上的树层次结构的快照的方法,其中所述克隆引用被存储在自适应替换高速缓存、二级自适应替换高速缓存、或所述云对象存储库中。

5.如权利要求1所述的创建远离块存储系统的云对象存储库上的树层次结构的快照的方法,其中所述快照被周期性地自动生成。

6.如权利要求1所述的创建远离块存储系统的云对象存储库上的树层次结构的快照的方法,其中所述快照被周期性地删除。

7.如权利要求6所述的创建远离块存储系统的云对象存储库上的树层次结构的快照的方法,其中能够从所述快照引用访问的所述多个数据块和所述多个元数据块被标记为已释放。

8.一个或多个非瞬态有形计算机可读存储介质,存储用于执行计算机处理的计算机可执行指令,所述计算机处理用于创建远离计算系统上的块存储系统的云对象存储库上的树层次结构的快照,所述云对象存储库是基于对象的系统,所述块存储系统是基于文件的系统,所述计算机处理包括：

在所述块存储系统的应用层处通过所述块存储系统的接口层的系统调用接口接收用于存储或修改文件的第一请求,所述第一请求包括文件数据；

在所述块存储系统的事务对象层处生成多个数据块,所述多个数据块中的每个数据块与所述文件数据的至少一部分对应；

在所述块存储系统的所述事务对象层处生成与所述多个数据块对应的多个元数据块,所述多个元数据块被配置为分层地指向在与所述文件相关联的树层次结构中位于比所述多个元数据块更低级别处的块,从而与用于所述文件的所述树层次结构的至少一部分对应,其中：

所述多个元数据块中的每个元数据块包括一个或多个地址指针，所述一个或多个地址指针中的每个地址指针指向所述多个数据块中的数据块或指向所述多个元数据块中的元数据块；

所述多个数据块中的每个数据块由所述多个元数据块中的至少一个元数据块指向；

所述多个元数据块包括位于用于所述文件的所述树层次结构的顶部处的根块和一个或多个非根元数据块；以及

所述多个元数据块中的每个非根元数据块由用于所述文件的所述树层次结构的所述多个元数据块中的至少一个元数据块指向；

通过将所述多个数据块和所述多个元数据块发送到混合云存储系统，使得一组云存储对象被存储在所述云对象存储库中，所述混合云存储系统管理所述云对象存储库中的数据存储，使得一组云存储对象被存储在所述云对象存储库中包括：

经由云接口设备，将数据块和相应的元数据转化为一组云存储对象；

通过所述云接口设备，将所述一组云存储对象传送到远程的所述云对象存储库；

通过多个虚拟设备中的所述云接口设备，创建第一逻辑树的所述树层次结构中的每个逻辑块到远程的所述云对象存储库中的相应云存储对象的映射；

创建第二逻辑树，所述第二逻辑树是所述树层次结构的基于云的实例化，并包括指向所述云存储对象的地址指针；

向所述混合云存储系统发送对于一组地址的一个或多个第二请求，所述一组地址中的每个地址与所述多个数据块中的数据块或所述多个元数据块中的元数据块对应；

由所述云接口设备将所述一个或多个第二请求转换为对象接口请求，用于执行针对所述树层次结构的存储在所述云对象存储库中的所述基于云的实例化的操作；

在所述块存储系统处，经由所述云接口设备从所述混合云存储系统接收对所述一个或多个第二请求的一个或多个响应，所述一个或多个响应中的每个响应识别与所述多个数据块中的数据块或所述多个元数据块中的元数据块对应的地址，所述地址识别所述云对象存储库中的存储位置；

在所述块存储系统的应用层处通过所述块存储系统的所述接口层的所述系统调用接口接收用于生成所述树层次结构的快照的第三请求；以及

通过以下操作在所述块存储系统的数据和快照层处生成所述树层次结构的所述快照：

存储对活的根块的快照引用，其中所述活的根块是在接收到所述第三请求时开始的时间段内处于使用中的根块；

将能够从所述快照引用访问的所述多个数据块和所述多个元数据块标记为具有只读能力；以及

将所述树层次结构的所述快照存储在所述映射中。

9.如权利要求8所述的一个或多个非瞬态有形计算机可读存储介质，存储用于执行计算机处理的计算机可执行指令，所述计算机处理用于创建远离计算系统上的块存储系统的云对象存储库上的树层次结构的快照，其中通过以下操作从所述快照创建克隆：

在所述块存储系统的应用层处通过所述块存储系统的所述接口层的所述系统调用接口接收用于生成所述克隆的第四请求；

通过以下操作在所述数据和快照层生成克隆：

保存对所述活的根块的克隆引用；

将所述克隆引用链接到所述快照引用，使得在删除所述克隆引用之前无法删除所述快照引用；以及

将能够从所述克隆引用访问的所述多个数据块和所述多个元数据块标记为具有读取和写入能力。

10.如权利要求8所述的一个或多个非瞬态有形计算机可读存储介质，存储用于执行计算机处理的计算机可执行指令，所述计算机处理用于创建远离计算系统上的块存储系统的云对象存储库上的树层次结构的快照，其中所述快照引用被存储在自适应替换高速缓存、二级自适应替换高速缓存、或所述云对象存储库中。

11.如权利要求9所述的一个或多个非瞬态有形计算机可读存储介质，存储用于执行计算机处理的计算机可执行指令，所述计算机处理用于创建远离计算系统上的块存储系统的云对象存储库上的树层次结构的快照，其中所述克隆引用被存储在自适应替换高速缓存、二级自适应替换高速缓存、或所述云对象存储库中。

12.如权利要求8所述的一个或多个非瞬态有形计算机可读存储介质，存储用于执行计算机处理的计算机可执行指令，所述计算机处理用于创建远离计算系统上的块存储系统的云对象存储库上的树层次结构的快照，其中所述快照被周期性地自动生成。

13.如权利要求8所述的一个或多个非瞬态有形计算机可读存储介质，存储用于执行计算机处理的计算机可执行指令，所述计算机处理用于创建远离计算系统上的块存储系统的云对象存储库上的树层次结构的快照，其中所述快照被周期性地删除。

14.如权利要求13所述的一个或多个非瞬态有形计算机可读存储介质，存储用于执行计算机处理的计算机可执行指令，所述计算机处理用于创建远离计算系统上的块存储系统的云对象存储库上的树层次结构的快照，其中能够从所述快照引用访问的所述多个数据块和所述多个元数据块被标记为已释放。

15.一种用于创建远离块存储系统的云对象存储库上的树层次结构的快照的基于处理器的系统，所述云对象存储库是基于对象的系统，所述块存储系统是基于文件的系统，所述基于处理器的系统执行包括以下各项的操作：

在所述块存储系统的应用层处通过所述块存储系统的接口层的系统调用接口接收用于存储或修改文件的第一请求，所述第一请求包括文件数据；

在所述块存储系统的事务对象层处生成多个数据块，所述多个数据块中的每个数据块与所述文件数据的至少一部分对应；

在所述块存储系统的所述事务对象层处生成与所述多个数据块对应的多个元数据块，所述多个元数据块被配置为分层地指向在与所述文件相关联的树层次结构中位于比所述多个元数据块更低级别处的块，从而与用于所述文件的所述树层次结构的至少一部分对应，其中：

所述多个元数据块中的每个元数据块包括一个或多个地址指针，所述一个或多个地址指针中的每个地址指针指向所述多个数据块中的数据块或指向所述多个元数据块中的元数据块；

所述多个数据块中的每个数据块由所述多个元数据块中的至少一个元数据块指向；

所述多个元数据块包括位于用于所述文件的所述树层次结构的顶部处的根块和一个

或多个非根元数据块;以及

所述多个元数据块中的每个非根元数据块由用于所述文件的所述树层次结构的所述多个元数据块中的至少一个元数据块指向;

通过将所述多个数据块和所述多个元数据块发送到混合云存储系统,使得一组云存储对象被存储在所述云对象存储库中,所述混合云存储系统管理所述云对象存储库中的数据存储,使得一组云存储对象被存储在所述云对象存储库中包括:

经由云接口设备,将数据块和相应的元数据转化为一组云存储对象;

通过所述云接口设备,将所述一组云存储对象传送到远程的所述云对象存储库;

通过多个虚拟设备中的所述云接口设备,创建第一逻辑树的所述树层次结构中的每个逻辑块到远程的所述云对象存储库中的相应云存储对象的映射;

创建第二逻辑树,所述第二逻辑树是所述树层次结构的基于云的实例化,并包括指向所述云存储对象的地址指针;

向所述混合云存储系统发送对于一组地址的一个或多个第二请求,所述一组地址中的每个地址与所述多个数据块中的数据块或所述多个元数据块中的元数据块对应;

由所述云接口设备将所述一个或多个第二请求转换为对象接口请求,用于执行针对所述树层次结构的存储在所述云对象存储库中的所述基于云的实例化的操作;

在所述块存储系统处,经由所述云接口设备从所述混合云存储系统接收对所述一个或多个第二请求的一个或多个响应,所述一个或多个响应中的每个响应识别与所述多个数据块中的数据块或所述多个元数据块中的元数据块对应的地址,所述地址识别所述云对象存储库中的存储位置;

在所述块存储系统的应用层处通过所述块存储系统的所述接口层的所述系统调用接口接收用于生成所述树层次结构的快照的第三请求;以及

通过以下操作在所述块存储系统的数据和快照层处生成所述树层次结构的所述快照:

存储对活的根块的快照引用,其中所述活的根块是在接收到所述第三请求时开始的时间段内处于使用中的根块;

将能够从所述快照引用访问的所述多个数据块和所述多个元数据块标记为具有只读能力;以及

将所述树层次结构的所述快照存储在所述映射中。

16.如权利要求15所述的用于创建远离块存储系统的云对象存储库上的树层次结构的快照的基于处理器的系统,其中通过以下操作从所述快照创建克隆:

在所述块存储系统的应用层处通过所述块存储系统的所述接口层的所述系统调用接口接收用于生成所述克隆的第四请求;

通过以下操作在所述数据和快照层生成克隆:

保存对所述活的根块的克隆引用;

将所述克隆引用链接到所述快照引用,使得在删除所述克隆引用之前无法删除所述快照引用;以及

将能够从所述克隆引用访问的所述多个数据块和所述多个元数据块标记为具有读取和写入能力。

17.如权利要求15所述的用于创建远离块存储系统的云对象存储库上的树层次结构的

快照的基于处理器的系统，其中所述快照引用被存储在自适应替换高速缓存、二级自适应替换高速缓存、或所述云对象存储库中。

18.如权利要求16所述的用于创建远离块存储系统的云对象存储库上的树层次结构的快照的基于处理器的系统，其中所述克隆引用被存储在自适应替换高速缓存、二级自适应替换高速缓存、或所述云对象存储库中。

19.如权利要求15所述的用于创建远离块存储系统的云对象存储库上的树层次结构的快照的基于处理器的系统，其中所述快照被周期性地自动生成。

20.如权利要求15所述的用于创建远离块存储系统的云对象存储库上的树层次结构的快照的基于处理器的系统，其中所述快照被周期性地删除。

21.如权利要求20所述的用于创建远离块存储系统的云对象存储库上的树层次结构的快照的基于处理器的系统，其中能够从所述快照引用访问的所述多个数据块和所述多个元数据块被标记为已释放。

22.一种包括用于执行如权利要求1-7中任一项所述的方法的部件的装置。

用于ZFS快照生成和存储的云网关

[0001] 对相关申请的交叉引用

[0002] 本申请是于2017年1月6日提交的标题为“FILE SYSTEM HIERARCHIES AND FUNCTIONALITY WITH CLOUD OBJECT STORAGE”的美国临时申请No.62/443,391的非临时申请,并援引35U.S.C.119(e)要求其权益和优先权,该临时申请的全部内容通过引用并入本文,用于所有目的。

技术领域

[0003] 本公开一般而言涉及数据存储的系统和方法,并且更具体地涉及在对象接口上将文件系统功能分层。

背景技术

[0004] 互联网的不断扩展,以及计算网络和系统的扩展和复杂化,已经导致通过互联网存储和可访问的内容的激增。这进而推动了对大型复杂数据存储系统的需求。随着对数据存储的需求不断增加,正在设计和部署更大和更复杂的存储系统。许多大规模数据存储系统利用包括物理存储介质阵列的存储器件。这些存储器件能够存储大量数据。例如,在这个时候,Oracle的SUN ZFS Storage ZS5-4装备可以存储高达6.9PB的数据。而且,多个存储装备可以联网在一起以形成存储池,这可以进一步增加存储的数据的体量。

[0005] 通常,诸如这些之类的大型存储系统可以包括用于存储和访问文件的文件系统。除了存储系统文件(操作系统文件、设备驱动器文件等)之外,文件系统还提供用户数据文件的存储和访问。如果这些文件中的任何一个(系统文件和/或用户文件)包含关键数据,那么采用备份存储方案以确保在文件存储设备发生故障时不会丢失该关键数据是有利的。

[0006] 常规的基于云的存储是基于对象的,并提供弹性和规模。然而,云对象存储存在许多问题。云对象存储提供基于取出(get)和放置(put)整个对象的接口。云对象存储提供有限的搜索能力,并且通常具有高延时。有限的基于云的接口不符合本地文件系统应用的需求。将遗留应用转换成使用对象接口将是昂贵的并且可能不实际或甚至不可能。云对象存储加密使制作加密密钥的数据更加脆弱且更不安全。

[0007] 因此,需要解决上述问题的系统和方法,以便在对象接口上提供文件系统功能的分层。本公开解决了这一需求和其它需求。

发明内容

[0008] 本公开的某些实施例一般而言涉及数据存储的系统和方法,并且更具体地涉及用于在对象接口上将文件系统功能分层的系统和方法。

[0009] 本文描述了各种技术(例如,系统、方法、在非瞬态机器可读存储介质中有形地实施的计算机程序产品等),用于在对象接口上提供文件系统功能的分层。在某些实施例中,文件系统功能可以在云对象接口上分层以提供基于云的存储,同时允许遗留应用预期的功能。例如,可移植操作系统接口(POSIX)接口和语义可以在基于云的存储上分层,同时以与

对名称层次结构(name hierarchy)中的数据组织的基于文件的访问一致的方式提供对数据的访问。各种实施例还可以提供数据的存储器映射,使得存储器映射改变被反映在持久存储装置中,同时确保存储器映射改变和写入之间的一致性。例如,通过将ZFS文件系统基于盘的存储转换成ZFS基于云的存储,ZFS文件系统获得了云存储的弹性。

[0010] 根据下文提供的详细描述,本公开的其它应用领域将变得清楚。应当理解的是,详细描述和具体示例在指示各种实施例的时候仅旨在说明的目的,而不旨在必然限制本公开的范围。

附图说明

[0011] 通过结合以下附图参考说明书的其余部分,可以实现对根据本公开的实施例的本质和优点的进一步理解。

[0012] 图1图示了可以根据本公开的某些实施例使用的一个示例存储网络。

[0013] 图2图示了根据本公开某些实施例的可以在存储环境中执行的文件系统的实例。

[0014] 图3A-图3F图示了根据本公开某些实施例的用于文件系统的写时复制(copy-on-write)处理。

[0015] 图4是图示根据本公开某些实施例的混合云存储系统的示例的高级图。

[0016] 图5图示了根据本公开某些实施例的混合云存储系统的示例网络文件系统的实例。

[0017] 图6是图示根据本公开某些实施例的混合云存储系统的云接口装备的附加方面的图。

[0018] 图7A-图7F是图示根据本公开某些实施例的示例方法的框图,该示例方法针对用于混合云存储系统的COW处理的某些特征,包括数据服务、快照和克隆。

[0019] 图8是图示根据本公开某些实施例的处理增量修改的云接口装备的示例的高级别图。

[0020] 图9是图示根据本公开某些实施例的示例方法的框图,该示例方法针对混合云存储系统的某些特征,这些特征确保云中的完整性以及根据最终一致的对象模型(eventually consistent object model)的始终一致的语义(always-consistent semantics)。

[0021] 图10是图示根据本公开某些实施例的处理校验的云接口装备的示例的高级别图。

[0022] 图11是根据本公开某些实施例的进一步图示混合云存储系统的特征的简化示例的图。

[0023] 图12是根据本公开某些实施例的进一步图示混合云存储系统的特征的简化示例的图。

[0024] 图13是图示根据本公开某些实施例的示例方法的框图,该示例方法针对混合云存储系统的用于高速缓存管理和云延时掩盖(cloud latency masking)的某些特征。

[0025] 图14图示了根据本公开某些实施例的促进同步镜像的混合云存储系统的示例网络文件系统的实例。

[0026] 图15是图示根据本公开某些实施例的示例方法的框图,该示例方法针对混合云存储系统的用于同步镜像和云延时掩盖的某些特征。

[0027] 图16描绘了用于实现根据本公开某些实施例的分布式系统的简化图。

[0028] 图17是根据本公开某些实施例的系统环境的一个或多个部件的简化框图，由系统的一个或多个部件提供的服务可以通过该系统环境作为云服务提供。

[0029] 图18图示了示例性计算机系统，在该示例性计算机系统中可以实现本发明的各种实施例。

[0030] 在附图中，类似的部件和/或特征可以具有相同的参考标签。另外，相同类型的各种部件可以通过在参考标签之后跟随短划线和区分相似部件的第二标签来区分。如果在说明书中仅使用第一参考标签，那么该描述适用于具有相同第一参考标签的任何一个类似部件，而与第二参考标签无关。

具体实施方式

[0031] 随后的描述仅提供优选的(一个或多个)示例性实施例，并且不旨在限制本公开的范围、适用性或配置。而是，随后对优选的(一个或多个)示例性实施例的描述将为本领域技术人员提供用于实现本公开的优选示例性实施例的实现性描述。应该理解的是，在不脱离如所附权利要求中阐述的本公开的精神和范围的情况下，可以对元素的功能和布置进行各种改变。

[0032] 如上所述，基于云的存储提供弹性和规模，但存在许多问题。云对象存储提供基于取出和放置整个对象的接口。云对象存储提供有限的搜索能力，并且通常具有高延时。有限的基于云的接口不符合本地文件系统应用的需求。将遗留应用转换成使用对象接口将是昂贵的并且可能不实际或甚至不可能。因此，需要解决方案，使得为了直接访问云对象存储装置不必对文件系统应用进行改变，因为它肯定复杂且昂贵。

[0033] 解决方案应当允许保留本地应用接口，而不引入各种类型的适配层以将本地存储系统的数据映射到云中的对象存储装置。因而，根据本公开某些实施例可以在云对象接口上对文件系统功能进行分层以提供基于云的存储，同时允许遗留应用预期的功能。例如，非基于云的遗留应用可以访问主要作为文件的数据，并且可以针对POSIX接口和语义进行配置。从遗留应用的角度来看，预期能够在不重写文件的情况下修改文件的内容。同样，预期能够在名称层次结构中组织数据。

[0034] 为了适应这种预期，某些实施例可以在基于云的存储装置上对POSIX接口和语义进行分层，同时从用户的角度来看以与对名称层次结构中的数据组织的基于文件的访问一致的方式来提供对数据的访问。另外，某些实施例可以提供数据的存储器映射，使得存储器映射改变被反映在持久存储装置中，同时确保存储器映射改变和写入之间的一致性。通过将ZFS文件系统基于盘的存储转换成基于ZFS云的存储，ZFS文件系统获得了云存储的弹性。通过将“盘块”映射到云对象，对ZFS文件系统的存储要求只是实际使用中的“块”。系统可以始终是瘦配置(thinly provisioned)的，没有后备存储被耗尽的风险。相反，云存储获得ZFS文件系统语义和服务。可以向云客户端提供完整的POSIX语义，以及由ZFS文件系统提供的任何附加数据服务(诸如压缩、加密、快照等)。

[0035] 某些实施例可以提供将数据迁移到云和从云迁移的能力，并且可以通过混合云存储系统的方式提供本地数据与云中的数据共存，混合云存储系统提供存储弹性和规模同时在云存储上将ZFS文件系统功能分层。通过扩展ZFS文件系统以允许在云对象存储库中存储

对象，可以向传统对象存储提供桥接，同时除了保留所有ZFS文件系统数据服务之外还保留ZFS文件系统功能。对传统本地文件系统与在各种云对象存储库中存储数据的能力之间的差距(gap)进行桥接有助于显著改善性能。

[0036] 此外，本发明的实施例使得能够结合混合云存储来使用传统的ZFS数据服务。作为示例，压缩、加密、去重(deduplication)、快照和克隆各自在本发明的某些实施例中可用，并且紧接着在下面简要描述。在本发明中，当将存储扩展到云时，用户可以继续无缝地使用由ZFS文件系统提供的所有数据服务。例如，Oracle密钥管理器或等同物对在用户本地的密钥进行管理，从而允许在存储到云时使用本地管理的密钥进行端到端安全加密。用于在盘存储上压缩、加密、去重、拍摄快照以及创建克隆的相同命令被用于到云的存储。因此，用户继续受益于ZFS压缩、加密、去重、快照和克隆提供的效率和安全性。

[0037] 压缩通常被开启，因为它减少存储和发送数据所需的资源。计算资源在压缩处理中被消耗并且通常在该处理的逆转(解压)中被消耗。数据压缩受到空间-时间复杂度权衡的影响。例如，压缩方案可能要求足够快的密集处理解压缩，以便在正在被解压缩时被消耗。数据压缩方案的设计涉及各种因素之间的权衡，包括压缩程度以及压缩和解压缩数据所需的计算资源。

[0038] ZFS加密启用具有本地保存的加密密钥的端到端安全数据块系统，从而提供附加的安全层。ZFS加密本身并不阻止数据块被盗用，但会拒绝将消息内容给拦截器(interceptor)。在加密方案中，使用加密算法对预期数据块进行加密，生成只有在解密时才能被读取的密文。出于技术原因，加密方案通常使用由算法生成的伪随机加密密钥。原理上有可能在不拥有密钥的情况下解密消息，然而，对于精心设计的加密方案，需要大的计算资源和技巧。ZFS数据块使用AES(高级加密标准)加密，密钥长度为128、192和256。

[0039] 数据块复制是一种专门的数据压缩技术，用于消除重复数据块的重复副本。数据块去重用于提高存储装置利用率，并且也可以应用于网络数据传送，以减少必须被发送以存储到存储器的数据块的数量。在去重处理中，在分析处理期间识别并存储唯一(unique)数据块。随着分析的继续，将其它数据块与存储的副本进行比较，并且每当匹配发生时，冗余的数据块就被替换为指向所存储的存储数据块的小的引用。鉴于相同的数据块模式可能会发生数十次、数百次甚至数千次，使用去重大大减少了必须存储或传送的数据块的数量。

[0040] ZFS存储到云对象存储库的快照在ZFS系统中无缝地创建。快照冻结某些数据和元数据块，以便在需要备份到快照时不会在这些数据和元数据块上写入。树层次结构可以具有许多快照，并且每个快照都将被保存直到被删除为止。快照可以存储在本地或存储在云对象存储库中。并且快照在ZFS系统中是“自由的”，因为除创建快照所指向的根块以外，它们不需要任何额外的存储能力。当从针对根块的快照引用进行访问时，根块和从根块开始的所有后续块对写时复制操作是不可用的。在拍摄快照后的下一次进展中——新的根块成为活动的根块。

[0041] 克隆是从快照创建的并且，与快照不同，使用针对根块的克隆引用访问的块对写时复制操作是可用的。克隆允许在系统上进行开发和故障排除，而不会损坏活动的根块和树。克隆被链接到快照，并且如果链接到快照块的克隆仍然存在，那么就无法删除快照。在一些情况下，克隆可以被提升为活动的分层树。

[0042] 现在将参考附图更详细地讨论各种实施例，从图1开始。

[0043] 图1图示了可以用于实现根据本公开的某些实施例的一个示例存储网络100。图1中描绘的硬件设备的选择和/或布置仅作为示例示出，并不旨在进行限制。图1提供了通过一个或多个交换电路122连接的多个存储装备120。交换电路122可以将多个存储装备120连接到多个I/O服务器136，这些I/O服务器136进而可以为客户端设备(诸如本地计算机系统130、通过网络132可用的计算机系统和/或云计算系统134)提供对多个存储装备120的访问。

[0044] 每个I/O服务器136可以执行多个独立的文件系统实例，每个实例可以负责管理整个存储容量的一部分。如下面将更详细描述的，这些文件系统实例可以包括Oracle ZFS文件系统。I/O服务器136可以包括刀片和/或独立服务器，刀片和/或独立服务器包括主机端口124以通过接收读取和/或写入数据访问请求来与客户端设备通信。主机端口124可以与外部接口提供器126通信，外部接口提供器126识别用于服务于每个I/O请求的正确的数据存储控制器128。数据存储控制器128可以各自专门管理下面描述的存储装备120中的一个或多个存储装备中的数据内容的一部分。因此，每个数据存储控制器128可以访问存储池的一逻辑部分并通过访问它们自己的数据内容来满足从外部接口提供器126接收的数据请求。通过数据存储控制器128的重定向可以包括将每个I/O请求从主机端口124重定向到在I/O服务器136上执行并负责所请求的块的文件系统实例(例如，ZFS实例)。例如，这可以包括从一个I/O服务器136-1上的主机端口124-1到另一个I/O服务器136-n上的ZFS实例的重定向。这种重定向可以允许从任何主机端口124到达可用存储容量的任何部分。然后，ZFS实例可以向存储池中的任何存储设备发出必要的直接I/O事务以完成请求。然后，确认和/或数据可以通过始发主机端口124转发回客户端设备。

[0045] 低延时的存储器映射网络可以将主机端口124、任何文件系统实例和存储装备120绑定在一起。这个网络可以使用一个或多个交换电路122(诸如Oracle的Sun Data Center InfiniBand Switch 36)来实现，以提供可扩展的高性能集群。总线协议(诸如PCI Express总线)可以在存储网络内路由信号。I/O服务器136和存储装备120可以作为对等体进行通信。重定向流量和ZFS存储器流量都可以使用相同的交换架构。

[0046] 在各种实施例中，存储装备120的许多不同配置可以用在图1的网络中。在一些实施例中，可以使用Oracle ZFS存储装备系列。ZFS存储装备使用下面描述的Oracle的ZFS文件系统(“ZFS”)提供基于Oracle Solaris内核的存储。处理核心114处理实现任何所选择的数据保护(例如，镜像、RAID-Z等)、数据减少(例如，内联压缩、复制等)以及任何其它实现的数据服务(例如，远程复制)所需的任何操作。在一些实施例中，处理核心可以包括2.8GHz **Intel®**Xeon处理器的8x15核心。处理核心还处理所存储数据在DRAM和闪存112中的高速缓存。在一些实施例中，DRAM/闪存高速缓存可以包括3TB DRAM高速缓存。

[0047] 在一些配置中，存储装备120可以包括I/O端口116以从数据存储控制器128接收I/O请求。每个存储装备120可以包括整体的机架安装单元，该单元具有其自己的内部冗余电源和冷却系统。集中器板110或其它类似的硬件设备可以用于互连多个存储设备。诸如存储器板、集中器板110、电源和冷却设备之类的有源部件可以是热插拔的。例如，存储装备120可以包括闪存存储器102、非易失性RAM(NVRAM)104、各种配置的硬盘驱动器105、带驱动器、盘驱动器的RAID阵列108等。这些存储单元可以被设计用于高可用性，具有热插拔和存储卡、电源、冷却和互连的内部冗余。在一些实施例中，可以通过在断电时将RAM备份到专用闪

存而使RAM变为非易失性的。闪存和NVRAM卡的混合可以是可配置的，并且两者都可以使用相同的连接器和板简档。

[0048] 虽然没有明确示出，但是每个I/O服务器136可以执行全局管理处理或数据存储系统管理器，其可以以伪静态“低触摸”方式监督存储系统的操作，当必须在ZFS实例之间重新分配容量时进行干预，用于全局闪存耗损均衡(global Flash wear leveling)、用于配置改变和/或用于故障恢复。在各个ZFS实例之间划分容量的“分而治之”策略可以实现性能、连接性和容量的高可扩展性。通过水平添加更多I/O服务器136，然后为每个ZFS实例指派更少的容量和/或为每个I/O服务器136指派更少的ZFS实例，可以实现附加性能。还可以使用更快的服务器垂直扩展性能。可以通过填充I/O服务器136中的可用插槽然后添加附加服务器来添加附加主机端口。还可以通过添加附加存储装备120并将新容量分配给新的或现有的ZFS实例来实现附加的容量。

[0049] 图2图示了根据本公开某些实施例的可以在存储环境(包括图1的存储环境)中执行的示例网络文件系统200的实例。例如，文件系统200可以包括Oracle ZFS文件系统(“ZFS”)，Oracle ZFS文件系统提供非常大的容量(128位)、数据完整性、始终一致的盘上格式、自优化性能，以及实时远程复制。除了其它方式之外，ZFS与传统的文件系统的不同在于至少不需要单独的卷管理器。代替地，ZFS文件系统共享存储设备的公共存储池，并充当卷管理器和文件系统二者。因此，ZFS完全了解物理盘和卷(包括它们的状况、状态和到卷中的逻辑安排，连同存储在它们上的所有文件)。随着文件系统容量要求随时间改变，可以从池中添加或删除设备，以按照需要动态增长和缩小而无需对底层存储池重新分区。

[0050] 在某些实施例中，系统200可以通过操作系统与应用202交互。操作系统可以包括与文件系统交互的功能，文件系统进而与存储池接口。操作系统通常经由系统调用接口208与文件系统200接口。系统调用接口208提供传统的文件读取、写入、打开、关闭等操作，以及特定于VFS体系架构的VNODE操作和VFS操作。系统调用接口208可以充当用于与作为文件系统的ZFS交互的主要接口。这一层位于数据管理单元(DMU)218之间，并呈现存储在其中的文件和目录的文件系统抽象。系统调用接口208可以负责桥接文件系统接口和底层DMU 218接口之间的差距。

[0051] 除了系统调用接口208的POSIX层之外，文件系统200的接口层还可以提供分布式文件系统接口210，用于与集群/云计算设备204交互。例如，可以提供Lustre®接口以提供用于计算机集群的文件系统，计算机集群的尺寸的范围从小工作组集群到大规模多站点集群。卷模拟器212还可以提供用于创建可以用作块/字符设备的逻辑卷的机制。卷模拟器212不仅允许客户端系统区分块与字符，而且还允许客户端系统指定期望的块尺寸，从而在称为“瘦配置”的处理中创建较小的稀疏的卷。卷模拟器212提供到外部设备的原始访问206。

[0052] 接口层下面是事务对象层。这一层提供意图日志214，意图日志214被配置为记录每数据集的事务历史，事务历史可以在系统崩溃时重放。在ZFS中，意图日志214将改变文件系统的系统调用的事务记录和足够的信息保存在存储器中，以便能够重放这些系统调用。这些被存储在存储器中，直到DMU 218将它们提交到存储池，并且它们可以被丢弃或者被刷新。在电源故障和/或盘故障的情况下，可以重放意图日志214事务以使存储池保持最新和一致。

[0053] 事务对象层还提供属性处理器216，属性处理器216可以用于通过在对象内进行任意{键，值}关联来实现系统调用接口208的POSIX层内的目录。属性处理器216可以包括位于DMU 218顶部的模块并且可以对ZFS中被称为“ZAP对象”的对象进行操作。ZAP对象可以用于存储数据集的特性、导航文件系统对象和/或存储存储池特性。ZAP对象可以有两种形式：“microzap”对象和“fatzap”对象。Microzap对象可以是fatzap对象的轻量级版本，并且可以为少量属性条目提供简单且快速的查找机制。Fatzap对象可能更适合包含大量属性的ZAP对象，诸如更大的目录、更长的键、更长的值等。

[0054] 事务对象层还提供数据集和快照层220，数据集和快照层220聚合分层名称空间中的DMU对象并提供用于描述和管理对象集的特性之间的关系的机制。这允许特性的继承，以及存储池中的配额和预留强制实施。DMU对象可以包括ZFS文件系统对象，克隆对象，CFS卷对象和快照对象。因此，数据和快照层220可以管理快照和克隆。

[0055] 快照是文件系统或卷的只读副本。快照是当文件系统处于特定时间点时的文件系统的视图。ZFS的快照与其它一些文件系统的快照一样有用：通过备份快照，您会具有一致的、不变的目标以供备份程序使用。通过从快照复制损坏的文件，快照还可以用于从最近的错误中恢复。几乎可以即时地创建快照，并且它们最初不会在池内消耗附加的盘空间。然而，当活动的数据集内的数据改变时，快照通过继续引用旧数据而占用盘空间，从而阻止盘空间被释放。只有删除快照时才会释放包含旧数据的块。拍摄快照是定时操作。快照的存在不会减慢任何操作。删除快照将花费与删除将释放的块的数量成比例的时间，并且效率非常高。ZFS快照包括以下特征：它们跨系统重新启动保持不变；快照的理论最大数量为2^{64}；它们不使用单独的后备存储库；它们直接从与创建它们的文件系统或卷相同的存储池中消耗盘空间；递归快照作为一个原子操作被快速创建；以及它们被一起创建(一次全部)或根本不创建。原子的快照操作的好处是，即使在后代文件系统中，快照数据也总是在一个一致的时间被拍摄。快照无法被直接访问，但是它们可以被克隆、备份、回滚等等。快照可以用于在时间上“回滚”到拍摄快照时的点。

[0056] 克隆是初始内容与创建它的数据集相同的可写卷或文件系统。在ZFS系统中，始终从快照创建克隆。与快照一样，创建克隆几乎是即时的，最初不会消耗附加的盘空间。此外，可以对克隆进行快照。克隆只能从快照创建。在快照被克隆时，会在克隆和快照之间创建隐式依赖关系。即使克隆是在数据集层次结构中的其它位置创建的，只要克隆存在，就不能销毁原始快照。克隆不会继承创建它的数据集的特性。克隆最初与原始快照共享其所有盘空间。随着对克隆做出改变，它会使用更多盘空间。克隆对于分支并进行开发或故障排除是有用的——并且可以被提升以代替活(live)文件系统。克隆还可以用于在多个机器上复制文件系统。

[0057] DMU 218呈现在由存储池呈现的平坦地址空间的顶部上构建的事务对象模型。上述模块经由对象集、对象和事务与DMU 218交互，其中对象是来自存储池的存储片，诸如数据块的集合。通过DMU 218的每个事务包括作为组提交给存储池的一系列操作。这是在文件系统中维护盘一致性的机制。换句话说，DMU 218从接口层获取指令并将它们转化(translate)成事务批次。不是请求数据块和发送单个读取/写入请求，而是DMU 218可以将这些操作组合成基于对象的事务的批次，这些基于对象的事务的批次可以在任何盘活动发生之前被优化。一旦完成此操作，事务的批次就被移交到存储池层，以对检索/写入所请求

的数据块所需的原始I/O事务进行调度和聚合。如下面将描述的，这些事务是在写时复制(COW)的基础上编写的，这消除了对事务日志(transaction journaling)的需要。

[0058] 存储池层或简称为“存储池”可以被称为存储池分配器(SPA)。SPA提供公共接口来操纵存储池配置。这些接口可以创建、销毁、导入、导出和池化各种存储介质，并管理存储池的名称空间。在一些实施例中，SPA可以包括自适应替换高速缓存(ARC)222，自适应替换高速缓存(ARC)222充当用于SPA的存储器管理的中心点。传统上，ARC为高速缓存管理提供基本的最近最少使用(LRU)对象替换算法。在ZFS中，ARC 222包括可以基于I/O工作负载进行调整的自调整高速缓存。此外，ARC 222定义DMU 218使用的数据虚拟地址(DVA)。在一些实施例中，ARC 222具有由于存储器压力而从高速缓存中驱逐存储器缓冲器以维持高吞吐量的能力。

[0059] SPA还可以包括I/O管道224或“I/O管理器”，其将来自ARC 222的DVA转化成下面描述的每个虚拟设备(VDEV)226中的逻辑位置。I/O管道224跨活动VDEV驱动动态条带化、压缩、校验和能力、和数据冗余。虽然未在图2中明确示出，但I/O管道224可以包括可以由SPA用于从存储池读取数据和/或将数据写入存储池的其它模块。例如，I/O管道224可以包括但不限于压缩模块、加密模块、校验和模块和metaslab分配器。例如，可以使用校验和来确保数据没有被损坏。在一些实施例中，SPA可以使用metaslab分配器来管理存储池中的存储空间的分配。

[0060] 压缩是通常通过利用数据块本身中的冗余来减小数据块(可与叶子节点或数据节点互换地指代)的数据尺寸的处理。ZFS使用许多不同的压缩类型。在启用压缩时，可以为每个数据块分配较少的存储。可以使用以下压缩算法。LZ4——在创建特征标志后添加的算法。它明显优于LZJB。LZJB是用于ZFS的原始默认压缩算法。它的创建是为了满足适用于文件系统的压缩算法的期望。具体而言，它提供公平的压缩、具有高压缩速度、具有高解压速度并且快速检测不可压缩数据检测。GZIP(1至9在经典的Lempel-Ziv实现方案中实现)。它提供高压缩，但它常常使IO是CPU受限的(CPU-bound)。ZLE(零长度编码)——一种只压缩零的非常简单的算法。在这些情况中的每种情况下，都存在压缩率与压缩和解压缩数据块所涉及的延时量之间的权衡。通常，数据压缩得越多，压缩和解压缩所需的时间就越长。

[0061] 加密是通过用密钥加密地对数据块进行编码来向数据块添加端到端安全性的处理。只有拥有密钥的用户才能解密数据块。在ZFS系统中使用时，ZFS池可以支持加密和未加密的ZFS数据集(文件系统和ZVOL)的混合。数据加密对应用是完全透明的，并提供用于保护静止数据(data at rest)的非常灵活的系统，并且它不需要任何应用改变或资格。此外，ZFS加密从密码短语(passphrase)或AES密钥随机生成本地加密密钥，并且所有密钥都存储在客户端本地——而不是像传统文件系统那样存储在云对象存储库404中。当被开启时，加密对于云对象存储库404的应用和存储是透明的。ZFS使得加密数据和管理数据加密是容易的。您可以在同一存储池中拥有加密和未加密的文件系统二者。您还可以为不同的系统使用不同的加密密钥，并且您可以在本地或远程地管理加密——但是随机生成的加密密钥始终保持在本地。ZFS加密对于后代文件系统是可继承的。在CCM和GCM操作模式下，使用密钥长度为128、192和256的AES(高级加密标准)来加密数据。

[0062] 去重是辨识要存储在文件系统中的数据块已经作为现有数据块存储在文件系统上并指向该现有数据块而不是再次存储数据块的处理。ZFS提供块级的去重，因为这是对通

用存储系统有意义的最精细的粒度。块级去重也自然地映射到ZFS的256位块校验和，只要校验和功能在加密方面很强(例如SHA256)，ZFS的256位块校验和就为存储池中的所有块提供唯一的块签名。去重是同步的，并且在数据块被发送到云对象存储库404时执行。如果数据块不是复制的，那么启用去重会增加开销而不会带来任何好处。如果存在重复的数据块，那么启用去重将既节省空间又提高性能。节省空间是显而易见的；性能提高是由于在存储重复数据时消除了存储写入，以及由于许多应用共享相同存储器页面而导致存储器占用减少。大多数存储环境包含大多独特的数据和大多被复制的数据的混合。ZFS去重是针对每个数据集的，并且可以在它可能有帮助时被启用。

[0063] 在ZFS中，存储池可以由VDEV的集合组成。在某些实施例中，存储池的至少一部分可以被表示为自描述的Merkle树，即，其中数据和元数据都通过逻辑树的VDEV存储的逻辑树。有两种类型的虚拟设备：称为叶子VDEV的物理虚拟设备，以及称为内部VDEV的逻辑虚拟设备。物理VDEV可以包括可写介质块设备，诸如硬盘或闪存驱动器。逻辑VDEV是物理VDEV的概念化分组。VDEV可以布置在树中，其中物理VDEV作为树的叶子存在。存储池可以具有称为“根VDEV”的特殊逻辑VDEV，它是树的根。根VDEV的所有直接子节点(物理或逻辑)都被称为“顶级”VDEV。一般而言，VDEV实现数据复制、镜像和体系架构(诸如RAID-Z和RAID-Z2)。每个叶子VDEV表示实际存储由文件系统提供的数据的一个或多个物理存储设备228。

[0064] 在一些实施例中，文件系统200可以包括基于对象的文件系统，其中数据和元数据都被存储为对象。更具体而言，文件系统200可以包括将数据和对应的元数据存储在存储池中的功能。执行特定操作(即，事务)的请求从操作系统经由系统调用接口208转发到DMU 218，DMU 218将对对象执行操作的请求直接转化成在存储池内的物理位置处执行读取或写入操作的请求(即，I/O请求)。SPA从DMU 218接收请求，并使用COW过程将块写入存储池。可以为对文件的数据写请求来执行COW事务。不是在写入操作时覆写现有的块，而是写请求使得为修改后的数据分配新的片段。因此，在提交数据块和元数据的修改版本之前，永远不会覆写检索到的数据块和对应的元数据。因此，DMU 218将所有修改后的数据块都写入存储池内未使用的片段，并且随后将对应的块指针写入存储池内未使用的片段。为了完成COW事务，SPA发出I/O请求以引用修改后的数据块。

[0065] 图3A-图3D图示了根据本公开某些实施例的用于文件系统(诸如文件系统200)的COW处理。例如，上述ZFS系统使用COW事务模型，其中文件系统内的所有块指针可以包含目标块的256位校验和，该256位校验和在块被读取时被验证。如上所述，包含活动数据的块不会被原地覆写。相反，新的块被分配，修改后的数据被写到新的块，然后引用它的任何元数据块被简单地读取、重新分配和重写。

[0066] 图3A图示了根据一些实施例的与作为逻辑树300的一个或多个文件对应的数据和元数据的文件系统存储的简化图。逻辑树300以及本文描述的其它逻辑树可以是自描述的Merkle树，其中数据和元数据被存储为逻辑树300的块。根块302可以表示逻辑树300的根或“超级块”。逻辑树300可以通过导航通过根302的每个子节点304、306来遍历文件和目录。每个非叶子节点表示目录或文件，诸如节点308、310、312和314。在一些实施例中，可以为每个非叶子节点指派其子节点的值的散列。每个叶子节点316、318、320、322表示文件的数据块。

[0067] 图3B图示了在写入操作的初始阶段之后的逻辑树300-1的示例。在这个示例中，由节点324和326表示的数据块已由文件系统200写入。不是覆写节点316和318中的数据，而是

为节点324和326分配新的数据块。因此，在这个操作之后，节点316和318中的旧数据与节点324和326中的新数据一起保留在存储器中。

[0068] 图3C图示了随着写入操作继续的逻辑树300-2的示例。为了引用节点324和326中新写入的数据块，文件系统200确定引用旧节点316和318的节点308和310。新节点328和330被分配为引用节点324、326中的新数据块。递归地向上通过文件系统层次结构重复相同的处理，直到引用改变后的节点的每个节点都被重新分配为指向新节点。

[0069] 当指针块在层次结构中的新节点中被分配时，每个节点中的地址指针被更新为指向所分配的子节点在存储器中的新位置。此外，每个数据块包括校验和，该校验和通过由地址指针引用的数据块来计算。例如，使用节点324中的数据块计算节点328中的校验和。这种布置意味着校验和与从其计算该校验和的数据块分开存储。这防止所谓的“代写(ghost write)”，在代写中永远不会写入新数据，但是与数据块一起存储的校验和将指示该块是正确的。可以通过遍历逻辑树300并基于子节点计算每个级别的校验和来快速校验逻辑树300的完整性。

[0070] 为了完成写入操作，可以重新分配和更新根302。图3D图示了写入操作结束时逻辑树300-3的示例。当根302准备好被更新时，可以分配并初始化新的超级块根336以指向新分配的子节点332和334。然后，根336可以在原子操作中成为逻辑树300-3的根，以最终确定逻辑树300-3的状态。

[0071] 快照是文件系统或卷的只读副本。快照是当文件系统处于特定时间点时的文件系统的视图。ZFS的快照与其它一些文件系统的快照一样有用：通过备份快照，您会具有一致的、不变的目标以供备份程序使用。通过从快照复制损坏的文件，快照还可以用于从最近的错误中恢复。几乎可以即时地创建快照，并且它们最初不会在池内消耗附加的盘空间。然而，当活动的数据集内的数据改变时，快照通过继续引用旧数据而占用盘空间，从而阻止盘空间被释放。只有删除快照时才会释放包含旧数据的块。拍摄快照是定时操作。快照的存在不会减慢任何操作。删除快照将花费与删除将释放的块的数量成比例的时间，并且效率非常高。ZFS快照包括以下特征：它们跨系统重新启动保持不变；快照的理论最大数量为264；它们不使用单独的后备存储库；它们直接从与创建它们的文件系统或卷相同的存储池中消耗盘空间；递归快照作为一个原子操作被快速创建；以及它们被一起创建(一次全部)或根本不创建。原子的快照操作的好处是，即使在后代文件系统中，快照数据也总是在一个一致的时间被拍摄。快照无法被直接访问，但是它们可以被克隆、备份、回滚等等。快照可以用于在时间上“回滚”到拍摄快照时的点。图3E描绘了ZFS中的快照数据服务的示例，其中快照是在图3A-图3D中描述的COW处理将根块336置为新的活的根块之前拍摄的。活的根块是下一次数据进展在执行COW时将从其开始的根块。快照根和“活”根被示出。活根是将在下一次存储操作中将对其进行操作的根。由快照根指向的所有块(302-322)都被置为“只读”，这意味着它们被放置在无法释放以供存储系统进一步使用的块的列表上，直到快照被删除为止。

[0072] 应当注意的是，贯穿整个申请可以互换地使用一些术语。例如，叶子节点、叶子块和数据块在某些情况下可以是相同的，特别是在引用本地树实例时。另外，非叶子节点、元数据和元数据块在某些情况下可以互换使用，特别是在引用本地树实例时。类似地，根节点和根块可以在某些实例中互换使用，特别是在引用本地树实例时。另外，应当注意的是，对叶子节点、非叶子节点和根节点的引用可以类似地应用于与至少部分地基于本地树实例生

成的逻辑树的云版本对应的云存储对象。

[0073] 当创建块时，它们被赋予“出生”时间，该“出生”时间表示创建块的COW的迭代或进展。图3F展示了这个想法。在图3F中，如365中所示——图3A的出生时间19。366示出了图3D的出生时间25。367示出了由块372、378、379、384、386和392创建的新树所示的出生时间37，并且表示出生时间25的12次迭代后树上的数据事务。因此，回滚或备份到快照将仅留下如图3A所示的块。因此，使用出生时间层次结构，ZFS系统可以从树的根的快照生成并回滚到整个树结构的出生时间中的任何点。实质上，这允许具有在快照之后的出生时间的所有新块在存储池中被置为可用，只要它们未被任何其它快照或元数据块链接。

[0074] 克隆是初始内容与创建它的数据集相同的可写卷或文件系统。在ZFS系统中，始终从快照创建克隆。与快照一样，创建克隆几乎是即时的，最初不会消耗附加的盘空间。此外，可以对克隆进行快照。克隆只能从快照创建。在快照被克隆时，会在克隆和快照之间创建隐式依赖关系。即使克隆是在数据集层次结构中的其它位置创建的，只要克隆存在，就不能销毁原始快照。克隆不会继承创建它的数据集的特性。克隆最初与原始快照共享其所有盘空间。随着对克隆做出改变，它会使用更多盘空间。克隆对于分支并进行开发或故障排除是有用的——并且可以被提升以代替实时(live)文件系统。克隆还可以用于在多个机器上复制文件系统。

[0075] 本文描述的实施例可以在上面在图1-图3F中描述的系统中实现。例如，该系统可以包括图1的各种存储装备、交换电路和/或服务器的一个或多个处理器。指令可以存储在系统的一个或多个存储器设备中，使得一个或多个处理器执行影响文件系统的功能的各种操作。各种方法的步骤可以由图1-图2中的系统的处理器、存储器设备、接口和/或电路系统执行。

[0076] 现在转到图4，图4是图示根据本公开某些实施例的混合云存储系统400的示例的高级别图。混合云存储系统400可以将网络文件系统(诸如ZFS文件系统)变换为支持云的文件系统，在支持云的文件系统中，文件系统的功能(包括文件系统数据服务)在远离文件系统的云对象存储库上被分层。如在所描绘的图中，混合云存储系统400可以包括网络文件系统200(在本文也称为“本地文件系统200”)。本地文件系统200可以通信地耦合到云对象存储404。在一些实施例中，云对象存储404可以与图2中指示的集群/云204对应。本地文件系统200可以通过云接口装备402通信地耦合到云对象存储404。云接口装备402可以由本地文件系统200用作针对云对象存储404的接入点。

[0077] 混合云存储系统400提供克服云对象存储的传统限制的解决方案。传统的云对象协议限于受限的数据/对象访问语义。传统上，云对象存储具有有限的接口和原语，并且与POSIX不兼容。例如，一旦对象被写入，此后就不能对其进行修改；它只能被删除并替换为新创建的对象。作为另一个示例，传统云对象存储具有名称空间限制，使得名称空间被简化并且仅限于顶级容器。然而，混合云存储系统400不仅可以将数据迁移到云对象存储库404以及从云对象存储库404迁移数据，而且还可以将本地文件系统200的文件系统功能在到云对象存储库404的云对象接口上分层，以提供基于云的存储。

[0078] 本地文件系统200可以被配置用于POSIX接口和语义。例如，本地文件系统200可以向用户提供对作为文件的数据的访问，从而允许修改文件的内容而不重写文件。本地文件系统200还可以提供名称层次结构中的数据的组织，如对于ZFS文件系统是典型的那样。ZFS

文件系统的所有功能可以对本地文件系统200的用户可用。云接口装备402可以允许在云对象协议之上对文件系统语义进行分层——例如，提供构造名称空间、创建文件、创建目录等的能力——并且相对于迁移到云对象存储404和从云对象存储404迁移的数据扩展这样的能力。云接口装备402可以促进即插即用对象存储解决方案，以改进本地文件系统200，同时支持ZFS文件系统数据服务。

[0079] 云接口装备402可以被配置为提供对象API(应用编程接口)。在一些实施例中，云接口装备402可以被配置为使用多个API转化简档。根据某些实施例，API转化简档可以集成模块和功能(例如，数据服务和模块)、POSIX接口和语义，以及可能不被本地设计为与云存储交互的其它部件。在一些实施例中，API转化简档可以(例如，通过API调用)转化文件系统200的协议、格式和例程以允许与云数据存储库404的交互。用于这种集成的信息可以存储在API转化数据存储库中，该API转化数据存储库可以与云接口装备402共同定位或以其它方式通信地耦合到云接口装备402。云接口装备402可以利用该信息来内聚地集成POSIX接口和语义以与云数据存储库404接口，同时保留这些语义。

[0080] 混合云存储系统400可以允许本地文件系统200将云对象存储404用作“驱动器”。在各种情况下，文件410都可以作为具有元数据对象的数据对象存储，和/或作为具有相关联的元数据的数据块。云接口装备402可以从本地文件系统200接收文件410并将文件410传送到本地文件系统200。在各种实施例中，本地文件系统200可以经由NFS(网络文件系统)协议、SMB(服务器消息块协议)等接收和/或发送文件410。在一些实施例中，云接口装备402可以将文件410转化成对象412。文件410的转化可以包括转化数据块和关联元数据和/或与元数据对象相关联的数据对象，其中任何一个都可以与文件410对应。在一些实施例中，转化可以包括云接口装备402利用多个API转化简档执行API转化。根据一些实施例，转化可以包括云接口装备402从文件410、对象和/或块提取数据和/或元数据。云接口装备402可以至少部分地通过使用提取出的数据将文件410、对象和/或块转化成云存储对象。在一些实施例中，云接口装备402可以利用提取出的数据来创建对应的云存储对象，其中提取出的数据被嵌入在指向云对象存储库404的放置请求中。同样，利用由云接口装备402实现的用以向外接口到云数据存储库404的任何转化，在一些实施例中，云接口装备402可以反转转化处理以与本地文件系统200的本地部件接口。

[0081] 云接口装备402可以向云对象存储404传送对象412以及从云对象存储404接收对象412。在一些实施例中，云接口装备402可以经由HTTPS等来收发对象412。在一些实施例中，云接口装备402可以与本地文件系统200共同定位。在其它实施例中，云接口装备402可以远离本地文件系统200定位，诸如，与促进云对象存储库404的至少一些装备一起或在某个其它通信耦合的站点。

[0082] 如本文进一步公开的，本地文件系统200中的文件可以被存储为“盘块”，其中与文件对应的数据对象和元数据被存储为逻辑树300的虚拟存储块(例如，自描述的Merkle树，其中数据和元数据被存储为块)。云接口装备402可以创建数据的树300中的每个逻辑块直接到云对象存储库404中的云对象414的映射406。一些实施例可以采用一对一的块到对象映射。附加地或可替代地，其它实施例可以采用任何其它合适的块与云对象的比率，例如，将多个块映射到一个云对象。在此类实施例的一些实例中，可以将块的整个逻辑树映射到单个云对象。在其它情况下，块的逻辑树的仅一部分可以被映射到单个云对象。

[0083] 在一些实施例中，当块被转换成云对象时，更新地址指针。当使用一对一的块到对象转换方案将块转换成云对象时，可以更新地址指针，使得层次结构中的非叶子云对象指向云对象存储库404中的子云对象。举例来说，地址指针可以与子云对象的对象名称和路径规范对应，其可以包括诸如对象名称、桶规范(bucket specification)等参数。相应地，一些实施例可以将逻辑树300的块转化成逻辑树300A的云对象。利用一些实施例，这种转化可以使云接口装备402能够利用云对象的地址指针遍历云对象的逻辑树300A。

[0084] 在一些实施例中，当使用多对一的块到对象转换方案将块转换成云对象使得逻辑树300的一部分被转换成一个云对象时，包括逻辑300A的云对象的地址指针可以类似地被更新，但是是在更粗粒度程度上，使得层次结构中的非叶子云对象指向云对象存储库404中的子云对象。这种转化可以使云接口装备402能够利用云对象的地址指针以比通过一对一的块到对象转化方案促进的遍历更粗粒度但更快的方式遍历云对象的逻辑树300A。此外，在一些实施例中，可以利用转换处理来更新校验和。用于各个云对象的校验和可以被更新并分开存储在父云对象中。在采用多对一的块到对象转换方案的转换中，可以针对与一组块对应的云对象来计算单个校验和。

[0085] 相应地，映射406的实现允许通过一个或多个网络到云对象存储库404的通信，并且与云对象存储库404的接口可以是基于对象的而不是基于块的。如本文进一步公开的，利用本地文件系统200和云对象存储库404之间的云接口装备402，混合云存储系统400可以拥有与传统ZFS文件系统不同的特点和故障模式。云接口装备402可以在客户端侧转化本地文件系统202的文件系统接口，并且可以能够经由对象协议协调到云对象存储库404以读取和写入数据。通过云接口装备402，云对象414可以保持为能够由本地文件系统200通过NFS、SMB等访问。

[0086] 利用云对象414作为逻辑块的映射406，可以将云对象414的集合分组以形成将ZFS存储池托管为自包含集合的驱动器。驱动器内容可以是弹性的，使得可以仅为已经分配的逻辑块创建云对象。在一些实施例中，云接口装备402可以具有(例如，针对不同数据类型)指派可变对象尺寸的能力，以允许更大的存储灵活性。数据不必限于特定的字节尺寸。可以通过修改元数据来根据需要扩展存储尺寸。在一些实施例中，可以在任何服务器上导入基于云的池。一旦被导入，基于云的池就可以表现为本地存储，并且对基于云的池支持所有ZFS服务。可以将基于云的池指示为新类型的存储池。然而，从用户的角度来看，来自基于云的池的数据可以看起来与本地池无法区分。

[0087] 图5图示了根据本公开某些实施例的混合云存储系统400的示例网络文件系统200-1的实例。文件系统200-1可以与文件系统200对应，但是云设备管理直接集成到ZFS控制栈中。除了关于文件系统200公开的内容之外，文件系统200-1还可以包括云接口设备502，云接口设备502促进充分利用云对象存储库404作为文件系统200-1的存储介质。云接口设备502可以至少部分地通过将云存储映射到设备抽象来促进云驱动器。

[0088] 在一些实施例中，云接口设备502可以与ZFS文件系统体系架构内的另一个VDEV类型的设备驱动器接口的一个或多个VDEV对应。ZFS可以直接与云接口设备502通信。云接口设备502可以位于文件系统200-1的驱动器层直接上方的虚拟设备层处。云接口设备502的一些实施例可以与ZFS体系架构内的设备驱动器接口的抽象对应。文件系统200-1的其它部件可以与云接口设备502通信，就好像它是另一个设备类型的另一个VDEV(诸如VDEV 226)

一样。为了能够通过云接口设备502相对于通过其它VDEV 226传递更大量的信息，与云接口设备502相关联的接口可以更宽以使更多信息传递通过I/O管道224和云接口设备502，向外到达云对象数据存储库404。

[0089] 在一些实施例中，云接口设备502可以转化客户端上的文件系统接口。在一些实施例中，为了提供完整的POSIX文件系统语义，云接口设备502可以将文件系统接口请求转换成针对云对象存储库404的对象接口请求。在一些实施例中，云接口设备502可以能够经由对象协议向外与云对象存储库404通信以读取和写入数据。

[0090] 图6是图示根据本公开某些实施例的混合云存储系统400-1的云接口装备402-1的附加方面的图。如所描绘的示例中所指示的，云接口装备402的一些实施例可以包括虚拟存储池602和云接口守护进程604。虚拟存储池602可以位于文件系统200-1的内核处，并且云接口守护进程604可以位于文件系统200-1的用户空间上。在各种实施例中，云接口守护进程604可以与图5中指示的应用202和/或集群/云204的云接口部件对应。

[0091] 在一些实施例中，虚拟存储池602可以包括至少一个云接口设备502、意图日志214-2和高速缓存222-1。上面关于图1-图2以及下面关于图7描述意图日志214-2和高速缓存222-1。云接口设备502可以与云接口守护进程604交互，以至少部分地基于映射406来协调关于云对象数据存储库404的操作。云接口守护进程604可以包括云客户端接口608，以与云对象数据存储库404接口。在一些实现方案中，云客户端接口608可以包括提供与云对象数据存储库404的Swift/S3兼容性的端点。云客户端接口608的操作可以至少部分地基于取出和放置整个数据对象412，以便促进对云对象数据存储库404的读取和写入访问。

[0092] 返回去参考图4和图5，在根据一些实施例的操作中，可以通过文件系统200-1的接口层的系统调用接口208从文件系统200-1的应用层处的应用202接收针对一个或多个文件执行一个或多个事务的请求。请求可以是POSIX兼容的，并且可以由文件系统200-1的一个或多个部件转换成针对存储在云对象存储库404中的逻辑树300的基于云的实例化300A执行一个或多个操作的一个或多个对象接口请求。例如，在一些实施例中，云接口装备402可以将兼容POSIX的请求或由兼容POSIX的请求造成的中间请求转换成对应的对象接口请求。在一些实施例中，DMU 218可以将兼容POSIX的请求转化成执行I/O操作的I/O请求，并且云接口装备402可以将I/O请求转化成对应的对象接口请求，从而使用映射406协调对象接口请求。

[0093] 在一些情况下，事务可以与导致文件存储在本地的操作对应。在一些实例中，文件系统200-1可以将数据对象和对应的元数据存储在由VDEV 226中的一个或多个和一个或多个物理存储设备228提供的系统存储池416中。数据对象可以与一个或多个文件对应。如上面所公开的，与一个或多个文件对应的数据对象和元数据可以存储为逻辑树300。因此，逻辑树300的存储可以本地存储在系统存储池416中。

[0094] 在根据一些实施例的进一步操作中，文件系统200-1可以使得在云对象存储库404中存储逻辑树300的数据对象和对应元数据。虽然在一些实施例中逻辑树300可以在迁移到云存储之前首先存储在本地存储池中，但是在其它实施例中，逻辑树300可以在被存储在云对象存储库404中之前不存储在本地存储池中。例如，一些实施例可以在高速缓存中创建逻辑树300的至少一部分，然后将它迁移到云对象存储库404。因此，应当认识到的是，各种实施例都是可能的。

[0095] 为了将逻辑树300的数据对象和对应元数据存储在云对象存储库404中，云接口设备502可以创建逻辑树300中的每个逻辑块到云对象存储库404中相应的云对象414的映射406。在一些实施例中，DMU 218可以从系统存储池416(例如，从本地池的RAIDZ或RAIDZ2)读取数据以将数据提供给云接口设备502，作为用于创建映射406的基础。在一些实施例中，云接口设备502可以直接或间接地与另一个VDEV 226通信以读取数据作为用于映射406的基础。在一些实施例中，映射406可以将对象直接映射到物理驱动器中表示的块。映射406可以比将文件部分映射到对象更精细；它可以在更低的级别进行映射。相应地，映射406可以是每对象映射406。映射406可以将虚拟存储块映射到云对象存储库404中的对象，使得逻辑树300在云对象存储库404中表示，如逻辑树300A所示。当本地文件系统200-1与云对象存储库404中的数据对象416接口时，逻辑树300A符合本地文件系统200-1能够与之通信的新设备类型。

[0096] 可以利用对云对象存储库404的每个I/O操作或者仅利用写入操作来更新映射406。在一些实施例中，映射406可以包括索引所有云对象414的对象目录。云对象状态可以保持在索引、表、索引组织的表等等当中，其可以基于每个对象被索引。在一些实施例中，映射406可以包括仅索引云对象414中的一些的对象目录。例如，此类实施例可以仅索引与超级块对应的云对象414。在一些实施例中，可以索引每个对象相对于每个叶子路径的云对象状态。在一些实施例中，对象目录可以通过地址指针来引用云对象414，地址指针可以与对象名称和路径规范对应。在一些实施例中，对象目录可以通过URL来引用云对象414。

[0097] 在映射406中被索引的云对象状态可以用于路由对象请求。利用索引，云接口设备502可以至少部分地基于超级块来请求云对象。根据第一种方法，此类请求可能需要请求与特定超级块相关联的云对象集合，以便整个逻辑树300A由响应于该请求而传送的云对象集合表示。根据第二种方法，此类请求可能需要对与特定超级块相关联的云对象的子集的迭代请求，以便迭代地遍历整个逻辑树300A，直到从云对象存储库404读取期望的一个或多个云对象为止。对于某些实施例中，云接口设备502可以至少部分地基于表示各个逻辑树300A的云对象的尺寸来选择性地使用这两种方法中的一种。例如，当云对象的尺寸小于聚合尺寸阈值时，云接口设备502可以使用一种方法，并且当云对象的尺寸满足或超过聚合尺寸阈值时，云接口设备502可以过渡到另一种方法。

[0098] 一些实施例可以采用另一种方法，其中对象目录可以在每个对象的基础上来索引云对象，并且可以用于直接请求云对象而无需在云级别进行树遍历。一些实施例可以保留逻辑树300A的元数据的本地快照。此类实施例可以利用本地快照直接或间接地请求云对象。此外，一些实施例可以在对象目录或本地快照中保留逻辑树300A的校验和，该校验和可以用于验证从云数据存储库404检索到的云对象。

[0099] 因此，文件系统200-1可以维护数据的树并将该树映射到云对象存储库404上。树300A的名称空间可以与存储在树300A的节点内的元数据对应。文件系统200-1可以继续使用分层树表示，但是将分层树表示映射到云对象存储库中作为存储数据的方式。

[0100] 再次参考图6，为了实现针对云对象存储库404的I/O操作，云接口设备502可以向云接口守护进程604发送请求。例如，在一些实现方案中，云接口设备502可以通过文件系统200-1的事务对象层和接口层向云接口守护进程604发送请求。云接口设备502发送的请求可以至少部分地基于经由应用202接收的兼容POSIX的请求和/或至少部分地基于由DMU

218(例如,响应于兼容POSIX的请求)创建的I/O请求,云接口设备502可以将这些请求转换成对云接口守护进程604的请求。

[0101] 在一些实施例中,由云接口设备502发送到云接口守护进程604的请求可以被转化成针对云客户端接口608的取出请求和放置请求。在一些实施例中,由云接口设备502发送的请求可以是取出请求和放置请求;在其它实施例中,云接口守护进程604可以将由云接口设备502发送的请求转化成取出请求和放置请求。在任何情况下,响应于由云接口设备502发送的请求,云接口守护进程604可以经由一个或多个网络上的对象协议与云对象存储库404通信,以执行针对数据对象414的对应I/O操作。

[0102] 例如,到云对象存储库404的通信可以包括至少部分地基于映射406来指定逻辑树300A的数据对象和对应元数据在云对象存储库404中的存储。在一些实施例中,该通信可以例如针对不同的数据类型指定不同的对象尺寸。因此,云接口装备402可以指定某个对象尺寸来存储被识别为具有一定数据类型的某些数据对象,并且可以指定不同的对象尺寸来存储被识别为具有不同数据类型的其它数据对象。

[0103] 图7A是图示根据本公开某些实施例示例方法700的框图,示例方法700针对混合云存储系统400的COW处理的某些特征。根据某些实施例,方法700可以如框702所指示的那样开始。然而,本公开的教导可以以各种配置实现。照此,包括方法700和/或本文公开的其它方法的某些步骤的次序可以以任何合适的方式混洗或组合,并且可以取决于所选择的实现方案。而且,虽然为了描述可以分离以下步骤,但是应当理解的是,某些步骤可以同时或基本上同时进行。

[0104] 如框702所指示的,可以从应用202接收执行(一个或多个)特定操作(即,(一个或多个)事务)的(一个或多个)兼容POSIX的请求。这种操作可以与写入和/或修改数据对应。如框704所指示的,(一个或多个)兼容POSIX的请求可以经由系统调用接口208从操作系统转发到DMU 218。在各种实施例中,通过DMU 218实现的事务可以包括作为一个组提交给系统存储池416和云对象存储库404中的一者或两者的一系列操作。这些事务可以基于COW编写。

[0105] 如框706所指示的,DMU 218可以将对数据对象执行操作的请求直接转化成执行针对系统存储池416和/或云对象存储库404内的物理位置的写入操作的请求(即,I/O请求)。在一些模式中,可以首先对本地存储的数据对象执行这些操作,然后可以通过以下方式将对数据对象的改变传播到对应的云存储的数据对象,即,由DMU 218引导具体的改变或者由DMU 218指导云接口装备402直接或利用另一个VDEV 226间接读取这些改变。在其它模式中,可以同时或基本上同时对本地存储的数据对象和对应的云存储的数据对象执行操作。在还有其它模式中,可以仅对云存储的数据对象执行操作。例如,一些实现方案可以不具有本地树300,并且可以仅具有逻辑树300A的基于云的版本。各种实施例可以被配置为允许用户选择一个或多个模式。

[0106] 如框708所指示的,SPA可以从DMU 218接收I/O请求。并且,响应于请求,SPA可以发起使用COW过程来将数据对象写入系统存储池416和/或云对象存储库404。如框710所指示的,在数据对象写入云对象存储库404之前或同时在本地存储的数据对象上执行数据对象的写入的模式中,上面(例如,鉴于图3A-图3D)公开的COW过程可以相对于系统存储池416继续进行。

[0107] 如框712所指示的，云接口装备402可以接收I/O请求并识别对逻辑树300A的增量修改。增量修改可以与由COW处理产生的新树部分对应以实现写请求。云接口装备402可以将I/O请求转化成对应的对象接口请求。通过具有来自I/O请求或来自读取对本地存储的数据对象的改变的修改后的数据，云接口设备502可以使用云对象存储库404中的云存储对象414的映射406来协调对象接口请求。

[0108] 例如，在一些实施例中，可以至少部分地基于对本地存储的数据对象的改变来确定增量修改，以反映对逻辑树300的改变。在一些实例中，云接口装备402可以读取逻辑树300或读取至少对逻辑树300的改变，以便确定增量修改。这种数据可以由文件系统的另一个部件(诸如DMU 218或镜像VDEV)传递到云接口装备402。对于一些实施例，增量修改可以被传送到云接口装备402。然而，在一些实施例中，云接口装备402可以至少部分地基于鉴于逻辑树300的副本或快照和/或逻辑树300A的快照分析写请求来确定增量修改。对于云接口装备402使用逻辑树300A的快照的实施例，在一些实施例中，快照可以保留在映射406中或者以其它方式存储在存储器和/或物理层中。

[0109] 再次更具体地参考图7A，如框714所指示的，增量修改的确定可以包括创建新的叶子节点(例如，叶子节点324、326)。在写入操作的初始阶段之后，已经在存储器中分配了新的数据块(即，叶子节点)，并且每个写入操作的数据已经由云接口装备402写入新的数据块，而先前的数据和数据块也同样保留在存储器中。

[0110] 如框729所指示的，在框714中创建叶子节点(数据块)之后，确定是否已经开启或请求任何ZFS数据服务。这些包括但不限于压缩、加密、去重、快照和克隆。如果不需要任何这些数据服务——那么如框716所指示的，可以创建新的非叶子节点(例如，非叶子节点328、330)。随着写入操作继续，云接口装备402可以确定引用节点的先前版本的非叶子节点。为了引用新写入的数据，将新的非叶子节点分配为引用叶子节点中的新数据块。如快照301所反映的，可以递归地向上通过逻辑树300A的层次结构重复同一处理，直到引用改变的节点的每个非叶子节点被重新分配为指向新节点。当指针块被分配在层次结构中的新节点中时，可以更新每个节点中的地址指针以指向所分配的子节点在存储器中的新位置。如框718所指示的，为了完成写入操作，可以重新分配和更新根节点(例如，根节点336)。当准备好更新根节点时，可以分配和初始化新的根节点(超级块)以指向新根节点之下的新分配的子节点。

[0111] 作为写入操作的一部分，利用校验求和来更新事务写入操作中涉及的树的所有部分的元数据。创建的每个节点都包括校验和，该校验和是使用由地址指针引用的节点计算的。这种布置意味着校验和与从其计算该校验和的节点分开存储。通过将每个节点的校验和存储在该节点的父节点指针中而不存储在该节点本身中，树中的每个节点都包含用于其所有子节点的校验和。关于图3A-图3D的示例进一步详细说明这一点。通过这样做，每个树都是自动进行自我验证的，并且始终有可能检测与读取操作的不一致。

[0112] 如果在框729处需要数据服务，那么下一个框719进到图7B上框730。数据服务包括但不限于压缩、加密、去重(必须按那个次序完成)、快照和克隆。

[0113] 压缩是通常通过利用数据块本身中的冗余来减小数据块(可与叶子节点或数据节点互换地指代)的数据尺寸的处理。ZFS使用许多不同的压缩类型。在启用压缩时，可以为每个数据块分配较少的存储。可以使用以下压缩算法。LZ4——在创建特征标志后添加的算

法。它明显优于LZJB。LZJB是用于ZFS的原始默认压缩算法。它的创建是为了满足适用于文件系统的压缩算法的期望。具体而言，它提供公平的压缩、具有高压缩速度、具有高解压速度并且快速检测不可压缩数据检测。GZIP(1至9在经典的Lempel-Ziv实现方案中实现)。它提供高压缩，但它常常使IO是CPU受限的(CPU-bound)。ZLE(零长度编码)——一种只压缩零的非常简单的算法。在这些情况中的每种情况下，都存在压缩率与压缩和解压缩数据块所涉及的延时量之间的权衡。通常，数据压缩得越多，压缩和解压缩所需的时间就越长。

[0114] 加密是通过用密钥加密地对数据块进行编码来向数据块添加端到端安全性的处理。只有拥有密钥的用户才能解密数据块。在ZFS系统中使用时，ZFS池可以支持加密和未加密的ZFS数据集(文件系统和ZVOL)的混合。数据加密对应用是完全透明的，并提供用于保护静止数据(data at rest)的非常灵活的系统，并且它不需要任何应用改变或资格。此外，ZFS加密从密码短语或AES密钥随机生成本地加密密钥，并且所有密钥都存储在客户端本地——而不是像传统文件系统那样存储在云对象存储库404中。当被开启时，加密对于云对象存储库404的应用和存储是透明的。ZFS使得加密数据和管理数据加密是容易的。您可以在同一存储池中拥有加密和未加密的文件系统二者。您还可以为不同的系统使用不同的加密密钥，并且您可以在本地或远程地管理加密——但是随机生成的加密密钥始终保持在本地。ZFS加密对于后代文件系统是可继承的。在CCM和GCM操作模式下，使用密钥长度为128、192和256的AES(高级加密标准)来加密数据。

[0115] 去重是辨识要存储在文件系统中的数据块已经作为现有数据块存储在文件系统上并指向该现有数据块而不是再次存储数据块的处理。ZFS提供块级的去重，因为这是对通用存储系统有意义的最精细的粒度。块级去重也自然地映射到ZFS的256位块校验和，只要校验和功能在加密方面很强(例如SHA256)，ZFS的256位块校验和就为存储池中的所有块提供唯一的块签名。去重是同步的，并且在数据块被发送到云对象存储库404时执行。如果数据块不是复制的，那么启用去重会增加开销而不会带来任何好处。如果存在重复的数据块，那么启用去重将既节省空间又提高性能。节省空间是显而易见的；性能提高是由于在存储重复数据时消除了存储写入，以及由于许多应用共享相同存储器页面而导致存储器占用减少。大多数存储环境包含大多独特的数据和大多被复制的数据的混合。ZFS去重是针对每个数据集的，并且可以在它可能有帮助时被启用。

[0116] 快照是文件系统或卷的只读副本。快照是当文件系统处于特定时间点时的文件系统的视图。ZFS的快照与其它一些文件系统的快照一样有用：通过备份快照，您会具有一致的、不变的目标以供备份程序使用。通过从快照复制损坏的文件，快照还可以用于从最近的错误中恢复。几乎可以即时地创建快照，并且它们最初不会在池内消耗附加的盘空间。然而，当活动的数据集内的数据改变时，快照通过继续引用旧数据而占用盘空间，从而阻止盘空间被释放。只有删除快照时才会释放包含旧数据的块。拍摄快照是定时操作。快照的存在不会减慢任何操作。删除快照将花费与删除将释放的块的数量成比例的时间，并且效率非常高。ZFS快照包括以下特征：它们跨系统重新启动保持不变；快照的理论最大数量为264；它们不使用单独的后备存储库；它们直接从与创建它们的文件系统或卷相同的存储池中消耗盘空间；递归快照作为一个原子操作被快速创建；以及它们被一起创建(一次全部)或根本不创建。原子的快照操作的好处是，即使在后代文件系统中，快照数据也总是在一个一致的时间被拍摄。快照无法被直接访问，但是它们可以被克隆、备份、回滚等等。快照可以用于

在时间上“回滚”到拍摄快照时的点。

[0117] 克隆是初始内容与创建它的数据集相同的可写卷或文件系统。在ZFS系统中，始终从快照创建克隆。与快照一样，创建克隆几乎是即时的，最初不会消耗附加的盘空间。此外，可以对克隆进行快照。克隆只能从快照创建。在快照被克隆时，会在克隆和快照之间创建隐式依赖关系。即使克隆是在数据集层次结构中的其它位置创建的，只要克隆存在，就不能销毁原始快照。克隆不会继承创建它的数据集的特性。克隆最初与原始快照共享其所有盘空间。随着对克隆做出改变，它会使用更多盘空间。克隆对于分支并进行开发或故障排除是有用的——并且可以被提升以代替活(live)文件系统。克隆还可以用于在多个机器上复制文件系统。

[0118] 现在返回去参考图7B，流程图700-2示出了用于确定数据服务并将数据服务应用于数据块的方法。在判定框731处，如果开启或请求压缩，那么下一个框是740。在判定框732处，如果开启或请求加密，那么下一个框是750。在判定框733处，如果开启或请求去重，那么下一个框是738。在判定框734处，如果要拍摄快照，那么下一个框是775。框735返回到框716。图7B还图示了任何请求的数据服务的所需排序。必须首先执行压缩，然后执行加密、去重以及快照和克隆。当读取数据块时，必须采用相反的次序。

[0119] 如框720所指示的，与增量修改对应的数据对象和元数据可以存储在云对象存储库404中。在各种实施例中，云接口装备402可以创建、读取、转发、定义和/或以其它方式指定与增量修改对应的数据对象和元数据。如框722所指示的，在一些实施例中，数据对象和元数据的存储可以至少部分地由云接口设备502向云接口守护进程604发送请求而造成。如框724所指示的，由云接口设备502发送到云接口守护进程604的请求可以被转化成针对云客户端接口608的放置请求。在一些实施例中，云接口设备502发送的请求可以是放置请求；在其它实施例中，云接口守护进程604可以将云接口设备502发送的请求转化成放置请求。

[0120] 如框726所指示的，响应于由云接口设备502发送的请求，云接口守护进程604可以经由一个或多个网络上的对象协议与云对象存储库404通信，以使得与增量修改对应的数据对象和元数据被存储为新的云对象。如框728所指示的，云接口装备402可以鉴于存储在云对象存储库404中的与增量修改对应的数据对象和元数据来更新映射406。

[0121] 现在参考图7C，从框740开始在框742处描绘流程图700-3。图7C描绘了压缩数据块以保留存储空间的流程图。压缩是在如图2中所示DMU 218处的事务对象层中执行的。压缩通常被开启，因为它减少存储(云对象存储库404)和发送数据所需的资源。计算资源在压缩处理中在DMU 218中消耗，并且通常在该处理的逆转(解压)中被消耗。数数据压缩受到空间-时间复杂度权衡的影响。例如，压缩方案可能要求足够快的密集处理解压缩，以便在正在被解压缩时被消耗。数据压缩方案的设计涉及各种因素之间的权衡，包括压缩程度以及压缩和解压缩数据所需的计算资源。在框742处由DMU 218接收或检索压缩类型以压缩数据块。ZFS支持许多不同类型的压缩，包括但不限于LZ4 LZJB、GZIP和ZLE。在框744处，DMU 218使用压缩类型来压缩数据块。在判定框746处，确定是否存在来自将被写入云对象存储库404的树层次结构的更多需要被压缩的数据块。如果是这样——那么重复框744，直到DMU 218使用压缩类型压缩了所有块。一旦所有块都被压缩，框748就返回到图7B的框732。

[0122] 图7D描绘了在请求或开启加密的情况下加密数据块的处理的流程图700-4。在框752处，检索密码短语或AES密钥或将其提供给DMU 218。ZFS使用“卷绕式(wraparound)”加

密密钥系统，该系统使用本地存储的密码短语或AES密钥，然后随机生成用于加密数据块的加密密钥，如框754中所示。密码短语或AES密钥可以存储在任何本地存储装置中，包括ARC 224。加密本身并不会阻止数据块被盗用，而是会拒绝将消息内容给拦截器。在加密方案中，使用加密算法对预期数据块进行加密，生成只有在解密时才能被读取的密文。出于技术原因，加密方案通常使用由算法生成的伪随机加密密钥。原理上有可能在不拥有密钥的情况下解密消息，然而，对于精心设计的加密方案，需要大的计算资源和技巧。ZFS数据块使用AES(高级加密标准)加密，密钥长度为128、192和256。在框756处使用随机生成的加密密钥对数据块进行加密。在判定框758处，确定是否存在更多数据块需要加密，并且如果存在，那么在框756处对它们进行加密，直到不再有数据块要加密为止。然后，框759返回到图7B的框733，以确定数据块是否需要更多的数据服务处理。

[0123] 图7E描绘了对云对象存储库404中的数据块进行去重的流程图700-5。数据块复制是一种专门的数据压缩技术，用于消除重复数据块的重复副本。数据块去重用于提高存储装置利用率，并且也可以应用于网络数据传送，以减少必须被发送以存储在COW存储器的数据块的数量。在去重处理中，在分析处理期间识别并存储唯一(unique)数据块。随着分析的继续，将其它数据块与存储的副本进行比较，并且每当匹配发生时，冗余的数据块就被替换为指向所存储的存储数据块的小的引用。鉴于相同的数据块模式可能会发生数十次、数百次甚至数千次，使用去重大大减少了必须存储或传送的数据块的数量。这种类型的去重不同于针对图7C讨论的标准文件压缩所执行的去重。该压缩识别各个数据块内的短的重复子串，基于存储的去重的目的是校验大量数据并识别相同的整个数据块，以便仅存储它的一个副本。例如，考虑典型的电子邮件系统可能包含100个相同的1MB(兆字节)文件附件的实例，如果保存该附件的所有100个实例，那么需要100MB的存储空间。通过去重，实际只存储附件的一个实例；后续实例将向后引用保存的副本，去重比率约为100比1。因此，数据块去重可以减少云对象存储库404中所需的存储空间，并减少将数据块传送到云对象存储库404的网络上的压力。

[0124] 在图7D中，在框762处，第一处理是使用名称生成协议为数据块生成名称。在ZFS中，这包括使用校验和算法(诸如SHA256)。当对数据块执行校验和算法时，它生成对数据块的内容唯一的校验和。因此，如果任何其它数据块具有完全相同的内容，那么使用相同算法或名称协议的校验和将完全相同。这是ZFS数据去重的关键。在判定框764处，确定是否存在具有相同名称的现有数据块。这可以通过多种方式完成。一种是保持现有名称的本地表。然而，这会将数据块去重限制为仅对从本地发起的数据块进行去重。现有名称的表可以存储在对象存储库404上。对象存储库404上的表可以存储在客户端本地数据池中或全局存储并且对所有客户端可用。数据存储在对象存储库中的位置将影响可通过数据块去重实现的数据块压缩的量。例如，如果具有现有名称的表是全局的，那么即使存在使用云对象存储库404的多个客户端，也只需要将上面讨论的1MB文件的一个副本存储在云对象存储库404上。即使该1MB附件通过电子邮件“病毒化”并最终成为数千封电子邮件的附件，情况也是如此。利用云对象存储库404上的全局现有名称表，该1MB文件将仅在云对象存储库404上存储一次，但是可以由指向它的数千个元数据块引用。为此，在框766处，当利用数据块进行进一步的存储计算时，忽略具有与现有名称相同的名称的数据块，并且使用指向现有块的指针来创建元数据块以根据图7A中的框716和718的方法在框768处生成树。判定框770通过返回到

框762而使得对树中的尽可能多的数据块重复该处理。在判定框773处，如果已经请求了快照，那么在框774处，下一个框是图7F中的775。如果尚未请求快照，那么在772处的下一个框是图7A中的框720。

[0125] 图7F描绘了用于ZFS快照和克隆到云对象存储库404的方法的流程图700-6。ZFS快照是ZFS系统必不可少的备份机制。ZFS快照是树的“生命周期”中的某一时刻处的ZFS树层次结构的“照片”。如图3F中所讨论的，对于快照，时间基于根块(与根节点和超级块交替使用)的出生时间，该出生时间是针对该根块是在哪个进展中创建来表述的。每次生成数据块以进行存储时，都会发生进展。最终，在拍摄快照之前生成完整的树层次结构，如图3A-图3D和图7A的框716和718所述。如框780所描绘的，存储对根块的引用。根块和在该进展中活动的所有块都是快照的一部分，因为快照引用指向该根块，而根块指向通过每个较低级块的所有块，但它们不被存储为重复块。更确切地说，如框782中所示，树中可从树根块访问的所有块被标记为“不释放”，这在ZFS语法中将它们指定为“只读”。在ZFS存储COW系统的正常进展中，一旦块不活动——换句话说，它不被任何更高级别的块引用——它就可以被释放以用于其它存储需求。快照块必须保持不变，以使快照完全有用。当块被快照所引用的根块引用时，在删除该快照之前不能“释放”该块。这使得有可能备份到制作快照的点，直到快照被删除。快照可以存储在云对象存储库404、ARC 222、L2ARC 222-3或诸如系统存储装置228之类的任何其它本地存储装置中。使用ZFS文件系统，可以在分层树的进展中的特定实例处请求快照，并且可以在特定的周期性时间点自动生成快照。旧快照不会在创建新快照时被自动删除，因此即使创建了新快照，只要先前的快照尚未删除，就可以发生到先前的快照的备份。因此，快照启用增量备份，因为不必复制整个文件系统，实际上整个备份已经存在云对象存储库404中，如快照所引用的根块所指向的那样。快照引用所指向的根块将成为活动根块，并且所有后续根块和在快照之后的出生时间创建的块都可以被释放以供其它存储使用。

[0126] 在判定框784处，确定是否已请求克隆。如果尚未请求克隆，那么在框799处，快照处理结束，接着是图7A中的框720。如果已请求克隆(它是从快照创建的，并且如框786中所示)，那么克隆引用指向该快照指向的相同根块。始终从快照生成克隆，使得不能在删除克隆之前删除快照，如框788中所示。克隆用于各种目的——从同一数据存储集进行多个开发、实例化新虚拟机、解决问题等。为此，克隆必须能够在从克隆引用访问时作为COW。在这方面，克隆与快照不同——因为从快照不会发生COW。克隆在生成时不需要额外的存储能力，但随着克隆树上进行的进展，克隆将使用更多的存储。可以从克隆制作快照，就像可以从活动树制作快照一样，并且出于完全相同的原因。克隆不会继承根块数据的特性。克隆最终也可以被提升为活动树。在框799处，完成克隆生成，接着是图7A中的框720。

[0127] 图8是图示根据本公开某些实施例的处理增量修改的云接口装备402的示例的高级别图。除了从云中最佳恢复(不应用多个增量更新)的能力之外，某些实施例还可以提供非常高效的总是增量(incremental-always)(也称为永久增量(incremental forever))备份能力。业界的传统备份方法涉及将副本推送到云中。然而，根据本公开的某些实施例允许基于云的写时复制文件系统，其中仅将新数据对象写入云存储。只需要发送新数据和修改后的数据的云存储为从云提供商进行备份和还原提供了极为有效的解决方案。不需要修改旧数据对象。这些实施例与使用事务组的一致性模型一起促进具有永久增量备份的云存

储,其中始终可以确认一致性。

[0128] 为了说明,图8将云接口装备402描绘为已经创建了逻辑树300的本地实例的初始备份(例如,与备份逻辑树300A对应的云数据对象414)。在一些实施例中,云接口装备402可以利用活动树像(image)301或完整副本来创建备份。在初始地创建逻辑树300的本地实例的完全备份之后,可以利用各种事务对逻辑树300进行多个修改303。在图8中,图示了与增量修改303对应的数据对象和元数据的存储。作为示例,增量修改303被描绘为具有新的叶子节点324、326;新的非叶子节点328、330、332、334;以及新的根节点336。

[0129] 云接口装备402可以被配置为潜在地无限地创建增量备份。为此,云接口装备402的某些实施例可以利用快照COW处理(例如,如以上关于图3E和图7F所公开的)。在图8中描绘的示例中,可以利用像304来创建增量修改。在一些实施例中,像304可以与活动树像对应;在一些实施例中,像304可以与快照对应。利用像304,云接口装备402可以导致增量修改303A的存储,增量修改303A可以与云存储对象414A对应。根据基于云的COW处理,为与增量修改303对应的数据对象和元数据分配新的云对象414A,其中增量修改303的基于云的实例化被指示为增量修改303A。然后,新的根节点可以通过存储操作被置为修改后的逻辑树300A-1的根,以最终确定修改后的逻辑树300A-1的状态。修改后的逻辑树300A-1可以与被增量修改303A修改的逻辑树300A对应。

[0130] 通过存储增量修改303A,包含活动数据的逻辑树300A(被保存为云数据对象414)的块可以不被原地覆写。可以分配新的云数据对象414A,并且可以将修改后的/新的数据和元数据写入云数据对象414A。可以保留数据的先前版本,从而允许维护逻辑树300A的快照版本。在一些实施例中,可以在修改后的逻辑树300A-1及其相关联的快照之间共享任何未改变的数据。

[0131] 因此,利用诸如示例性快照301之类的快照和诸如与根336、336-1对应的超级块之类的超级块,可以仅针对自上次快照以来发生更改的节点的子集来更新云对象存储库404。并且增量303A的更新可以经由根节点挂接到云中的树的版本300A,使得可以访问整个修改后的逻辑树300A-1的任何部分。

[0132] 通过将增量快照发送到云数据存储库404,元数据可以被保留在数据实例内,这允许遍历树并获得在快照时的树的照片。同时,这允许非常精简的数据表示,其仅保留被指定/请求为要被快照保留的所引用的块的数据,使得仅需要将最少量的数据存储在树中以保留树的时间点像。每个增量可以与先前存储在云数据存储库404中的树300A合并,使得在发送每个增量之后,总是存在完整的当前数据表示。附加增量与树300A的每次合并可以导致单个数据实例,使得不需要完整的备份操作。

[0133] 虽然从性能角度来看是不必要的,但是在一些实施例中,可以周期性地进行完整备份以避免在客户端不期望的情况下必须汇集大量增量。对于一些实现方案,如果期望,那么可以保留多个版本的完整备份。因此,某些实施例提供对间隔、快照和备份的完全控制。有利地,某些实施例可以被配置为动态地自我调整增量间隔。例如,一些实施例可以最初根据第一间隔操作。在各种情况下,间隔可以是对本地树的每次写入操作、每半小时、每天等等。当流失率(例如,由混合云存储系统400监视的度量,其指示本地树的变化率)超过(或减小到)某个流失阈值时,混合云存储系统400可以自动过渡到不同的间隔。例如,如果流失率超过第一流失阈值,那么混合云存储系统400可以自动过渡到增量间隔的更大时间间隔。同

样，如果流失率减小到第一流失阈值或另一个流失阈值，那么混合云存储系统400可以过渡到增量间隔的更小时间间隔。某些实施例可以采用多个流失阈值来限制每个分级方案的增量频率。类似地，某些实施例可以至少部分地基于这样的流失阈值和/或增量尺寸和/或数量阈值来动态地自我调整完整备份间隔，这些阈值和/或增量尺寸和/或数量阈值在一些情况下可以是客户端定义的。

[0134] 将对象写入云并随后从云中读取对象的主要问题之一是不能保证读取对象的完整性。云存储存在数据运营商(incumbent)降级的风险(例如，存储丢失、传输失败、位衰减等等)。此外，由于涉及多个版本的对象，存在读取不期望版本的对象的风险。例如，可能读取期望的对象的先前版本，诸如在对象的更新未完成的情况下读取最近的版本。

[0135] 传统的对象存储体系架构依赖于数据的副本，其中最初满足法定数量(即，2个副本)并且异步地更新附加副本(例如，第三副本)。这提供了客户端可以在更新所有副本之前接收数据副本的可能性，从而获得数据的不一致视图。对象存储一致性的典型解决方案是确保在数据可用之前制作所有副本，但是使用该解决方案确保一致性通常是不现实的或不可实现的。其中对象的校验和存储在与对象一起的对象的元数据中或其它地方的引用数据库中的基于对象的方法将允许验证对象内容，但不会验证对象的正确版本。而且，可能使用对象版本控制并从单个位置进行校验的解决方案在某种程度上会损坏云存储的目的和意图。

[0136] 然而，根据本公开的某些实施例可以提供一致性模型，该一致性模型可以确保云中有保证的完整性并且可以确保根据最终一致的对象模型的始终一致的语义。例如，某些实施例可以通过使用本文公开的逻辑树来提供故障隔离和一致性。利用对云对象存储库的所有事务写入操作，可以更新自描述的Merkle树的元数据中的校验和。如上所述，与从其计算校验和的节点分开存储校验和确保每个树自动进行自我验证。在每个树层处，下面的节点由包括校验和的节点指针引用。因此，当从云中读出对象时，可以确定该对象是否正确。

[0137] 图9是图示根据本公开某些实施例的针对混合云存储系统400的某些特征的示例方法900的框图，这些特征确保云中有保证的完整性以及根据最终一致的对象模型的始终一致的语义。根据某些实施例，方法900可以如框902所示开始。然而，如上文所阐明的，本公开的教导可以以各种配置来实现，使得本文公开的方法的某些步骤的次序可以以任何合适的方式混洗或组合，并且可以取决于所选择的实现方案。而且，虽然为了描述可以分离以下步骤，但是应当理解的是，某些步骤可以同时或基本上同时执行。

[0138] 如框902所指示的，可以从应用202接收执行一个或多个特定操作(即，一个或多个事务)的兼容POSIX的请求。这种操作可以与读取或以其它方式访问数据对应。如框904所指示的，兼容POSIX的请求可以经由系统调用接口208从操作系统转发到DMU 218。如框906所指示的，DMU 218可以将对数据对象执行操作的请求直接转化成执行针对云对象存储库404的一个或多个读取操作的请求(即，一个或多个I/O请求)。如框908所指示的，SPA可以从DMU 218接收该(一个或多个)I/O请求。响应于该(一个或多个)请求，SPA可以发起从云对象存储库404读取一个或多个数据对象。

[0139] 如框910所指示的，云接口装备402可以接收(一个或多个)I/O请求，并且可以将对应的(一个或多个)云接口请求发送到云对象存储库404。在一些实施例中，云接口装备402可以将I/O请求转化成对应的对象接口请求。云接口设备502可以使用云对象存储库404中

的云存储对象414的映射406来协调对象接口请求。例如，云接口装备402可以识别并请求根据逻辑树被存储为云数据对象的文件的全部或一部分。

[0140] 如框912所指示的，云接口装备402可以响应于对象接口请求而接收(一个或多个)数据对象。如框914所指示的，可以利用来自逻辑树中的(一个或多个)父节点的(一个或多个)校验和来校验该(一个或多个)数据对象。在各种实施例中，校验可以由云接口装备402和/或I/O管道224执行。

[0141] 图10是图示根据本公开某些实施例的处理校验的云接口装备402的示例的高级别图。同样，在其它实施例中，系统400的另一个部件可以执行该校验。然而，在图10中，云接口装备402被描绘为已经访问了云对象存储库404以读取根据逻辑树300A-1存储的数据。在所描绘的示例中，云存储装备402被示为已经访问了与逻辑树300A-1对应的云对象414A。云存储装备402通过非叶子节点326-2、334-2的地址和根节点336-2访问了叶子节点324-2。

[0142] 如本文所述，当从逻辑树中读出节点时，使用逻辑树中较高级别的节点的指针来读取该节点。该指针包括预期要读取的数据的校验和，以便当从云中提取数据时，可以用校验和来校验该数据。可以计算数据的实际校验和，并将其与已经由云接口装备402和/或I/O管道224获得的预期校验和进行比较。在所描绘的示例中，非叶子节点326-2包括用于叶子节点324-2的校验和。

[0143] 在一些实施例中，可以与一个对象一起从云数据存储库接收校验和，并且与不同的对象一起接收要使用该校验和校验的数据。根据一些实施例，可以在接收具有要被校验的数据的该不同的对象之前，与分开的对象一起从云数据存储库接收校验和。一些实施例可以采用迭代对象接口请求处理，使得响应于特定对象接口请求，接收具有校验和的对象，并且响应于后续对象接口请求，接收具有要被校验的数据的对象。另外，对于一些实施例，可以使用与该特定对象接口请求一起接收的寻址信息来制作该后续对象接口请求，并且指向具有要被校验的数据的对象。对于一些实施例，不是迭代处理，而是从云数据存储库接收多个对象，之后可以对该多个对象执行校验。通过遍历逻辑树并基于父节点计算每个级别处的子节点的校验和，可以快速校验逻辑树的完整性。在替代实施例中，云接口装备402和/或I/O管道224可以在发起读取操作之前获得校验和和/或可以从另一个源获得校验和。例如，在这样的替代实施例中，校验和可以保留在映射406、逻辑树的快照和/或本地数据存储库中。

[0144] 再次参考图9，如框916所指示的，可以确定是否通过来自逻辑树中的(一个或多个)父节点的(一个或多个)校验和与(一个或多个)数据对象的(一个或多个)实际校验和来验证(一个或多个)数据对象。如框918所指示的，在(一个或多个)数据对象被验证的情况下，系统400的读取和/或进一步处理操作可以继续进行，因为已经确定数据未被损坏并且不是不正确的版本。在必要时，校验可以继续使用其它对象，直到所有对象都通过父对象指针元数据进行校验求和并验证为止。

[0145] 在(一个或多个)数据对象的实际校验和与预期校验和不匹配的情况下，可以识别错误状况。这是图10中所示的情况，其中对象D'(其与叶子节点324-2对应)的实际校验和与父节点326-2(其对应于对象C')指定的校验和不匹配。在错误状况的情况下，图9的处理流程可以继续前进到框920。实际校验和与预期校验和的不匹配可能与获得错误的不是最新的数据版本、由于云存储故障、位衰减和/或传输损失等导致的数据降级的情况对应。如框

920所指示的，可以发起补救。在一些实施例中，补救可以包括将一个或多个云接口请求重新发布到云对象存储库404。响应于每个重新发布的请求，处理流程可以返回到框912，在那里云接口装备402可以接收一个或多个数据对象，并且可以进行校验处理的另一次迭代。

[0146] 重新发布一个或多个云接口请求可以与云对象存储库404更努力地尝试找到所请求的对象的正确版本的请求对应。对于一些实施例，云接口装备402可以迭代地通过云节点/设备，直到检索到正确的版本。一些实施例可以迭代地校验可以存储在云对象存储库404中的树部分的先前版本的快照。一些实施例可以采用阈值，使得云接口装备402可以在满足阈值之后进行其它补救措施。阈值可以与一个或多个云接口请求的重新发布的次数对应。可替代地或附加地，阈值可以与对节点和/或快照的搜索范围的限制对应。例如，阈值可以控制在云接口装备402转向不同的补救措施之前要搜索多少节点和/或多少快照。

[0147] 云接口装备402可以采用的另一个补救措施是在多云实现方案中从另一个云对象存储库请求正确版本的数据，如框924所指示的。本文进一步描述了多云存储的某些实施例。可以校验第二云对象存储库的记录以确定数据的副本是否已经存储在第二云对象存储库中。在一些实施例中，该记录可以与特定于第二云对象存储库的另一个映射406对应。在确定数据的副本已经存储在第二云对象存储库中之后，云接口装备402可以向第二云对象存储库发起一个或多个对象接口请求以便检索期望的数据。

[0148] 利用这样的实施例，在已经满足对云对象存储库404的重新发出的请求的阈值而没有成功接收到正确版本之后，可以对第二云对象存储库做出对感兴趣的对象的请求。可替代地，作为第一默认值，一些实现方案可以求助于第二云对象存储库，而不是重新发出对云对象存储库404的请求。在任何情况下，响应于对第二云对象存储库的请求，处理流程都可以返回到框912，在那里云接口装备402可以接收一个或多个数据对象，并且可以进行校验处理的另一次迭代。在从第二云对象存储库检索出正确数据并通过验证的情况下，系统400的读取和/或进一步处理操作可以继续，如框918所指示的。此外，在已经从第二云对象存储库接收到正确数据以后，系统400可以将正确数据写入云对象存储库404，如框926所指示的。如关于图7所公开的，可以利用COW过程来实现正确数据的写入以应用增量修改。

[0149] 在一些实施例中，如果通过补救处理不能检索到正确的数据，那么可以返回错误消息和最近版本的数据，如框928所指示的。在一些情况下，最近版本的数据可能是无法检索的。例如，应当指向该数据的元数据可能被损坏，使得不能引用和检索最近版本的数据。在这些情况下，可以返回错误消息，而没有最近版本的数据。然而，当最近版本的数据可检索到时，可以也返回最近版本的数据。因而，利用端到端校验和模型，某些实施例可以覆盖从客户端到云并再次返回的端到端遍历。某些实施例不仅可以提供对整个树的校验求和并检测错误，而且还提供通过重新同步和数据清理来校正树的部分的能力。

[0150] 图11是根据本公开某些实施例的进一步图示混合云存储系统400-2的特征的简化示例的图。混合云存储系统400-2图示了混合存储池1100如何至少部分地由系统存储池416和虚拟存储池602-1形成。针对系统存储池416和虚拟存储池602-1中的每一个图示了用于读取操作和写入操作的流程。

[0151] 在一些实施例中，ARC 222可以包括ARC 222-2。ARC 222-2和系统存储块228-1可以促进系统存储池416的读取操作。如所指示的，ARC 222-2的某些实施例可以用DRAM实现。如还指示的，系统存储块228-1的某些实施例可以具有基于SAS/SATA的实现方案。对于一些

实施例,可以通过可以扩展高速缓存尺寸的L2ARC设备222-3来进一步促进系统存储池416的读取操作。对于一些实施例,ARC 222可以被引用为包括L2ARC设备222-3。如图11中所指示的,L2ARC设备222-3的某些实施例可以具有基于SSD的实现方案。系统存储块228-1和意图日志214-3可以促进系统存储池416的写入操作。如所指示的,意图日志214-3的某些实施例可以具有基于SSD的实现方案。

[0152] 对于本地系统,虚拟存储池602-1可以出现并且表现为逻辑盘。类似于系统存储池416,ARC 222-4和云存储对象块414-1可以促进虚拟存储池602-1的读取操作。对于一些实施例,ARC 222可以被引用为包括ARC 222-4,尽管ARC 222-4可以在一些实施例中用一个或多个分开的设备实现。如图11所指示的,ARC 222-4的某些实施例可以用DRAM实现。在一些实施例中,ARC 222-4可以是与ARC 222-2相同的高速缓存;在其它实施例中,ARC 222-4可以是虚拟存储池602-1的不同于ARC 222-2的分开的高速缓存。如所指示的,促进云存储对象块414-1的某些实施例可以具有基于HTTP的实现方案。

[0153] 对于一些实施例,L2ARC设备222-5还可以促进虚拟存储池602-1的读取操作。对于一些实施例,ARC 222可以被引用为包括L2ARC设备222-5。在一些实施例中,LSARC设备222-5可以是与L2ARC设备222-3相同的高速缓存;在其它实施例中,LSARC设备222-5可以是虚拟存储池602-1的不同于L2ARC设备222-3的分开的高速缓存设备。如所指示的,L2ARC设备222-5的某些实施例可以具有基于SSD或基于HDD的实现方案。

[0154] 可以通过意图日志214-4来促进虚拟存储池602-1对云存储对象块414-1的写入操作。在一些实施例中,意图日志214-4可以与意图日志214-3相同;在其它实施例中,意图日志214-4可以是与意图日志214-3分开且不同的。如所指示的,意图日志214-4的某些实施例可以具有基于SSD或基于HDD的实现方案。

[0155] 向云存储的过渡提供了若干优点(例如,成本、规模和地理位置),然而,当常规使用时,云存储具有一些限制。当应用客户端位于本机(on premise)而非共存于云中时,延时常常是常规技术的重要问题。然而,混合云存储系统400可以消除该问题。混合云存储系统400的某些实施例可以提供镜像特征,以促进性能、迁移和可用性。

[0156] 利用混合云存储系统400,可以最小化系统存储池416的读取操作的延时与虚拟存储池602-1的读取操作之间的差异。在一些实施例中,两个池的ARC和L2ARC设备可以是本地的实现方案。两个池的ARC和L2ARC设备处的延时可以等同或基本等同。例如,ARM的读取操作的典型延时可以是0.01ms或更小,并且L2ARC的典型延时可以是0.10ms或更小。来自云存储对象块414-1的读取操作的延时可以更高,但是混合云存储系统400可以智能地管理两个池以最小化该更高的延时。

[0157] 某些实施例可以提供具有文件系统语义的低延时、直接云访问。某些实施例可以促进从云存储运行,同时保留应用语义。某些实施例可以实现本地存储读取性能,同时保留云中所有数据的复制副本。为了提供这些特征,某些实施例可以利用本机高速缓存设备并充分利用混合存储池高速缓存算法。通过为云存储提供有效的高速缓存和文件系统语义,云存储可以被用于不止备份和恢复。混合存储系统可以使用“活”云存储。换句话说,通过智能地使用本机高速缓存设备,可以向本地系统提供全部性能的益处,而不必在本地保持完整的或多个副本。

[0158] 图12是根据本公开某些实施例的进一步图示混合云存储系统400-2的特征的简化

示例的图。利用有效的高速缓存设备和算法，混合云存储系统400-2可以在块级而不是文件级进行高速缓存，因此对大对象(例如，大型数据库)的访问不需要高速缓存整个对象。在各种实施例中，所描绘的示例可以与虚拟存储池602-1和系统存储池416中的一个或组合对应。

[0159] 混合云存储系统400-2可以采用自适应I/O暂存(stage)来捕获系统操作所需的大多数对象。混合云存储系统400-2可以配置多个高速缓存设备以提供自适应I/O暂存。在所描绘的示例中，自适应I/O暂存是利用ARC 222-5实现的。然而，在各种实施例中，混合云存储系统400-2可以被配置为使用多个ARC 222和/或L2ARC 222来提供自适应I/O暂存。虽然以下描述使用ARC 222作为示例，但是应当理解的是，各种实施例可以使用多个ARC 222以及一个或多个L2ARC 222来提供所公开的特征。

[0160] 在一些实施例中，ARC 222-5可以是自调谐(self-tuning)的，使得ARC 222-5可以基于I/O工作负载进行调整。举例来说，在混合云存储系统400-2在活模式下使用云存储的实施例中(不仅仅用于备份和迁移)，ARC 222-5还可以提供根据优先次序对对象进行暂存的高速缓存算法。用于高速缓存的优先次序可以与最近使用的(MRU)对象、最常用(MFU)对象、最不常用(LFU)对象和最近最少使用(LRU)对象对应。对于每个I/O操作，ARC 222-5可以确定是否需要对暂存的数据对象进行自我调整。要注意的是，在某些实施例中，L2ARC 225-5可以与ARC 222-5一起工作以促进一个或多个暂存。举例来说，可能具有比ARC 222-5更高的延时的L2ARC 225-5可以用于一个或多个较低排名的暂存，诸如LRU和/或LFU暂存。在一些实施例中，混合云存储系统400-2的另一个部件可以根据这些实施例引起高速缓存。举例来说，云存储装备402可以协调读取请求和写入请求的高速缓存和服务。另外，根据一些实施例，云存储装备402可以包括(一个或多个)ARC 222-5、(一个或多个)L2ARC 222-5和/或意图日志214-4。

[0161] 对于每个I/O操作，ARC 222-5可以调整先前在某种程度上暂存的一个或多个对象的暂存。至少，该调整可以包括更新对至少一个对象的访问的跟踪。调整可能包括降级到较低的暂存、驱逐，或提升到更高的暂存。用于提升和降级的过渡标准对于从当前暂存到另一个暂存或到驱逐的每个过渡可以是不同的。如本文所公开的，ARC 222-5可以具有由于存储器压力而从高速缓存驱逐存储器缓冲器以维持高吞吐量和/或满足使用率阈值的能力。

[0162] 对于给定的I/O操作，如果与I/O操作对应的一个或多个对象尚未被暂存为MRU对象，那么可以将该一个或多个对象新暂存为MRU对象。然而，如果与I/O操作对应的一个或多个对象已经被暂存为MRU对象，那么ARC 222-5可以将过渡标准应用于该一个或多个对象以确定是否将该一个或多个对象过渡到不同的暂存。如果未满足过渡标准，那么在服务于该I/O操作时不需要暂存的改变。

[0163] 图13是图示根据本公开某些实施例的针对混合云存储系统400-3的用于高速缓存管理和云延时掩盖的某些特征的示例方法1300的框图。根据某些实施例，方法1300可以如框1302所指示的那样开始。然而，如上所述，本文公开的方法的某些步骤可以以任何合适的方式混洗、组合和/或同时或基本上同时执行，并且可取决于所选择的实现方案。

[0164] 如框1302所指示的，可以从应用202接收执行(一个或多个)特定操作(即，(一个或多个)事务)的(一个或多个)兼容POSIX的请求。这种操作可以与读取、写入和/或修改数据对应。如框1304所指示的，该(一个或多个)兼容POSIX的请求可以经由系统调用接口208从

操作系统转发到DMU 218。如框1306所指示的，DMU 218可以将对对象执行操作的请求直接转化成执行针对云对象存储库404内的物理位置的I/O操作的请求。DMU 218可以将I/O请求转发到SPA。

[0165] 如框1308所指示的，SPA可以从DMU 218接收I/O请求。并且，响应于请求，SPA可以发起执行I/O操作。如框1310所指示的，在写入操作的情况下，SPA可以使用COW过程来发起将对象写入云对象存储库404。例如，云接口装备402可以关于云对象存储库404继续上面(例如，鉴于图7)公开的COW过程。如框1312所指示的，在读取操作的情况下，SPA可以发起对象的读取。可以检查ARC 222-5以查找一个或多个请求的对象。在一些实施例中，如框1314所指示的，可以确定与(一个或多个)I/O请求对应的一个或多个通过验证的数据对象是否存在于ARC 222-5中。这可以包括SPA首先确定与(一个或多个)读取请求对应的一个或多个对象是否可从ARC 222-5中检索。然后，如果这样的一个或多个对象是可检索到的，那么可以利用来自逻辑树中一个或多个父节点的一个或多个校验和来校验该(一个或多个)对象。在各种实施例中，校验可以由ARC 222-5、云存储装备402和/或I/O管道224执行。如框1316所指示的，在(一个或多个)数据对象通过验证的情况下(或者，在一些未采用验证的实施例中，在命中的简单情况下)，系统400的读取和/或进一步处理操作可以继续。

[0166] 在一些实施例中，如框1318所指示的，在(一个或多个)数据对象未通过验证的情况下(或者，在一些未采用验证的实施例中，在没有命中的简单情况下)，SPA可以发起从本地存储228读取一个或多个对象。对于一些实施例，指向一个或多个对象的指针可以被高速缓存并用于从本地存储228读取一个或多个对象。如果没有高速缓存这种指针，那么一些实现方案可以不检查本地存储228以查找一个或多个对象。如果一个或多个对象是可检索到的，那么可以使用来自逻辑树中一个或多个父节点的一个或多个校验和来校验该一个或多个对象。再次，在各种实施例中，校验可以由ARC 222-5、云存储装备402和/或I/O管道224执行。如框1320所指示的，在(一个或多个)对象通过验证的情况下，处理流程可以过渡到框1316，并且系统400的读取和/或进一步处理操作可以继续。

[0167] 如框1320所指示的，在(一个或多个)数据对象未通过验证的情况下(或者，在一些未采用验证的实施例中，在没有命中的简单情况下)，处理流程可以过渡到框1322。如框1322所指示的，SPA可以发起从云对象存储库404读取一个或多个数据对象。在各种实施例中，可以通过DMU 218、ARC 222-5、I/O管道224、和/或云存储装备402中的一个或组合来执行读取的发起。从云对象存储库404读取一个或多个对象可以包括先前在本文(例如，关于图9)公开的步骤。此类步骤可以包括上面详细描述的向云接口装备402发布(一个或多个)I/O请求、使用云存储对象414的映射406向云对象存储库404发送对应的(一个或多个)云接口请求、响应于对象接口请求而接收(一个或多个)数据对象等当中的一个或组合。如框1324所指示的，可以确定是否已经从云对象存储库404检索到通过验证的对象。再次，这可以涉及先前在本文(例如，关于图9)公开的步骤，其中确定是否通过来自逻辑树中的(一个或多个)父节点的(一个或多个)校验和来验证(一个或多个)数据对象。

[0168] 在(一个或多个)数据对象的实际校验和与预期校验和不匹配的情况下，处理流程可以过渡到块1326，其中可以发起补救处理。这可以涉及先前在本文(例如，关于图9)公开的步骤，其中公开了补救处理。然而，在数据通过验证的情况下，处理流程可以过渡到框1316，并且系统400的读取和/或进一步处理操作可以继续。

[0169] 如框1328所指示的，在检索到一个或多个对象之后，可以调整高速缓存暂存。在某些情况下，高速缓存暂存的调整可以包括将一个或多个对象新高速缓存为MRU对象，如框1330所指示的。如果与给定I/O操作对应的一个或多个对象尚未被暂存为MRU对象，那么可以将该一个或多个对象新暂存为MRU对象。

[0170] 然而，在某些情况下，当与I/O操作对应的一个或多个对象已经被暂存为MRU、MFU、LFU或LRU对象时，ARC 222-5可以将过渡标准应用于一个或多个对象，以确定是否将该一个或多个对象过渡到不同的暂存，如框1332所指示的。然而，如果不满足过渡标准，那么在服务于I/O操作时可能不需要暂存的改变。

[0171] 在一些实施例中，对象的暂存可以至少部分地是对象的访问的新近度(recency)的函数。如框1334所指示的，在一些实施例中，高速缓存暂存的调整可以包括更新一个或多个新近度属性。ARC 222-5可以为一个或多个新对象定义新近度属性，以便跟踪一个或多个对象的访问的新近度。新近度属性可以对应于时间参数和/或顺序参数，时间参数(例如，通过绝对时间、系统时间、时间差等)指示与一个或多个对象对应的最后访问时间，顺序参数指示与其它对象的新近度属性可以与其进行比较的一个或多个对象对应的访问计数。

[0172] 在各种实施例中，过渡标准可以包括为了使对象有资格从当前阶段过渡而定义的一个或多个新近度阈值。例如，ARC 222-5可以至少部分地基于指派给一个或多个对象的新近度属性的值来确定是否应当将一个或多个对象过渡到LFU或LRU暂存(或驱逐)。在一些实施例中，新近度阈值可以是动态阈值，其根据为一个或多个暂存中的其它对象定义的新近度属性来调整。例如，当为暂存的对象定义的新近度属性的值按升序或降序排序时，新近度阈值可以是为已经被暂存为MFU对象的任何对象定义的任何新近度属性的最低值的函数。

[0173] 附加地或可替代地，在一些实施例中，对象的暂存可以至少部分地是对对象的访问频率的函数。如框1336所指示的，在一些实施例中，高速缓存暂存的调整可以包括更新一个或多个频率属性。对于特定的I/O操作，ARC 222-5可以递增为一个或多个对象定义的频率属性，以便跟踪一个或多个对象的访问频率。频率属性可以指示在任何合适的时间段上的访问次数，该时间段可以是绝对时间段、基于活动的时间段(例如，用户会话，或者自满足最小活动阈值的最后访问活动量以来的时间)等等。

[0174] 在各种实施例中，过渡标准可以包括为了使对象有资格从当前级过渡而定义的一个或多个频率阈值。例如，在频率属性的值改变之后，ARC 222-5可以确定一个或多个对象是否应当被暂存为MFU对象(或者作为另一暂存中的对象)。可以至少部分地基于将更新后的频率属性与频率阈值进行比较来做出这样的确定。在一些实施例中，频率阈值可以是动态阈值，其根据为其它暂存的对象(例如，暂存为MFU对象或另一暂存中的对象)定义的频率属性进行调整。例如，当为暂存的对象定义的频率属性的值按升序或降序排序时，频率阈值可以是为已经被暂存为MFU对象的任何对象定义的任何频率属性的最低值的函数。

[0175] 如框1338所指示的，附加地或可替代地，根据一些实施例，高速缓存暂存的调整可以包括指定一个或多个其它暂存属性。暂存属性可以指示操作类型。对于一些实施例，对象的暂存可以至少部分地是操作类型的函数。例如，暂存算法可以采用写入操作与读取操作的区分，使得只有通过读取操作访问的对象可以被暂存为MFU对象。在这样的实施例中，用写入操作引用的对象可以最初被维护为MRU对象，并且此后根据LFU暂存标准进行降级。可替代地，在此类实施例中，用写入操作引用的对象可以最初被维护为MRU对象，并且此后根

据LRU暂存标准进行降级，然后受到驱逐，从而有效地跳过LFU暂存。作为另一个替代方案，用写入操作引用的对象可以最初被维护为MRU对象并且然后受到驱逐，从而有效地跳过LFU暂存和LRU暂存。对于此类替代方案，ARC 222-5将用于将云对象提交到云对象存储库404的写入操作区分为对于后续读取操作而言是不太可能需要的，因此如果操作出现，那么允许这种潜在操作引起云访问延时。

[0176] 暂存属性可以指示数据类型。附加地或可替代地，在一些实施例中，对象的暂存可以至少部分地是数据类型的函数。例如，一些实施例可以相对于数据给予元数据更高的优先级。该更高的优先级可以包括保留所有元数据对象，以及使数据对象被暂存。可替代地，该更高的优先级可以包括对于元数据对象与数据对象应用用于暂存过渡(从当前暂存的提升和/或降级)的不同标准。例如，为数据对象定义的有资格降级的阈值(例如，新近度、频率和/或类似阈值)可以比为元数据对象定义的有资格降级的阈值更低(并且因此更容易被满足)。可替代地，其它实施例可以相对于元数据给予数据更高的优先级。对于一些实施例，可以将云对象的一部分定义为始终被高速缓存，而不管使用频率如何。

[0177] 暂存属性可以指示操作特点。附加地或可替代地，在一些实施例中，对象的暂存可以至少部分地是读取操作特点的函数。例如，一些实施例可以给予具有尺寸特点(使得读取的对象的尺寸满足尺寸阈值)的读取操作更高的优先级。附加地或可替代地，可以赋予具有序列特点(使得读取的对象序列满足序列阈值)的读取操作更高的优先级。因而，对于高速缓存，大的顺序流式读取操作可以被赋予比较小的、更加孤立的读取操作更高的优先级。以这种方式，避免了大的顺序流式读取操作的更高的云访问延时。

[0178] ARC 222-5的某些实施例可以针对每个暂存采用不同的功能、过渡标准和/或阈值。在一些实施例中，ARC 222-5可以采用暂存评分系统。一些实施例可以用数字表达式(例如，暂存得分)对对象进行评分。暂存得分可以反映对象相对于任何合适标准(诸如过渡标准)的资格。例如，对象的给定暂存得分可以是根据诸如访问频率、访问新近度、操作类型、数据类型、操作特点、对象尺寸等标准的累积评分。可以针对每个标准对给定对象进行评分。例如，诸如频率属性、新近度属性等属性的相对较大的值可以被赋予更大的得分。同样，可以鉴于其它标准和优先级指派得分。可以使用对象的累积暂存得分以及存储在高速缓存中的其它对象的暂存得分来根据优先次序对对象进行排名。再次，优先次序可以用于将对象过渡到不同的暂存和/或驱逐。

[0179] ARC 222-5可以调整暂存和存储在其中的对象以满足一个或多个高速缓存使用率和容量约束。例如，给定比如说1TB DRAM的高速缓存设备容量，ARC 222-5可以调整暂存和存储在其中的对象以维持80％的最大高速缓存使用率。此外，一些实施例可以调整暂存和存储在其中的对象以满足一个或多个速度约束。例如，ARC 222-5可以监视吞吐量以在给定本地访问和云访问的情况下维持可接受的访问延时量(例如，平均访问延时)，以便确定是否应当采用更多或更少的高速缓存来满足一个或多个延时容忍度。鉴于这种调整，暂存和存储在其中的对象的调整可以包括为每个暂存应用不同的函数和阈值，以便按优先次序对对象进行排序。ARC 222-5可以利用优先次序将存储的对象移向驱逐，以便满足调整约束。

[0180] 如框1340所指示的，在一些实施例中，高速缓存暂存调整可以包括过渡到不同的高速缓存模式。一些实施例可以动态地改变操作模式，以便在满足使用率和延时约束的同时进行负载平衡。初始或默认操作模式可以与以活方式从云进行操作对应，使得首先从高

速缓存访问对象，然后如果必要的话就从云访问对象。ARC 222-5的一些实施例可以最初(例如，在会话或时间段内)高速缓存利用I/O操作访问的所有对象，然后当高速缓存使用率满足一个或多个阈值时过渡到采用暂存。过渡到暂存可以是一个或多个辅助操作模式的增量。例如，暂存可以最初降级到MRU和MFU暂存，然后在满足一个或多个高速缓存使用率阈值(其可以初始是并且低于最大高速缓存使用率阈值)时扩展到一个或多个其它暂存。

[0181] 鉴于高速缓存使用率接近使用率约束并且满足一个或多个使用率阈值，可以使用一个或多个附加操作模式递增地应用某些过渡标准。例如，最初可能不会区分与写入操作对应的对象。然而，当高速缓存使用率接近使用率约束时，可以在满足一个或多个使用率阈值之后应用该区分标准。

[0182] 作为另一个示例，随着高速缓存使用率进一步接近使用率约束并且满足一个或多个使用率阈值，对于较低排名的暂存(例如，LRU和/或LFU暂存)，混合云存储系统400-2可以开始利用具有一个或多个L2ARC设备的扩展高速缓存(其可以与一个或多个低延时侧对应)。作为又一个示例，随着高速缓存使用率进一步接近使用率约束并满足一个或多个使用率阈值，混合云存储系统400-2可以开始利用本地存储228以便符合关于一个或多个第三操作模式的延时容忍度。通过更具体的示例，不是驱逐大的顺序流式读取操作而不为将来的低延时访问提供供应，而是该操作的扩展和对应对象的访问频率可以足以满足尺寸和频率阈值，使得这些对象被过渡到本地存储228。通过这种方式，混合云存储系统400-2可以保持对象可用于本地延时读取操作，同时释放高速缓存容量用于其它低延时访问(它们可能需要其它大的顺序的读取操作)并且如果再次调用大的顺序读取操作的话就避免花费云延时。在对象级别对本地存储228和云存储的这种选择性利用还可以促进在大部分时间使用高速缓存的同时掩盖云延时，以及促进云存储和本地存储228之间的负载平衡，以便在延时容忍度内操作。在各种实施例中，这种三叉式(trifurcated)存储调整可以作为退守(fallback)操作模式或者作为用于具有某些特点的某些类型的操作的初始默认被发起。

[0183] 因而，某些实施例可以至少部分地基于对象访问的特点来更改高速缓存模式和技术。某些实施例可以充分利用高速缓存特征、云存储和本地存储228来掩盖基于云的操作的延时。在此类实施例中，大多数操作可以从高速缓存得到服务，其高速缓存命中率通常超过90％或更多，这导致大多数时间是本地延时。如果任何本地对象丢失或损坏，那么对象的云副本可以被访问。对于一些实施例，可以仅当没有高速缓存命中并且不能从本地存储228服务读取请求时，才需要从云对象存储库404进行读取。

[0184] 在一些实施例中，代替上面公开的ARC校验、本地存储校验和/或云对象存储校验，映射406可以被用于识别一个或多个感兴趣的对象的位置。如上所述，映射406可以包括对象目录，并且可以维护随着每个I/O操作更新的对象状态。云对象状态可以保存在基于每个对象进行索引的索引、表、索引组织的表等等当中。对象状态可以包括对象高速缓存状态。对象高速缓存状态可以指示ARC、L2ARC、自适应暂存，本地存储和/或云存储中的任何一个或组合中的对象的位置。通过利用映射406，云接口设备402可以直接识别一个或多个感兴趣的对象的位置。在一些实施例中，云接口设备402可以仅在根据ARC校验没有命中的情况下利用映射406。

[0185] 在一些实施例中，作为高速缓存的补充或替代，智能池管理包括保持与云对象存储404连续同步的镜像。至少部分地通过将云对象数据存储库支持为虚拟设备，某些实施例

可以提供本地存储和云存储之间的镜像。利用本地存储对云存储进行镜像可以实现本地存储读性能，同时在云中保留所有数据的复制副本。通过使用镜像，可以向本地系统提供全部性能的好处，而不必在本地保留多个副本。如果任何本地数据丢失或损坏，那么可以访问数据的云副本。同步镜像云和本地设备可以促进更高级别的性能。

[0186] 为了促进这种同步镜像，某些实施例可以包括镜像VDEV。图14图示了根据本公开某些实施例的混合云存储系统400的促进同步镜像的示例网络文件系统200-2的实例。文件系统200-2可以与文件系统200-1对应，但是具有直接集成到ZFS控制栈中的镜像管理。除了关于文件系统200-1公开的内容之外，文件系统200-2还可以包括镜像VDEV 1402，镜像VDEV 1402促进数据的云副本，但是具有读取数据的本地访问时间/速度。

[0187] 在一些实施例中，镜像VDEV 1402可以与ZFS文件系统体系架构内的另一个VDEV类型的设备驱动器接口的一个或多个VDEV。ZFS可以直接与镜像VDEV 1402通信，镜像VDEV 1402可以位于文件系统200-2的驱动器层直接上方的虚拟设备层处，并且在一些实施例中，与ZFS体系架构内的设备驱动器接口的抽象对应。文件系统200-2可以将镜像VDEV 1402创建为用于I/O操作的漏斗(funnel)。在一些实施例中，镜像VDEV 1402可以是文件系统200-2的其它部件可以主要与之通信的点。例如，在一些实施例中，来自事务对象层的通信可以通过镜像VDEV 1402到达物理层。更具体而言，来自DMU 218的通信可以被指引到I/O管道224和镜像VDEV 1402。响应于此类通信，镜像VDEV 1402可以将通信指引到其它VDEV(诸如VDEV 226)和云接口设备502。照此，镜像VDEV 1402可以针对本地存储228和云对象存储404协调I/O操作。

[0188] 在一些实施例中，镜像VDEV 1402可以针对本地存储228和云对象存储404仅协调写入操作，使得读取操作不需要通过镜像VDEV 1402。对于此类实施例，诸如VDEV 226之类的其它VDEV和云接口设备502可以绕过镜像VDEV 1402以进行读取操作。在替代实施例中，镜像VDEV 1402可以协调所有I/O操作。

[0189] 有利地，镜像VDEV 1402可以协调写入操作，使得经由一个或多个VDEV 226针对本地存储228以及经由云接口设备502针对云对象存储库404同步地执行每个写入操作。每个I/O操作的这种同步镜像是在对象级别而不是文件级别执行的。每次I/O操作的数据复制使得混合云存储系统400能够实现掩盖云访问延时的本地存储读取性能。作为默认，混合云存储系统400可以从本地存储228读取，以避免为绝大多数读取操作付出云延时。仅当混合云存储系统400确定本地数据丢失或损坏时，混合云存储系统400才需要访问云对象存储库404以读取期望数据的云副本。可以基于对象执行此类例外(exception)，以便最小化云访问的延时。

[0190] 图15是图示根据本公开某些实施例的针对混合云存储系统400的用于同步镜像和云延时掩盖的某些特征的示例方法1500的框图。根据某些实施例，方法1500可以如框1502所示开始。然而，如上所述，本文公开的方法的某些步骤可以以任何合适的方式混洗、组合和/或同时或基本上同时执行，并且可以取决于所选择的实现方案。

[0191] 如框1502所指示的，可以从应用202接收执行(一个或多个)特定操作(即，(一个或多个)事务)的(一个或多个)兼容POSIX的请求。这种操作可以与写入和/或修改数据。如框1504所指示的，(一个或多个)兼容POSIX的请求可以经由系统调用接口208从操作系统转发到DMU 218。如框1506所指示的，DMU 218可以将对数据对象执行操作的请求直接转化成执

行针对系统存储池416和云对象存储库404内的物理位置的写入操作的请求(即,I/O请求)。DMU 218可以将I/O请求转发到SPA。

[0192] SPA的镜像VDEV 1402可以接收I/O请求(例如,通过ARC 222和I/O管道224)。如框1508所指示的,镜像VDEV 1402可以以每I/O操作为基础发起具有同步复制的数据对象的写入。镜像VDEV 1402的一部分可以指向本地存储,并且镜像VDEV 1402的一部分可以指向云存储装备402的云接口设备502。如框1510所指示的,镜像VDEV 1402可以将写入操作的第一实例指引到VDEV 226中的一个或多个。在一些实施例中,如框1514所指示的,上面(例如,鉴于图3A-图3D)公开的COW过程可以继续,以便将数据对象写入本地系统存储。

[0193] 如框1512所指示的,镜像VDEV 1402可以将写入操作的第二实例指引到云存储装备402的云接口设备502。在一些实施例中,如框1516所指示的,上面公开的COW过程可以继续,以便将数据对象写入本地系统存储。例如,方法1500可以过渡到框712或方法700的另一个步骤。

[0194] 随着关于本地存储228和关于云对象存储库404同步执行的每个写入操作,混合云存储系统400此后可以智能地协调读取操作,以便实现掩盖云访问的延时的本地存储读取性能。在完成复制数据存储之后的任何合适的时间,混合云存储系统400可以协调此类读取操作。再次参考图15,如框1518所指示的,可以从应用202接收执行一个或多个特定操作(即,一个或多个事务)的兼容POSIX的请求。这种操作可以与读取或以其它方式访问数据对应。如框1520所指示的,兼容POSIX的请求可以经由系统调用接口208从操作系统转发到DMU 218。如框1522所指示的,DMU 218可以将对数据对象执行操作的请求直接转化成执行针对本地存储228的一个或多个读取操作的请求(即,一个或多个I/O请求)。SPA可以从DMU 218接收(一个或多个)I/O请求。如框1524所指示的,响应于该(一个或多个)请求,SPA可以发起从本地存储228读取一个或多个数据对象。在一些实施例中,可以首先校验ARC 222以查找一个或多个数据对象的高速缓存版本,并且,如果没有找到高速缓存版本,那么可以尝试从本地存储228读取一个或多个数据对象。

[0195] 如框1526所指示的,可以确定是否存在与(一个或多个)I/O请求对应的一个或多个通过验证的数据对象。这可以包括SPA首先确定与一个或多个I/O请求对应的一个或多个对象是否可从本地存储228检索。然后,如果这样的一个或多个对象是可检索到的,那么可以使用来自逻辑树中的(一个或多个)父节点的校验和来校验该(一个或多个)对象。在各种实施例中,校验可以由VDEV 226、镜像VDEV 1402和/或I/O管道224中的一个或多个执行。如框1528所指示的,在(一个或多个)数据对象通过验证的情况下,系统400的读取和/或进一步处理操作可以继续进行,因为已经确定数据未被损坏并且不是不正确的版本。

[0196] 然而,在SPA确定不存在与(一个或多个)I/O请求对应的通过验证的数据对象的情况下,处理流程可以过渡到框1530。这种确定可以与不存在与一个或多个I/O请求对应的可检索数据对象对应,在这种情况下,可以识别错误状况。类似地,这种确定可以与与一个或多个I/O请求对应的(一个或多个)数据对象的实际校验和与预期校验和的不匹配对应,在这种情况下也可以识别错误状况。在任一情况下,SPA都可以发起从云对象存储库404读取一个或多个数据对象,如框1530所指示的。在各种实施例中,可以通过DMU 218、镜像VDEV 1402、I/O管道224和/或云接口装备402中的一个或组合来执行读取的发起。

[0197] 从云对象存储库404读取一个或多个数据对象可以包括先前在本文(例如,关于图

9)公开的步骤。此类步骤可以包括上面详细描述的向云接口装备402发布(一个或多个)I/O请求、使用云存储对象414的映射406向云对象存储库404发送对应的(一个或多个)云接口请求、响应于对象接口请求而接收(一个或多个)数据对象等当中的一个或组合。如框1532所指示的,可以确定是否已经从云对象存储库404检索到通过验证的数据对象。再次,这可以涉及先前在本文(例如,关于图9)公开的步骤,其中确定是否通过来自逻辑树中的(一个或多个)父节点的(一个或多个)校验和来验证(一个或多个)数据对象。

[0198] 在数据通过验证的情况下,处理流程可以过渡到框1528,并且系统400的读取和/或进一步处理操作可以继续。另外,如框1534所指示的,可以将正确的数据写入本地系统存储。在各种实施例中,校正处理可以由DMU 218、镜像VDEV 1402、I/O管道224和/或云接口装备402中的一个或组合来执行。在一些实施例中,处理流程可以过渡到框1514,其中可以继续COW过程以便将正确的数据写入本地系统存储。

[0199] 然而,在(一个或多个)数据对象的实际校验和与预期校验和不匹配的情况下,处理流程可以过渡到框1536,其中可以发起补救处理。这可以涉及先前在本文(例如,关于图9)公开的步骤,其中公开了补救处理。例如,补救处理可以包括重新发布一个或多个云接口请求、从另一个云对象存储库请求正确版本的数据,等等。

[0200] 有利地,当由混合云存储系统400维护的数据量超过某个量时,从上面公开的镜像模式过渡到根据关于图12和图13公开的实施例的高速缓存模式会变得更具成本效益。某些实施例可以自动进行这种过渡。混合云存储系统400可以以镜像技术开始,一直继续到达到一个或多个阈值。可以鉴于存储使用率来定义此类阈值。例如,当本地存储容量的使用率达到阈值百分比(或绝对值、相对值等)时,混合云存储系统400可以过渡到自适应I/O暂存。以这种方式,混合云存储系统400可以通过改变操作模式来平衡施加在本地存储上的负载。然后,混合云存储系统400可以是最相关的X量的数据(例如,10TB等等),并将其余数据分流到云存储。这种负载平衡可以允许更少的存储设备,同时适应不断增加的数据存储量。

[0201] 图16绘出了用于实现实施例之一的分布式系统1600的简化图。在所示实施例中,分布式系统1600包括一个或多个客户端计算设备1602、1604、1606和1608,其被配置为通过一个或多个网络1610执行和操作客户端应用,诸如web浏览器、专有客户端(例如,Oracle Forms)等。服务器1612可以经由网络1610与远程客户端计算设备1602、1604、1606和1608通信地耦合。

[0202] 在各种实施例中,服务器1612可以适于运行由系统的一个或多个部件提供的一个或多个服务或软件应用。在一些实施例中,这些服务可以作为基于web的服务或云服务被提供,或者在软件即服务(SaaS)模型下被提供给客户端计算设备1602、1604、1606和/或1608的用户。操作客户端计算设备1602、1604、1606和/或1608的用户又可以利用一个或多个客户端应用来与服务器1612交互以利用由这些部件提供的服务。

[0203] 在图中绘出的配置中,系统1600的软件部件1618、1620和1622被示出为在服务器1612上实现。在其它实施例中,系统1600的一个或多个部件和/或由这些部件提供的服务也可以由客户端计算设备1602、1604、1606和/或1608中的一个或多个来实现。然后,操作客户端计算设备的用户可以利用一个或多个客户端应用来使用由这些部件提供的服务。这些部件可以用硬件、固件、软件或其组合来实现。应该认识到的是,各种不同的系统配置是可能的,其可能与分布式系统1600不同。图中所示的实施例因此是用于实现实施例系统的分布

式系统的一个示例，而不是要进行限制。

[0204] 客户端计算设备1602、1604、1606和/或1608可以是便携式手持设备(例如，iPhone®、蜂窝电话、iPad®、计算平板电脑、个人数字助理(PDA))或可穿戴设备(例如，Google Glass®头戴式显示器)，运行诸如Microsoft Windows Mobile®和/或各种移动操作系统(诸如iOS、Windows Phone、Android、BlackBerry、Palm OS等)的软件，并且启用互联网、电子邮件、短消息服务(SMS)、Blackberry®或其它通信协议。客户端计算设备可以是通用个人计算机，作为示例，包括运行各种版本的Microsoft Windows®、Apple Macintosh®和/或Linux操作系统的个人计算机和/或膝上型计算机。客户端计算设备可以是运行任何各种可商业获得的UNIX®或类UNIX操作系统(包括但不限于各种GNU/Linux操作系统，诸如例如Google Chrome OS)的工作站计算机。替代地或附加地，客户端计算设备1602、1604、1606和1608可以是能够通过(一个或多个)网络1610通信的任何其它电子设备，诸如瘦客户端计算机、启用互联网的游戏系统(例如，具有或不具有Kinect®手势输入设备的微软Xbox游戏控制台)和/或个人消息传送设备。

[0205] 虽然示例性分布式系统1600被示出为具有四个客户端计算设备，但是可以支持任何数量的客户端计算设备。其它设备(诸如具有传感器的设备等)可以与服务器1612交互。

[0206] 分布式系统1600中的(一个或多个)网络1610可以是本领域技术人员熟悉的、可以利用任何各种可商业获得的协议支持数据通信的任何类型的网络，其中协议包括但不限于TCP/IP(传输控制协议/网际协议)、SNA(系统网络体系架构)、IPX(互联网报文交换)、AppleTalk，等等。仅仅作为示例，(一个或多个)网络1610可以是局域网(LAN)，诸如基于以太网、令牌环等的LAN。(一个或多个)网络1610可以是广域网和互联网。它可以包括虚拟网络，包括但不限于虚拟专用网(VPN)、内联网、外联网、公共交换电话网(PSTN)、红外网络、无线网络(例如，依据电子电气学会(IEEE)802.11协议套件、Bluetooth®和/或任何其它无线协议当中任意一种操作的网络)；和/或这些和/或其它网络的任意组合。

[0207] 服务器1612可以由一个或多个通用计算机、专用服务器计算机(作为示例，包括PC(个人计算机)服务器、UNIX®服务器、中档服务器、大型计算机、机架安装的服务器等)、服务器场、服务器集群或任何其它适当的布置和/或组合组成。在各种实施例中，服务器1612可以适于运行在前述公开中所描述的一个或多个服务或软件应用。例如，服务器1612可以与用于执行以上根据本公开的实施例描述的处理的服务器对应。

[0208] 服务器1612可以运行包括以上讨论的操作系统当中任意一种的操作系统，以及任何可商业获得的服务器操作系统。服务器1612还可以运行任何各种附加的服务器应用和/或中间层应用，包括HTTP(超文本传输协议)服务器、FTP(文件传输协议)服务器、CGI(公共网关接口)服务器、JAVA®服务器、数据库服务器，等等。示例性数据库服务器包括但不限于从Oracle、Microsoft、Sybase、IBM(国际商业机器)等可商业获得的那些数据库服务器。

[0209] 在一些实现方案中，服务器1612可以包括一个或多个应用，以分析和整合从客户端计算设备1602、1604、1606和1608的用户接收到的数据馈送和/或事件更新。作为示例，数据馈送和/或事件更新可以包括，但不限于，Twitter®馈送、Facebook®更新或者从一

个或多个第三方信息源接收到的实时更新和连续数据流，其可以包括与传感器数据应用、金融报价机、网络性能测量工具(例如，网络监视和流量管理应用)、点击流分析工具、汽车交通监视等相关的实时事件。服务器1612还可以包括一个或多个应用，以经由客户端计算设备1602、1604、1606和1608的一个或多个显示设备显示数据馈送和/或实时事件。

[0210] 分布式系统1600还可以包括一个或多个数据库1614和1616。数据库1614和1616可以驻留在各种位置中。作为示例，数据库1614和1616中的一个或多个可以驻留在服务器1612本地的(和/或驻留在服务器1612中的)非瞬态存储介质上。替代地，数据库1614和1616可以远离服务器1612，并且经由基于网络的连接或专用的连接与服务器1612通信。在一组实施例中，数据库1614和1616可以驻留在本领域技术人员熟悉的存储区域网络(SAN)中。类似地，用于执行服务器1612所具有的功能的任何必要的文件都可以适当地本地存储在服务器1612上和/或远程存储。在一组实施例中，数据库1614和1616可以包括适于响应于SQL格式的命令而存储、更新和检索数据的关系数据库，诸如由Oracle提供的数据库。

[0211] 图17是根据本公开的实施例的系统环境1700的一个或多个部件的简化框图，通过该系统环境1700，由实施例系统的一个或多个部件提供的服务可以作为云服务提供。在所示实施例中，系统环境1700包括可以由用户使用以与提供云服务的云基础设施系统1702交互的一个或多个客户端计算设备1704、1706和1708。客户端计算设备可以被配置为操作客户端应用，诸如web浏览器、专有客户端应用(例如，Oracle Forms)或某种其它应用，这些应用可以由客户端计算设备的用户用来与云基础设施系统1702交互以使用由云基础设施系统1702提供的服务。

[0212] 应该认识到的是，图中描绘的云基础设施系统1702可以具有除了所描绘的那些之外的其它部件。另外，图中所示的实施例仅是可以结合本发明的实施例的云基础设施系统的一个示例。在一些其它实施例中，云基础设施系统1702可以具有比图中所示更多或更少的部件、可以组合两个或更多个部件、或者可以具有不同的部件配置或布置。

[0213] 客户端计算设备1704、1706和1708可以是与上面针对1602、1604、1606和1608所描述的设备类似的设备。虽然示例性系统环境1700被示出具有三个客户端计算设备，但是任何数量的客户端计算设备可以被支持。诸如具有传感器的设备等的其它设备可以与云基础设施系统1702交互。

[0214] (一个或多个)网络1710可以促进客户端1704、1706和1708与云基础设施系统1702之间的数据通信和交换。每个网络可以是本领域技术人员所熟悉的可以使用各种商业上可获得的协议(包括上面针对(一个或多个)网络1610所描述的那些协议)中的任何一种支持数据通信的任何类型的网络。云基础设施系统1702可以包括一个或多个计算机和/或服务器，其可以包括上面针对服务器1612所描述的那些计算机和/或服务器。

[0215] 在某些实施例中，由云基础设施系统提供的服务可以包括按需对云基础设施系统的用户可用的许多服务，诸如在线数据存储和备份解决方案、基于Web的电子邮件服务、被托管的办公室(office)套件和文档协作服务、数据库处理、受管理的技术支持服务等。由云基础设施系统提供的服务可以动态扩展以满足云基础设施系统的用户的需要。由云基础设施系统提供的服务的具体实例化在本文中被称为“服务实例”。一般而言，从云服务提供商的系统经由通信网络(诸如互联网)对用户可用的任何服务被称为“云服务”。通常，在公共云环境中，构成云服务提供商的系统的服务器和系统与客户自己的本地服务器和系统不

同。例如，云服务提供商的系统可以托管应用，并且用户可以经由诸如互联网的通信网络按需订购和使用应用。

[0216] 在一些示例中，计算机网络云基础设施中的服务可以包括对存储装置、被托管的数据库、被托管的Web服务器、软件应用或由云供应商向用户提供的其它服务的受保护的计算机网络访问，或者如本领域中另外已知的那样。例如，服务可以包括通过互联网对云上的远程存储装置进行密码保护的访问。作为另一个示例，服务可以包括基于Web服务的被托管的关系数据库和脚本语言中间件引擎，以供联网的开发人员私有使用。作为另一个示例，服务可以包括对在云供应商的网站上托管的电子邮件软件应用的访问。

[0217] 在某些实施例中，云基础设施系统1702可以包括以自助服务、基于订阅、弹性可扩展、可靠、高度可用和安全的方式递送给客户的应用、中间件和数据库服务产品的套件。这种云基础设施系统的示例是由本受让人提供的Oracle公共云。

[0218] 在各种实施例中，云基础设施系统1702可以适于自动供应、管理和跟踪客户对由云基础设施系统1702供给的服务的订阅。云基础设施系统1702可以经由不同的部署模型来提供云服务。例如，可以依据公共云模型提供服务，其中云基础设施系统1702被销售云服务的组织拥有(例如，被Oracle拥有)，并且服务对一般公众或不同行业的企业可用。作为另一个示例，可以依据私有云模型来提供服务，其中云基础设施系统1702仅针对单个组织操作，并且可以为该组织内的一个或多个实体提供服务。还可以依据社区云模型来提供云服务，其中云基础设施系统1702和由云基础设施系统1702提供的服务由相关社区中的若干组织共享。云服务还可以依据混合云模型被提供，该混合云模型是两个或更多个不同模型的组合。

[0219] 在一些实施例中，由云基础设施系统1702提供的服务可以包括在软件即服务(SaaS)类别、平台即服务(PaaS)类别、基础设施即服务(IaaS)类别或包括混合服务的其它服务类别下提供的一个或多个服务。客户经由订阅订单可以订购由云基础设施系统1702提供的一个或多个服务。云基础设施系统1702然后执行处理以提供客户的订阅订单中的服务。

[0220] 在一些实施例中，由云基础设施系统1702提供的服务可以包括但不限于应用服务、平台服务和基础设施服务。在一些示例中，应用服务可以由云基础设施系统经由SaaS平台提供。SaaS平台可以被配置为提供落入SaaS类别的云服务。例如，SaaS平台可以提供在集成开发和部署平台上构建和递送按需应用套件的能力。SaaS平台可以管理和控制用于提供SaaS服务的底层软件和基础设施。通过利用由SaaS平台提供的服务，客户可以利用在云基础设施系统上执行的应用。客户可以获取应用服务，而无需客户购买单独的许可和支持。可以提供各种不同的SaaS服务。示例包括但不限于为大型组织提供销售绩效管理、企业集成和业务灵活性的解决方案的服务。

[0221] 在一些实施例中，平台服务可以由云基础设施系统经由PaaS平台提供。PaaS平台可以被配置为提供落入PaaS类别的云服务。平台服务的示例可以包括但不限于使组织(诸如Oracle)能够在共享的公共体系架构上整合现有应用以及充分利用平台提供的共享服务来构建新应用的能力的服务。PaaS平台可以管理和控制用于提供PaaS服务的底层软件和基础设施。客户可以获取由云基础架构系统提供的PaaS服务，而无需客户购买单独的许可和支持。平台服务的示例包括但不限于Oracle Java云服务(JCS)、Oracle数据库云服务

(DBCS)等。

[0222] 通过利用由PaaS平台提供的服务,客户可以采用由云基础设施系统支持的编程语言和工具,并且还控制所部署的服务。在一些实施例中,由云基础设施系统提供的平台服务可以包括数据库云服务、中间件云服务(例如,Oracle融合中间件服务)和Java云服务。在一个实施例中,数据库云服务可以支持共享服务部署模型,该模型使得组织能够汇集数据库资源并且以数据库云的形式向客户供应数据库即服务。在云基础设施系统中,中间件云服务可以为客户提供开发和部署各种业务应用的平台,并且Java云服务可以为客户提供部署Java应用的平台。

[0223] 各种不同的基础设施服务可以由云基础设施系统中的IaaS平台提供。基础设施服务促进底层计算资源(诸如存储装置、网络和其它基础计算资源)的管理和控制,以供客户利用由SaaS平台和PaaS平台提供的服务。

[0224] 在某些实施例中,云基础设施系统1702还可以包括基础设施资源1730,用于向云基础设施系统的客户提供用于提供各种服务的资源。在一个实施例中,基础设施资源1730可以包括预先集成和优化的硬件(诸如服务器、存储装置和联网资源)的组合,以执行由PaaS平台和SaaS平台提供的服务。在一些实施例中,云基础设施系统1702中的资源可以由多个用户共享并且根据需要动态重新分配。此外,可以将资源分配给在不同时区的用户。例如,云基础设施系统1730可以使在第一时区中的第一组用户能够在指定的小时数内利用云基础设施系统的资源,并且然后使相同资源能够被重新分配给位于不同时区的另一组用户,从而使资源的利用率最大化。

[0225] 在某些实施例中,可以提供由云基础设施系统1702的不同部件或模块以及由云基础设施系统1702提供的服务共享的多个内部共享服务1732。这些内部共享服务可以包括但不限于:安全和身份服务、集成服务、企业储存库服务、企业管理器服务、病毒扫描和白名单服务、高可用性、备份和恢复服务、启用云支持的服务、电子邮件服务、通知服务、文件传输服务等。在某些实施例中,云基础设施系统1702可以提供云基础设施系统中的云服务(例如,SaaS、PaaS和IaaS服务)的综合管理。在一个实施例中,云管理功能可以包括用于供应、管理和跟踪由云基础设施系统1702接收到的客户订阅等的能力。

[0226] 在某些实施例中,如图中所绘出的,云管理功能可以由一个或多个模块提供,诸如订单管理模块1720、订单编排模块1722、订单供应模块1724、订单管理和监视模块1726,以及身份管理模块1728。这些模块可以包括一个或多个计算机和/或服务器或者使用一个或多个计算机和/或服务器来提供,这些计算机和/或服务器可以是通用计算机、专用服务器计算机、服务器场、服务器集群或任何其它适当的布置和/或组合。

[0227] 在示例性操作1734中,使用客户端设备(诸如客户端设备1704、1706或1708)的客户可以通过请求由云基础设施系统1702提供的一个或多个服务并且下订阅由云基础设施系统1702供应的一个或多个服务来的订单来与云基础设施系统1702交互。在某些实施例中,客户可以访问云用户界面(UI)(云UI 1712、云UI 1714和/或云UI 1716)并经由这些UI下订阅订单。云基础设施系统1702响应于客户下订单而接收到的订单信息可以包括识别客户以及客户想要订阅的云基础设施系统1702供应的一个或多个服务的信息。

[0228] 在客户下订单之后,经由云UI 1712、1714和/或1716接收订单信息。在操作1736处,订单存储在订单数据库1718中。订单数据库1718可以是由云基础设施系统1718操作和

与其它系统元件一起操作的若干数据库之一。在操作1738处，订单信息被转发到订单管理模块1720。在一些情况下，订单管理模块1720可以被配置为执行与订单相关的计费和记账功能，诸如验证订单、以及在验证后预订订单。

[0229] 在操作1740处，将关于订单的信息传送到订单编排模块1722。订单编排模块1722可以利用订单信息为客户下的订单编排服务和资源的供应。在一些情况下，订单编排模块1722可以使用订单供应模块1724的服务来编排资源的供应以支持所订阅的服务。

[0230] 在某些实施例中，订单编排模块1722使得能够管理与每个订单相关联的业务过程并应用业务逻辑来确定订单是否应该进行到供应。在操作1742处，在接收到新订阅的订单时，订单编排模块1722向订单供应模块1724发送分配资源并配置履行订阅订单所需的那些资源的请求。订单供应模块1724使得能够为客户订购的服务分配资源。订单供应模块1724提供在由云基础设施系统1700提供的云服务和用于供应用于提供所请求的服务的资源的物理实现层之间的抽象层。因此，订单编排模块1722可以与实现细节(诸如服务和资源是否实际上即时供应或预先供应并仅在请求后才分配/指派)隔离。

[0231] 在操作1744处，一旦供应了服务和资源，就可以通过云基础设施系统1702的订单供应模块1724向客户端设备1704、1706和/或1708上的客户发送所提供的服务的通知。在操作1746处，订单管理和监视模块1726可以管理和跟踪客户的订阅订单。在一些情况下，订单管理和监视模块1726可以被配置为收集订阅订单中的服务的使用统计信息，诸如，所使用的存储量、传输的数据量、用户的数量，以及系统运行时间量和系统停机时间量。

[0232] 在某些实施例中，云基础设施系统1700可以包括身份管理模块1728。身份管理模块1728可以被配置为提供身份服务，诸如云基础设施系统1700中的访问管理和授权服务。在一些实施例中，身份管理模块1728可以控制关于希望利用由云基础设施系统1702提供的服务的客户的信息。这样的信息可以包括认证这些客户的身份的信息以及描述这些客户被授权相对于各种系统资源(例如，文件、目录、应用、通信端口、存储器段等)执行哪些动作的信息。身份管理模块1728还可以包括对关于每个客户的描述性信息以及关于如何和由谁来访问和修改这些描述性信息的管理。

[0233] 图18示出了其中可以实现本发明的各种实施例的示例性计算机系统1800。系统1800可以用于实现上述任何计算机系统。如图所示，计算机系统1800包括经由总线子系统1802与多个外围子系统通信的处理单元1804。这些外围子系统可以包括处理加速单元1806、I/O子系统1808、存储子系统1818和通信子系统1824。存储子系统1818包括有形计算机可读存储介质1822和系统存储器1810。

[0234] 总线子系统1802提供用于让计算机系统1800的各种部件和子系统按意图彼此通信的机制。虽然总线子系统1802被示意性地示出为单条总线，但是总线子系统的替代实施例可以利用多条总线。总线子系统1802可以是若干种类型的总线结构中的任何一种，包括存储器总线或存储器控制器、外围总线、以及使用任何各种总线体系架构的局部总线。例如，这种体系架构可以包括工业标准体系架构(ISA)总线、微通道体系架构(MCA)总线、增强型ISA(EISA)总线、视频电子标准协会(VESA)局部总线和外围部件互连(PCI)总线，其可以被实现为按IEEE P1386.1标准制造的Mezzanine总线。

[0235] 可以被实现为一个或多个集成电路(例如，常规微处理器或微控制器)的处理单元1804控制计算机系统1800的操作。一个或多个处理器可以被包括在处理单元1804中。这些

处理器可以包括单核或多核处理器。在某些实施例中，处理单元1804可以被实现为一个或多个独立的处理单元1832和/或1834，其中在每个处理单元中包括单核或多核处理器。在其它实施例中，处理单元1804也可以被实现为通过将两个双核处理器集成到单个芯片中形成的四核处理单元。

[0236] 在各种实施例中，处理单元1804可以响应于程序代码执行各种程序并且可以维护多个并发执行的程序或进程。在任何给定的时间，要被执行的程序代码中的一些或全部代码可以驻留在(一个或多个)处理器1804中和/或存储子系统1818中。通过适当的编程，(一个或多个)处理器1804可以提供上述各种功能。计算机系统1800可以附加地包括处理加速单元1806，其可以包括数字信号处理器(DSP)、专用处理器，等等。在一些实施例中，处理加速单元1806可以包括如本文所公开的加速引擎或者与加速引擎一起工作以改进计算机系统功能。

[0237] I/O子系统1808可以包括用户接口输入设备和用户接口输出设备。用户接口输入设备可以包括键盘、诸如鼠标或轨迹球的定点设备、结合到显示器中的触摸板或触摸屏、滚动轮、点击轮、拨盘、按钮、开关、键盘、具有语音命令识别系统的音频输入设备、麦克风以及其它类型的输入设备。用户接口输入设备可以包括，例如，运动感测和/或手势识别设备，诸如的Microsoft **Kinect®**运动传感器，其使得用户能够使用手势和语音命令通过自然用户接口来控制诸如的Microsoft **Xbox®** 360游戏控制器的输入设备并与之交互。用户接口输入设备也可以包括眼睛姿势识别设备，诸如从用户检测眼睛活动(例如，当拍摄照片和/或做出菜单选择时的“眨眼”)并且将眼睛姿势转换为到输入设备(例如，Google **Glass®**)中的输入的Google **Glass®**眨眼检测器。此外，用户接口输入设备可以包括使用户能够通过语音命令与语音识别系统(例如，**Siri®**导航器)交互的语音识别感测设备。

[0238] 用户接口输入设备也可以包括但不限于三维(3D)鼠标、操纵杆或指向棒、游戏面板和绘图板，以及音频/视频设备，诸如扬声器、数码相机、数码摄像机、便携式媒体播放器、网络摄像头、图像扫描仪、指纹扫描仪、条形码阅读器3D扫描仪、3D打印机、激光测距仪和视线跟踪设备。此外，用户接口输入设备可以包括，例如，医学成像输入设备，诸如计算机断层扫描、磁共振成像、正电子发射断层摄影术、医疗超声设备。用户接口输入设备也可以包括，例如，诸如MIDI键盘、数字乐器等的音频输入设备。

[0239] 用户接口输出设备可以包括显示子系统、指示灯，或者诸如音频输出设备的非可视显示器，等等。显示子系统可以是阴极射线管(CRT)、诸如使用液晶显示器(LCD)或等离子显示器的平板设备、投影设备、触摸屏，等等。一般而言，术语“输出设备”的使用意在包括用于从计算机系统1800向用户或其它计算机输出信息的所有可能类型的设备和机制。例如，用户接口输出设备可以包括，但不限于，可视地传达文本、图形和音频/视频信息的各种显示设备，诸如监视器、打印机、扬声器、耳机、汽车导航系统、绘图仪、语音输出设备，以及调制解调器。

[0240] 计算机系统1800可以包括包含软件元件、被示为当前位于系统存储器1810中的存储子系统1818。系统存储器1810可以存储可加载并且可在处理单元1804上执行的程序指令，以及在这些程序的执行期间所产生的数据。取决于计算机系统1800的配置和类型，系统存储器1810可以是易失性的(诸如随机存取存储器(RAM))和/或非易失性的(诸如只读存储

器(ROM)、闪存存储器,等等)。RAM通常包含可被处理单元1804立即访问和/或目前正被处理单元1804操作和执行的数据和/或程序模块。在一些实现方案中,系统存储器1810可以包括多种不同类型的存储器,例如静态随机存取存储器(SRAM)或动态随机存取存储器(DRAM)。在一些实现方案中,诸如包含有助于在启动期间在计算机系统1800的元件之间传送信息的基本例程的基本输入/输出系统(BIOS),通常可以被存储在ROM中。作为示例,但不是限制,系统存储器1810也示出了可以包括客户端应用、web浏览器、中间层应用、关系数据库管理系统(RDBMS)等的应用程序1812,程序数据1814,以及操作系统1816。作为示例,操作系统1816可以包括各种版本的Microsoft**Windows®**,Apple**Macintosh®**和/或Linux操作系统、各种可商业获得的**UNIX®**或类UNIX操作系统(包括但不限于各种GNU/Linux操作系统、Google**Chrome®**操作系统等)和/或诸如iOS、**Windows®** Phone、**Android®** OS、**BlackBerry®** 10OS和**Palm®** OS操作系统的移动操作系统。

[0241] 存储子系统1818也可以提供用于存储提供一些实施例的功能的基本编程和数据结构的有形计算机可读存储介质。当被处理器执行时提供上述功能的软件(程序、代码模块、指令)可以被存储在存储子系统1818中。这些软件模块或指令可以被处理单元1804执行。存储子系统1818也可以提供用于存储根据本发明被使用的数据的储存库。

[0242] 存储子系统1800也可以包括可被进一步连接到计算机可读存储介质1822的计算机可读存储介质读取器1820。与系统存储器1810一起并且,可选地,与其相结合,计算机可读存储介质1822可以全面地表示用于临时和/或更持久地包含、存储、发送和检索计算机可读信息的远程、本地、固定和/或可移除存储设备加存储介质。

[0243] 包含代码或代码的部分的计算机可读存储介质1822也可以包括本领域已知或使用的任何适当的介质,包括存储介质和通信介质,诸如但不限于,以用于信息的存储和/或传输的任何方法或技术实现的易失性和非易失性、可移除和不可移除介质。这可以包括有形的计算机可读存储介质,诸如RAM、ROM、电可擦除可编程ROM(EEPROM)、闪存存储器或其它存储器技术、CD-ROM、数字多功能盘(DVD)或其它光学储存器、磁带盒、磁带、磁盘储存器或其它磁存储设备,或者其它有形的计算机可读介质。这也可以包括非有形的计算机可读介质,诸如数据信号、数据传输,或者可以被用来发送期望信息并且可以被计算系统1800访问的任何其它介质。

[0244] 作为示例,计算机可读存储介质1822可以包括从不可移除的非易失性磁介质读取或写到其的硬盘驱动器、从可移除的非易失性磁盘读取或写到其的磁盘驱动器、以及从可移除的非易失性光盘(诸如CD ROM、DVD和Blu-**Ray®**盘或其它光学介质)读取或写到其的光盘驱动器。计算机可读存储介质1822可以包括,但不限于,**Zip®**驱动器、闪存卡、通用串行总线(USB)闪存驱动器、安全数字(SD)卡、DVD盘、数字音频带,等等。计算机可读存储介质1822也可以包括基于非易失性存储器的固态驱动器(SSD)(诸如基于闪存存储器的SSD、企业闪存驱动器、固态ROM等)、基于易失性存储器的SSD(诸如固态RAM、动态RAM、静态RAM)、基于DRAM的SSD,磁阻RAM(MRAM)SSD,以及使用基于DRAM和闪存存储器的SSD的组合的混合SSD。盘驱动器及其关联的计算机可读介质可以为计算机系统1800提供计算机可读指令、数据结构、程序模块及其它数据的非易失性存储。

[0245] 通信子系统1824提供到其它计算机系统和网络的接口。通信子系统1824用作用于

从其它系统接收数据和从计算机系统1800向其它系统发送数据的接口。例如，通信子系统1824可以使计算机系统1800能够经由互联网连接到一个或多个设备。在一些实施例中，通信子系统1824可以包括用于访问无线语音和/或数据网络的射频(RF)收发器部件(例如，使用蜂窝电话技术，诸如18G、4G或EDGE(用于全球演进的增强型数据速率)的先进数据网络技术，WiFi(IEEE 802.11系列标准)，或其它移动通信技术，或其任意组合)、全球定位系统(GPS)接收器部件和/或其它部件。在一些实施例中，作为无线接口的附加或者替代，通信子系统1824可以提供有线网络连接(例如，以太网)。

[0246] 在一些实施例中，通信子系统1824也可以代表可以使用计算机系统1800的一个或多个用户接收结构化和/或非结构化数据馈送1826、事件流1828、事件更新1830等形式的输入通信。作为示例，通信子系统1824可被配置为实时地从社交网络和/或其它通信服务的用户接收数据馈送1826，诸如Twitter®馈送、Facebook®更新、诸如丰富站点摘要(RSS)馈送的web馈送和/或来自一个或多个第三方信息源的实时更新。

[0247] 此外，通信子系统1824也可被配置为接收连续数据流形式的数据，这可以包括本质上可以是连续的或无界的没有明确终止的实时事件的事件流1828和/或事件更新1830。产生连续数据的应用的示例可以包括，例如，传感器数据应用、金融报价机、网络性能测量工具(例如，网络监视和流量管理应用)、点击流分析工具、汽车流量监视，等等。通信子系统1824也可被配置为向一个或多个数据库输出结构化和/或非结构化数据馈送1826、事件流1828、事件更新1830，等等，这一个或多个数据库可以与耦合到计算机系统1800的一个或多个流式数据源计算机通信。

[0248] 计算机系统1800可以是各种类型之一，包括手持便携式设备(例如，iPhone®蜂窝电话、iPad®计算平板电脑、PDA)、可穿戴设备(例如，GoogleGlass®头戴式显示器)、PC、工作站、大型机、信息站、服务器机架、或任何其它数据处理系统。由于计算机和网络的不断变化的本质，在图中绘出的计算机系统1800的描述仅仅要作为具体的示例。具有比图中绘出的系统更多或更少部件的许多其它配置是可能的。例如，定制的硬件也可以被使用和/或特定的元素可以用硬件、固件、软件(包括applets)或其组合来实现。另外，也可以采用到诸如网络输入/输出设备之类的其它计算设备的连接。基于本文提供的公开内容和示教，本领域普通技术人员将认识到实现各种实施例的其它方式和/或方法。

[0249] 在前面的描述中，出于解释的目的，阐述了许多具体细节以便提供对本发明的各种实施例的透彻理解。然而，对于本领域技术人员清楚的是，可以在没有这些具体细节中的一些的情况下实践本发明的实施例。在其它情况下，以框图形式示出了众所周知的结构和设备。

[0250] 以上描述仅提供示例性实施例，并且不旨在限制本公开的范围、适用性或配置。更确切地说，示例性实施例的前述描述将为本领域技术人员提供用于实现示例性实施例的使能描述。应当理解的是，在不脱离所附权利要求中阐述的本发明的精神和范围的情况下，可以对元件的功能和布置进行各种改变。

[0251] 在前面的描述中给出了具体细节以提供对实施例的透彻理解。然而，本领域普通技术人员将理解的是，可以在没有这些具体细节的情况下实践这些实施例。例如，电路、系统、网络、处理和其它部件可能已经以框图形式示出为部件，以免用不必要的细节模糊实施例。在其它情况下，可能已经示出了众所周知的电路、处理、算法、结构和技术而没有不必要

的细节,以避免模糊实施例。

[0252] 而且,要注意的是,各个实施例可能已被描述为处理,该处理被描绘为流程图、数据流程图、结构图或框图。虽然流程图可能已将操作描述为顺序处理,但许多操作可以并行或同时执行。此外,可以重新布置操作的次序。处理在其操作完成时终止,但可以有其它未包含在图中的步骤。处理可以与方法、函数、过程、子例程、子程序等对应。当处理与函数对应时,其终止可以与函数返回到调用函数或主函数对应。

[0253] 术语“计算机可读介质”包括但不限于便携式或固定存储设备、光学存储设备、无线信道以及能够存储、包含或携带(一个或多个)指令和/或数据的各种其它介质。代码段或机器可执行指令可以表示过程、函数、子程序、程序、例程、子例程、模块、软件包、类或指令、数据结构或程序语句的任何组合。代码段可以通过传递和/或接收信息、数据、自变量、参数或存储器内容而耦合到另一个代码段或硬件电路。信息、自变量、参数、数据等可以经由任何合适的手段传递、转发或发送,包括存储器共享、消息传递、令牌传递、网络传输等。

[0254] 此外,实施例可以通过硬件、软件、固件、中间件、微代码、硬件描述语言或其任何组合来实现。当在软件、固件、中间件或微代码中实现时,用于执行必要任务的程序代码或代码段可以存储在机器可读介质中。处理器可以执行必要的任务。

[0255] 在前述说明书中,参考其具体实施例对本发明的各方面进行了描述,但是本领域技术人员将认识到的是,本发明不限于此。上述公开的各个特征和方面可以被单独使用或联合使用。另外,在不脱离本说明书的更广泛精神和范围的情况下,实施例可以在除本文所述的那些环境和应用之外的任何数目的环境和应用中被使用。相应地,本说明书和附图应当被认为是说明性的而不是限制性的。

[0256] 此外,出于说明的目的,以特定顺序描述了方法。应该认识到的是,在替代实施例中,可以以与所描述的顺序不同的顺序执行方法。还应该认识到的是,上述方法可以由硬件部件执行,或者可以以机器可执行指令的序列被实施,机器可执行指令可以用于使机器(诸如编程有指令的通用或专用处理器或逻辑电路)执行方法。这些机器可执行指令可以被存储在一个或多个机器可读介质上,诸如CD-ROM或其它类型的光盘、软盘、ROM、RAM、EPROM、EEPROM、磁卡或光卡、闪存、或适合于存储电子指令的其它类型的机器可读介质。替代地,可以通过硬件和软件的组合来执行方法。

[0257] 而且,权利要求中的术语具有其普通的普遍含义,除非专利权人另有明确和清晰的定义。如权利要求中所使用的,不定冠词“一”或“一个”在本文中被定义为表示特定物品引入的元件中的一个或多于一个;而定冠词“该”的后续使用并不是要否定那个含义。此外,阐明权利要求中的不同元素的诸如“第一”、“第二”等序数术语的使用并不旨在向已向其应用该序数术语的元素赋予系列中的特定位置、或任何其它顺序项或次序。

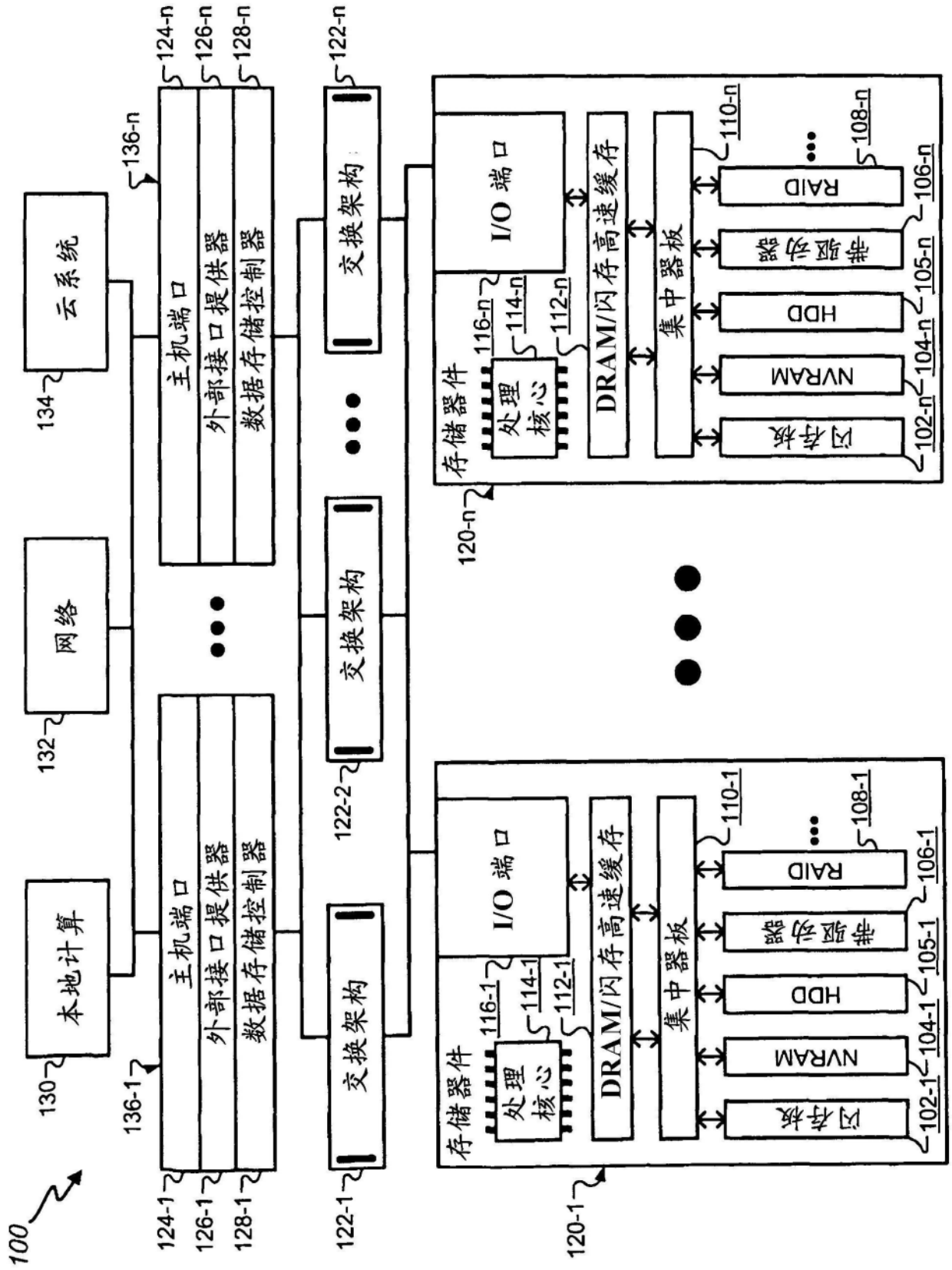

100
本地计算
130
网络
132
云系统
134
主机端口
外部接口提供器
数据存储控制器
136-1
124-1
126-1
128-1
136-n
124-n
126-n
128-n
交换架构
122-1
122-2
122-n
存储器件
处理核心
116-1
114-1
112-1
I/O 端口
DRAM/闪存高速缓存
集中器板
110-1
RAID
108-1
HDD
NVRAS
闪存
102-1
104-1
105-1
106-1
120-1
116-n
114-n
112-n
110-n
108-n
104-n
105-n
106-n
102-n
120-n

图1

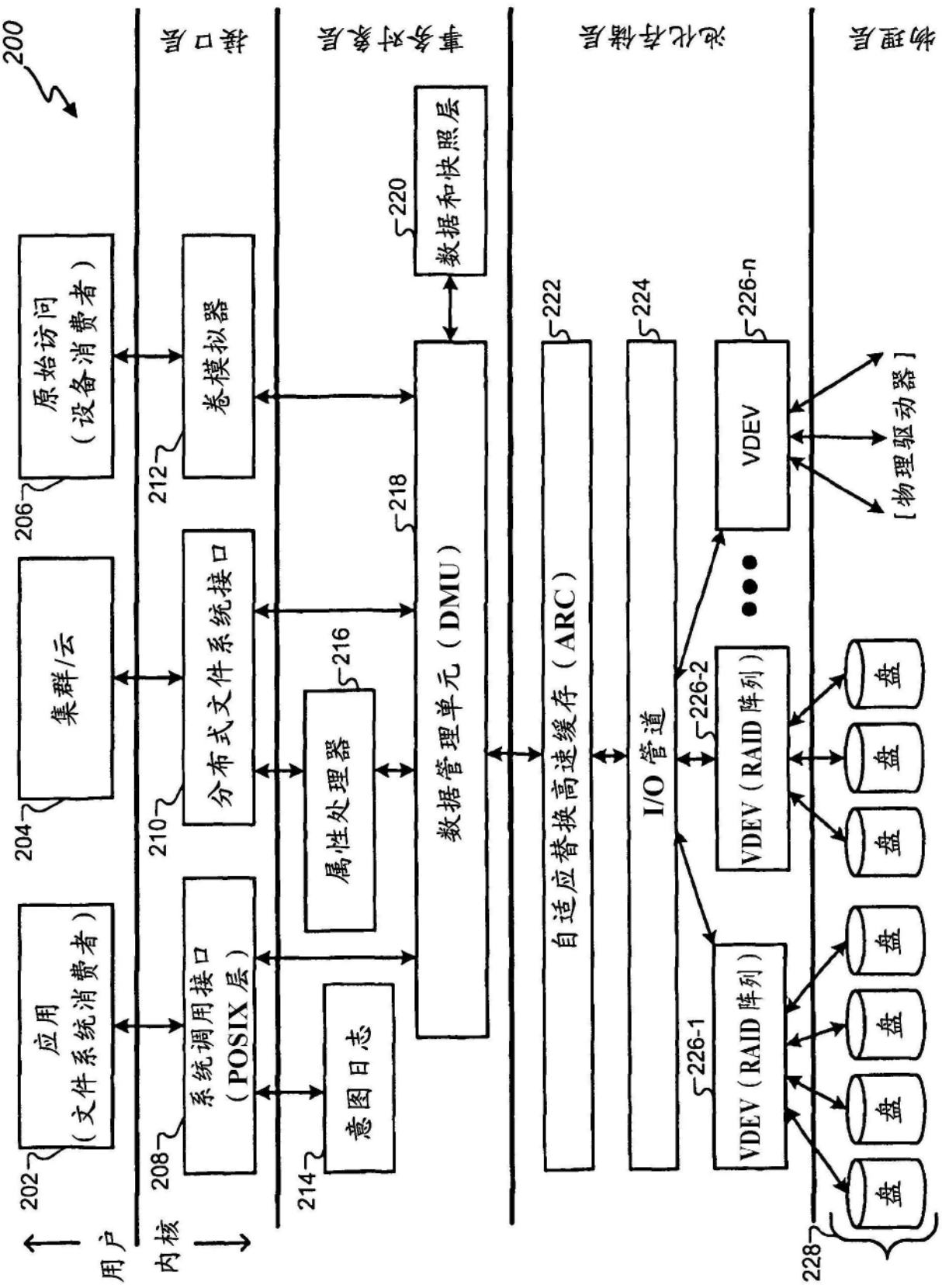

200
接口层
事务对象层
虚化存储层
物理层
用户
内核
202
应用
（文件系统消费者）
204
集群/云
206
原始访问
（设备消费者）
208
系统调用接口
（POSIX层）
210
分布式文件系统接口
212
卷模拟器
214
意图日志
216
属性处理器
218
数据管理单元（DMU）
220
数据和快照层
222
自适应替换高速缓存（ARC）
224
I/O 管道
226-1
VDEV（RAID 阵列）
226-2
VDEV（RAID 阵列）
226-n
VDEV
[物理驱动器]
228
盘

图2

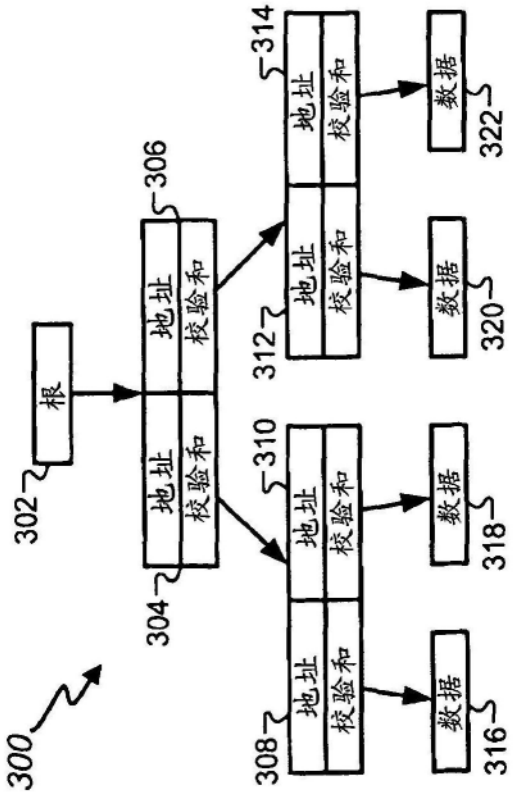

300
302
根
304
地址
校验和
306
地址
校验和
308
地址
校验和
310
地址
校验和
312
地址
校验和
314
地址
校验和
316
数据
318
数据
320
数据
322
数据

图3A

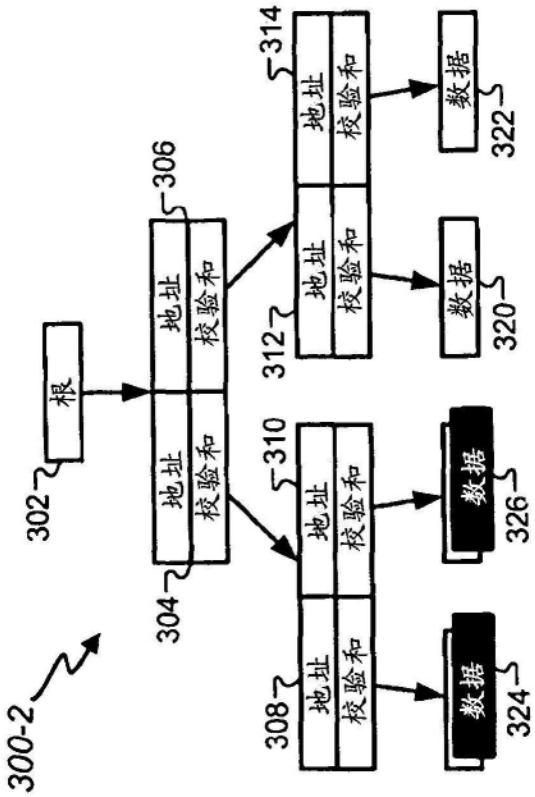

300-2
302
根
304
地址
校验和
306
地址
校验和
308
地址
校验和
数据
310
地址
校验和
数据
312
地址
校验和
数据
314
地址
校验和
数据
320
322
324
326

图3B

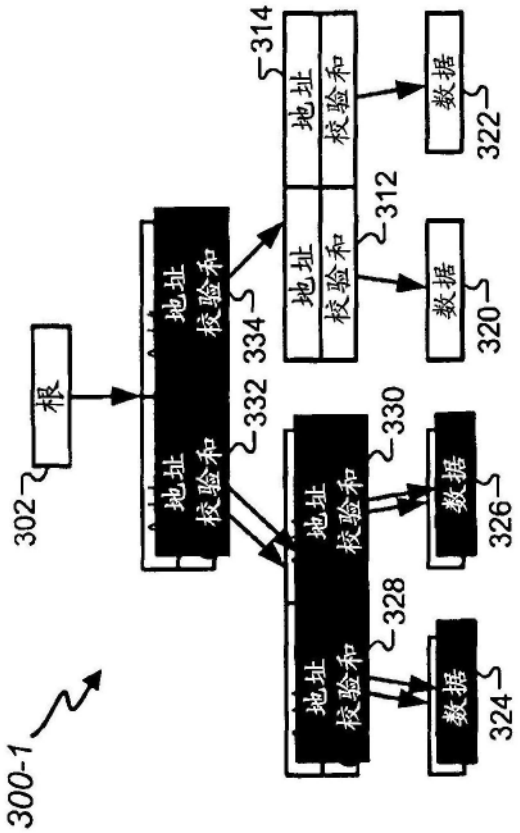


图3C

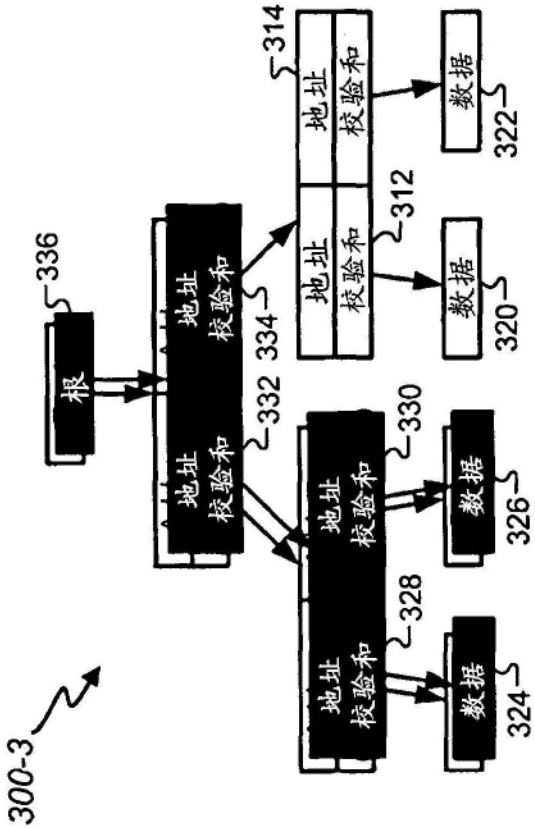

300-3
336
根
地址
校验和
332
334
312
314
地址
校验和
数据
320
322
330
328
326
324

图3D

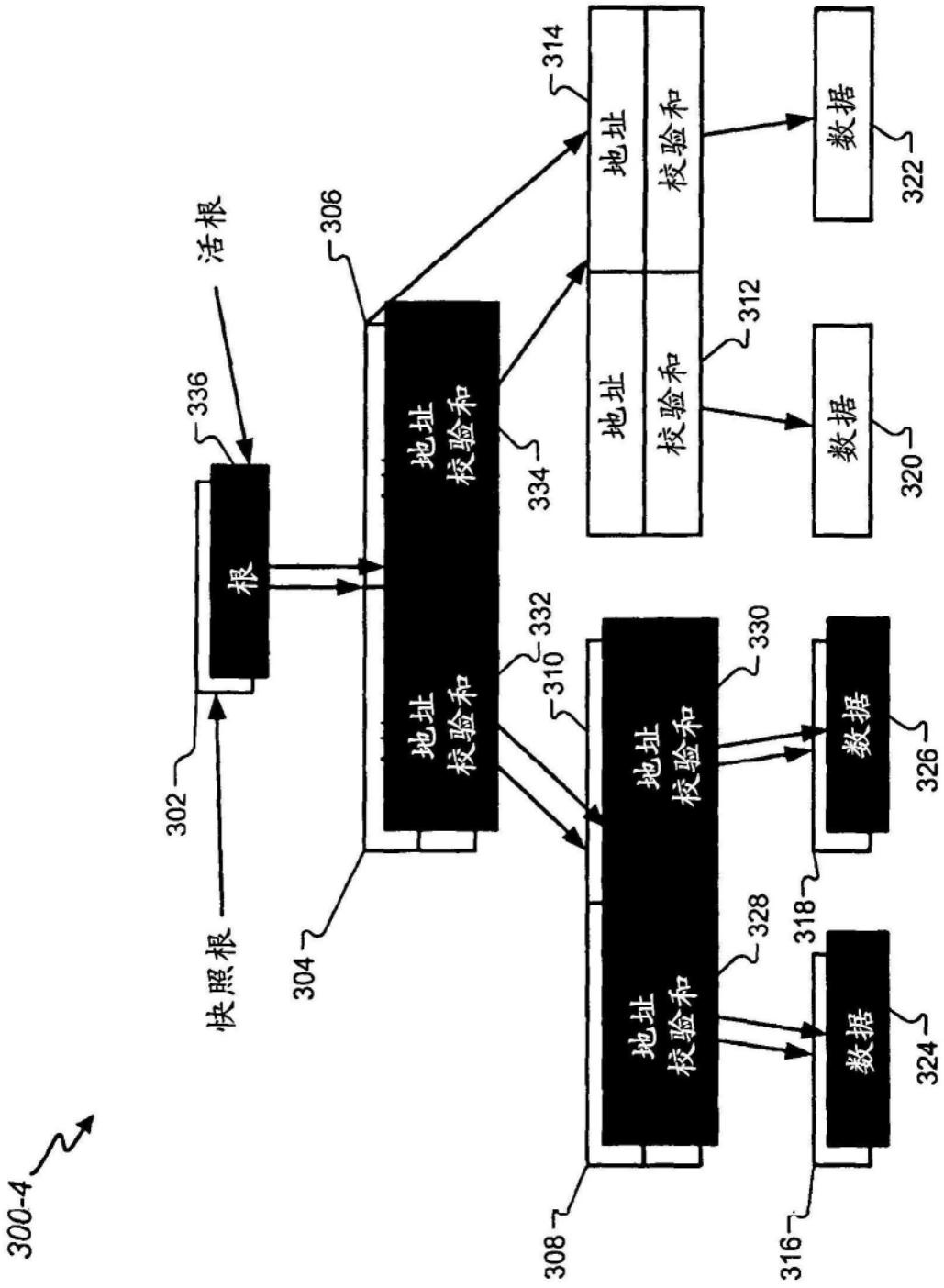

300-4
活根
336
根
302
快照根
304
306
地址
校验和
334
314
地址
校验和
322
数据
312
地址
校验和
320
数据
332
310
地址
校验和
330
326
数据
328
地址
校验和
318
324
数据
308
316

图3E

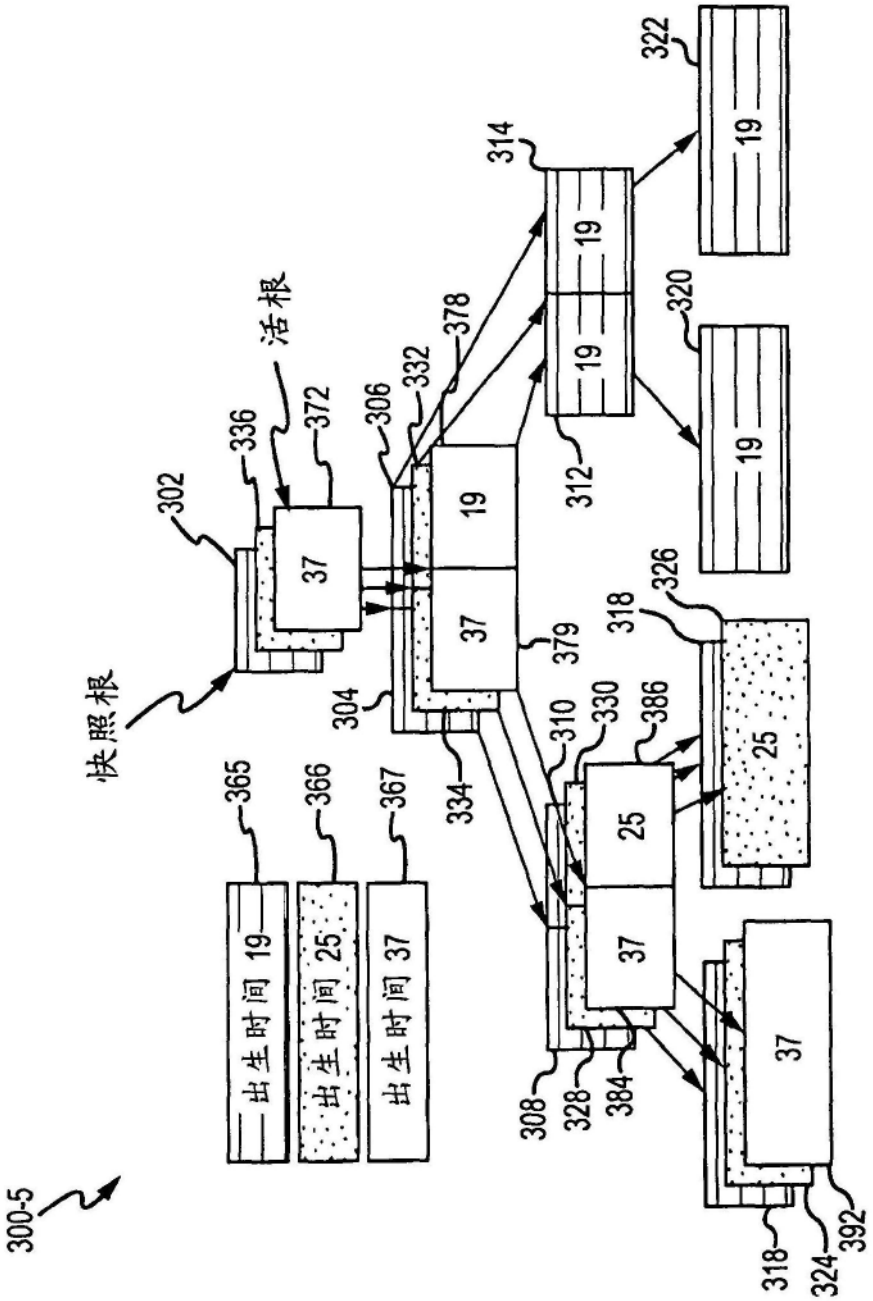

300-5
快照根
活根
302
304
306
308
310
312
314
318
320
322
324
326
328
330
332
334
336
365
366
367
372
378
379
384
386
392
19
25
37
出生时间 19
出生时间 25
出生时间 37

图3F

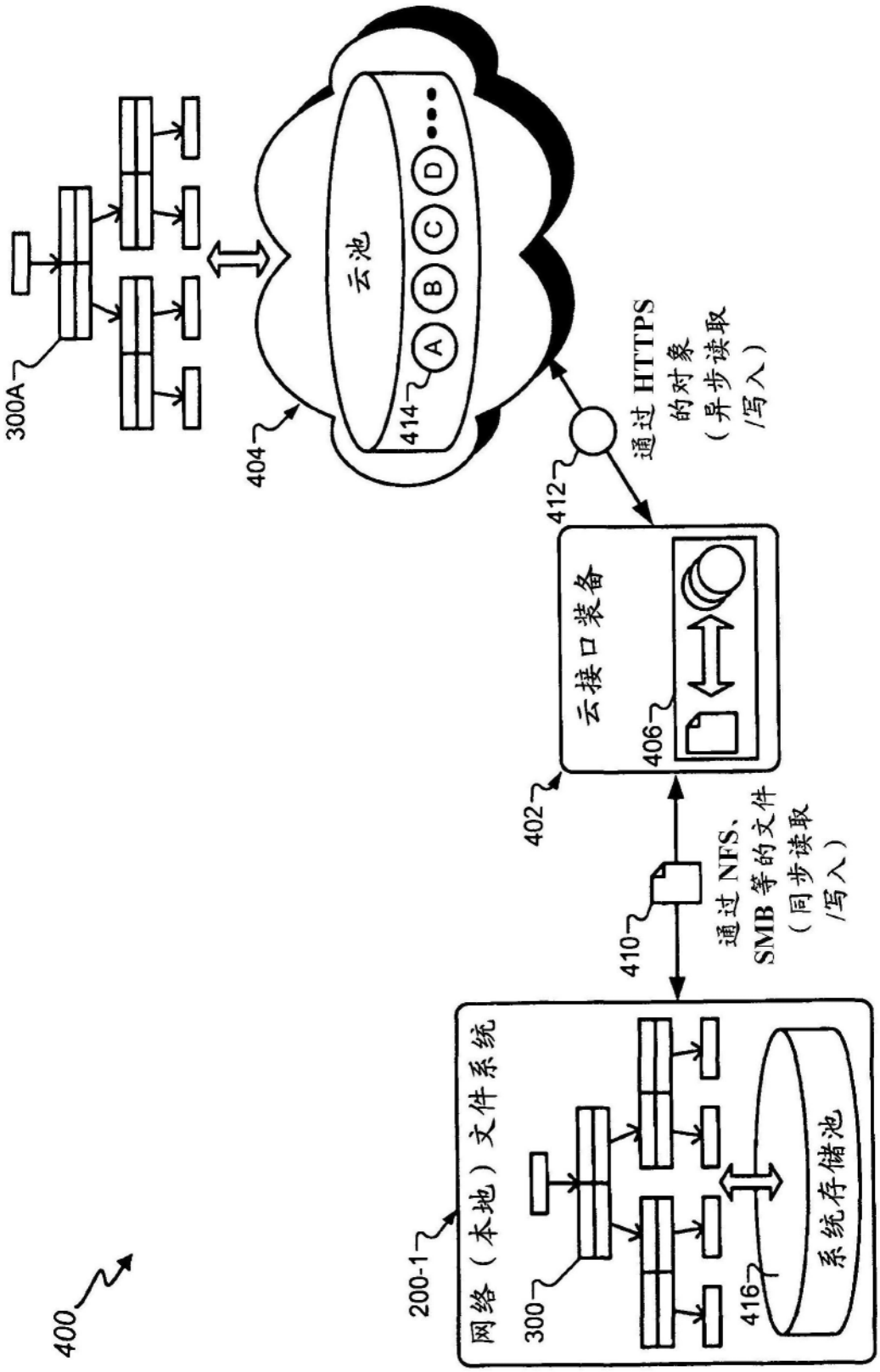

400
200-1
网络（本地）文件系统
300
416
系统存储池
410
通过 NFS、SMB 等的文件（同步读取/写入）
402
云接口装备
406
412
通过 HTTPS 的对象（异步读取/写入）
404
云池
414
A
B
C
D
300A

图4

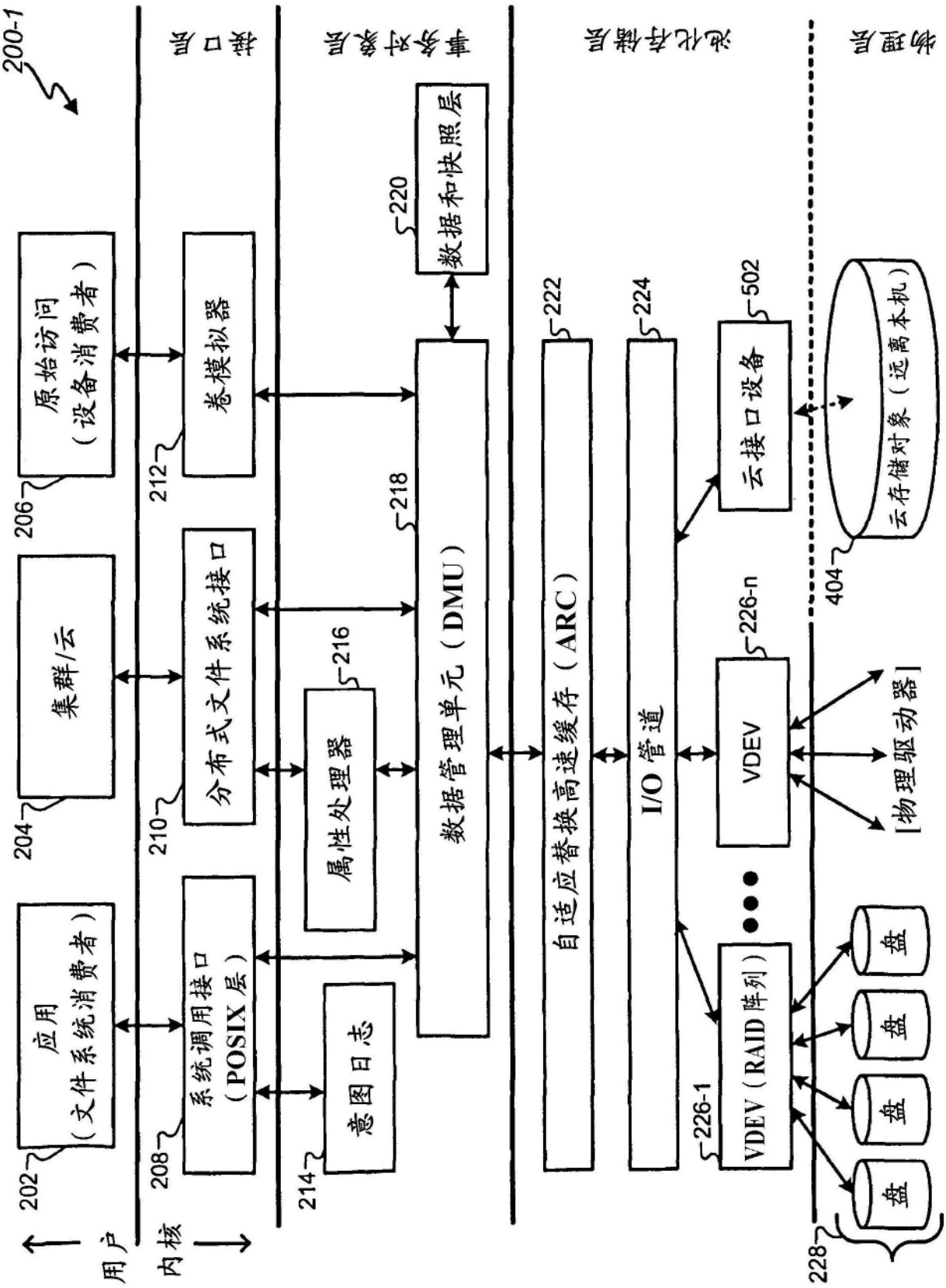

200-1
接口层
事务处理层
高速存储层
物理层
用户
内核
202
应用
（文件系统消费者）
204
集群/云
206
原始访问
（设备消费者）
208
系统调用接口
（POSIX层）
210
分布式文件系统接口
212
卷模拟器
214
意图日志
216
属性处理器
218
数据管理单元（DMU）
220
数据和快照层
222
自适应替换高速缓存（ARC）
224
I/O管道
226-1
VDEV（RAID阵列）
226-n
VDEV
502
云接口设备
404
云存储对象（远离本机）
物理驱动器
228
盘
盘
盘
盘

图5

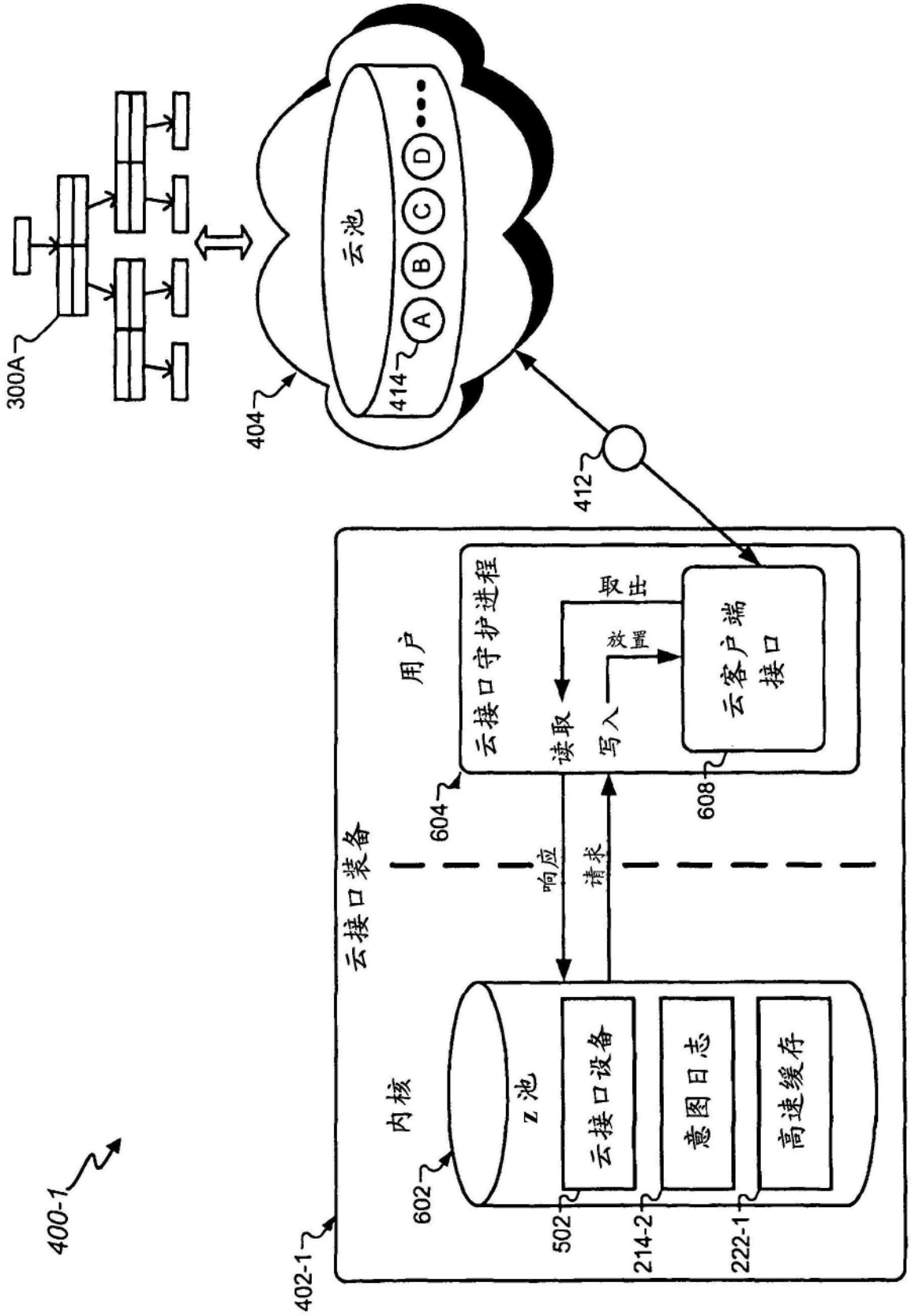

400-1
300A
404
云池
A
B
C
D
414
412
402-1
云接口装备
用户
内核
602
Z 池
云接口设备
意图日志
高速缓存
502
214-2
222-1
604
云接口守护进程
读取
写入
取出
放置
608
云客户端接口
响应
请求

图6

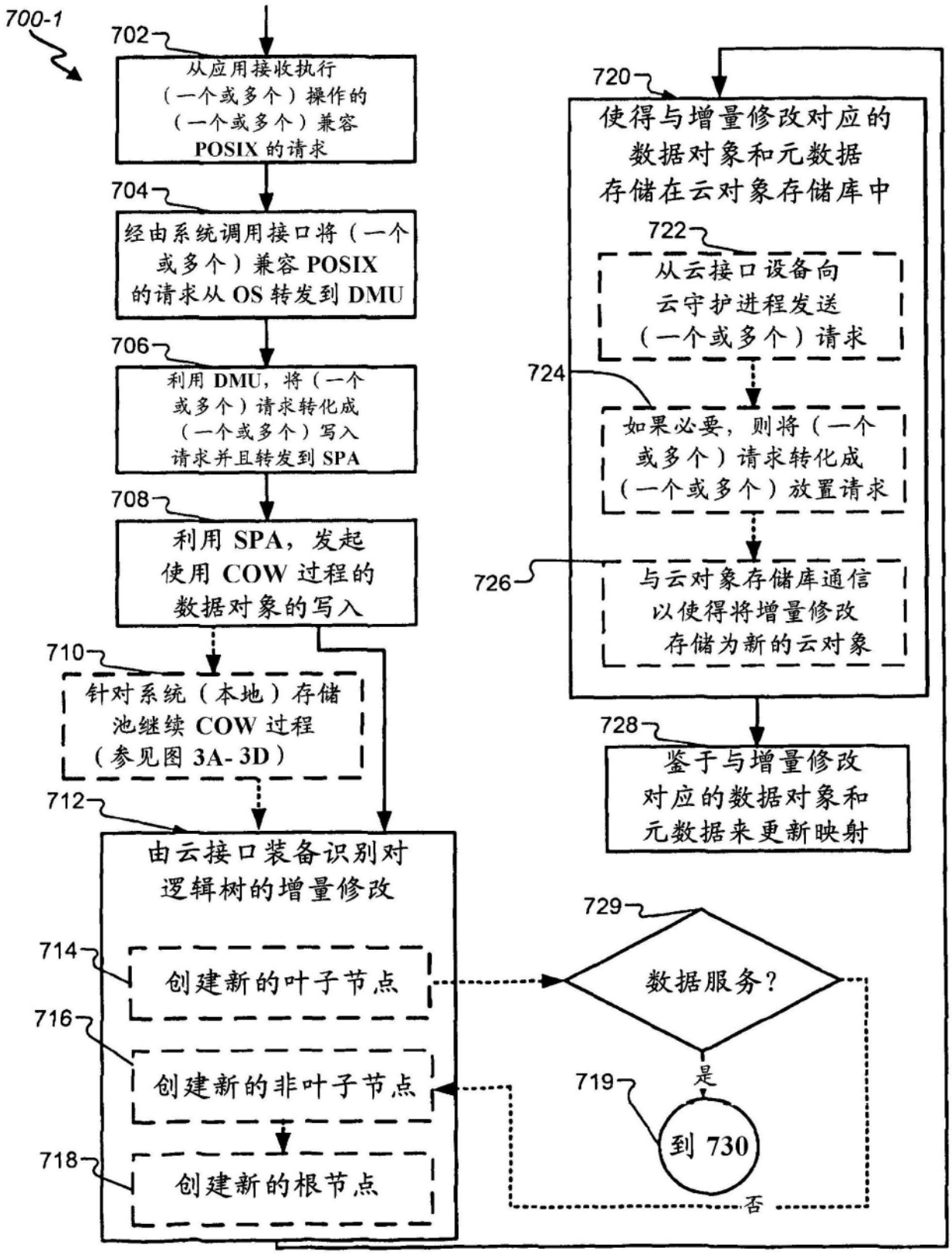

700-1
702
从应用接收执行（一个或多个）操作的（一个或多个）兼容POSIX的请求
704
经由系统调用接口将（一个或多个）兼容POSIX的请求从OS转发到DMU
706
利用DMU，将（一个或多个）请求转化成（一个或多个）写入请求并且转发到SPA
708
利用SPA，发起使用COW过程的数据对象的写入
710
针对系统（本地）存储池继续COW过程（参见图3A-3D）
712
由云接口装备识别对逻辑树的增量修改
714
创建新的叶子节点
716
创建新的非叶子节点
718
创建新的根节点
720
使得与增量修改对应的数据对象和元数据存储在云对象存储库中
722
从云接口设备向云守护进程发送（一个或多个）请求
724
如果必要，则将（一个或多个）请求转化成（一个或多个）放置请求
726
与云对象存储库通信以使得将增量修改存储为新的云对象
728
鉴于与增量修改对应的数据对象和元数据来更新映射
729
数据服务？
是
719
到730
否

图7A

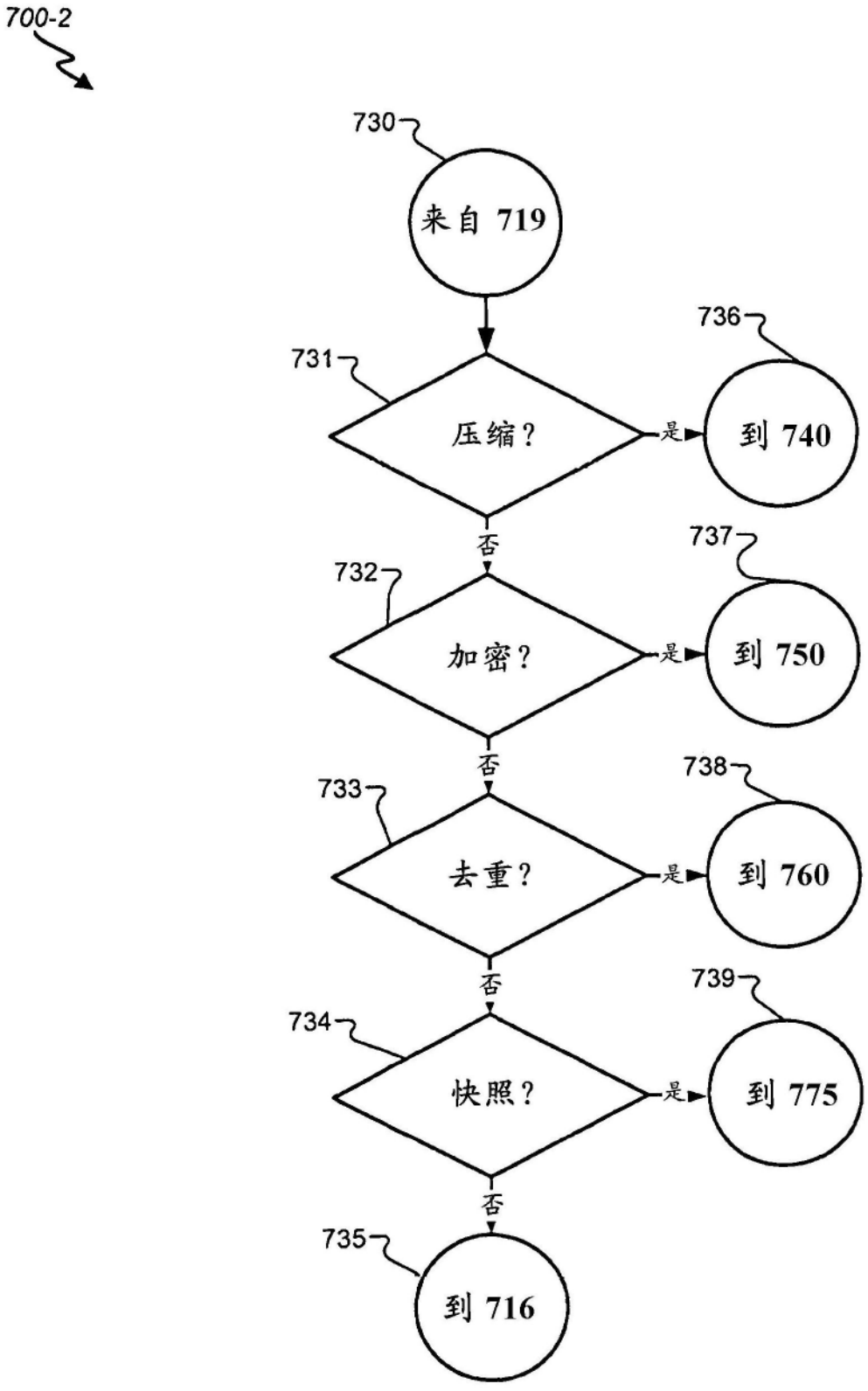

700-2
730
来自 719
731
压缩？
是
736
到 740
否
732
加密？
是
737
到 750
否
733
去重？
是
738
到 760
否
734
快照？
是
739
到 775
否
735
到 716

图7B

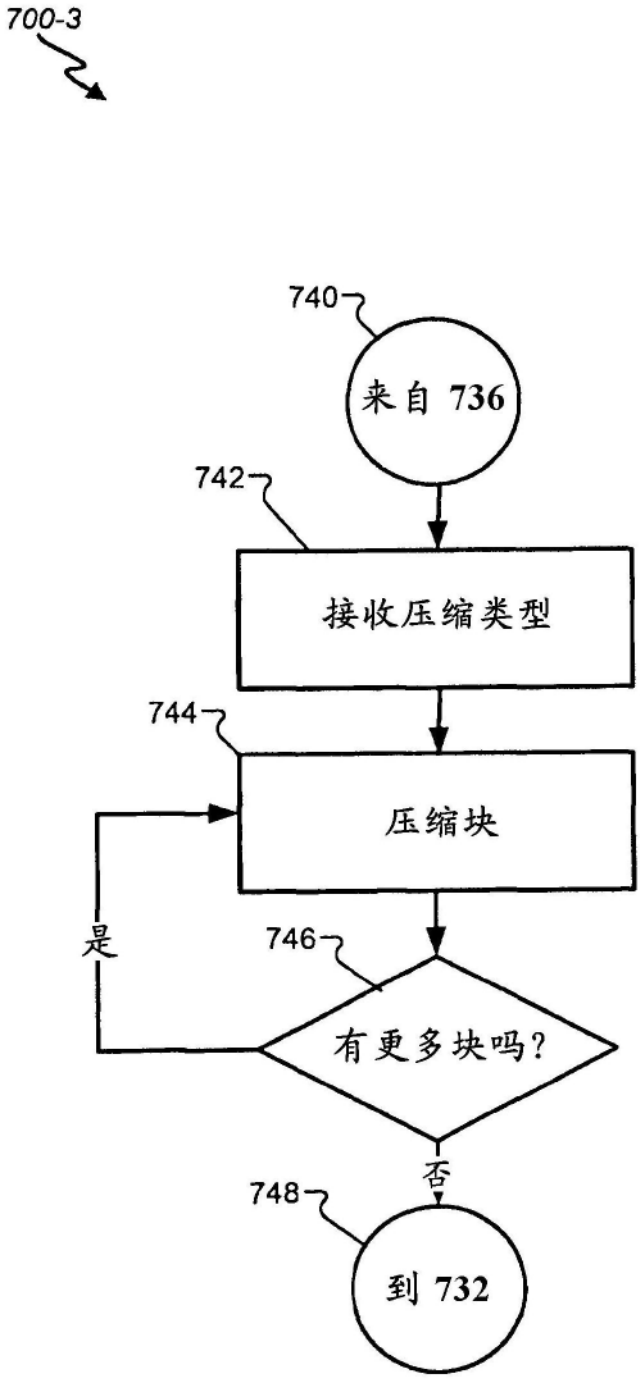


图7C

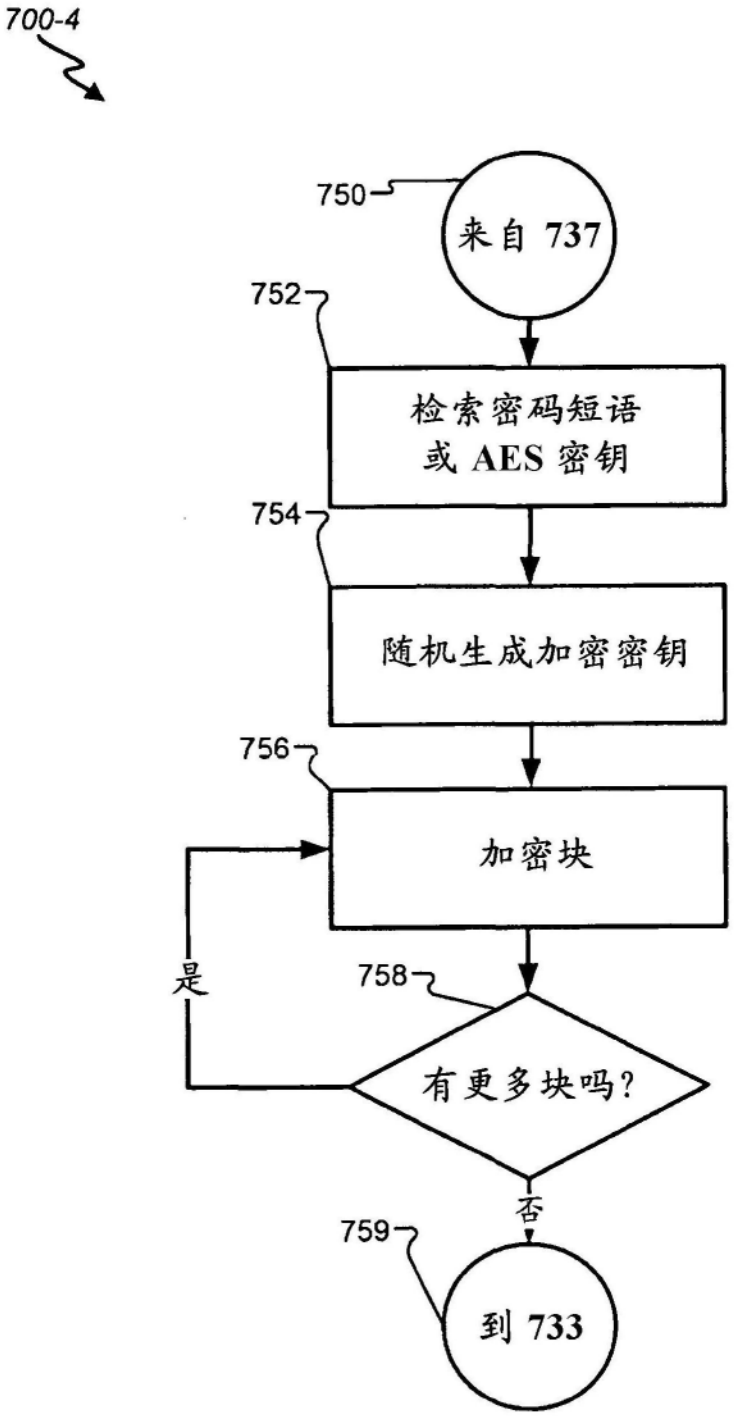

700-4
750
来自 737
752
检索密码短语
或 AES 密钥
754
随机生成加密密钥
756
加密块
是
758
有更多块吗?
否
759
到 733

图7D

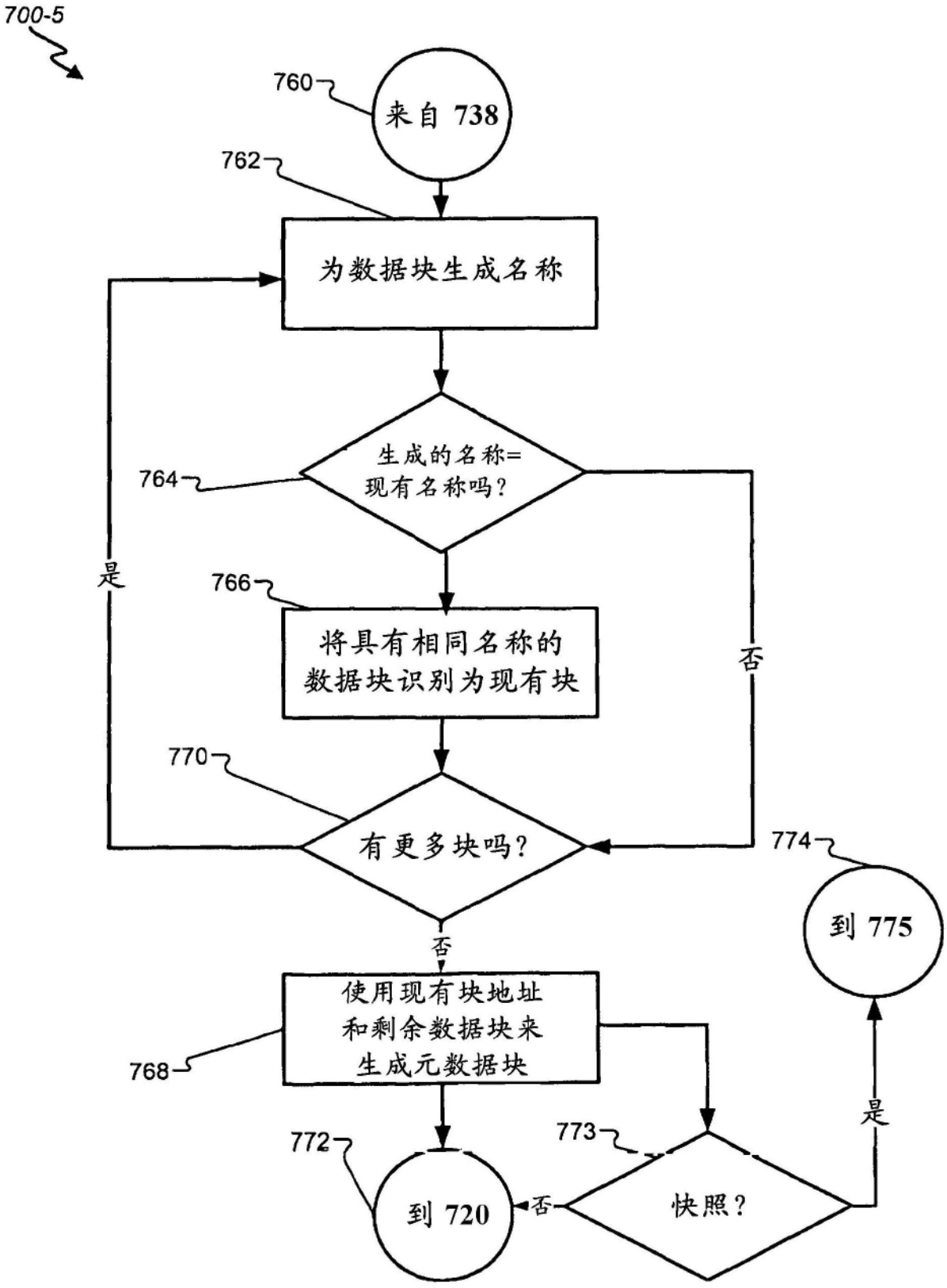

700-5
760
来自 738
762
为数据块生成名称
生成的名称=
现有名称吗？
764
是
否
766
将具有相同名称的
数据块识别为现有块
770
有更多块吗？
否
774
到 775
使用现有块地址
和剩余数据块来
生成元数据块
768
是
773
772
到 720
否
快照？

图7E

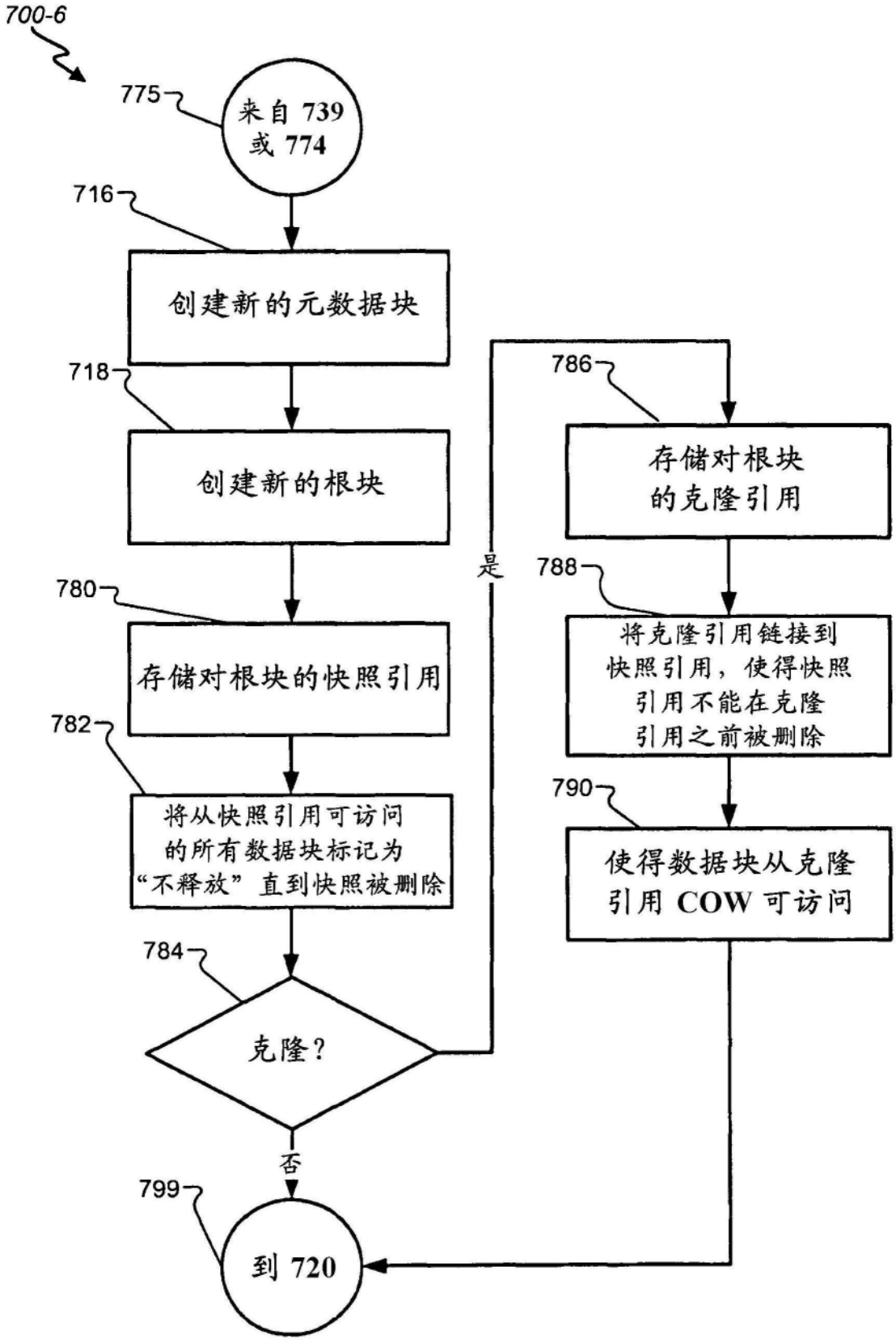

700-6
775
来自 739
或 774
716
创建新的元数据块
718
创建新的根块
780
存储对根块的快照引用
782
将从快照引用可访问
的所有数据块标记为
“不释放”直到快照被删除
784
克隆?
是
否
786
存储对根块
的克隆引用
788
将克隆引用链接到
快照引用，使得快照
引用不能在克隆
引用之前被删除
790
使得数据块从克隆
引用 COW 可访问
799
到 720

图7F

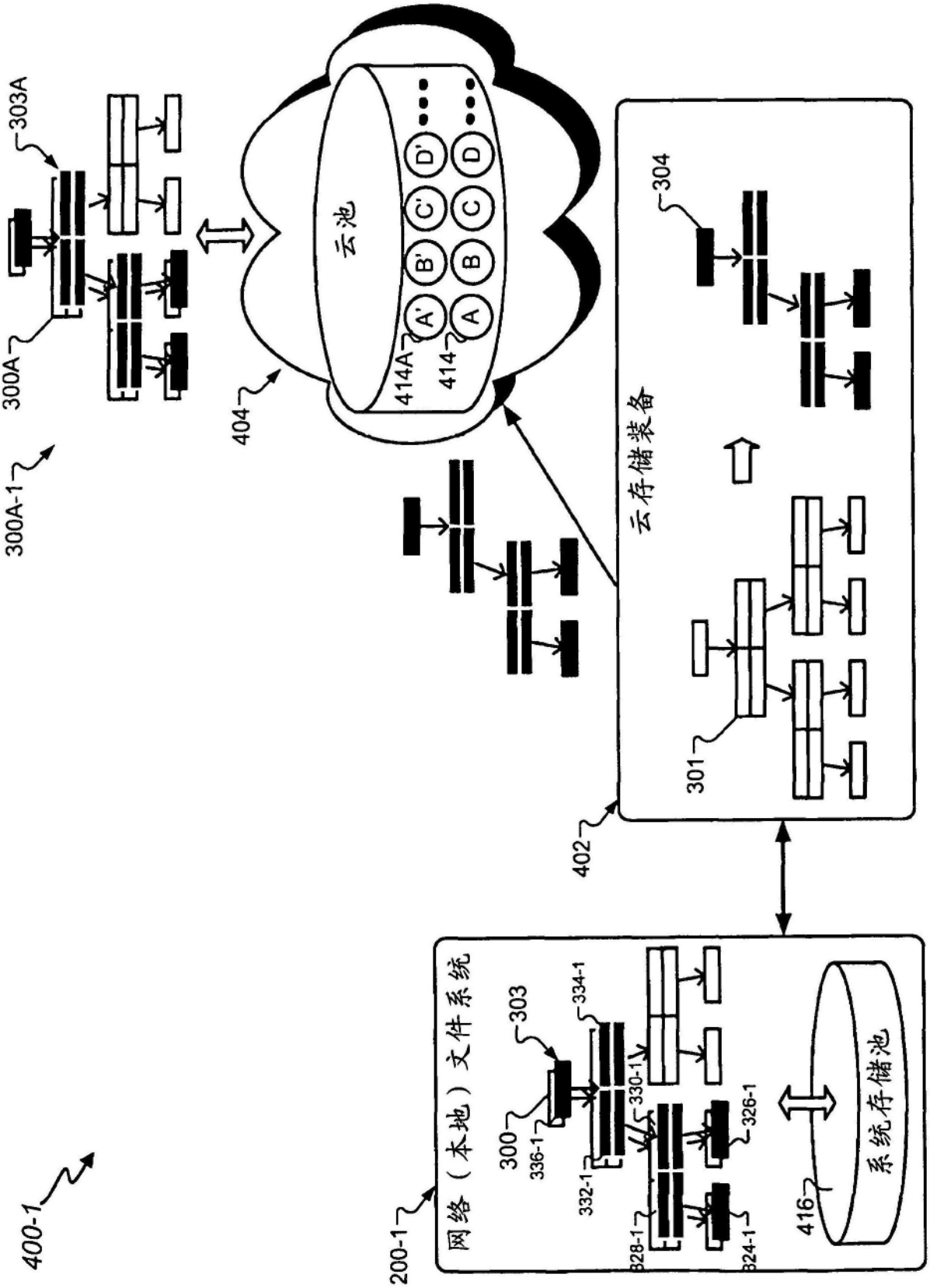

400-1
200-1
网络（本地）文件系统
300
303
334-1
330-1
326-1
328-1
324-1
332-1
336-1
416
系统存储池
402
云存储装备
304
301
404
云池
414A
414
A'
A
B'
B
C'
C
D'
D
300A-1
300A
303A

图8

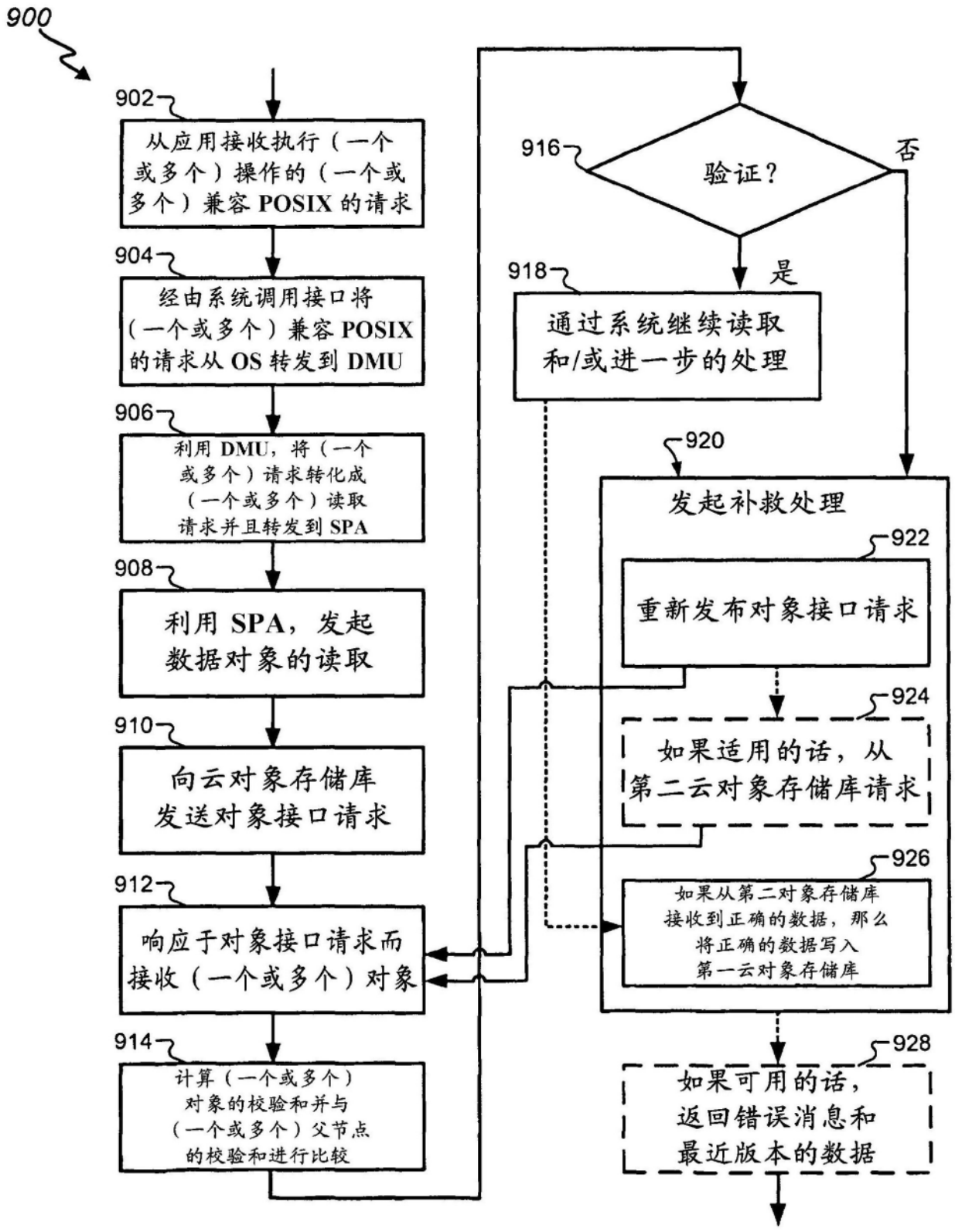

900
902
从应用接收执行（一个或多个）操作的（一个或多个）兼容 POSIX 的请求
904
经由系统调用接口将（一个或多个）兼容 POSIX 的请求从 OS 转发到 DMU
906
利用 DMU，将（一个或多个）请求转化成（一个或多个）读取请求并且转发到 SPA
908
利用 SPA，发起数据对象的读取
910
向云对象存储库发送对象接口请求
912
响应于对象接口请求而接收（一个或多个）对象
914
计算（一个或多个）对象的校验和并与（一个或多个）父节点的校验和进行比较
916
验证？
否
是
918
通过系统继续读取和/或进一步的处理
920
发起补救处理
922
重新发布对象接口请求
924
如果适用的话，从第二云对象存储库请求
926
如果从第二对象存储库接收到正确的数据，那么将正确的数据写入第一云对象存储库
928
如果可用的话，返回错误消息和最近版本的数据

图9

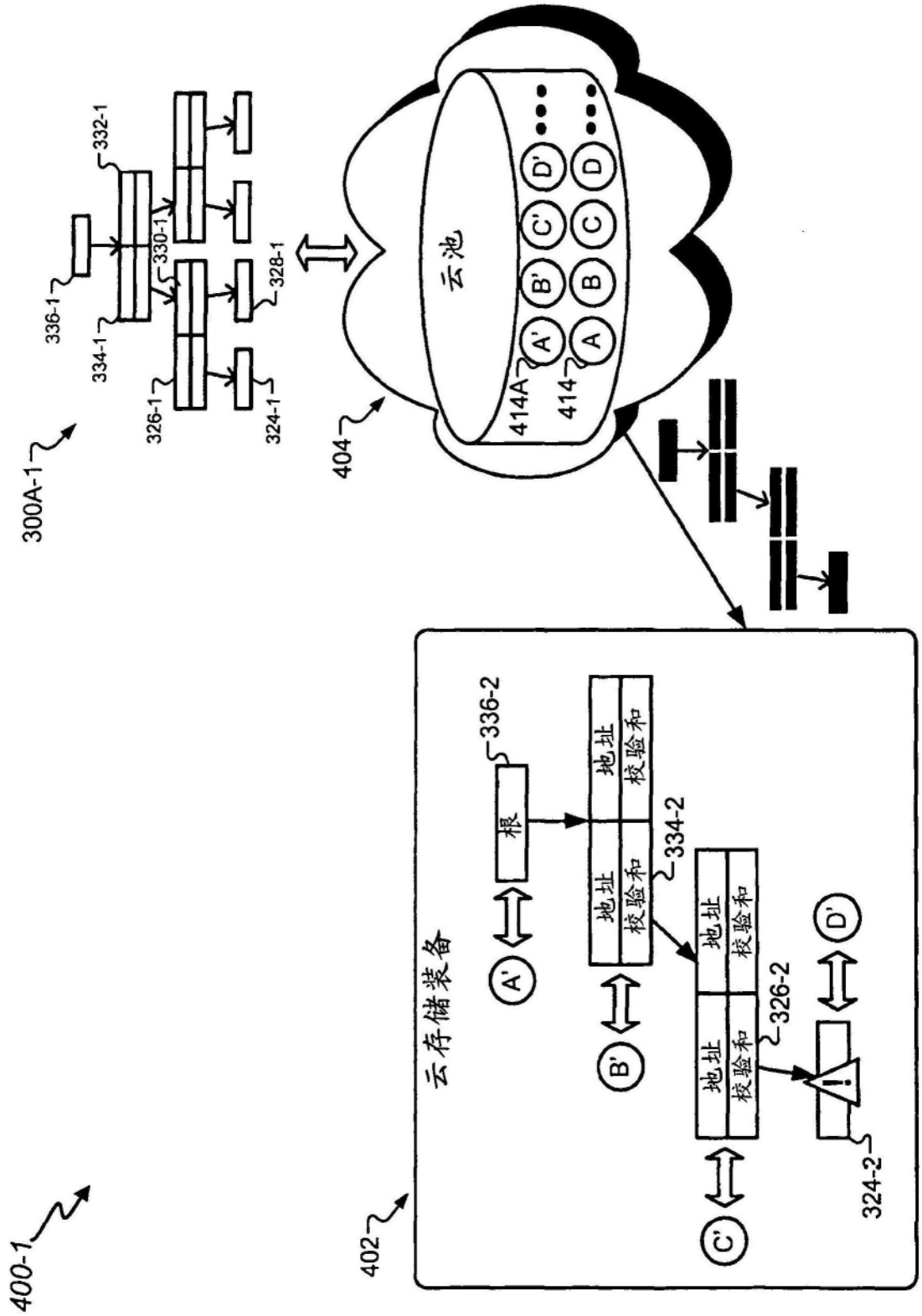

400-1
300A-1
332-1
330-1
336-1
334-1
328-1
326-1
324-1
404
云池
414A
414
A'
A
B'
B
C'
C
D'
D
402
云存储装备
336-2
根
地址
校验和
334-2
326-2
324-2

图10

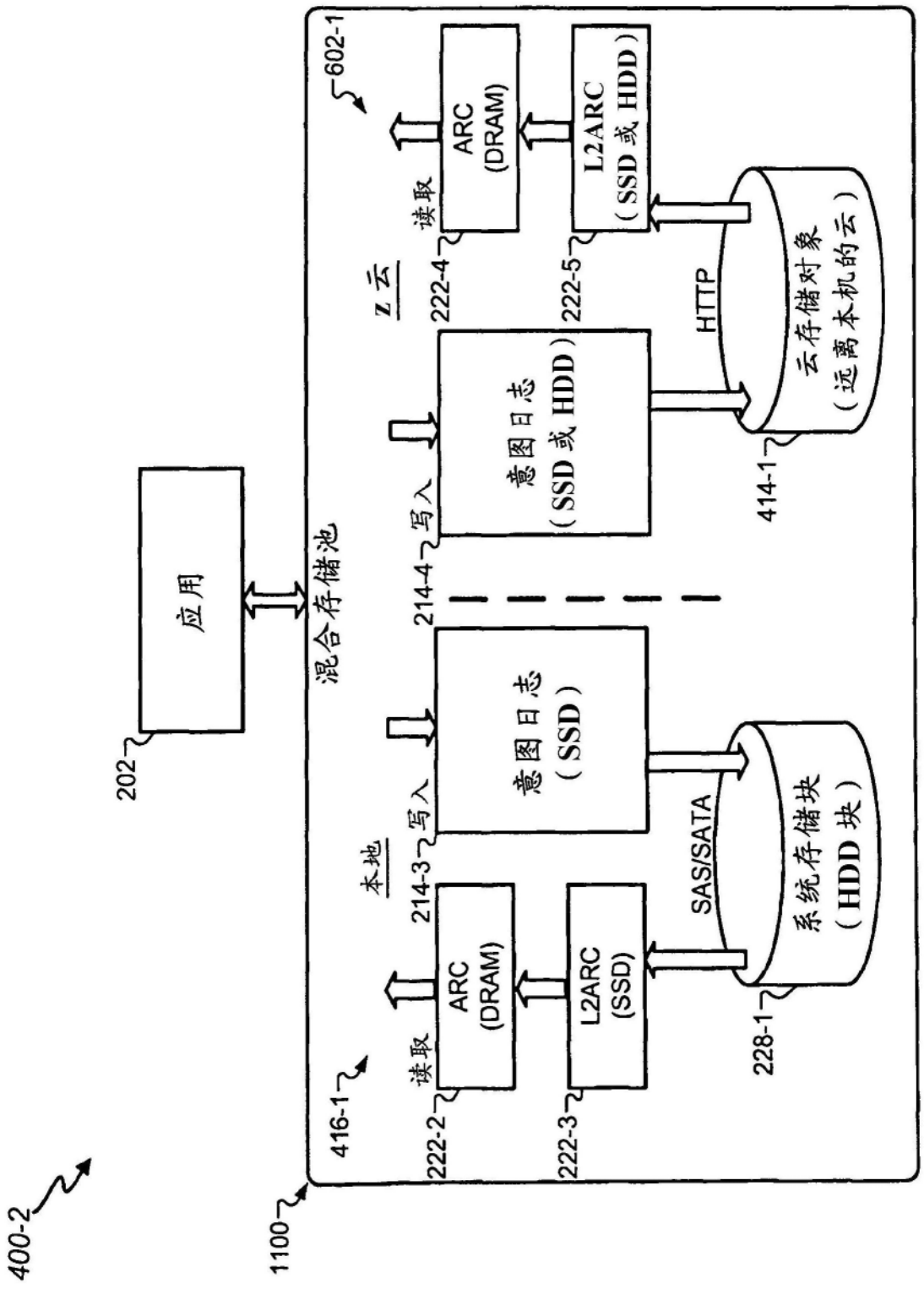

400-2
1100
混合存储池
602-1
202
应用
本地
Z云
416-1
读取
ARC (DRAM)
222-2
L2ARC (SSD)
222-3
写入
214-3
意图日志 (SSD)
写入
214-4
意图日志 (SSD 或 HDD)
读取
ARC (DRAM)
222-4
L2ARC (SSD 或 HDD)
222-5
SAS/SATA
系统存储块 (HDD 块)
228-1
HTTP
云存储对象 (远离本机的云)
414-1

图11

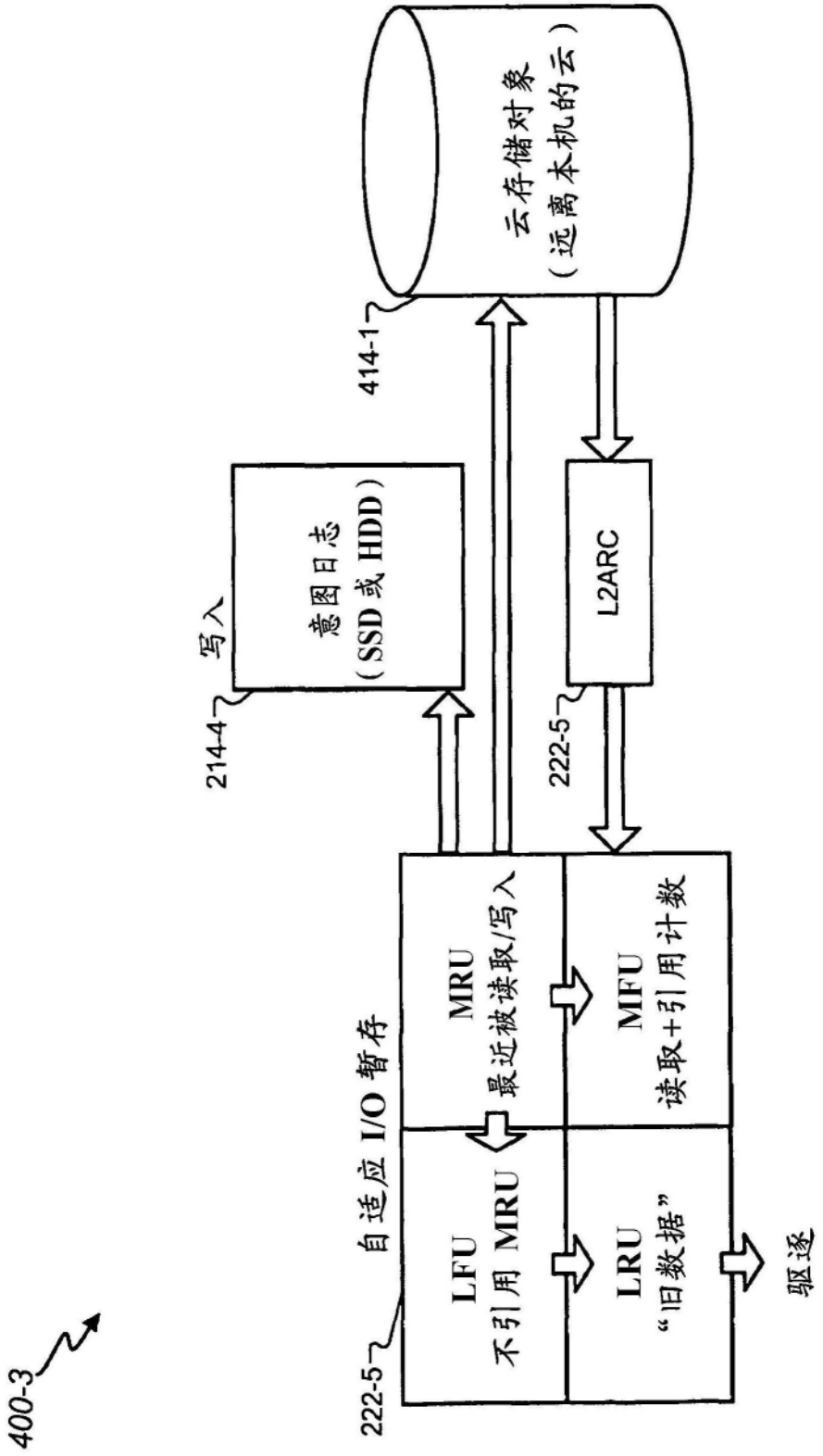

400-3
自适应 I/O 暂存
LFU
不引用 MRU
MRU
最近被读取/写入
LRU
“旧数据”
MFU
读取+引用计数
驱逐
222-5
222-5
214-4
写入
意图日志
(SSD 或 HDD)
L2ARC
414-1
云存储对象
(远离本机的云)

图12

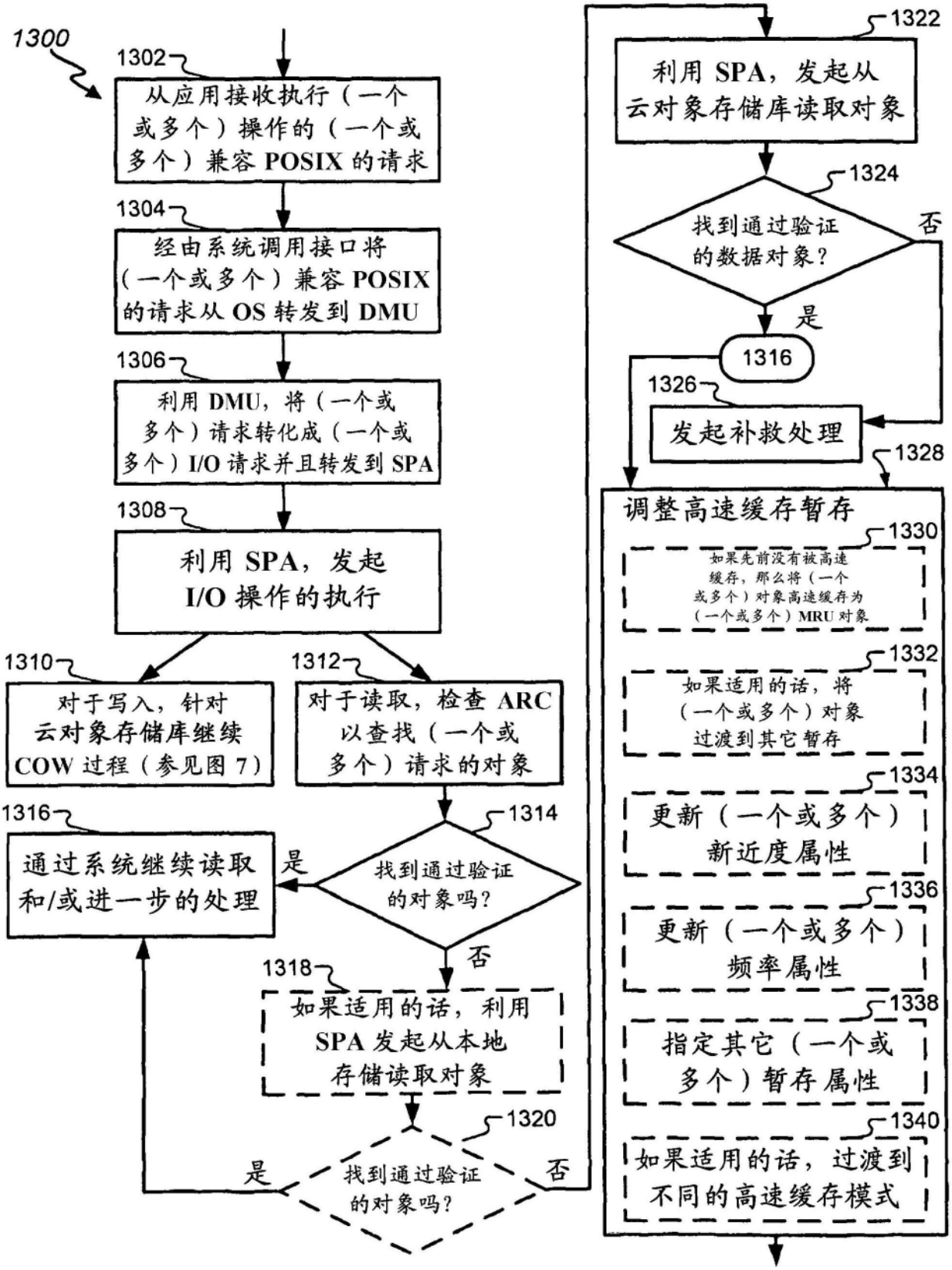

1300
1302
从应用接收执行（一个或多个）操作的（一个或多个）兼容 POSIX 的请求
1304
经由系统调用接口将（一个或多个）兼容 POSIX 的请求从 OS 转发到 DMU
1306
利用 DMU，将（一个或多个）请求转化成（一个或多个）I/O 请求并且转发到 SPA
1308
利用 SPA，发起 I/O 操作的执行
1310
对于写入，针对云对象存储库继续 COW 过程（参见图 7）
1312
对于读取，检查 ARC 以查找（一个或多个）请求的对象
1314
找到通过验证的对象吗？
是
否
1316
通过系统继续读取和/或进一步的处理
1318
如果适用的话，利用 SPA 发起从本地存储读取对象
1320
找到通过验证的对象吗？
是
否
1322
利用 SPA，发起从云对象存储库读取对象
1324
找到通过验证的数据对象？
否
是
1316
1326
发起补救处理
1328
调整高速缓存暂存
1330
如果先前没有被高速缓存，那么将（一个或多个）对象高速缓存为（一个或多个）MRU 对象
1332
如果适用的话，将（一个或多个）对象过渡到其它暂存
1334
更新（一个或多个）新近度属性
1336
更新（一个或多个）频率属性
1338
指定其它（一个或多个）暂存属性
1340
如果适用的话，过渡到不同的高速缓存模式

图13

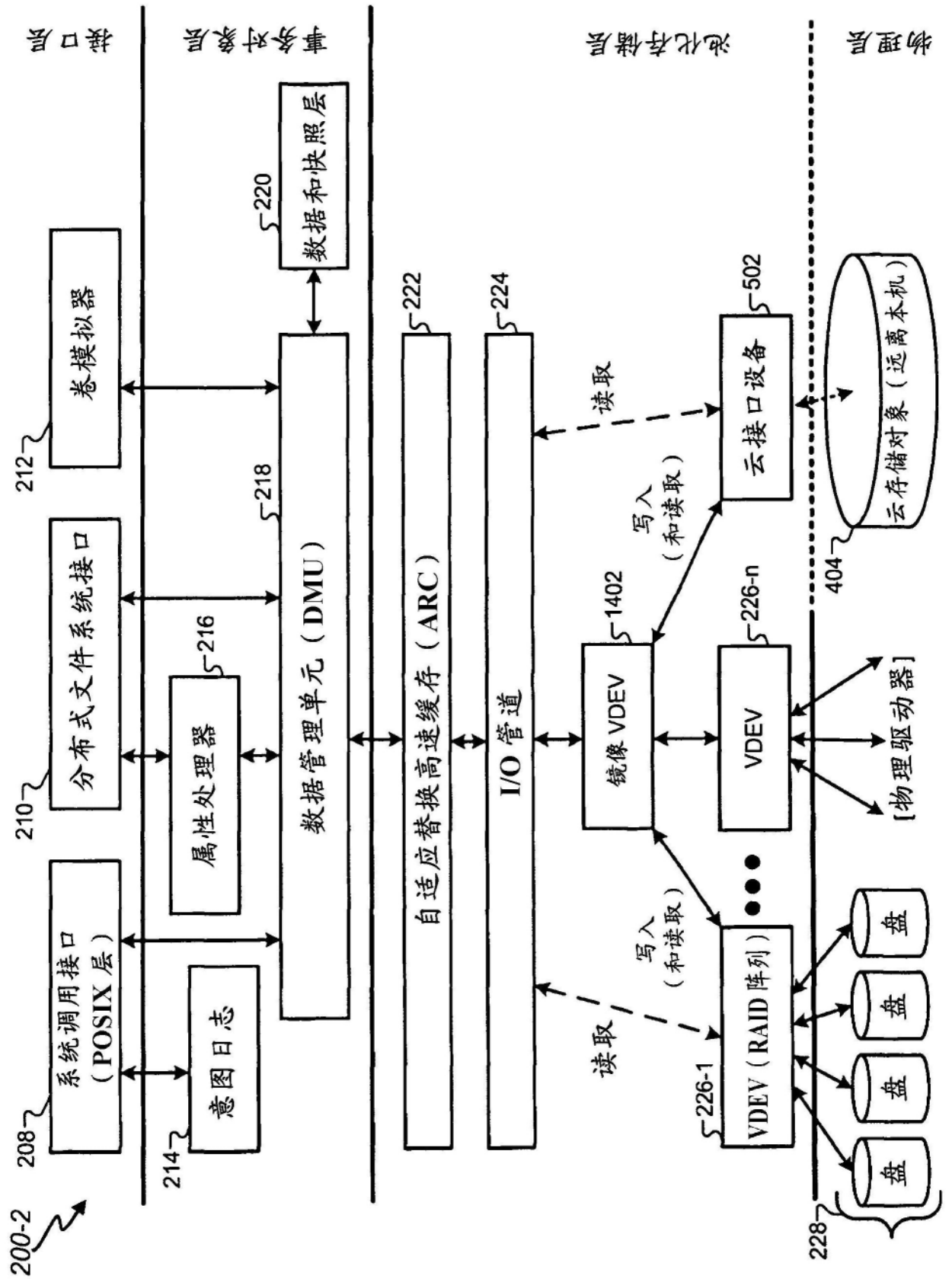

接口层
事务处理层
动态存储层
物理层
200-2
208
系统调用接口（POSIX层）
210
分布式文件系统接口
212
卷模拟器
214
意图日志
216
属性处理器
218
数据管理单元（DMU）
220
数据和快照层
222
自适应替换高速缓存（ARC）
224
I/O管道
读取
写入（和读取）
1402
镜像VDEV
226-1
VDEV（RAID阵列）
226-n
VDEV
[物理驱动器]
盘
盘
盘
盘
228
502
云接口设备
404
云存储对象（远离本机）

图14

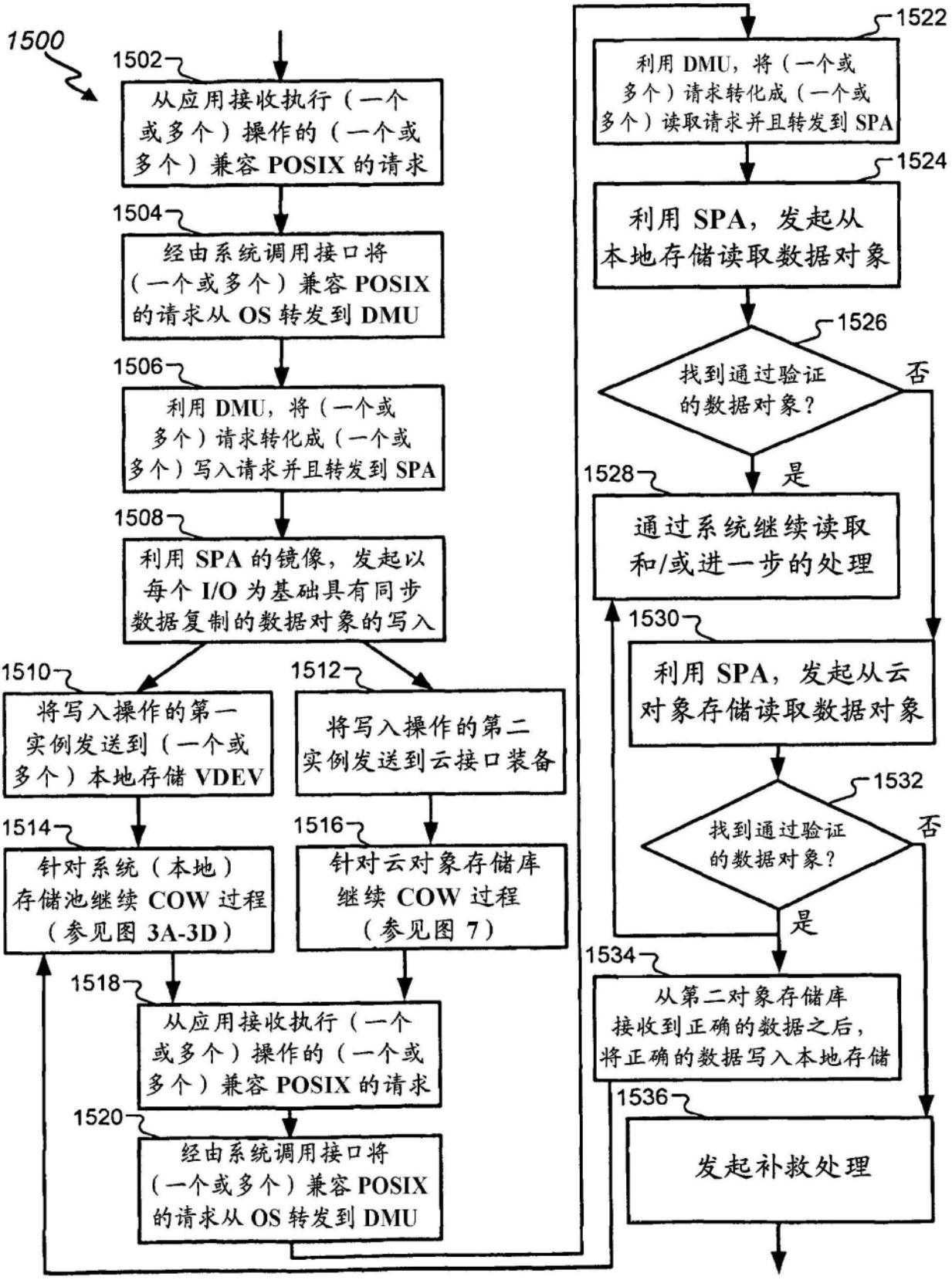

1500
1502
从应用接收执行（一个或多个）操作的（一个或多个）兼容 POSIX 的请求
1504
经由系统调用接口将（一个或多个）兼容 POSIX 的请求从 OS 转发到 DMU
1506
利用 DMU，将（一个或多个）请求转化成（一个或多个）写入请求并且转发到 SPA
1508
利用 SPA 的镜像，发起以每个 I/O 为基础具有同步数据复制的数据对象的写入
1510
将写入操作的第一实例发送到（一个或多个）本地存储 VDEV
1512
将写入操作的第二实例发送到云接口装备
1514
针对系统（本地）存储池继续 COW 过程（参见图 3A-3D）
1516
针对云对象存储库继续 COW 过程（参见图 7）
1518
从应用接收执行（一个或多个）操作的（一个或多个）兼容 POSIX 的请求
1520
经由系统调用接口将（一个或多个）兼容 POSIX 的请求从 OS 转发到 DMU
1522
利用 DMU，将（一个或多个）请求转化成（一个或多个）读取请求并且转发到 SPA
1524
利用 SPA，发起从本地存储读取数据对象
1526
找到通过验证的数据对象？
否
是
1528
通过系统继续读取和/或进一步的处理
1530
利用 SPA，发起从云对象存储读取数据对象
1532
找到通过验证的数据对象？
否
是
1534
从第二对象存储库接收到正确的数据之后，将正确的数据写入本地存储
1536
发起补救处理

图15

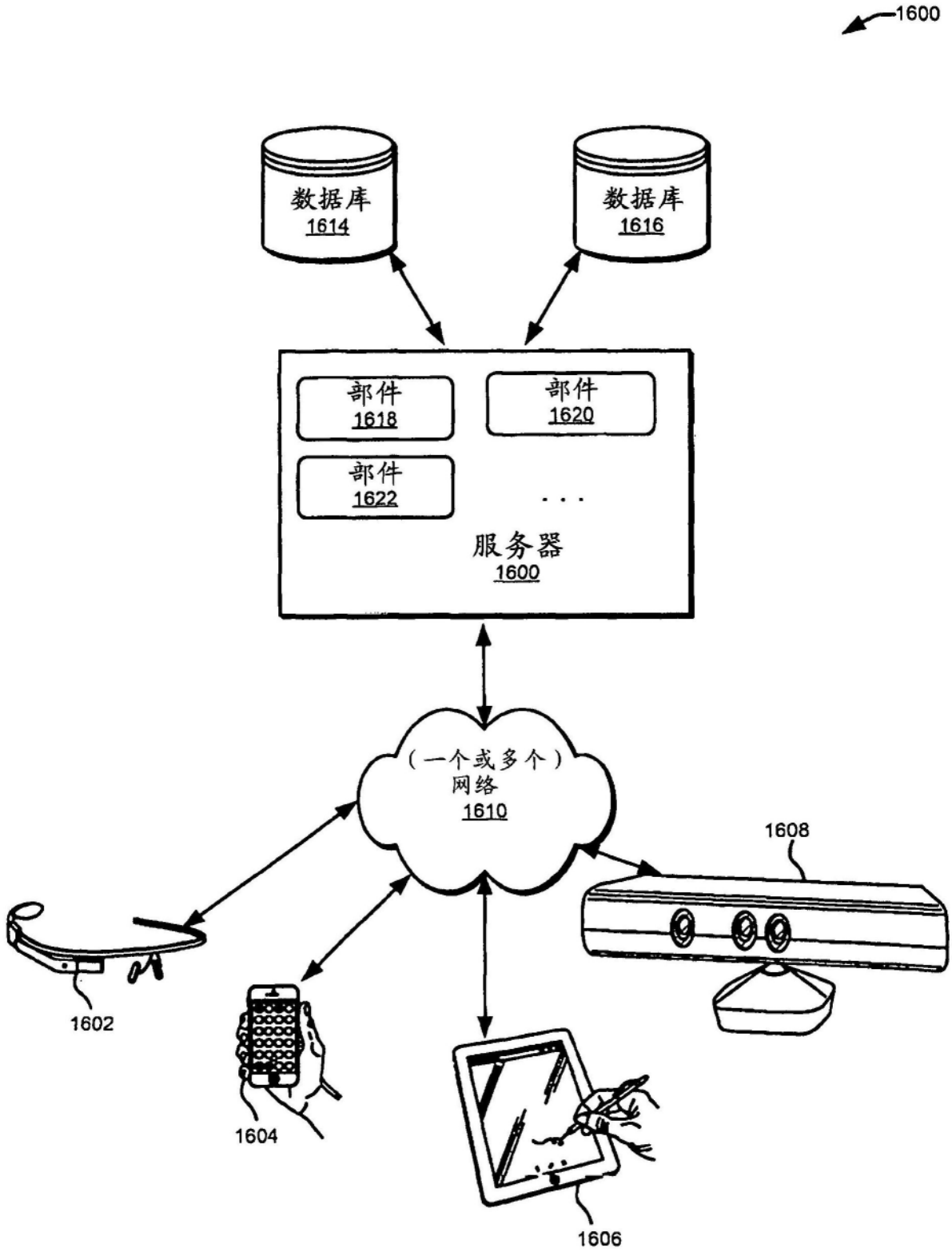

1600
数据库
1614
数据库
1616
部件
1618
部件
1620
部件
1622
. . .
服务器
1600
(一个或多个)
网络
1610
1608
1602
1604
1606

图16

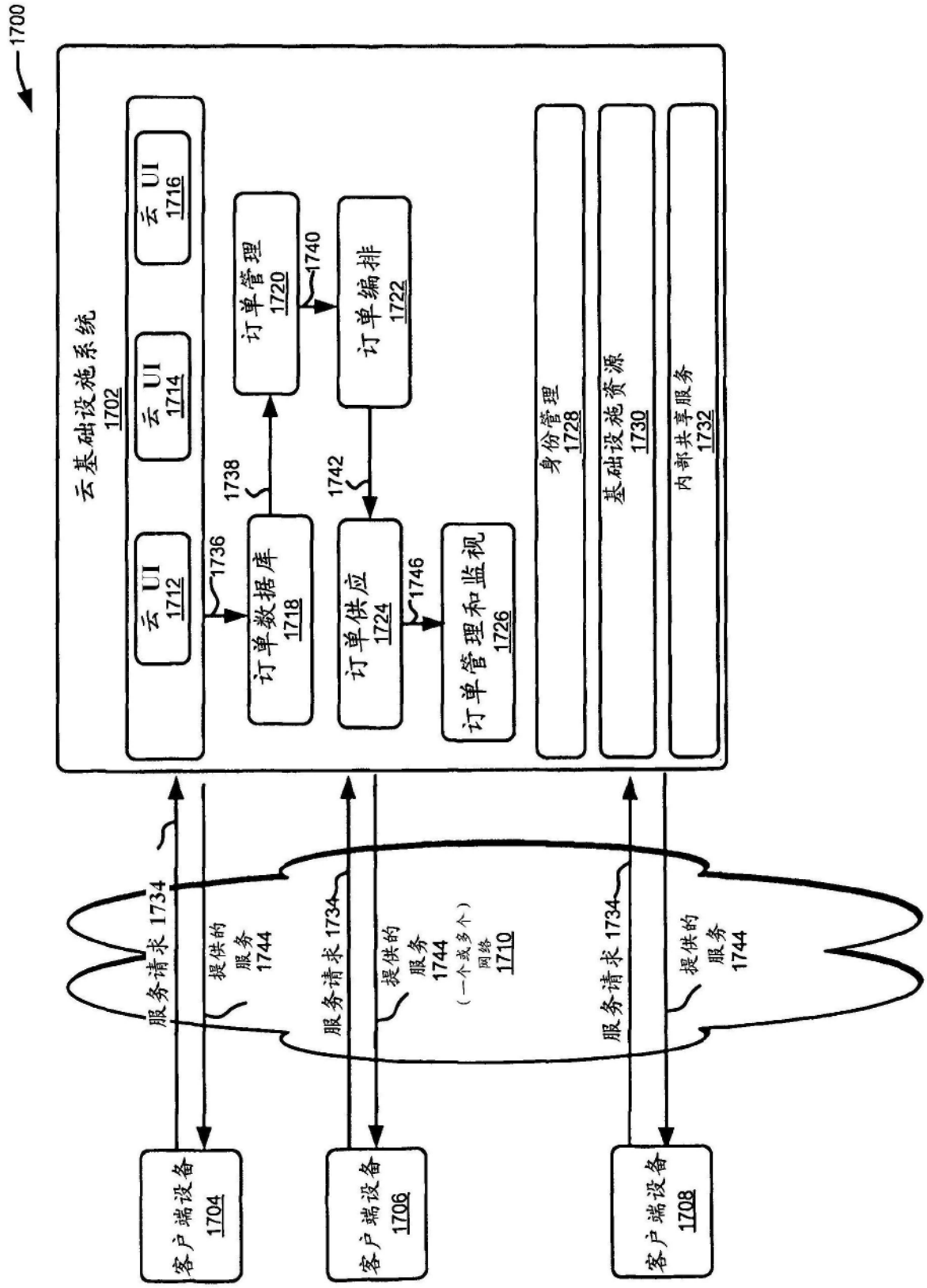

1700
云基础设施系统
1702
云 UI
1712
云 UI
1714
云 UI
1716
订单数据库
1718
订单管理
1720
订单编排
1722
订单供应
1724
订单管理和监视
1726
身份管理
1728
基础设施资源
1730
内部共享服务
1732
1736
1738
1740
1742
1746
服务请求 1734
提供的服务 1744
网络 1710
（一个或多个）
客户端设备 1704
客户端设备 1706
客户端设备 1708

图17

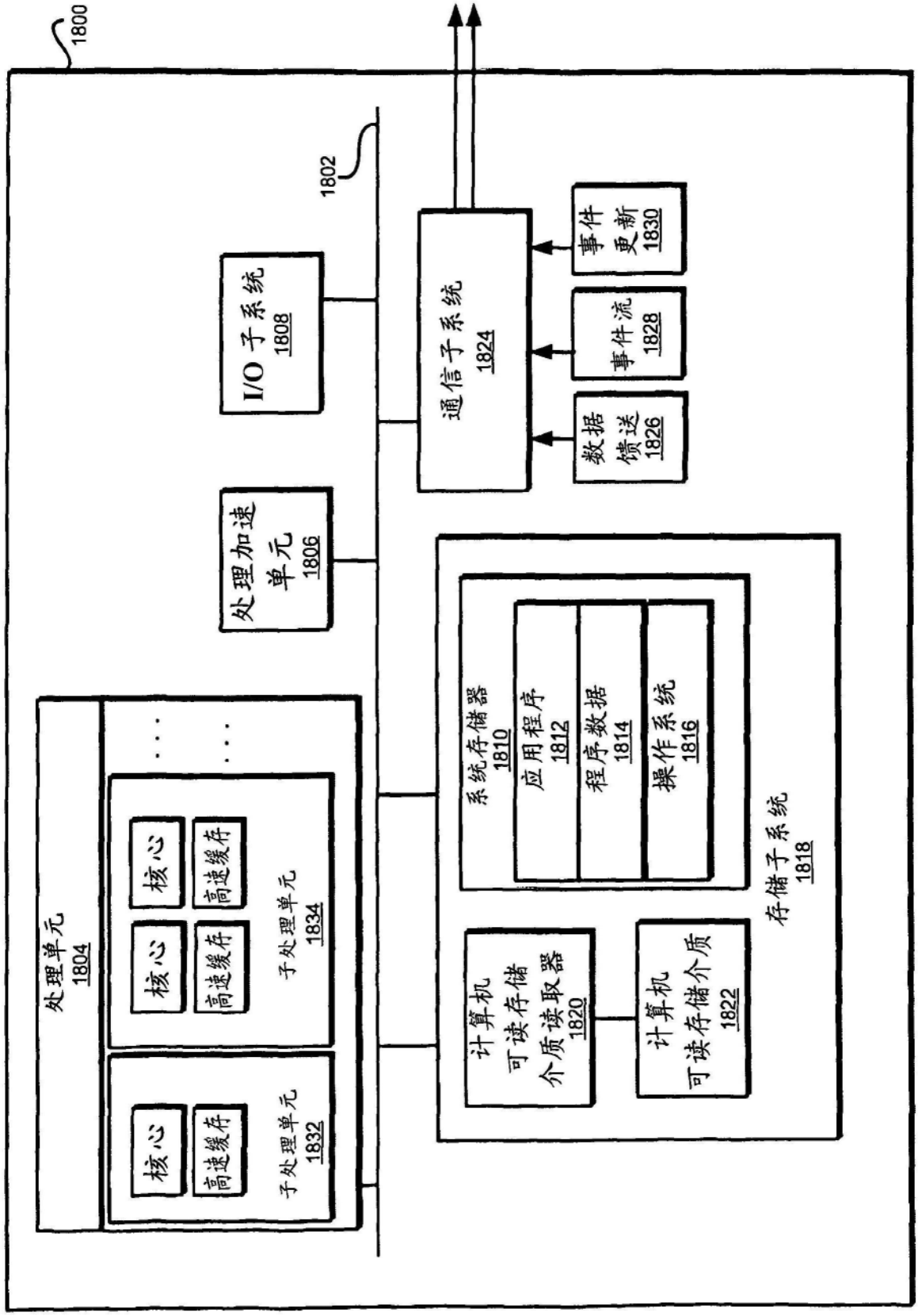

1800
1802
处理单元
1804
核心
高速缓存
子处理单元
1832
核心
高速缓存
核心
高速缓存
子处理单元
1834
I/O 子系统
1808
处理加速单元
1806
通信子系统
1824
事件更新
1830
事件流
1828
数据馈送
1826
存储子系统
1818
系统存储器
1810
应用程序
1812
程序数据
1814
操作系统
1816
计算机可读存储介质读取器
1820
计算机可读存储介质
1822

图18