(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 105393239 B
(45)授权公告日 2018.03.09

(21)申请号 201480040155.6

(22)申请日 2014.07.25

(65)同一申请的已公布的文献号
申请公布号 CN 105393239 A

(43)申请公布日 2016.03.09

(30)优先权数据
14/018,602 2013.09.05 US

(85)PCT国际申请进入国家阶段日
2016.01.14

(86)PCT国际申请的申请数据
PCT/US2014/048163 2014.07.25

(87)PCT国际申请的公布数据
WO2015/034584 EN 2015.03.12

(73)专利权人 谷歌公司
地址 美国加利福尼亚州

(72)发明人 J·T·阿德里亚恩斯
K·内斯比特 S·R·芬雷

(74)专利代理机构 北京市柳沈律师事务所 11105
代理人 邵亚丽

(51)Int.Cl.
G06F 15/16(2006.01)

审查员 徐晓艳

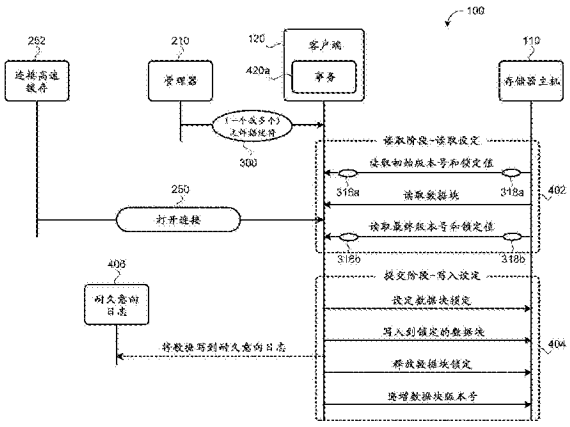
权利要求书4页 说明书22页 附图14页

(54)发明名称

隔离分布式存储系统的客户端

(57)摘要

分布式存储系统(100)包括存储器主机(110),每个存储器主机包括与存储器通信的非瞬态存储器(114)和网络接口控制器(116)并且服务来自客户端(120)的远程直接存储器访问请求(122)。存储器通过远程直接存储器访问从与存储器主机通信的每个客户端接收数据传送率(313)。每个存储器主机还包括与存储器和网络接口控制器通信的数据处理器(112)。数据处理器执行主机进程(118),该主机进程(118)读取每个所接收的客户端数据传送率,确定针对每个客户端的节流数据传送率(317),并且将每个节流数据传送率写入到由客户端通过远程直接存储器访问可访问的非瞬态存储器。



CN 105393239 B

1.一种分布式存储系统(100),包括:

存储器主机(110),每个存储器主机(110)包括:

非瞬态存储器(114);

网络接口控制器(116),所述网络接口控制器(116)与所述非瞬态存储器(114)通信并且服务来自客户端(120)的远程直接存储器访问请求(122),所述非瞬态存储器(114)通过远程直接存储器访问接收来自与所述存储器主机(110)通信的每个客户端(120)的数据传送率(313);以及

数据处理器(112),所述数据处理器(112)与所述非瞬态存储器(114)和所述网络接口控制器(116)通信,所述数据处理器(112)执行主机进程(118),所述主机进程(118):

读取每个所接收的客户端数据传送率(313);

接收隔离配置,所述隔离配置提供针对所述存储器主机的带宽容量和针对客户端的带宽预留的列表,每个带宽预留预留针对客户端的阈值数据传送率;

基于所述隔离配置,确定针对每个客户端(120)的节流数据传送率(317);以及

将每个节流数据传送率(317)写入到由所述客户端(120)通过远程直接存储器访问可访问的非瞬态存储器(114)。

2.根据权利要求1所述的分布式存储系统(100),其中在与客户端(120)建立通信连接(250)之后,所述数据处理器(112)实例化所述非瞬态存储器(114)中用于接收针对客户端(120)的所述数据传送率(313)的第一存储器区域(114n)和所述非瞬态存储器(114)中用于写入针对客户端(120)的所述节流数据传送率(317)的第二存储器区域(114m)。

3.根据权利要求2所述的分布式存储系统(100),其中所述主机进程(118)在确定针对每个客户端(120)的所述节流数据传送率(317)之前,周期性地读取针对每个客户端(120)的所述第一存储器区域(114n)。

4.根据权利要求2所述的分布式存储系统(100),其中所述主机进程(118)向所述网络接口控制器(116)注册所述非瞬态存储器(114)的远程直接存储器可访问区域(114n)的集合,所述主机进程(118)响应于接收来自所述客户端(120)的连接请求(254)而与所述客户端(120)建立能够远程直接存储器访问的连接(250)。

5.根据权利要求4所述的分布式存储系统(100),其中当所述客户端(120)在一段时间内未能坚持其对应的节流数据传送率(317)时,所述主机进程(118)单方面地破坏与客户端(120)的所述连接(250)。

6.根据权利要求1所述的分布式存储系统(100),其中在所述客户端(120)与所述存储器主机(110)之间的阈值数据量的每次传送之后,所述非瞬态存储器(114)接收客户端(120)的所述客户端数据传送率(313)。

7.根据权利要求6所述的分布式存储系统(100),其中所述主机进程(118)在从任何一个客户端(120)接收客户端数据传送率(313)之后,确定每个客户端(120)的所述节流数据传送率(317)。

8.根据权利要求1所述的分布式存储系统(100),其中存储器主机(110)的所述带宽容量(206)包括用于服务与带宽预留(208a-n)相关联的存储器访问请求(122)的预留带宽(152)和用于服务与任何带宽预留(208a-n)不关联的存储器访问请求(122)的弹性带宽(154)。

9.根据权利要求8所述的分布式存储系统(100),其中确定客户端(120)的所述节流数据传送率(317)包括:

针对所述客户端(120)的任何相关联的带宽预留(208a-n),分配等于跨所述存储器主机(110)的那些带宽预留(208a-n)的等分共享的预留带宽(152);以及

分配关于与所述存储器主机(110)通信的所有客户端(120)的弹性带宽(154)的等分共享。

10.根据权利要求9所述的分布式存储系统(100),其中确定客户端(120)的所述节流数据传送率(317)包括:将与客户端(120)的一个或多个带宽预留(208a-n)相关联的未使用的带宽(150)重新分布到其他客户端(120)。

11.根据权利要求1所述的分布式存储系统(100),其中所述主机进程(118)将隔离等级(160)与客户端(120)相关联,所述隔离等级(160)具有一个或多个相关联的存储器访问请求(122),所述主机进程(118):

基于所述存储器主机(110)的所述带宽容量(206),确定针对每个客户端(120)的分配的带宽(155);

基于针对每个客户端(120)的分配的所述带宽(155),确定针对每个客户端(120)的每个隔离等级(160)的配给的带宽(150);

基于对应的隔离等级(160)的所述带宽(150),确定针对与每个隔离等级(160)相关联的每个存储器访问请求(122)的带宽(150);以及

基于以下各项中的至少一项,确定针对每个客户端(120)的所述节流数据传送率(317):所述客户端(120)的分配的所述带宽(155)、针对每个隔离等级(160)的配给的所述带宽(150)或者针对每个存储器访问请求(122)的所述带宽(150)。

12.根据权利要求1所述的分布式存储系统(100),还包括与所述存储器主机(110)通信的管理器(210),所述管理器(210)配给针对文件(310)的数据块(320n_k)的所述存储器主机(110)中的存储器(114),其中响应于来自与所述存储器主机(110)和所述管理器(210)通信的客户端(120)的存储器访问请求(122),所述管理器(210)将文件描述符(300)返回给所述客户端(120),所述文件描述符(300)映射所述存储器主机(110)上的文件(310)的数据块(320n_k),以用于所述存储器主机(110)上的所述数据块(320n_k)的远程直接存储器访问。

13.根据权利要求9所述的分布式存储系统(100),其中文件描述符(300)包括针对文件的每个数据块(320n_k)的客户端密钥(321),每个客户端密钥(321)允许在其存储器主机(110)上对所对应的数据块(320n_k)的访问,管理器(210)拒绝在一段时间内未能坚持其对应的节流数据传送率(317)的客户端(120)对文件描述符(300)的访问。

14.一种分布式存储系统(100)中的隔离的方法,所述方法包括:

通过远程直接存储器访问将数据传送率(313)从与非瞬态存储器(114)通信的每个客户端(120)接收到所述非瞬态存储器(114)中;

将每个接收的客户端数据传送率(313)读取到与非瞬态存储器(114)通信的数据处理器(112)中;

接收隔离配置,所述隔离配置提供针对存储器主机的带宽容量和针对客户端的带宽预留的列表,每个带宽预留预留针对客户端的阈值数据传送率;

基于所述隔离配置,在所述数据处理器(112)处确定针对每个客户端(120)的节流数据

传送率(317);以及

将每个节流数据传送率(317)从所述数据处理器(112)写入到由所述客户端(120)通过远程直接存储器访问可访问的非瞬态存储器(114)。

15.根据权利要求14所述的方法,还包括在与客户端(120)建立通信连接(250)之后,实例化所述非瞬态存储器(114)中用于接收针对客户端(120)的所述数据传送率(313)的第一存储器区域(114n)和所述非瞬态存储器(114)中用于写入针对客户端(120)的所述节流数据传送率(317)的第二存储器区域(114m)。

16.根据权利要求15所述的方法,还包括在确定针对每个客户端(120)的所述节流数据传送率(317)之前,周期性地读取针对每个客户端(120)的所述第一存储器区域(114n)。

17.根据权利要求15所述的方法,还包括:

向网络接口控制器(116)注册所述非瞬态存储器(114)的远程直接存储器可访问区域(114n)的集合;

响应于接收来自所述客户端(120)的连接请求(254),与所述客户端(120)建立能够远程直接存储器访问的连接(250)。

18.根据权利要求17所述的方法,还包括当所述客户端(120)在一段时间内未能坚持其对应的节流数据传送率(317)时,单方面地破坏与客户端(120)的所述连接(250)。

19.根据权利要求14所述的方法,还包括在所述客户端(120)与所述非瞬态存储器(114)之间的阈值数据量的每次传送之后,将客户端(120)的所述客户端数据传送率(313)接收在所述非瞬态存储器(114)中。

20.根据权利要求19所述的方法,还包括在从任何一个客户端(120)接收客户端数据传送率(313)之后,确定每个客户端(120)的所述节流数据传送率(317)。

21.根据权利要求14所述的方法,其中存储器主机(110)的所述带宽容量(206)包括用于服务与带宽预留(208a-n)相关联的存储器访问请求(122)的预留带宽(152)和用于服务与任何带宽预留(208a-n)不关联的存储器访问请求(122)的弹性带宽(154)。

22.根据权利要求21所述的方法,其中确定客户端的所述节流数据传送率包括:

针对所述客户端(120)的任何相关联的带宽预留(208a-n),分配等于跨所述分布式存储系统(100)的所述存储器主机(110)的那些带宽预留(208a-n)的等分共享的预留带宽(152);以及

分配关于与所述存储器主机(110)通信的所有客户端(120)的弹性带宽(154)的等分共享。

23.根据权利要求22所述的方法,其中确定客户端(120)的所述节流数据传送率(317)包括:将与客户端(120)的一个或多个带宽预留(208a-n)相关联的未使用的带宽(150)重新分布到其他客户端(120)。

24.根据权利要求14所述的方法,还包括:

将隔离等级(160)与客户端(120)相关联,所述隔离等级(160)具有一个或多个相关联的存储器访问请求(122);

基于所述存储器主机(110)的所述带宽容量(206),确定针对每个客户端(120)的分配的带宽(155);

基于针对每个客户端(120)的分配的所述带宽(155),确定针对每个客户端(120)的每

个隔离等级(160)的配给的带宽(150)；

基于对应的隔离等级(160)的所述带宽(150)，确定针对与每个隔离等级(160)相关联的每个存储器访问请求(122)的带宽(150)；以及

基于以下各项中的至少一项，确定针对每个客户端(120)的所述节流数据传送率(317)：所述客户端(120)的分配的所述带宽(155)、针对每个隔离等级(160)的配给的所述带宽(150)或者针对每个存储器访问请求(122)的所述带宽(150)。

25.根据权利要求19所述的方法，还包括利用客户端存储器访问请求(122)接收密钥(321)以接收对所述非瞬态存储器(114)中的数据的访问。

隔离分布式存储系统的客户端

技术领域

[0001] 本公开内容涉及分布式存储系统。

背景技术

[0002] 分布式系统一般地包括许多松散耦合的计算机，其中的每个计算机通常包括计算资源(例如，一个或多个处理器)和存储资源(例如，存储器、闪速存储器和/或磁盘)。分布式存储系统重叠分布式系统的存储资源上的存储抽象(例如，键/值存储或者文件系统)。在分布式存储系统中，运行在一个计算机上的服务器进程可以将该计算机的存储资源输出到运行在其他计算机上的客户端进程。远程过程调用(RPC)可以将数据从服务器进程传送到客户端进程。

[0003] 远程过程调用是由执行在第一机器上的客户端软件所发起并且由执行在第二机器上的服务器软件所服务的双侧软件操作。服务软件中的存储系统请求(例如，读取数据)可以要求可用的处理器，其可以将大量的限制放置在分布式存储系统上。在分布式存储系统的情况中，这意味着除非远程计算机具有用于服务客户端的请求的可用的处理器，否则客户端进程不能访问远程计算机的存储资源。此外，针对分布式系统中的处理器资源和存储资源的需求经常不匹配。具体而言，计算资源(即，处理器)可以具有繁重和/或不可预测的使用模式，而存储资源可以具有轻量并且非常可预测的使用模式。

[0004] 隔离连接到相同服务器的用户的性能通常通过限制或者拒绝服务器处的用户操作而完成。如果存在处理请求的不足的资源，则服务器可以将其拒绝。例如，服务器侧资源管理涉及跟踪由服务器上的每个用户所消耗的资源的数量。当请求进入服务器时，服务器使用其对输入请求的全局知识来决定是否服务新请求。

发明内容

[0005] 当对于服务器的客户端请求是单侧操作时(例如，远程直接存储器访问(RDMA))，诸如单侧分布式存储系统中的那些操作，不存在对请求的服务器侧处理。消除服务器侧处理当其变得过载时不仅防止服务器拒绝请求，而且其防止服务器进程甚至检测其过载。因此，不能以传统的方式管理在访问相同服务器的用户/客户端之间共享的资源，诸如带宽。例如，服务器侧资源管理不针对RDMA请求工作，这是因为服务器进程不服务请求。由专用硬件直接服务请求。尽管每个客户端可以严格地限制对服务器做出请求的速率以便减少服务器处的资源使用，但是客户端缺乏其他客户端可以使用多少资源的全局知识。客户端之间的性能隔离确保行为不当的客户端过多地减少行为良好的客户端的性能并且，允许不同的服务质量等级在客户端之间被建立。

[0006] 本公开的一个方面提供包括存储器主机的分布式存储系统。每个存储器主机包括非瞬态存储器和网络接口控制器，其与存储器通信并且服务来自客户端的直接存储器访问请求。存储器通过远程直接存储器访问从与存储器主机通信的每个客户端接收数据传送率。每个存储器主机还包括与存储器和网络接口控制器通信的数据处理器。数据处理器执

行主机进程，该主机进程读取每个所接收的客户端数据传送率，确定针对每个客户端的节流数据传送率，并且将每个节流数据传送率写入到由客户端通过远程直接存储器访问可访问的非瞬态存储器。

[0007] 本公开的实现方式可以包括以下可选特征中的一个或多个特征。在一些实现方式中，在与客户端建立通信连接之后，数据处理器实例化非瞬态存储器中用于接收针对该客户端的数据传送率的第一存储器区域和非瞬态存储器中用于写入针对该客户端的节流率的第二存储器区域。主机进程在确定针对每个客户端的节流率之前，可以周期性地读取针对每个客户端的第一存储器区域。在一些示例中，主机进程向网络接口控制器注册存储器的远程直接内容可访问的区域的集合。主机进程响应于接收来自客户端的连接请求，主机进程与客户端建立能够远程直接存储器访问的连接。当客户端在一段时间期间未能坚持其对应的节流数据传送率时，主机进程可以单方面地破坏与客户端的连接。

[0008] 在一些实现方式中，在客户端与存储器主机之间的阈值数据量的每次传送之后，存储器接收客户端的客户端数据传送率。在从任何一个客户端接收客户端数据传送率之后，主机进程可以确定每个客户端的节流数据传送率。

[0009] 主机进程可以接收隔离配置，该隔离配置提供针对存储器主机的带宽容量和针对客户端的带宽预留的列表。每个带宽预留预留针对客户端的阈值数据传送率。主机进程基于隔离配置来确定客户端的节流数据传送率。存储器主机的带宽容量可以包括用于服务与带宽预留相关联的存储器访问请求的预留带宽和用于服务与任何带宽预留不关联的存储器访问请求的弹性带宽。当针对客户端的任何相关联的带宽预留而确定客户端的节流数据传送率时，主机进程可以分配等于跨主机进程的那些带宽预留的等分共享的预留带宽和分配关于与存储器主机通信的所有客户端的弹性带宽的等分共享。此外，当确定客户端的节流数据传送率时，主机进程可以将与客户端的一个或多个带宽预留相关联的未使用的带宽重新分布到其他客户端。

[0010] 在一些实现方式中，主机进程将隔离等级与客户端相关联。隔离等级具有一个或多个相关联的存储器访问请求。主机进程基于存储器主机的带宽容量，确定针对每个客户端的所分配的带宽；基于针对每个客户端的所分配的带宽，确定针对每个客户端的每个隔离等级的所配给(allot)的带宽；基于对应的隔离等级的带宽，确定针对与每个隔离等级相关联的每个存储器访问请求的带宽；以及基于以下中各项的至少一项，确定针对每个客户端的节流传送率：客户端的所分配的带宽；针对每个隔离等级的所配给的带宽；或者针对每个存储器访问请求的带宽。

[0011] 分布式存储系统可以包括与存储器主机通信的管理器(curator)。管理器配给针对文件的数据块的存储器主机中的存储器。响应于来自与存储器主机和管理器通信的客户端的存储器访问请求，管理器将文件描述符返回到客户端，该文件描述符映射存储器主机上的文件的数据块以用于存储器主机上的数据块的远程直接存储器访问。文件描述符包括针对文件的每个数据块的客户端密钥。每个客户端密钥允许对其存储器主机上的对应的数据块的访问。管理器拒绝对未能在一段时间期间坚持其对应的节流数据传送率的客户端的文件描述符的访问。

[0012] 本公开的另一方面提供一种分布式存储系统中的隔离的方法。该方法包括通过远程直接存储器访问将来自与存储器通信的每个客户端的数据传送率接收到非瞬态存储器

中并且将每个所接收的客户端数据传送率读取到与非瞬态存储器通信的数据处理器中。该方法还包括确定针对每个客户端的节流数据传送率和将每个节流数据传送率从数据处理器写入到由客户端通过远程直接存储器访问可访问的非瞬态存储器。该方面可以包括以下可选特征中的一个或多个特征。

[0013] 在一些实现方式中，该方法包括在与客户端建立通信连接之后，实例化非瞬态存储器中用于接收针对该客户端的数据传送率的第一存储器区域和非瞬态存储器中用于写入针对该客户端的节流率的第二存储器区域。该方法还可以包括在确定针对每个客户端的节流率之前，周期性地读取针对每个客户端的第一存储器区域。该方法可以包括向网络接口控制器注册存储器的远程直接存储器可访问的区域的集合，并且响应于接收来自客户端的连接请求而与客户端建立能够远程直接存储器访问的连接。如果客户端在一段时间期间未能坚持其对应的节流数据传送率，则该方法可以包括单方面地破坏与客户端的连接。

[0014] 该方法可以包括在客户端与存储器之间的阈值数据量的每次传送之后，将客户端的客户端数据传送率接收在存储器中。此外，该方法可以包括在从任何一个客户端接收客户端数据传送率之后，确定每个客户端的节流数据传送率。

[0015] 在一些实现方式中，该方法包括接收隔离配置，该隔离配置提供针对存储器主机的带宽容量和针对客户端的带宽预留的列表，并且基于隔离配置而确定客户端的节流数据传送率。每个带宽预留预留针对客户端的阈值数据传送率。存储器主机的带宽容量可以包括用于服务与带宽预留相关联的存储器访问请求的预留带宽和用于服务与任何带宽预留不关联的存储器访问请求的弹性带宽。

[0016] 确定客户端的节流数据传送率的步骤可以包括：针对客户端的任何相关联的带宽预留，分配等于跨分布式存储系统的存储器主机的那些带宽预留的等分共享的预留带宽和分配关于与存储器主机通信的所有客户端的弹性带宽的等分共享。该步骤还可以包括将与客户端的一个或多个带宽预留相关联的未使用的带宽重新分布到其他客户端。

[0017] 在一些实现方式中，该方法包括将具有一个或多个相关联的存储器访问请求的隔离等级与客户端相关联，并且基于存储器主机的带宽容量，确定针对每个客户端的所分配的带宽；基于针对每个客户端的所分配的带宽，确定针对每个客户端的每个隔离等级的所配给的带宽；基于对应的隔离等级的带宽，确定针对与每个隔离等级相关联的每个存储器访问请求的带宽；以及基于以下各项中的至少一项，确定针对每个客户端的节流传送率：客户端的所分配的带宽、针对每个隔离等级的所配给的带宽或者针对每个存储器访问请求的带宽。该方法可以包括接收具有客户端存储器访问请求的密钥以接收对存储器中的数据的访问。

[0018] 在附图和以下描述中阐述本公开的一个或多个实现方式的细节。其他方面、特征和优点将从说明书和附图并且从权利要求书而显而易见。

附图说明

[0019] 图1A是示例性分布式存储系统的示意图。

[0020] 图1B是具有由管理器所管理的存储器主机单元的示例性分布式存储系统的示意图。

[0021] 图1C是分布式存储系统的示例性单元的示意图。

[0022] 图1D是与客户端交互的示例性存储器主机的示意图。

[0023] 图2A是用于分布式存储系统的示例性管理器的示意图。

[0024] 图2B是分成复制的条纹的示例性文件的示意图。

[0025] 图2C是示例性文件描述符的示意图。

[0026] 图3A是建立客户端与示例性分布式存储系统的存储器主机之间的连接的示意图。

[0027] 图3B是将存储器访问请求发送给示例性分布式存储系统的存储器主机的客户端的示意图。

[0028] 图4A是示例性应用编程接口的示意图。

[0029] 图4B是将数据写入到存储在分布式存储系统中的文件的示例性事务的示意图。

[0030] 图4C是读取来自存储在分布式存储系统中的文件的数据的示例性事务的示意图。

[0031] 图4D是读取和写入示例性分布式存储系统中的数据的客户端的示意图。

[0032] 图5是用于分布式存储系统中的隔离的方法的操作的示例性布置的示意图。

[0033] 各附图中的相同参考标记指代相同元件。

具体实施方式

[0034] 参考图1A-1C，在一些实现方式中，分布式存储系统100包括松散耦合的存储器主机110、110a-n(例如，计算机或者服务器)，每个具有与可以被用于高速缓存数据的存储资源114(例如，存储器、闪速存储器、动态随机存取存储器(DRAM)、相变存储器(PCM)和/或磁盘)通信的计算资源112(例如，一个或多个处理器或者中央处理单元(CPU))。重叠在存储资源114上的存储抽象(例如，键/值存储或者文件系统)允许一个或多个客户端120、120a-n的存储资源114的可扩展的使用。客户端120可以通过网络130与存储器主机110通信(例如，经由RPC)。

[0035] 单侧分布式存储系统100可以消除对用于对响应于来自客户端120的远程过程调用(RPC)的任何服务器作业的需要以存储或者检索其对应的存储器主机110上的数据312，并且作为替代可以依赖于处理远程请求122的专用硬件。“单侧”是指在存储器主机110上处理的大部分请求可以以硬件而不是执行在存储器主机110的CPU112上的软件完成的方法。不是使存储器主机110(例如，服务器)的处理器112执行将对应的存储资源114(例如，非瞬态存储器)的访问输出给执行在客户端120上的客户端进程的服务器进程，客户端120可以通过存储器主机110的网络接口控制器(NIC)116直接访问存储资源114。换句话说，在不要求执行在计算资源112上的任何服务器进程的例程的执行的情况下，执行在客户端120上的客户端进程可以直接与一个或多个存储资源114对接。这提供单侧分布式存储架构，其提供相对高吞吐量和低延迟，这是因为客户端120可以在不与存储器主机110的计算资源112对接的情况下访问存储资源114。这具有将典型的双侧分布式存储系统承载的存储114和CPU周期的要求解耦的作用。无论在该存储器主机110上是否存在空闲CPU周期，单侧分布式存储系统100都可以利用远程存储资源114；此外，由于单侧操作不竞争服务器CPU资源112，因而甚至当存储器主机110以高CPU利用运行时，单侧系统可以利用非常可预测的低延迟服务高速缓存请求122。因此，单侧分布式存储系统100允许比传统双侧系统更高集群存储114和CPU资源112二者的利用。

[0036] 在一些实现方式中，分布式存储系统100包括存储逻辑部分102、数据控制部分104

和数据存储部分106。存储逻辑部分102可以包括事务应用编程接口(API)400(例如,单侧事务系统客户端库),其负责经由单侧操作访问底层数据。数据控制部分104可以利用任务管理对存储资源114的分配和访问,所述任务诸如分配存储资源114、向对应的网络接口控制器116注册存储资源114、建立客户端120与存储器主机110之间的(一个或多个)连接、处置机器故障的情况中的错误等。数据存储部分106可以包括松散耦合的存储器主机110、110a-n。

[0037] 在一些实现方式中,分布式存储系统100将数据312存储在动态随机存取存储器(DRAM)114中并且经由能够远程直接存储器访问(RDMA)的网络接口控制器116来服务来自远程主机110的数据312。网络接口控制器116(还被称为网络接口卡、网络适配器或者LAN适配器)可以是计算机硬件部件,其将计算资源112连接到网络130。网络接口控制器116使用诸如以太网、Wi-Fi或者令牌环的特定物理层(OSI第1层)和数据链路层(第2层)实现通信电路。这提供针对全网络协议栈的基础,其允许相同LAN上的小计算机组之间的通信和通过诸如因特网协议(IP)的可路由协议的大型网络通信。存储器主机110a-n和客户端120二者可以各自具有用于网络通信的网络接口控制器116。执行在存储器主机110的计算处理器112上的主机进程118向网络接口控制器116注册存储器114的远程直接存储器可访问的区域114a-n的集合。主机进程118可以向只读或者读/写的许可注册存储器114的远程直接存储器可访问的区域114a-n。存储器主机110的网络接口控制器116创建针对每个所注册的存储器区域114a-n的客户端密钥321。

[0038] 由网络接口控制器116所执行的单侧操作可以限于简单的读取、写入和比较并交换操作,其中没有一个可以足够复杂以充当用于由实现高速缓存请求和管理高速缓存策略的传统高速缓存服务器作业所实现的软件逻辑的简易替代者(drop-in replacement)。事务API 400将诸如查找或者插入数据命令的命令转译为基元网络接口控制器操作的序列。事务API 400与分布式存储系统100的数据控制和数据存储部分104、106对接。

[0039] 分布式存储系统100可以包括向网络接口控制器116注册用于远程访问的存储器114并且建立与客户端进程128的连接250(图3A和3B)的共同定位的软件过程。一旦建立连接250,则客户端进程128在没有来自对应的存储器主机110的本地CPU 112上的软件的参与的情况下,可以经由网络接口控制器116的硬件中的引擎访问所注册的存储器114。

[0040] 参考图1B和图1C,在一些实现方式中,分布式存储系统100包括多个单元200,每个单元200包括存储器主机110和与存储器主机110通信的管理器210。管理器210(例如,过程)可以执行在连接到网络130的计算处理器202(例如,服务器)上并且管理数据存储(例如,管理存储在存储器主机110上的文件系统)、控制数据放置和/或发起数据恢复。此外,管理器210可以跟踪存储器主机110上的数据的存在和存储位置。冗余的管理器210是可能的。在一些实现方式中,(一个或多个)管理器210跟踪跨多个存储器主机110的数据312的分段和针对冗余和/或性能的给定条纹的多个副本的存在和/或位置。在计算机数据存储中,数据加条纹是以以下方式逻辑地分割诸如文件310(图2B)的顺序数据312的技术:对不同的物理存储设备(例如,单元200和/或存储器主机110)做出顺序分段的访问。当处理设备请求访问数据312比存储设备可以提供访问更迅速时,加条纹是有用的。通过在多个设备上执行分段访问,可以同时访问多个分段。这提供更多数据访问吞吐量,其避免使得处理器空闲地等待数据访问。

[0041] 在一些实现方式中，事务API 400在客户端120(例如，利用客户端进程128)与管理器210之间对接。在一些示例中，客户端120通过一个或多个远程过程调用(RPC)与管理器210通信。响应于客户端请求122，事务API 400可以找到(一个或多个)存储器主机110上的特定数据的存储位置，并且获得允许对数据312的访问的密钥321。事务API 400(经由网络接口控制器116)与适当的存储器主机100直接通信以读取或者写入数据312(例如，使用远程直接存储器访问)。在存储器主机110是不可操作的，或者数据312移动到不同的存储器主机110的情况下，客户端请求122未能提示客户端120重新询问管理器210。

[0042] 参考图2A，在一些实现方式中，管理器210存储和管理文件系统元数据212。元数据212包括将文件310_{1-n}映射到文件描述符300_{1-n}的文件映射214。管理器210可以检查和修改其永久元数据212的表示。管理器210可以使用针对元数据212的三个不同的访问模式：只读、文件事务和条纹事务。只读访问允许管理器210利用最小竞争来检查元数据212的状态。只读请求返回文件310的最新状态，但是不与并发更新同步。只读访问可以被用于对来自客户端120的查找请求做出响应(例如，针对内部操作，诸如文件扫描)。

[0043] 还参考图2B和2C，在一些实现方式中，存储器主机110存储文件数据312。管理器210可以将每个文件310(和其数据312)分为条纹320a-n并且复制多个存储位置中的存储的条纹320a-n。条纹复制$320n_k$还被称为块或者数据块$320n_k$。可变文件310可以具有存储在(一个或多个)存储器主机110上的附加的元数据212，诸如锁字和版本号。锁字和版本号可以被用于实现分布式事务提交协议。

[0044] 由管理器210所存储的文件描述符300_{1-n}包含诸如文件映射214的元数据212，该元数据212将条纹320a-n映射到存储在存储器主机110上的数据块$320n_k$(即，条纹复制)。为了打开文件310，客户端120将请求122发送给管理器210，其返回文件描述符300。客户端120使用文件描述符300将文件数据库偏移转移到远程存储器位置114a-n。在客户端120加载文件描述符300之后，客户端120可以经由RDMA或者另一数据检索方法访问文件310的数据312。

[0045] 参考图3A和3B，RDMA是基于连接的进程对进程通信机制，因此RDMA连接自身通常不支持验证或者加密。因此，分布式存储系统100将RDMA连接250视为安全资源。为了客户端进程128通过RDMA访问主机进程118的存储器114，存储器主机110的网络接口控制器116执行与客户端进程128的网络接口控制器116的连接握手，以建立主机进程118与客户端进程128之间的能够RDMA的连接250。RDMA连接握手可以实现高层安全协议，该高层安全协议评价如在所信任的RDMA连接250的创建的时间处已知的主机和客户端进程118、128的身份。在建立能够RDMA的连接250之后，客户端进程128或者主机进程118可以单方面地破坏连接250。如果客户端进程128或者客户端进程118死亡，客户端120和/或存储器主机110(经由操作系统)可以拆除对应的(一个或多个)RDMA连接250。

[0046] 可以通过访问控制列表260控制对存储在远程存储器位置114a-n中的文件数据312(例如，数据块$320n_k$)的访问。每个访问控制列表260可以具有唯一名称、数据块$320n_k$的列表和客户端120a-n的列表，其具有对读取和写入与该访问控制列表260相关联的数据块$320n_k$的许可。在一些示例中，访问控制列表260提供针对每个相关联的客户端120或者每个相关联的数据块$320n_k$的访问许可级别。存储器主机110可以通过安全通信信道接收访问控制列表260并且可以由存储器主机110使用保护域270来实施。向每个存储器主机110的网络接口控制器116注册的每个RDMA访问的存储器区域114a-n与保护域270相关联。在一些实现

方式中，当管理器210配给针对数据块320n_k的存储器114时，其将数据块320n_k的所配给的存储器区域114a-n与一个或多个保护域270相关联。存储器主机110可以具有与其存储器114的各区域114a-n相关联的保护域270。每个保护域270还可以具有一个或多个相关联的连接250。

[0047] 当客户端120实例化针对存储在存储器主机110中的一个或多个上的文件310的存储器访问请求122时，客户端120请求来自管理器210的文件描述符300以标识哪一个或多个存储器主机110存储文件310的数据块320n_k。除了将文件310的数据块320n_k映射到存储器主机110的存储器区域114a-n外，文件描述符300还可以包括用于访问那些数据块320n_k的客户端密钥321。然后，客户端120搜索针对所标识的存储器主机110的任何能够开放RDMA的连接250的连接高速缓存252。如果每个存储器主机110未能具有与客户端120的打开连接250，所述客户端120在与所请求的(一个或多个)数据块320n_k相同的保护域270中，则客户端120将连接请求254发送给不具有必要的(一个或多个)开放连接250的任何存储器主机110。

[0048] 响应于接收来自客户端120的客户端进程128的用于访问数据块320n_k(例如，用于访问存储数据块320n_k的存储器区域114a-n)的连接请求254，当客户端120和所请求的数据块320n_k与由存储器主机110所接收的相同访问控制列表260相关联时，主机进程128可以与客户端进程128建立能够远程直接存储器访问的连接250。客户端进程128可以包括连接请求254中的访问控制列表260。主机进程118可以将所建立的开放连接250与保护域270相关联，并且客户端进程128可以存储连接高速缓存252中的开放连接250。连接250能够仅访问与其保护域270相关联的存储器区域114a-n。在接收到具有未注册的存储器的地址的RDMA请求时，存储器主机110的网络接口控制器116可以拆除连接250。

[0049] 在图3B中所示的示例中，第一和第二客户端120a、120b在第一和第二RDMA连接250a、250b上将存储器访问请求122发送给存储器主机110n。存储器主机110n具有与其存储器114相关联的第一和第二保护域270a、270b。第一保护域270a与第一和第二存储器区域114a、114b(例如，存储对应的第一和第二数据块320n_1、320n_2)和第一RDMA连接250a相关联，而第二保护域270b与第三存储器区域114c(例如存储对应的第三数据块320n_3)和仅第二RDMA连接250a相关联。

[0050] 第一客户端120a在第一RMDA连接250a上将第一和第二存储器访问请求122a、122b发送给存储器主机110n。第一存储器访问请求122a用于访问第二数据块320n_2的第二存储器区域114b，并且第二存储器访问请求122b用于访问针对第三数据块320n_3的第三存储器区域114c。第一存储器访问请求122a成功，这是因为第二存储器区域114b属于与第一连接250a相同的保护域270a。第二存储器访问请求122b失败，这是因为第三存储器区域114c属于不同的保护域270，第二保护域270b而不是第二存储器访问请求122b的保护域270(即，第一保护域270a)。

[0051] 第二客户端120b在第二RDMA连接上将第三和第四访问请求122c、122d发送给存储器主机110n。第三存储器访问请求122c用于访问针对第一数据块320n_1的第一存储器区域114a，并且第四存储器访问请求122d用于访问针对第三数据块320n_3的第三存储器区域114c。在这种情况下，存储器访问请求122c、122d二者成功，因为第二客户端120b的RDMA连接250b属于第一存储器区域114a和第三存储器区域114c二者的保护域270a、270b。

[0052] 当对存储器主机110(例如，服务器)的客户端请求122是单侧操作(例如，远程直接

存储器访问(RDMA))时，不存在请求的服务器侧处理。消除服务器侧处理不仅在其变得过载时防止存储器主机110拒绝请求122；而且其防止服务器进程甚至检测存储器主机110过载。因此，不能以传统的方式管理在客户端120之间共享的计算资源112和/或存储资源114的带宽。例如，服务器侧资源管理不针对RDMA请求122工作，这是因为主机进程118不服务请求122。通过专用硬件、网络接口控制器116直接服务请求122。虽然每个客户端120可以严格地限制对存储器主机110做出请求的速率以便减少存储器主机110处的资源使用，但是客户端120缺乏其他客户端120可以使用多少资源112、114的全局知识。客户端120之间的性能隔离确保行为不当的客户端120不必要地减少行为良好的客户端120的性能并且允许不同的服务质量等级在客户端120之间被建立。

[0053] 再次参考图1B和1D，在一些实现方式中，每个客户端120跟踪在其与每个存储器主机110之间传送的数据量312并且将所传送的数据量313(还被称为所传送的字节)写入到存储器主机110上的RDMA可访问的存储器区域114n。换句话说，每个客户端120保持读/写到存储器主机110的字节的总数目的当前和，并且周期性地将该和写入到存储器主机110。每个客户端120具有每个存储器主机110上的其自身的存储器区域114n。存储器主机110在连接建立时间处创建和发起存储器区域114n，并且在连接250的初始建立时，将存储器区域114n的位置发送给客户端120。在所传送的数据量312例如128字节中的阈值变化之后，客户端120写入存储器区域114n。该策略使得使用更多带宽并且更可能要求节流以更频繁地更新其所传送的字节313的客户端120和低带宽客户端120与存储器主机110较不频繁地通信。推送所传送的字节312的更新的阈值可以基于实际的实现方式、网络等而广泛地变化。要求客户端120将其所传送的字节313推送到存储器主机110简化服务器隔离逻辑，并且进而大大地减少CPU使用。

[0054] 存储器主机110周期性地扫描包含客户端的所传送的字节313(例如，每隔100毫秒)的存储器区域114n，计算带宽使用并且计算客户端带宽共享317(还被称为存储器主机110的节流率)。在一些实现方式中，存储器主机110周期性地读取所传送的字节数量313(例如，和)，将其与其读取的最后的和相比较并且根据差计算客户端120的数据速率315。扫描速率可以依赖实现方式。扫描之间的较短时间导致客户端120的极好的颗粒控制，但是折衷是较高的服务器CPU使用。存储器主机110将每个客户端的所计算的节流率317写入到另一本地存储器区域114m。当客户端120将所传送的字节313写入到存储器主机110时，客户端120例如经由RDMA读取来自该存储器主机110的该节流率317。客户端120例如经由RDMA将其数据速率315限制到从存储器主机110最新读取的节流率317。

[0055] 客户端120负责读取来自存储器主机110的其当前节流率317并且自我实施该节流率317。客户端120还负责跟踪并且周期性地将其自身的所传送的字节313写入到存储器主机110。在没有主机进程118跟踪硬件处理的每个RDMA请求122的情况下，这给予存储器主机110针对每个所连接的客户端120的数据速率315的所要求的全局知识。存储器主机110利用该信息，可以划分针对每个客户端120的带宽150并且计算适当的节流率317。存储器主机110的带宽150可以包括预留部分152(预留带宽)和弹性部分154(弹性带宽)。在一些实现方式中，弹性带宽是任何未使用的预留带宽152。

[0056] 默认情况下，每个客户端120可以从存储器主机110接收带宽150的同等共享。存储器主机110可以通过将任何未使用的带宽150分布在可以使用其的客户端120之间而连续工

作。在一些实现方式中，未相等地分布带宽150。第一，存储器主机110将预留带宽152分配给每个客户端120，并且任何未使用的带宽152被放置在弹性池中作为弹性带宽154。除不需要弹性带宽154的客户端120不将其从池取出，而是作为替代使其相同地在可以利用额外带宽150的客户端120之间划分外，存储器主机110可以将弹性池或者弹性带宽154相等地划分在客户端120之间。如果可用于客户端120的带宽数量150不是足够的，或者客户端120要求带宽保证(例如，因为同等共享带宽150可以随时间变化)，则客户端120可以请求将预留带宽152分配到隔离等级160。

[0057] 隔离等级160允许与客户端120相同运行的请求122接收分化的服务。客户端120可以具有多个相关联的隔离等级160。存储器主机110可以使用标识符来定义隔离等级160，诸如客户端名称加上任意的字符串。客户端120可以执行具有一个或多个客户端请求122的应用124。每个客户端120可以具有一个或多个相关联的隔离等级160，并且每个隔离等级160可以包含一个或多个客户端请求122。客户端标记可以确定请求122应当使用哪个隔离等级160。备选地，可以由在每请求基础上的客户端120指定隔离等级160，因此单个客户端120可以使用多个隔离等级160。运行为不同客户端120的请求122可以不共享相同隔离等级160，这是因为隔离等级160是客户端120的子代。备选实现方式可以具有跨多个客户端120的隔离等级160。可以对隔离等级160分配弹性带宽154加上预留带宽152。

[0058] 在一些实现方式中，客户端120、隔离等级160和客户端请求122形成层次关系。每个客户端120可以具有一个或多个相关联的隔离等级160，并且每个隔离等级160可以具有一个或多个相关联的客户端请求122。存储器主机110可以首先将其带宽150划分在客户端120之间。然后，对于每个客户端120，存储器主机110将针对相应的客户端120的所分配的带宽155划分在其相关联的隔离等级160之间。然后，对于每个隔离等级160，存储器主机110将对应的所分配的带宽155划分在相关联的客户端请求122之间。

[0059] 在每秒字节方面，每个单元200具有额定容量。原则上，单元200的额定容量是客户端120每秒可以从单元200读取和写到单元200的数据量312。实际上，单元200的额定容量可以均匀地划分在单元200中的存储器主机之上并且实施在每存储器主机基础上。例如，具有1000个存储器主机110和额定容量为1TB/S的单元200可能需要至少提供单元200中的每个存储器上的1GB/s的负载，以便服务1TB/s的数据312。存储器主机110的额定带宽容量206可以小于存储器主机110的网络接口控制器带宽，但是其不大于网络接口控制器带宽。

[0060] 存储器主机110根据单元隔离配置204(例如，被存储为文件)来访问和计算带宽150的共享。单元隔离配置204包括以兆字节每秒为单元的每个存储器主机110的额定带宽容量206和带宽预留208a-n的列表208。每个带宽预留208a-n包括客户端名称、隔离等级160和以兆字节每秒为单元所指定的带宽150。在一些示例中，隔离配置204不提供弹性带宽154，其可以是任何未使用的预留带宽152。

[0061] 在一些实现方式中，如由单元隔离配置204所阐述的隔离策略仅应用于有超过其额定带宽容量206的危险的存储器主机110。一旦被接合，则隔离策略旨在将存储器主机110的带宽150公平地分布在活跃地访问该存储器主机110的客户端120之间。隔离策略可以试图将带宽150均匀地分布到活跃的客户端120直到客户端120的所提供的负载。例如，具有1GB/s额定带宽容量206和具有.1、.2、.4和.8GB/s所提供的负载的四个活动客户端120的存储器主机110，那么公平带宽分布可以相应地是.1、.2、.35和.35GB/s。

[0062] 在一些示例中，客户端120可以访问来自数据中心内的多个过程的过载的存储器主机110。在这种情况下，隔离策略将客户端120的带宽150的公平共享均匀地分布在客户端120的隔离等级160和活跃地访问存储器主机110的任务之间。换句话说，对每个客户端120分配存储器主机110的带宽150，然后与该客户端120相关联的每个隔离等级160分割所分配的带宽150，并且然后隔离等级160内的每个客户端请求122分割隔离等级带宽165。

[0063] 需要超过其单元200的带宽150的公平共享的客户端120可以预留带宽150。带宽预留208a-n针对整个单元200根据每秒字节数。带宽预留208n均匀地分布在单元200中的所有存储器主机110之上。例如，如果单元200具有1000个存储器主机110并且客户端120预留500GB/s单元带宽150，那么保证客户端120从单元200中的每个存储器主机110接收至少.5GB/s的带宽150。如果客户端120不使用其预留带宽152，则存储器主机110可以将该客户端120的预留带宽152分布到可以使用带宽150的其他客户端120。

[0064] 带宽预留208a-n可以影响其他客户端120的公平共享带宽150。使用之前的示例，其中具有1GB/s额定带宽容量206和具有.1、.2、.4和.8GB/s的所提供的负载的四个活动客户端120的存储器主机110，如果具有.8GB/s所提供的负载的客户端120预留.2GB/s的存储器主机110的带宽150，那么存储器主机110的可用的弹性带宽154的池仅是.8GB/s。考虑到该带宽预留208n，隔离策略可以将.1、.2、.25和.45(.2预留+.25弹性)GB/s的带宽150相应地分布到客户端120。

[0065] 当存储器主机110检测到其在其额定带宽容量206之上时，存储器主机110节流使用超过其存储器主机带宽150的共享的客户端120。每个客户端120可以使用漏桶方案节流其对特定存储器主机110的访问。存储器主机110通过周期性地重新计算带宽共享和更新客户端的漏桶填充率来控制客户端的漏桶的填充率。在一些示例中，每个客户端数据通道具有拥有128kb的最大容量的漏桶，但是其他容量也是可能的并且可以依赖实现方式。漏桶的容量确定客户端120可以实现的最大突发速率。这允许其瞬时数据速率315暂时超过其节流率317。在发起RDMA操作之前，客户端120请求来自适当的漏桶的令牌。所请求的令牌的数目等于RDMA操作的有效载荷大小。如果存在可用的足够的令牌，则操作进行，如果不存在的话，则数据通道指示暂时误差已经发生，并且稍后应当重试操作。客户端120可以使逻辑在适当的位置以用于处置其他临时数据通道错误。漏桶的填充速率设定到由存储器主机110所分配的当前节流率317。

[0066] 存储器主机110还可以验证客户端120将遵守节流请求并且将不遵守节流请求的行为不当的客户端120列入黑名单。可以通过拆除存储器主机110与列入黑名单的客户端120之间的所有RDMA连接250来完成列入黑名单。

[0067] 存储器主机110向客户端120分配其预留带宽152和其存储器主机110的弹性带宽154的公平共享作为所分配的带宽155。如果客户端120的所提供的负载小于所分配的带宽155，则将预留带宽152的未使用的部分分布到其他客户端120。因此，所分配的带宽155(即，存储器主机的带宽150的客户端的共享)基于其客户端120的带宽使用而动态地改变。所分配的带宽共享155可以针对粗略地100ms是有效的，并且存储器主机110可以在另一100ms内重新计算客户端120的所分配的带宽共享155。

[0068] 在一些实现方式中，计算针对存储器主机110的客户端120的所分配的带宽共享155的算法是：

```
int ComputeFairShareBandwidth(int rated_BW, vector users) {
    available_BW = rated_BW
    for user in users
        available_BW -= user. reserved_BW
    fair_share_users = users.size
    fair_share_BW = available_BW / fair_share_users
    sorted_users = sort users from least user. BW to most
    for user in sorted_users
        unreserved_BW = user. BW - user. reserved_BW
        if unreserved_BW < fair_share_BW
            available_BW -= unreserved_BW
            --fair_share_users
            fair_share_BW = available_BW / fair_share_users
        else
            break // 剩余用户得到 fair_share_BW
    return fair_share_BW
}
```

[0069]

[0070]

[0071] 在一些实现方式中，对于隔离等级160和客户端进程128，存储器主机110计算如由存储器主机110的额定带宽206所约束的每个客户端120的所分配的带宽155、如由针对配给到客户端120的所分配的带宽155所约束的每个隔离等级160的带宽165、以及如由配给到其为成员的隔离等级160的带宽165所约束的每个客户端进程128的带宽155。在一些示例中，单独的客户端请求122可以具有或者可以不具有预留带宽152。

[0072] 计算存储器主机110的客户端120的所分配的带宽共享155的算法可以包括：

[0073]

```
user_fair_share_BW = ComputeFairShareBandwidth(rated_BW, users)
for user in users:
    user. rated_BW = user. reserved_BW + user_fair_share_BW
    class_fair_share_BW =
ComputeFairShareBandwidth(user.rated_BW, user.classes)
    for class in classes:
      class.rated_BW = class. reserved_BW + class_fair_share_BW
       task_fair_share_BW = ComputeFairShareBandwidth(class.
     rated_BW, class.tasks)
```

[0074] task.target_throttle_rate = task_fair_share_BW

[0075] 在计算针对每个客户端请求122的所分配的带宽155之后，存储器主机110调整针对每个客户端请求122的当前节流率317以接近所分配的带宽共享155。由于应用突发可以保持来自曾经实现其目标带宽共享155的应用124，因而存储器主机110可以调整节流率317以解释该突发并且更高效地使用存储器主机带宽150。

[0076] 在计算客户端带宽共享155之后，存储器主机110可以在客户端120的所测量的数据速率315小于其所分配的带宽155的情况下执行客户端节流率317的加法增加，或在客户端请求122的所测量的数据速率315大于其目标带宽共享155的情况下将客户端节流率317削减到所分配的带宽155。

[0077] 调整节流率317的示例性算法包括：

[0078]

```
for user in users
  for class in user.classes
    for task in class.tasks
      if task. BW > task.target_bandwidth_share
          task.throttle_rate = task.target_bandwidth_share
          task.throttle_adder = 1
      else
          task.throttle_rate += task.throttle_adder
           task.throttle_rate = min(class.rated_bw, task.throttle_rate)
          task.throttle_adder *= 2
```

[0079] 存储器主机110可以通过将节流率317写入到本地RDMA可访问的存储器区域114m来将节流率317传递给客户端120。例如，当客户端120将其所传送的字节313写入到存储器

主机110时(即,在每隔128KB的所传送的字节之后),客户端RDMA从存储器区域114m读取其节流率317。此外,这使得将使用更多带宽150和更可能要求节流的客户端120更频繁地更新其数据速率315。当其由于节流而不能读取或者写入时,客户端120还可以RDMA读取节流率317。该读取可以额定限于每100ms一次。

[0080] (一个或多个)管理器210可以将存储器主机节流信息合并到其负载平衡策略中,例如以归因于将该客户端120的太多数据块320n_k放置在单个存储器主机110上而最小化节流客户端120。管理器210可以接收来自包括节流信息的每个存储器主机110的状态,例如,存储器主机110是否超过其额定带宽206并且将哪些客户端120节流。如果在跨越单元200的许多存储器主机110上节流客户端120,则单元200可以警报客户端120其将使用太多带宽150。如果在单个存储器主机110(或者小数目的存储器主机100)上节流客户端120,则(一个或多个)管理器210可以将过载的(一个或多个)存储器主机110上的该客户端120的数据块320n_k迁移到单元200中的其他存储器主机110。如果该条件持续,则可以由热数据块320n_k引起节流。可以监视节流信息以检测何时单元200过载并且是否需要更多存储器主机100和带宽150添加到单元200。

[0081] 再次参考图2A和2C,在一些实现方式中,管理器210可以创建、复制、调整大小和删除文件310。其他操作也是可能的。为了服务来自客户端120的复制请求122cr,管理器210创建具有初始地设定到COPY_PENDING状态的新文件描述符300。管理器210可以设定/初始化以下字段中的一个或多个:大小、所有者、组、许可和/或备份文件。管理器210利用空条纹320n来填充文件描述符300的条纹阵列325(图3B),并且然后将文件描述符300提交给其文件映射214。将该信息提交给文件映射214允许在管理器210崩溃或者包含文件系统元数据212的平板电脑迁移到另一管理器210的情况下管理器210重新开始调整大小操作。一旦管理器210将文件描述符300提交给文件映射214,则管理器210通过通知已经发起复制操作的客户端120对客户端复制请求122cr做出响应。管理器210发起存储器主机拉数据块操作,其指示存储器主机110分配新数据块320n_k并且将备份文件的数据块320n_k读取到存储器主机110的存储器114中。当拉数据块操作成功地返回时,管理器210将新数据块320n_k添加到文件描述符300中的适当的条纹320n。管理器210将具有新数据块320n_k的条纹320n提交给文件映射214。

[0082] 在崩溃或者迁移的情况下,增量地更新文件描述符300允许新管理器210重新开始来自先前管理器210停止的位置的拷贝操作。这还允许客户端120通过检索文件描述符300(例如,经由查找方法)和检查利用数据块320n_k填充的文件描述符300中的条纹320n的数目来检查复制操作的状态。一旦所有数据块320n_k已经复制到存储器主机110,管理器210就将文件描述符的状态转换到READ并且将其提交给文件映射214。

[0083] 管理器210可以维持作为单元200的一部分的所有存储器主机110的状态信息。状态信息可以包括容量、自由空间、存储器主机110上的负载、来自客户端的视点的存储器主机110的延迟和当前状态。管理器210可以通过直接询问单元200中的存储器主机100和/或通过询问客户端120来获得该信息,以从客户端的视点采集延迟统计。在一些示例中,管理器210使用存储器主机状态信息来做出重新平衡、耗尽、恢复决策和分配决策。

[0084] (一个或多个)管理器210可以分配数据块320n_k以便处置用于文件310中的更多存储空间和用于重新平衡和恢复的客户端请求122。管理器210可以维持存储器主机负载和活

跃的负载映射216。在一些实现方式中，管理器210通过生成候选存储器主机110的列表来分配数据块320n_k并且将分配数据块请求发送给候选存储器主机110中的每个。如果存储器主机110过载或者不具有可用的空间，存储器主机110可以拒绝请求。在这种情况中，管理器210选择不同的存储器主机110。每个管理器210可以连续地扫描其文件命名空间的所指派的部分，大约每分钟检查所有元数据212。管理器210可以使用文件扫描检查元数据212的完整性、确定需要执行的工作和/或生成统计。文件扫描可以与管理器210的其他操作并发地操作。扫描自身可以不修改元数据212，但是调度要由系统的其他部件完成的工作并且计算统计。

[0085] 文件描述符300可以提供文件310的状态。文件310可以是以下状态之一：READ、READ_WRITE、DELETED、或者{CREATE,COPY,RESIZE}_PENDING。在READ状态中，客户端120可以读取文件310，但是不会写入到文件310。只读文件310针对文件310的整个寿命是只读的，即绝不直接向只读文件310写入。而是，可以将只读文件310从文件系统复制到另一文件系统。备份文件310可以被用于在存储器主机110崩溃时恢复数据312；因此，备份文件310持续文件310的整个寿命。在READ_WRITE状态中，具有适当的许可的客户端可以读取和写入可变文件的内容。可变文件310支持并发、精细粒度、随机写入。随机和顺序写入性能可以是可比较的。写入是强烈地一致的；即，如果任何客户端120可以观察写入的影响，那么所有客户端120可以观察到写入的影响。写入还可以批处理为事务。例如，客户端120可以发出由同步操作所跟随的一批异步写入。强一致性和事务处理语义确保如果任何客户端120可以观察事务中的任何写入，那么所有客户端120可以观察事务中的所有写入。在DELETED状态中，已经删除文件310。属于文件310的数据块320n_k被存储在所删除的数据块字段中并且等待垃圾回收。{CREATE,COPY,RESIZE}_PENDING状态表示文件310具有在文件上待决的创建、拷贝或者调整大小操作。

[0086] 由文件描述符300的文件编码协议缓冲器所指定的编码可以被用于文件310内的所有条纹320a-n。在一些示例中，文件编码包含以下字段："数据块"，其提供每条纹320n的若干数据块320n_k；"条纹长度"，其提供每条纹320n的若干字节；和"子条纹长度"，其提供每子条纹的若干字节。子条纹长度可以仅对READ_WRITE文件是有效的。可以通过文件描述符300的条纹协议缓冲器325的阵列来描述文件310的数据312。每个条纹320n表示由阵列内的索引所识别的文件的数据312的固定区域。条纹320n的内容可以包括数据块协议缓冲器327阵列，每个描述条纹320n内的数据块320n_k，包括数据块句柄、保持数据块320n_k的存储器主机110的标识和数据块320n_k的当前状态。出于RDMA目的，数据块协议缓冲器327还可以存储存储器主机110中的数据块320n_k的虚拟地址和客户端密钥321(例如，32位密钥)。客户端密钥321对于存储器主机110上的数据块320n_k是唯一的，并且用于RDMA读取该数据块320n_k。

[0087] 条纹320n还可以分被为与子条纹元数据324相关联的子条纹322n。每个子条纹322n可以包括子数据块326a-n阵列，每个具有对应的相关联的子数据块元数据328。

[0088] 参考图4A-4C，事务API 400可以促进具有原子性、一致性、隔离、耐久性(在某种程度上)的事务，使得事务关于其他事务是可串行化的。ACID(原子性、一致性、隔离、耐久性)是属性集，其保证可靠地处理数据库事务。在一些实现方式中，事务API 400包括读取器等级410和事务等级420。客户端120可以实例化继承读取器等级410的读取器410a以执行单元200中的存储器主机110上的读取或者读取批次。此外，客户端120可以实例化继承事务等级

420的事务420a以执行一个或多个读取和/或写入。事务420a中的读取和写入可以到单元200中的不同的文件310，但是在一些实现方式中，事务中的所有读取和写入必须到相同单元200中的文件310。所执行的读取可以是“快照一致的”，这意味着事务420a中的所有读取可以及时看到逻辑时刻处的文件310的快照。可以对写入进行缓冲，直到客户端120试图提交事务420a。

[0089] 参考图4B，响应于接收针对文件310的写入存储器访问请求122w，事务420a可以(充当写入器)写入或者修改(例如，数据块320n_k和/或子数据块326a-n的)文件310的数据312。在写操作之后，事务420a可以(例如，利用数据块320n_k和/或子数据块326a-n)计算经修改的数据312的校验和314并且将校验和314与经修改的数据312相关联。在一些示例中，事务420a存储针对经修改的子数据块326n的子数据块元数据328中的校验和314。事务420a可以执行诸如密码学散列函数的散列函数以计算校验和314。而且，散列函数可以被配置用于随机化。每个校验和314可以是具有至少64位的字。服务对应的存储器主机110上的远程直接存储器访问请求122的网络接口控制器116可以确定在其存储器主机110上访问的任何数据312的校验和314。

[0090] 当客户端120(例如，经由事务420a)将文件读取请求122r添加到读取器410a时，读取器410a将读取请求122r转译为RDMA读取网络操作并且将网络操作的状态存储在针对读取器410a所分配的存储器中。跨数据块界限的读取转译为多个RDMA操作。

[0091] 在一些实现方式中，为了将文件读取请求122r转译为RDMA读取网络操作，读取器410a根据读取请求122r的文件偏移来计算目标条纹数目。读取器410a可以使用条纹数目索引到数据块句柄高速缓存。数据块句柄高速缓存返回网络通道以访问对应的数据块320n_k和数据块320n_k的虚拟地址和r-密钥321。读取器410a将网络通道和r-密钥321直接存储在RDMA读取的操作状态中。读取器410a使用数据块320n_k和虚拟地址和文件偏移计算读取的数据块320n_k内的虚拟地址。读取器410a将偏移计算到由客户端120所提供的存储器框中(例如，接收针对每个RDMA读取操作的存储器块)。读取器410a可以然后初始化操作状态。

[0092] 当对新读取进行缓冲时，读取器410a可以计算和存储将被取回以完成读取的元数据的数量的累计和。这允许元数据缓冲空间在执行期间被分配在一个连续块中，这最小化分配开销。

[0093] 响应于接收来自客户端120的存储器访问请求122，事务420a可以检索来自管理器210的文件描述符300，其映射用于存储器主机110上的那些数据块320n_k的远程直接存储器访问的存储器主机110上的文件310的所请求的数据块320n_k。文件描述符300可以包括针对文件310的每个数据块320n_k的客户端密钥321。而且，每个客户端密钥312允许对其存储器主机110上的对应的数据块320n_k的访问。

[0094] 参考图4C，在一些实现方式中，读取器410a执行两个阶段中的读取操作。在第一阶段中，读取器410a读取文件310的数据312和相关联的元数据324、328。在第二阶段中，读取器410a验证在第一阶段中读取的数据312满足读取器410a的数据一致性约束。在第一阶段中，读取器410a识别与数据312相对应的一个或多个存储器位置并且发送其RDMA读取操作。当迭代通过并且发送RDMA读取时，读取器410a初始化和发送RDMA读取以读取子数据块元数据328并且读取计算子数据块326a-n(诸如未对齐的文件访问中的第一和最后的子数据块326a、326n)的校验和314所需要的数据312。在接收数据312和元数据328时，读取器410a可

以检查子数据块元数据328中的锁字以确保当读取数据312时未锁定子数据块326a-n。如果锁定子数据块326a-n，则读取器410a重新读取子数据块326a-n和其对应的元数据328。一旦读取器410a找到(读取)未锁定状态中的所有子数据块，则读取器410a计算子数据块校验和314并且将所计算的校验和314与从子数据块元数据328所读取的校验和314相比较。

[0095] 换句话说，为了检测读取/写入冲突，响应于接收存储在单元200的存储器主机110中的文件310的数据312的读取存储器访问请求122r，读取器410a可以计算数据312的第一校验和314a，将第一校验和314a与关联于(例如，在对应的子数据块326n的元数据328中所存储的)数据312的第二校验和314b相比较，并且当第一和第二校验和314a、314b匹配时允许对数据312的读取操作。读取器410a可以执行诸如密码学散列函数的散列函数以计算校验和314。读取器410a在接收读取/写入请求122之后并且在处理读取/写入请求122之前，可以读取数据312和与数据312相关联的元数据328。此外，读取器410a可以确定在读取数据312时是否锁定数据312，例如通过评价存储在元数据328中的锁字和/或版本号。当在先前读取数据312时锁定数据312时，读取器410a重新读取数据312和相关联的元数据328。

[0096] 虽然校验和314通常被用于避免硬件错误或者软件错误，但是使用其避免什么是实际上正常的操作提出某些附加的要求。由于冲突可能不是稀有事件，因而可以通过使校验和大小足够大以提供一致性匹配的相对小的概率，来最小化得到一致性匹配的校验和的机会。在一些示例中，64位校验和314是足够的，因为每纳秒检查随机坏校验和314可以产生低于每五个世纪一次的假阳性，其与其他类型的系统故障的速率相比，不频繁很多。此外，用于计算校验和314的散列函数可以产生针对数据312的所有共同修改的不同数目。例如，将所有数据312简单地加起来将不满足，这是因为简单地重新排序数据312中的一些数据312的改变将不改变校验和314。然而，密码学散列函数(其故意地不允许数据312的简单修改产生任何可预测的校验和314)可以是足够的。

[0097] 子数据块校验和314出于三个原因之一可能使比较失败：1)通过并发写入使所读取的数据312恶化；2)在发送到客户端时使数据312恶化；或者3)使存储在存储器主机110中的数据312恶化。情况1和2是暂态误差。通过重试子数据块读取解决暂态误差。情况3是永久误差，其可以要求客户端120通知恶化子条纹322n的管理器。

[0098] 为了在暂态误差与永久误差之间进行区分，客户端120可以重新读取子数据块数据312和子数据块元数据328。读取器410a然后检查子数据块锁字316和重新计算并比较子数据块校验和314。如果校验和误差仍然存在并且子数据块版本号318已经由于初始地读取子数据块326n而发生改变，那么由并发写入可能使得校验和比较失败，因此读取器410a重试子数据块读取。如果由于初始读取子数据块326n而版本号318尚未改变，那么误差是永久的，并且读取器410a通知管理器210，并且管理器210试图重建数据块$320n_k$的数据312。如果管理器210不能重建数据块数据，管理器210利用新未初始化的数据块$320n_k$来替换旧数据块$320n_k$。

[0099] 与锁定不同，用于检测读/写冲突的校验和比较方法实际上不关心冲突写入是否存在，只要数据312是一致的。例如，如果数据312利用相同数据312覆写，或者如果写入准备开始但是实际上尚未开始或者正好已经完成，则锁定方法将使得读取不必要地失败，而校验和比较将允许读取成功。由于锁定与解锁之间的时间可能比实际写入的持续时间大得多，因而这可以是显著的改进。

[0100] 读取器410a不知道其已经读取数据312的哪个版本318，并且其可以不重要。如果使读取事务获得版本号318是有利的，如果版本号318自身由校验和314覆盖，则这可以在没有附加的往返延迟惩罚的情况下完成。尽管计算校验和314可能招致处理器时间中的非平凡惩罚，但是对于读取器410a和写入器420a二者而言，校验和314可以无论如何必要以避免取决于实现方式的硬件误差。

[0101] 由于客户端120试图执行事务420a，但是在事务420a的提交协议期间崩溃，因而子数据块锁定可以变得阻塞。读取器410a可以通过重新读取子数据块锁字316和版本号318来检测阻塞锁定。如果子数据块锁字316和版本号318在一些超时时段期间不改变，那么子数据块锁定可能阻塞。当读取器410a检测阻塞锁定时，其通知管理器210阻塞锁定，并且管理器210恢复子条纹322n并且重置阻塞锁定。

[0102] 还参考图4A和4D，在一些实现方式中，在读取器410a验证每个子数据块锁字316和/或校验和314之后，读取器410a可以转到执行读取操作的第二阶段(即，验证阶段)。为了验证值，读取器410a重新读取子数据块元数据328并且重新检查子数据块锁字316是否未被锁定，并且由于在读取操作的第一阶段期间初始地读取版本号318，因而子数据块版本号318尚未改变。换句话说，读取器410a可以读取与事务420a的读取集402的每个数据块320n_k相关联的初始版本号318a和初始锁定值316a。在读取数据312之后，读取器410a读取与读取集402的每个数据块320n_k相关联的最终版本号318b和最终锁定值316b，并且当初始版本号318a匹配最终版本号318b并且初始锁定值316a匹配最终锁定值316b时，将读取数据312确定为有效的。

[0103] 如果读取器410a与事务420a相关联，读取器410a可以重新读取与由事务420a所读取的所有子数据块326n相关联的元数据328。如果单个子数据块版本号318误比较，则读取器410a返回错误。如果所有子数据块版本号318是相同的，则读取器410a丢弃读取器存储器块的前缀和后缀，以便修整所读取的额外数据312以计算读取中的第一和最后的子数据块326a、326n的校验和314。读取器410a可以将状态设定到OK并且返回到客户端120。

[0104] 如果读取器410a在读取数据块320n_k的数据312或者元数据212时遭遇网络信道上的错误，则读取器410a可以选择与数据块句柄高速缓存不同的数据块320n_k并且通知管理器210坏的存储器主机。如果没有读取器410a可以从其读取的其他良好的数据块320n_k，则读取器410a可以等待接收对其发送给管理器210的错误通知的响应。来自管理器210的响应可以包含更新的文件描述符300，其包含用于从其读取的新良好的数据块320n_k。

[0105] 在一些实现方式中，事务等级420使用验证集422跟踪哪些子条纹322n已经由事务420a读取。事务420a的每个读取将所有子条纹322n读取的版本号318添加到事务420a的验证集422。事务420a可以验证两个情况中的验证集422：1)作为提交协议的一部分和2)事务420a的读取的验证阶段。如果提交协议查明任何子条纹版本号318与验证集422中所记录的数目不同，则事务420a可能未能提交。在数据312返回到客户端120之前，全验证集422的验证允许预定事务420a的早期检测(例如，在提交阶段之前)。该验证还防止客户端120获得文件数据312的不一致的视图。

[0106] 事务420a可以提供同步、可串行化的读取操作(例如，使用读取器)。在一些示例中，将读取器410a实例化并且与事务420a相关联。读取器410a的读取结果返回最新的所提交的数据312。如此，未由该事务420a的读取看到相同事务420a的未提交的写入。

[0107] 事务420a可以缓冲针对稍后的事务提交的数据312。事务等级420将缓冲写入请求转译为一个或多个“准备写入”网络操作。针对由写入操作所接触的每个条纹320n需要一个网络操作。处理缓冲写入请求可以涉及准备“子条纹锁定”网络操作。针对由所请求的写入所接触的每个子条纹322n需要一个锁定操作。针对在事务提交期间的传输缓冲这些操作。事务420a可以将缓冲写入请求转译为网络操作并且执行标识或者合并影响文件310的相同区域的写入。事务420a可以通过针对所有数据块320n_k的存储器主机110以相同的顺序应用写入操作来确保所有副本是一致的。

[0108] 事务420a可以提供提交操作，其导致可调度为单个原子的可串行化的操作的事务420a中的所有读取和写入。在一些实现方式中，事务提交协议通过锁定相位、验证阶段、写入阶段和解锁阶段进行。在锁定阶段期间，发送响应于缓冲写入请求而创建的子条纹锁定网络操作。每个子条纹锁定操作执行所有副本320n_k中的锁字上的原子比较并交换操作。如果锁字的内容匹配所指定的比较数据312(例如，客户端标识符)，则利用所指定的交换数据312来写入锁字，并且返回字的先前的内容。如果客户端120在将其唯一客户端ID写入到元数据锁字中成功，则其已经成功地取得锁定。如果事务420a未能取得针对写入集中的任何子条纹322n的锁定，则提交失败并且中止。一旦保持所有子条纹锁定，则提交协议进行到验证阶段。

[0109] 在验证阶段期间，事务420a可以读取验证集中所引用的所有子条纹322n的元数据324的版本号318并且将版本号318与验证集中所记录的版本号318相比较。如果版本号318不匹配，则子条纹322n在其由该事务420a读取之后由另一事务420a写入，因此事务420a失败。在这种情况下，读取器410a释放其保持的锁定并且将事务冲突错误返回到客户端120。一旦已经验证验证集中的所有版本号318，则客户端120将事务420a的缓冲的写入数据312写入到每个副本320n_k并且更新在写入阶段期间与由事务420a所写入的每个子条纹322n相关联的元数据324。更新子条纹322n的元数据324可以包括计算和写入新检查字314、316并且递增子条纹322n的版本号318。一旦已经更新所有数据312和元数据324、328，事务420a就释放在解锁阶段期间其保持的锁定。

[0110] 对于事务420a的读取集402的数据块320n_k，执行事务420a的方法可以包括通过远程直接存储器访问来读取读取集402的数据块320n_k的数据312，并且通过评价读取集402的每个数据块320n_k的锁定316和版本318来确定读取数据312的有效性。对于事务420a的写入集404的数据块320n_k而言，该方法可以包括设定写入集404的数据块320n_k上的锁定316；通过远程直接存储器访问将数据312写入到锁定的数据块320n_k；释放锁定的数据块320n_k的锁定316；和递增每个所释放的数据块320n_k的版本号318。

[0111] 文件事务访问可以提供对文件描述符300的状态的排他性读/写访问。文件状态的更新可以在事务420a的末尾被应用并且是原子的。文件事务访问可以被用于诸如创建、最终化和删除文件310的操作。这些操作可以要求管理器210与诸如存储器主机110的其他部件通信，并且因此文件事务访问可以持续若干秒或更多。在活跃时，文件事务访问阻止需要修改文件描述符300的状态的任何其他操作。可以不阻止读取访问。

[0112] 为了减少竞争，条纹事务访问可以提供针对操作的相对更精细颗粒的同步，其仅需要利用文件描述符300来修改单个条纹320n的状态。该模式可以被用于诸如打开、关闭、重新平衡和恢复的条纹操作。可以存在针对文件310内的不同条纹320n的并发条纹事务，但

是条纹事务和文件事务是相互排斥的。在条纹事务内，管理器210可以检查条纹320n的状态和文件描述符300的各种字段，其针对诸如文件编码和实例标识符的事务420a的持续时间保持不变。条纹事务访问不提供对字段的访问，其可以改变下面的，诸如其他条纹320n的状态。操作可以一次仅保持一个活动事务以避免死锁。此外，事务420a可以仅对单个文件310原子地提交。

[0113] 图5提供用于分布式存储系统100中的隔离的方法的操作的示例性布置500。该方法包括通过远程直接存储器访问将来自与存储器114通信的每个客户端120的数据传送率接收502到非瞬态存储器114中，并且读取504每个所接收的客户端数据传送率313。该方法还包括确定506针对每个客户端120的节流数据传送率317并且将每个节流数据传送率317写入508到通过远程直接存储器访问由客户端120可访问的非瞬态存储器114。

[0114] 在一些实现方式中，方法包括在建立与客户端120的通信连接250之后，实例化非瞬态存储器114中用于接收针对该客户端120的数据传送率313的第一存储器区域114n和非瞬态存储器114中用于写入针对该客户端120的节流率的第二存储器区域114m。该方法还可以包括在确定针对每个客户端120的节流率317之前周期性地读取针对每个客户端120的第一存储器区域114n。该方法可以包括向网络接口控制器116注册存储器114的远程直接存储器可访问的区域114a-n的集合，并且响应于接收来自客户端120的连接请求254而建立与客户端120连接能够远程直接存储器访问的连接250。如果客户端120在一段时间期间未能坚持其对应的节流数据传送率317，则该方法可以包括单方面地破坏与客户端120的连接250。

[0115] 方法可以包括在客户端120与存储器114之间的阈值数据量的每次传送之后，将客户端的客户端数据传送率313接收在存储器114中。此外，方法可以包括在从任何一个客户端120接收客户端数据传送率313之后，确定每个客户端120的节流数据传送率317。

[0116] 在一些实现方式中，该方法包括接收隔离配置204，其提供针206对存储器主机110的带宽容量206和针对客户端120的带宽预留208a-n的列表208，并且基于隔离配置204来确定客户端120的节流数据传送率317。每个带宽预留208a-n针对客户端120的阈值数据传送率。存储器主机110的带宽容量206可以包括用于服务与带宽预留208a-n相关联的存储器访问请求122的预留带宽152和用于服务与任何带宽预留208a-n不相关联的存储器访问请求122的弹性带宽154。

[0117] 确定客户端120的节流数据传送率317的步骤可以包括：针对客户端120的任何相关联的带宽预留208a-n，分配152等于跨越存储器主机110的那些带宽预留208a-n的等分共享的预留带宽152和分配关于与存储器主机110通信的所有客户端120的弹性带宽154的等分共享。步骤还可以包括：确定客户端120的节流数据传送率317包括将与客户端120的一个或多个带宽预留208a-n相关联的未使用的带宽150重新分布到其他客户端120。

[0118] 在一些实现方式中，该方法包括将具有一个或多个相关联的存储器访问请求122的隔离等级160与客户端120相关联，并且基于存储器主机110的带宽容量206确定针对每个客户端120的所分配的带宽155；基于针对每个客户端120的所分配的带宽155，确定针对每个客户端120的每个隔离等级160的所配给的带宽150；基于对应的隔离等级160的带宽150，确定针对与每个隔离等级160相关联的每个存储器访问请求122的带宽150；以及基于以下各项中的至少一项，确定针对每个客户端120的节流传送率317：客户端120的所分配的带宽155、针对每个隔离等级160的所配给的带宽150或者针对每个存储器访问请求122的带宽

150。

[0119] 可以以数字电子电路、集成电路、专门设计的ASIC(专用集成电路)、计算机硬件、固件、软件和/或其组合实现本文所描述的系统和技术的各种实现方式。这些各种实现方式可以包括在包括至少一个可编程处理器的可编程系统上可执行和/或可解释的一个或多个计算机程序中的实现方式,该处理器(其可以是专用或者通用)被耦合为从存储系统、至少一个输入设备和至少一个输出设备接收数据和指令和将数据和指令发送给存储系统、至少一个输入设备和至少一个输出设备。

[0120] 这些计算机程序(还被称为程序、软件、软件应用或者代码)包括用于可编程处理器的机器指令,并且可以以高级程序和/或面向对象编程语言和/或汇编/机器语言实现。如本文所使用的,术语“机器可读介质”和“计算机可读介质”是指用于将机器指令和/或数据提供给可编程处理器的任何计算机程序产品、装置和/或设备(例如,磁盘、光盘、存储器、可编程逻辑设备(PLD)),包括接收机器指令作为机器可读信号的机器可读介质。术语“机器可读信号”是指用于将机器指令和/或数据提供给可编程处理器的任何信号。

[0121] 可以以数字电子电路或者以计算机软件、固件或者硬件实现本说明书中所描述的主题和功能操作的实现方式,包括本说明书中所公开的结构和其结构等同物或者其中的一个或多个的组合。此外,本说明书中所描述的主题可以实现为一个或多个计算机程序产品,即编码在计算机可读介质上以用于由数据处理装置执行或者控制数据处理装置的操作的计算机程序指令的一个或多个模块。计算机可读介质可以是机器可读存储设备、机器可读存储基板、存储器设备、实现机器可读传播信号的物质的组合物或者其中的一个或多个的组合。术语“数据处理装置”、“计算设备”和“计算处理器”涵盖用于处理数据的所有装置、设备和机器,以示例的方式包括可编程处理器、计算机或者多个处理器或者计算机。除硬件外,装置可以包括创建用于讨论中的计算机程序的执行环境的代码,例如构成处理器固件、协议栈、数据库管理系统、操作系统或者其中的一个或多个的组合的代码。传播信号是人工生成的信号,例如机器生成的电气、光学或者电磁信号,其被生成为编码用于传输到适合的接收器装置的信息。

[0122] 可以以包括编译或者解释语言的编程语言的形式书写计算机程序(还被称为应用、程序、软件、软件应用、脚本或者代码),并且其可以以任何形式部署,包括作为单独的程序或者模块、部件、子例程或者适于使用在计算环境中的其他单元。计算机程序不一定与文件系统中的文件相对应。程序可以被存储在保持其他程序或者数据的文件的一部分(例如,被存储在标记语言文档中的一个或多个脚本)、专用于讨论中的程序的单个文件或者多个协调文件(例如,存储一个或多个模块、子程序或者代码的部分的文件)中。计算机程序可以被部署为执行在一个计算机或者多个计算机上,其可以定位在一个地点或者跨多个地点分布并且通过通信网络互连。

[0123] 可以通过一个或多个可编程处理器执行本说明书中所描述的过程和逻辑流,其通过对输入数据进行操作并且生成输出来执行一个或多个计算机程序以执行功能。过程和逻辑流还可以通过专用逻辑电路执行并且装置还可以实现为专用逻辑电路,例如FPGA(现场可编程门阵列)或者ASIC(专用集成电路)。

[0124] 适于计算机程序的执行的处理器以示例的方式包括通用和专用微处理器二者以及任何种类的数字计算机中的任何一个或多个处理器。通常,处理器将从只读存储器或者

随机存取存储器或者二者接收指令和数据。计算机的基本元件是用于执行指令的处理器和用于存储指令和数据的一个或多个存储器设备。通常，计算机将还包括或者被操作性地耦合为从用于存储数据的一个或多个大容量存储设备(例如，磁盘、磁光盘或光盘)接收数据或者将数据传送到其或者二者。然而，计算机不是必需具有这样的设备。此外，计算机可以嵌入在另一设备中，例如移动电话、个人数字助理(PDA)、移动音频播放器、全球定位系统(GPS)接收器等。适于存储计算机程序指令和数据的计算机可读媒体包括所有形式的非易失性存储器、介质和存储器设备，以示例的方式包括半导体存储器设备，例如EPROM、EEPROM和闪速存储器设备；磁盘，例如内部硬盘或者可移动磁盘；磁光盘；和CD ROM和DVD-ROM磁盘。处理器和存储器可以通过专用逻辑电路补充或者被合并在专用逻辑电路中。

[0125] 为了提供与用户的交互，本公开内容的一个或多个方面可以实现在具有显示设备的计算机上，例如CRT(阴极射线管)、LCD(液晶显示器)监视器或者用于将信息显示给用户的触摸屏和可选地键盘和指点设备，例如鼠标或者轨迹球，用户可以通过其将输入提供给计算机。其他种类的设备也可以被用于提供与用户的交互；例如，提供给用户的反馈可以是任何形式的感觉反馈，例如视觉反馈、听觉反馈或者触觉反馈；并且可以以任何形式接收来自用户的输入，包括声音、语音或者触觉输入。另外，计算机可以通过将文档发送给由用户所使用的设备和从其接收文档而与用户交互；例如，通过响应于从网络浏览器所接收的请求而将网页发送给用户的客户端设备上的web浏览器。

[0126] 本公开内容的一个或多个方面可以实现在计算系统中，其包括例如作为数据服务器的后端部件、或者包括中间件部分，例如应用服务器、或者包括前端部件，例如具有用户可以通过的图形用户接口或者web浏览器与本说明书中所描述的主题的实现方式交互的客户端计算机、或者一个或多个这样的后端、中间件或者前端部件的任何组合。可以以任何形式或者数字数据通信(例如通信网络)介质相互连接系统的部件。通信网络的示例包括局域网(“LAN”)和广域网(“WAN”)、互连网络(例如，因特网)和对等网络(例如，自组织对等网络)。

[0127] 计算系统可以包括客户端和服务器。客户端和服务器通常彼此远离并且通常通过通信网络交互。客户端和服务器的关系借助于运行在相应的计算机上并且彼此具有客户端-服务器关系的计算机程序发生。在一些实现方式中，服务器将数据(例如，HTML页面)发送给客户端设备(例如，出于将数据显示给与客户端设备交互的用户并且从其接收用户输入的目的)。可以从服务器处的客户端设备接收在客户端设备处所生成的数据(例如，用户交互的结果)。

[0128] 尽管本说明书包含许多特性，但是这些不应当被解释为对本公开或者权利要求的范围的限制，而是特定于本公开的具体实现方式的特征的描述。不同实现方式的上下文中的本说明书中所描述的某些特征还可以实现在单个实现方式中的组合中。相反，单个实现方式的上下文中所描述的各种特征还可以分离地实现在多个实现方式中或者任何适合的子组合中。此外，尽管特征可以上文描述为在某些组合中作用并且甚至如此初始地要求保护，在一些情况下，可以从组合去除所要求保护的组合中的一个或多个特征，并且所要求保护的组合可以涉及子组合或者子组合的变型。

[0129] 类似地，尽管以特定的次序在附图中描绘了操作，但是这不应当理解为要求这样的操作以所示的特定次序或者以顺序次序执行或者所有所图示的操作被执行来实现期望

的结果。在某些情况下，多任务和并行处理可以是有利的。此外，上文所描述的实施例中的各种系统部件的分离不应当被理解为要求所有实施例中的这样的分离，并且应当理解，所描述的程序部件和系统可以一般地一起集成在单个软件产品或者被封装到多个软件产品中。

[0130] 已经描述若干实现方式。然而，应当理解，在不脱离本公开的精神和范围的情况下，可以做出各种修改。因此，其他实现方式在所附的权利要求的范围内。例如，可以以不同的次序执行权利要求中所记载的动作并且仍然实现期望的结果。

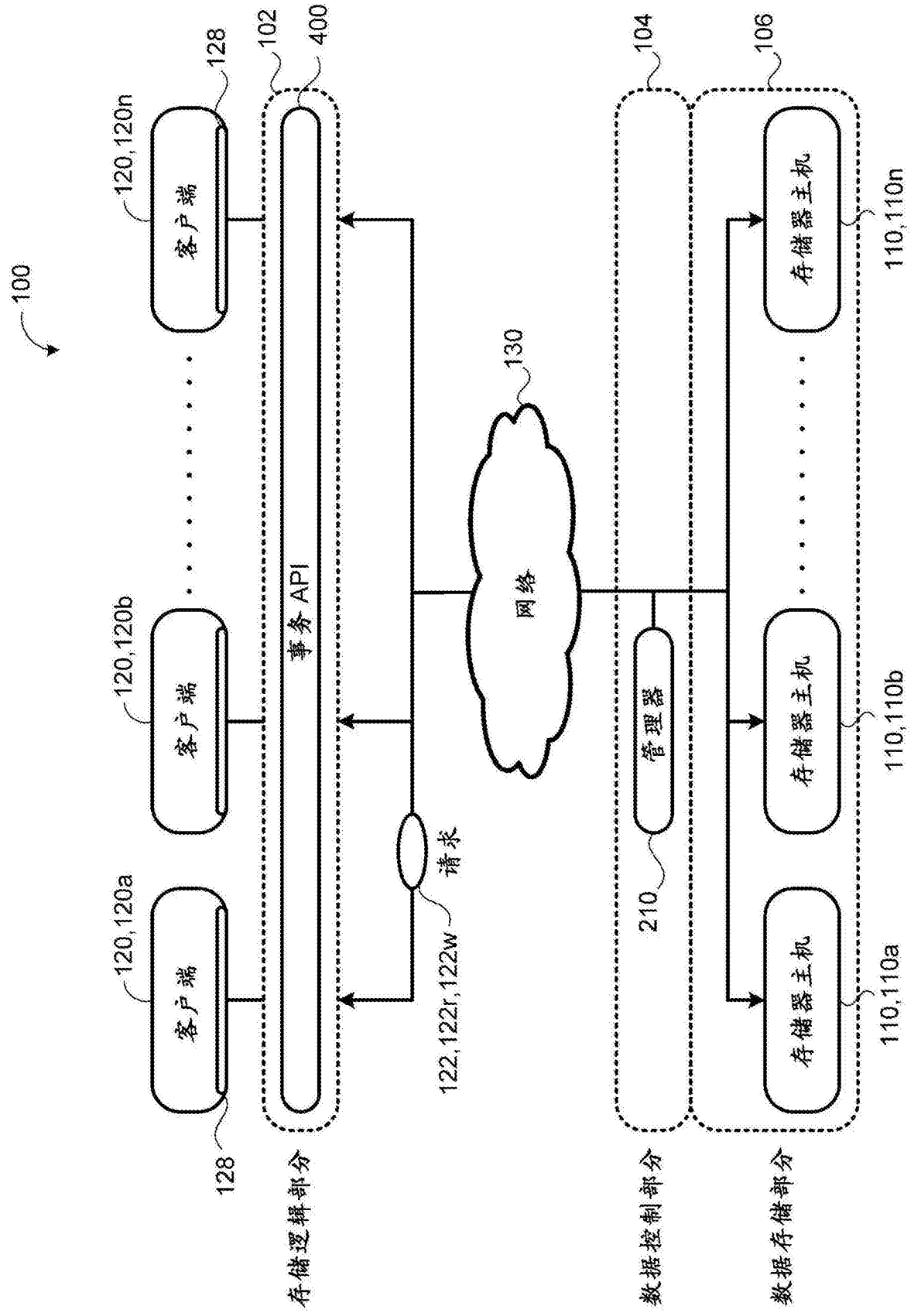

100
120,120n
客户端
128
102
400
104
106
存储器主机
110,110n
120,120b
客户端
事务 API
网络
130
管理器
存储器主机
110,110b
120,120a
客户端
请求
122,122r,122w
210
存储器主机
110,110a
128
存储逻辑部分
数据控制部分
数据存储部分

图1A

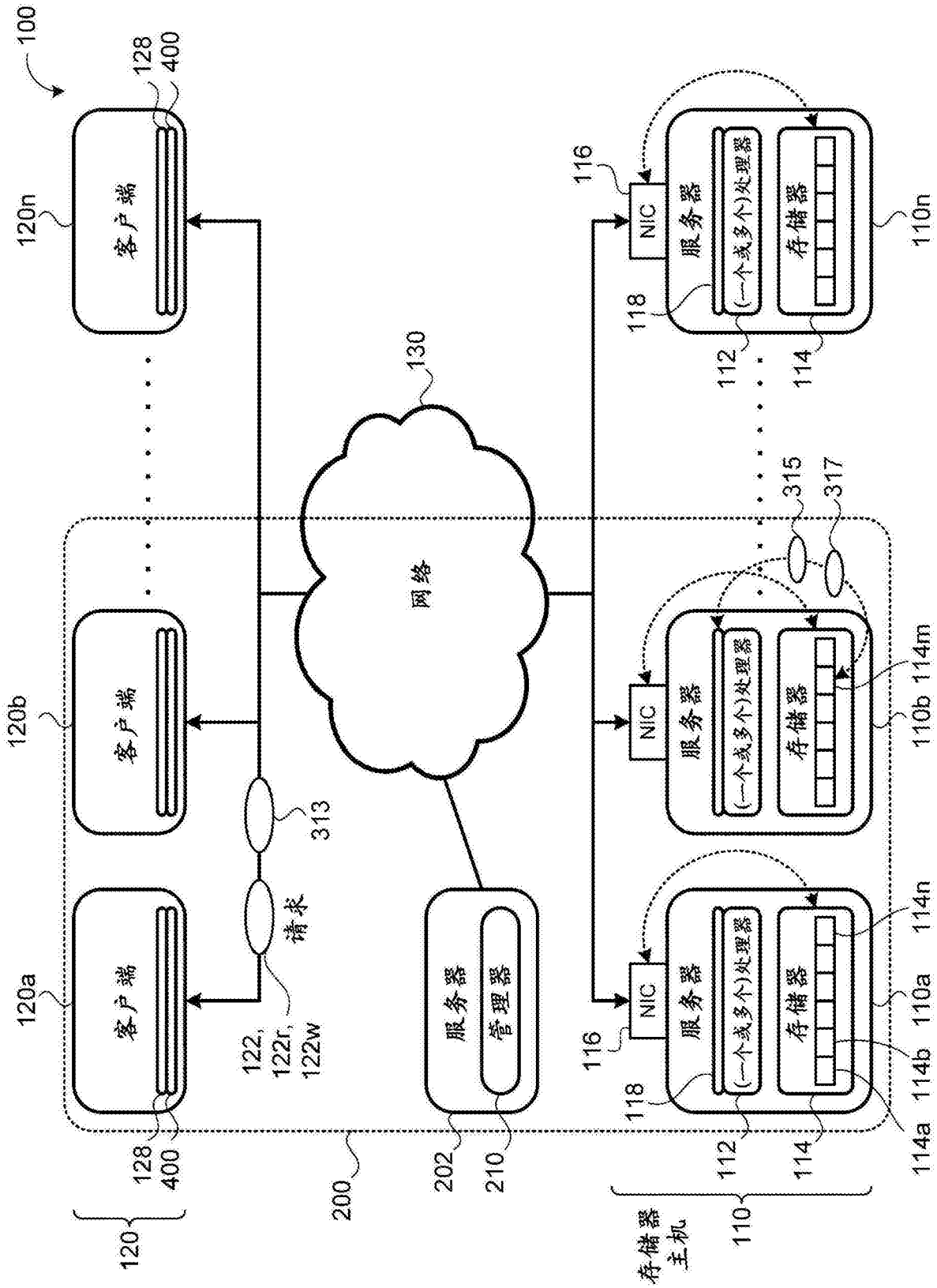

100
120n
客户端
128
400
120b
客户端
120a
客户端
128
400
120
122,
122r,
122w
请求
313
网络
130
200
服务器
管理器
202
210
116
NIC
服务器
(一个或多个)处理器
存储器
118
112
114
110n
315
317
110b
114m
110a
114n
114b
114a
存储器
主机
110

图1B

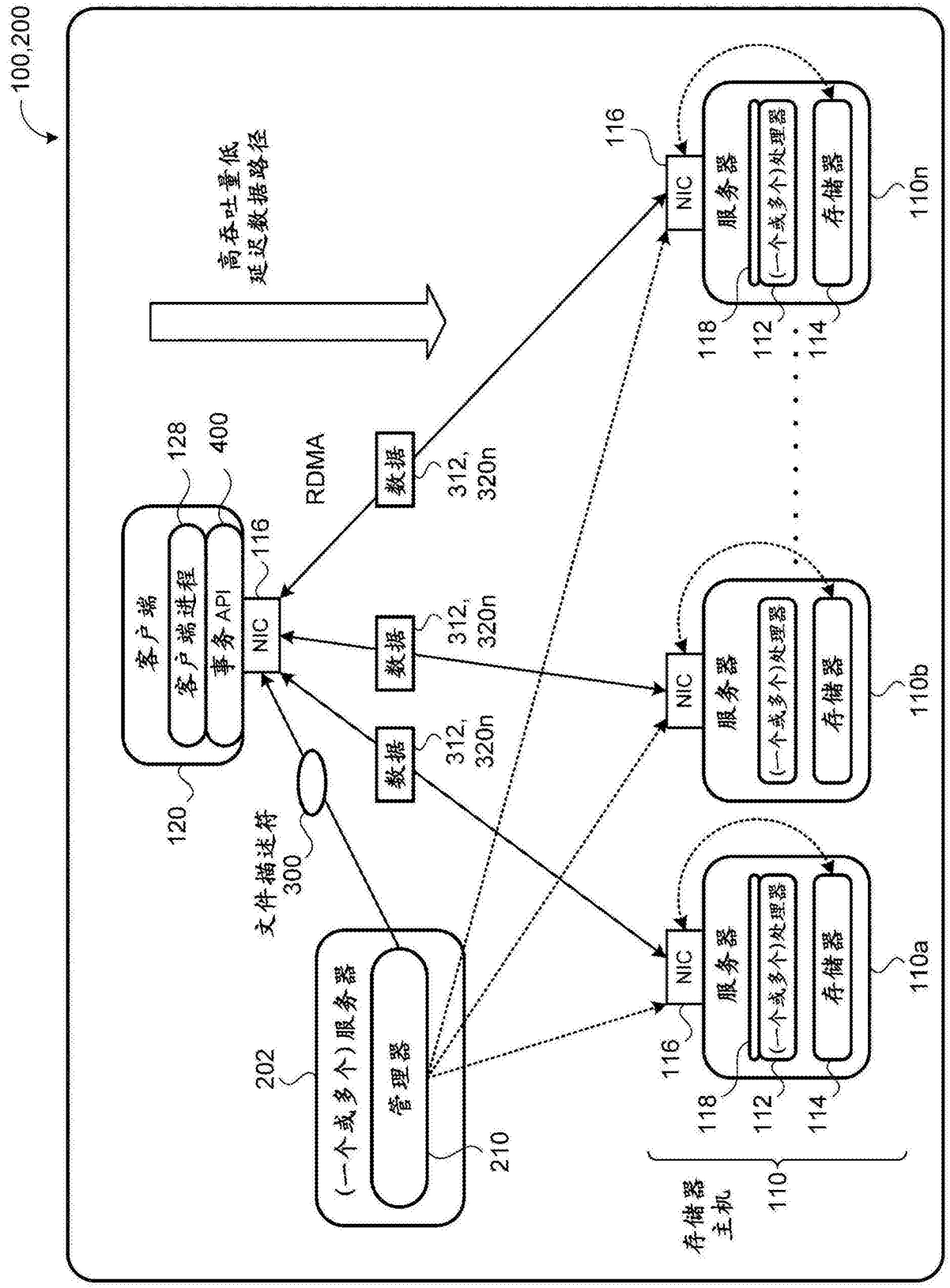

100,200
高吞吐量低延迟数据路径
客户端
客户端进程
事务 API
NIC
128
400
116
120
RDMA
数据
312, 320n
文件描述符
300
(一个或多个)服务器
管理器
202
210
服务器
(一个或多个)处理器
存储器
118
112
114
110a
110b
110n
存储器主机
110

图1C

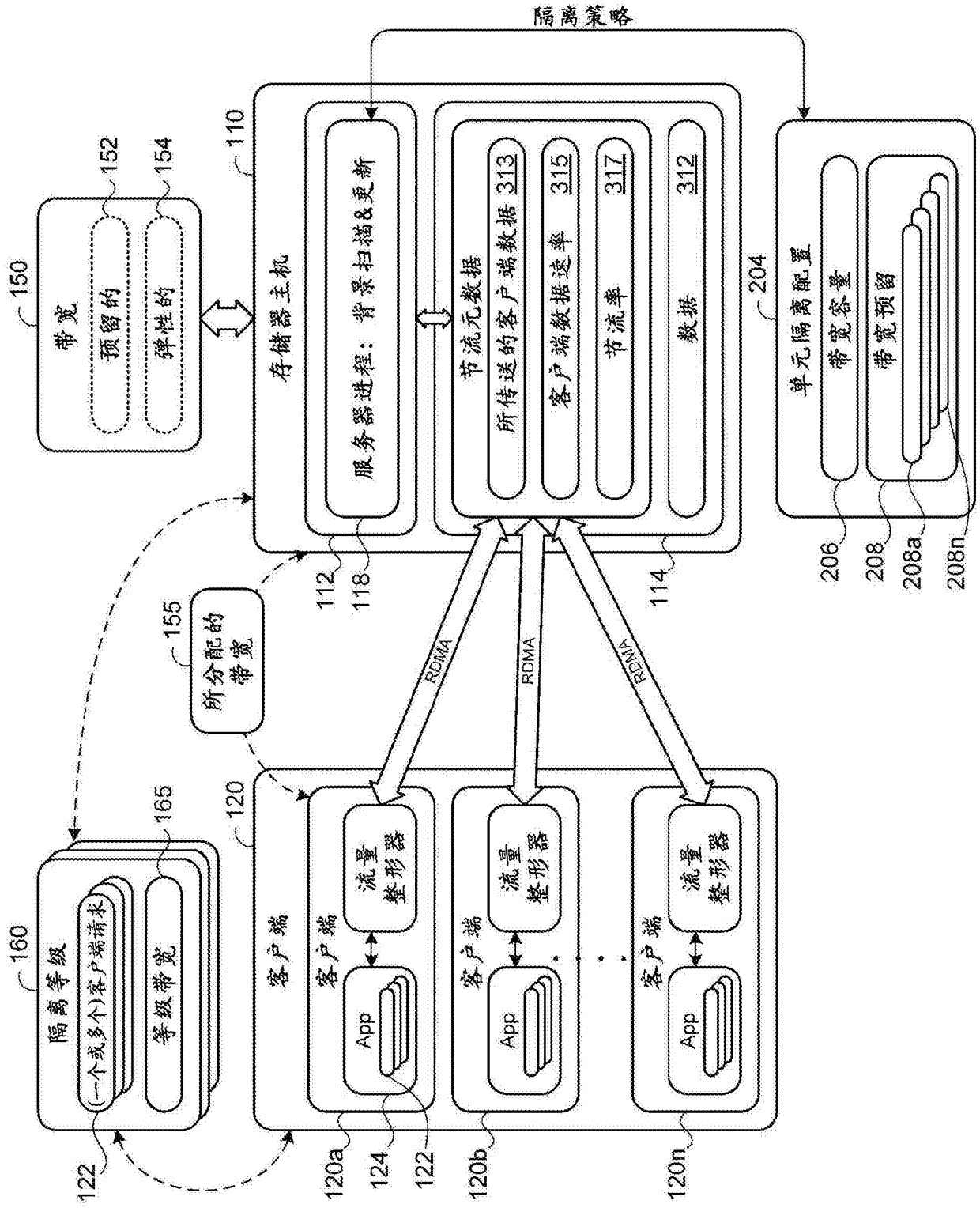

隔离策略
150
带宽
预留的
弹性的
152
154
110
存储器主机
服务器进程：背景扫描&更新
节流元数据
所传送的客户端数据 313
客户端数据速率 315
节流率 317
数据 312
单元隔离配置
带宽容量
带宽预留
204
206
208
208a
208n
112
118
114
155
所分配的带宽
RDMA
RDMA
RDMA
120
客户端
客户端
流量整形器
App
客户端
流量整形器
App
客户端
流量整形器
App
160
隔离等级
(一个或多个)客户端请求
等级带宽
165
122
120a
124
122
120b
120n

图1D

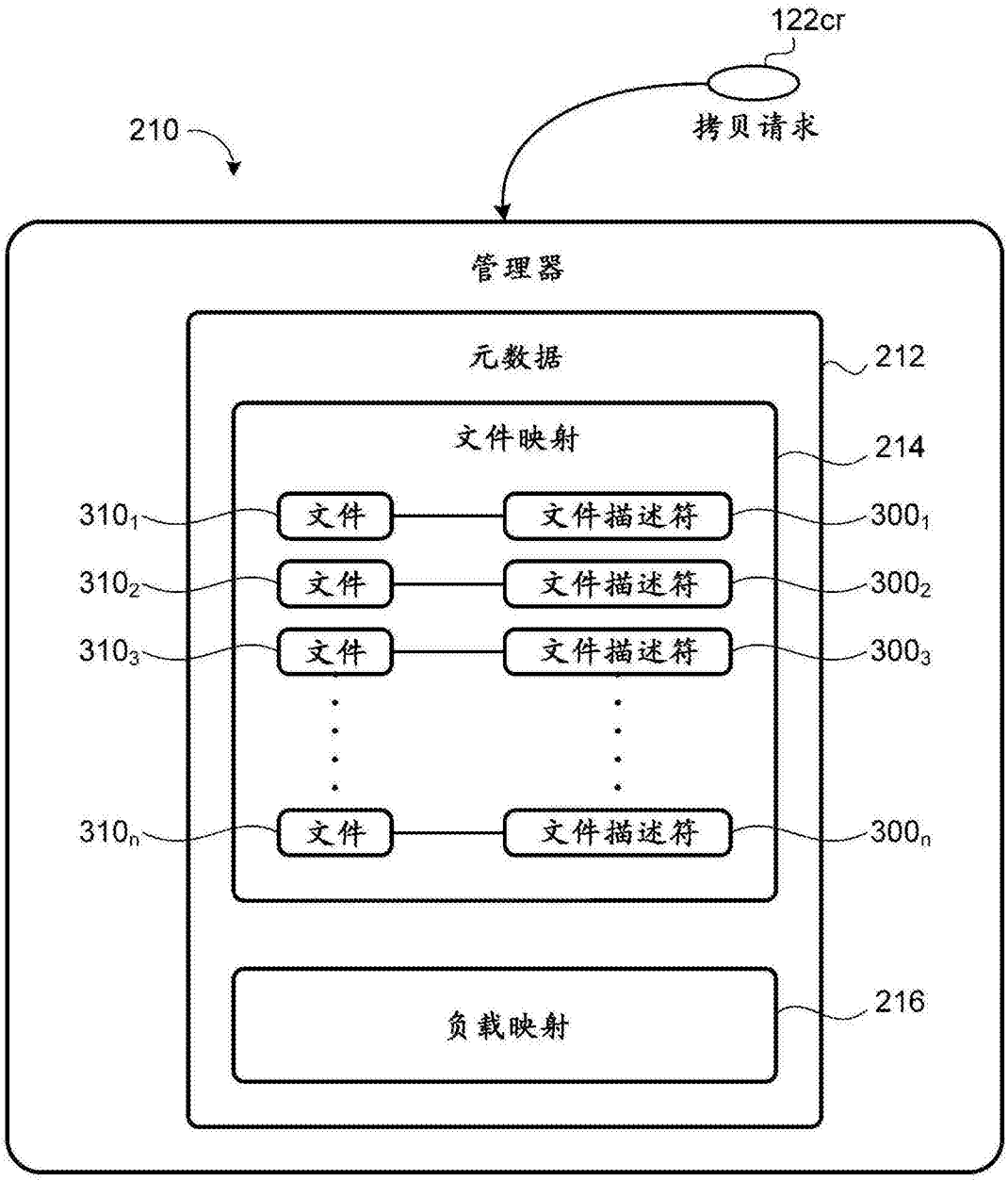

122cr
拷贝请求
210
管理器
元数据
212
文件映射
214
310_1
文件
文件描述符
300_1
310_2
文件
文件描述符
300_2
310_3
文件
文件描述符
300_3
310_n
文件
文件描述符
300_n
负载映射
216

图2A

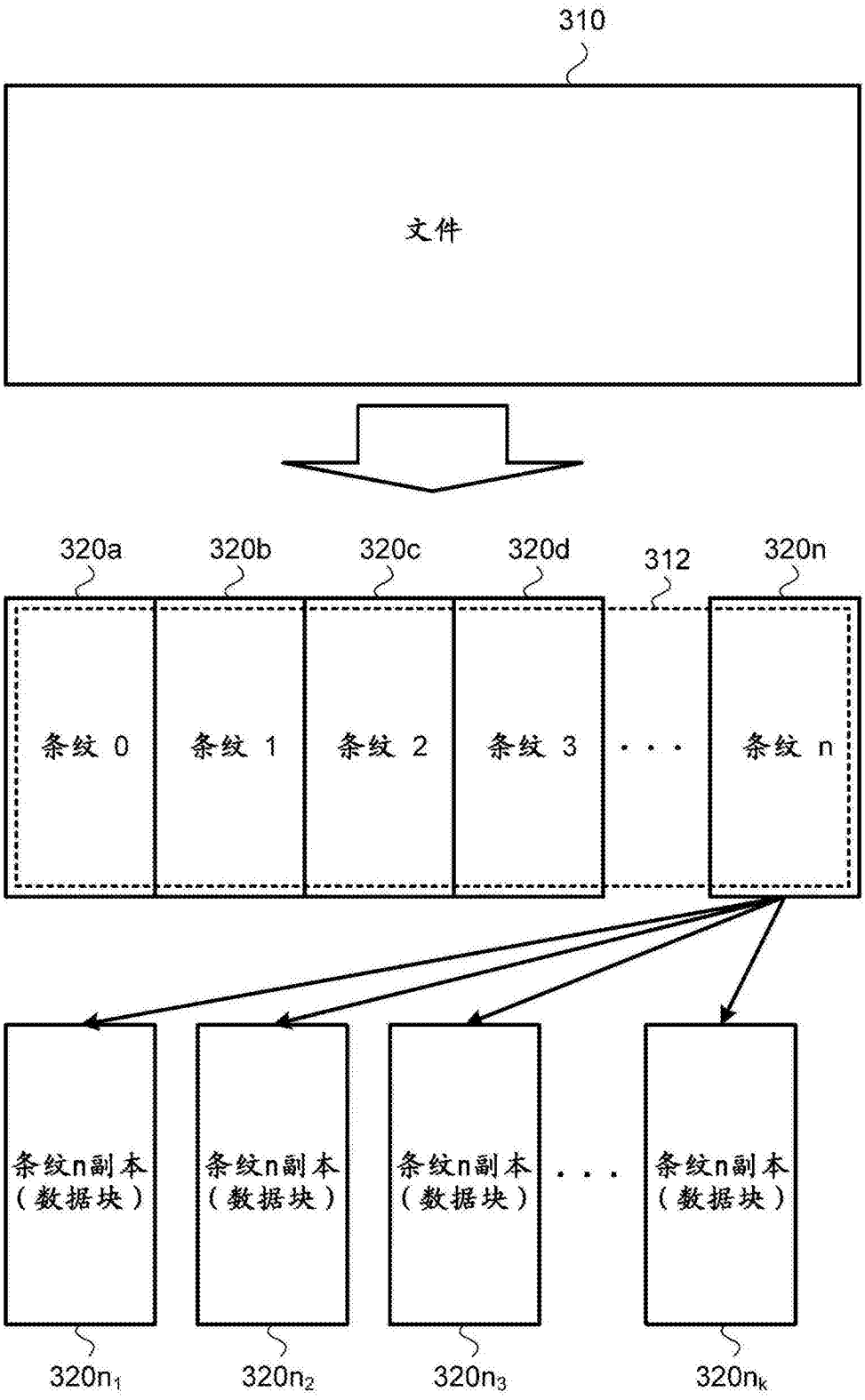

310
文件
320a
320b
320c
320d
312
320n
条纹 0
条纹 1
条纹 2
条纹 3
. . .
条纹 n
条纹n副本（数据块）
条纹n副本（数据块）
条纹n副本（数据块）
. . .
条纹n副本（数据块）
320n1
320n2
320n3
320nk

图2B

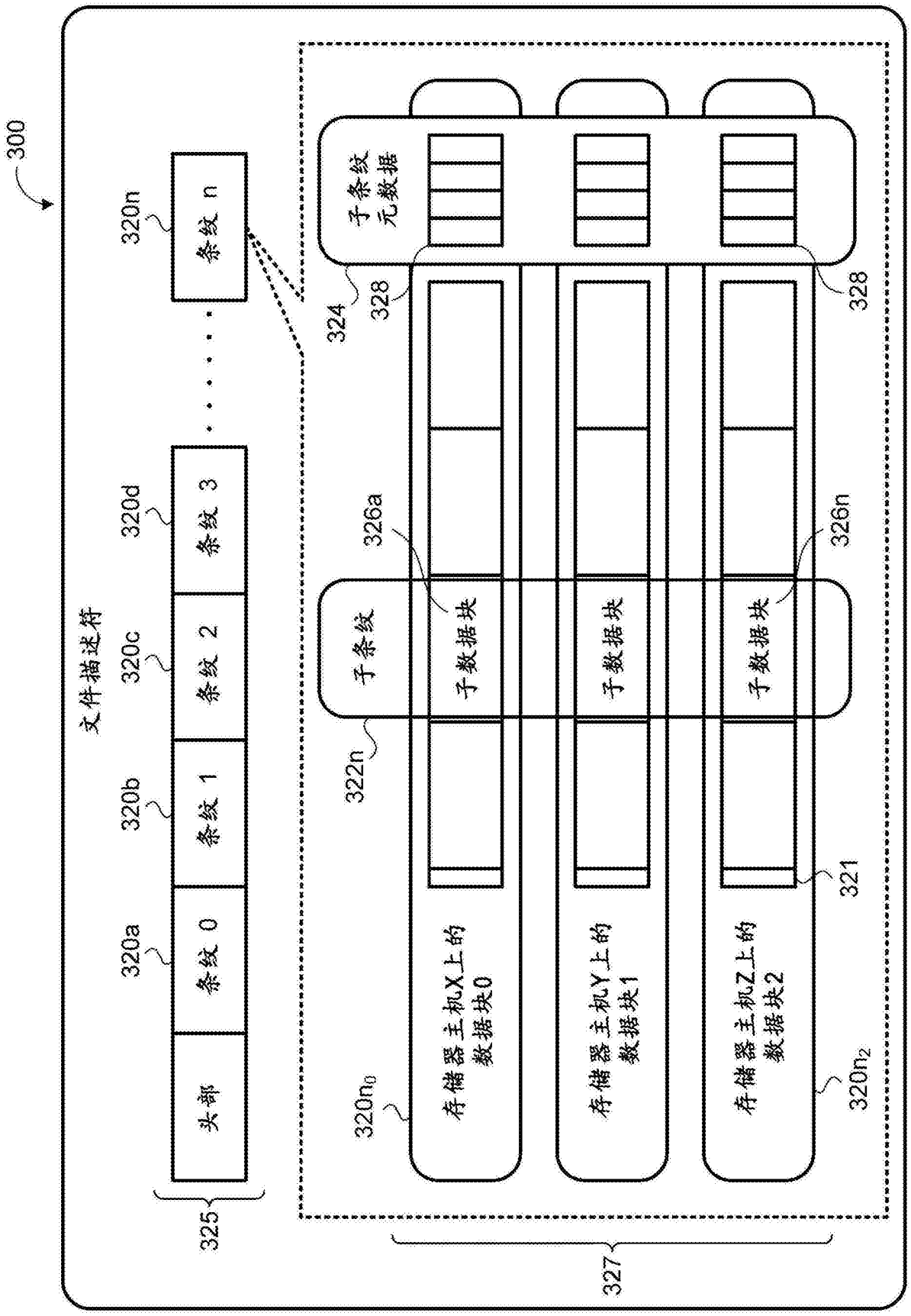

300
文件描述符
头部
条纹 0
条纹 1
条纹 2
条纹 3
条纹 n
320a
320b
320c
320d
320n
325
子条纹
元数据
子条纹
子数据块
子数据块
子数据块
324
328
328
326a
326n
322n
321
存储器主机X上的
数据块0
存储器主机Y上的
数据块1
存储器主机Z上的
数据块2
320n0
320n2
327

图2C

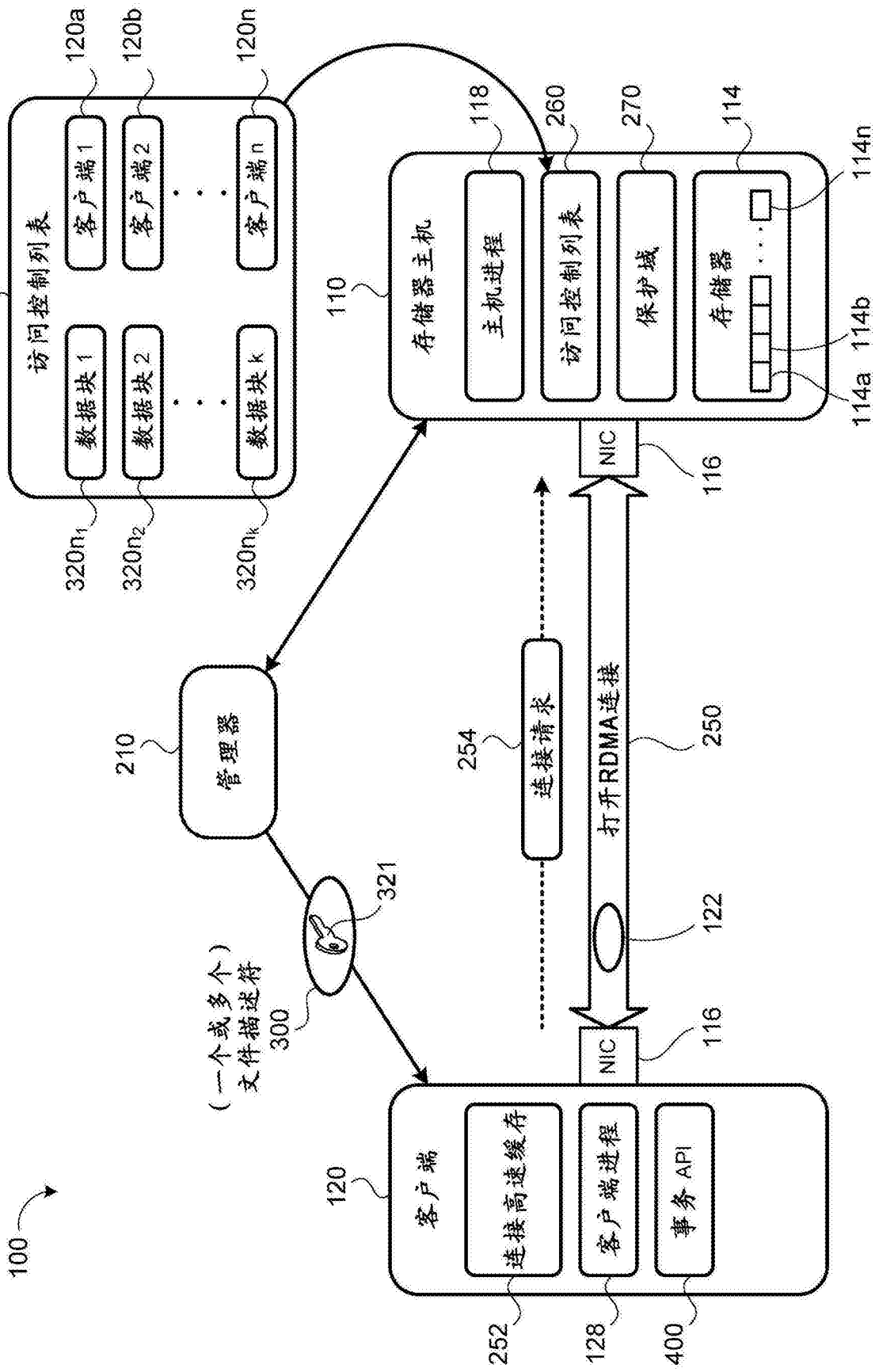

100
120
客户端
连接高速缓存
客户端进程
事务 API
252
128
400
NIC
116
打开RDMA连接
250
122
NIC
116
存储器主机
110
主机进程
118
访问控制列表
260
保护域
270
存储器
114
114n
114b
114a
连接请求
254
管理器
210
(一个或多个)
文件描述符
300
321
访问控制列表
260
客户端 1
客户端 2
客户端 n
120a
120b
120n
数据块 1
数据块 2
数据块 k
320n1
320n2
320nk

图3A

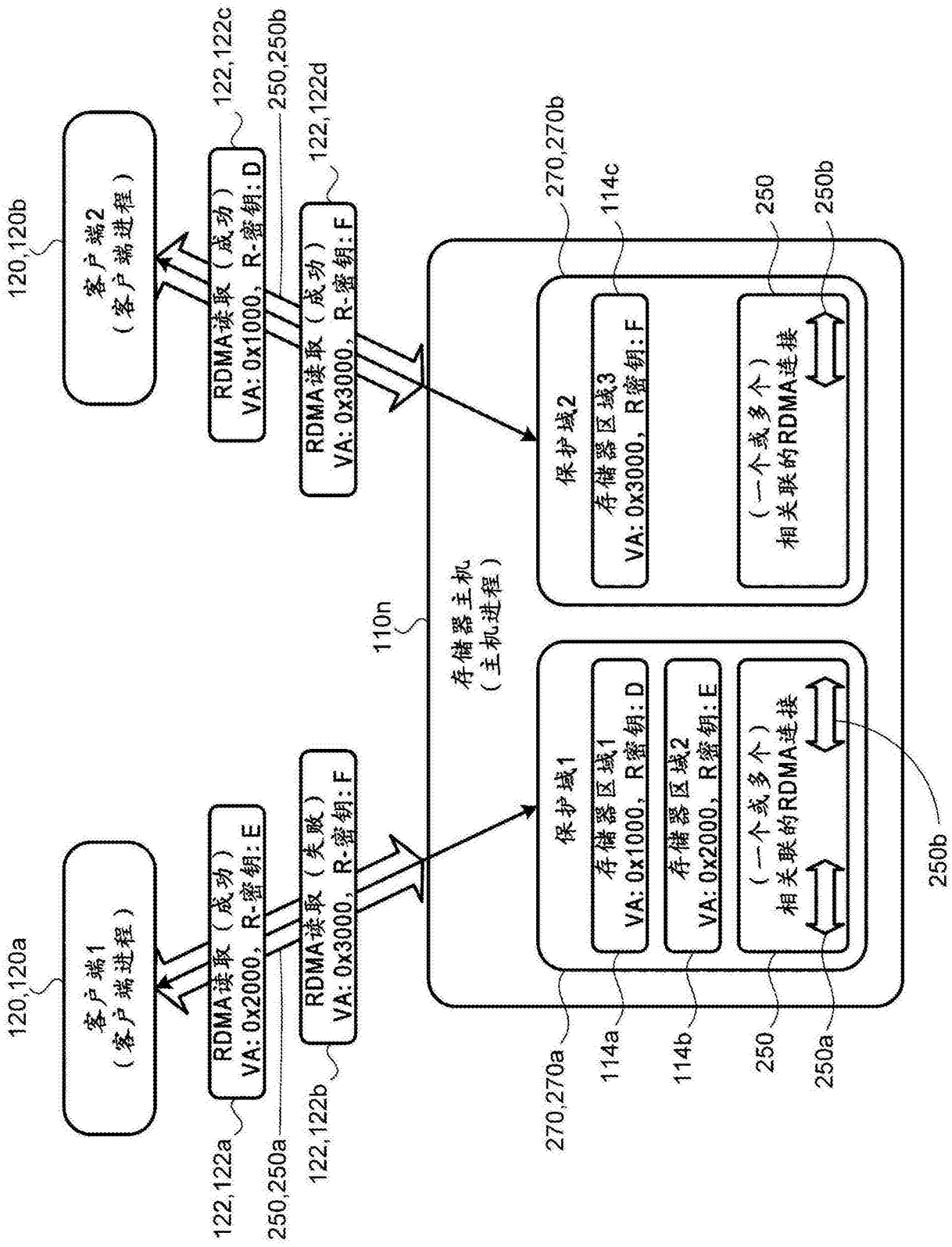

客户端1
（客户端进程）
120,120a
客户端2
（客户端进程）
120,120b
RDMA读取（成功）
VA: 0x2000, R-密钥: E
122,122a
250,250a
RDMA读取（失败）
VA: 0x3000, R-密钥: F
122,122b
RDMA读取（成功）
VA: 0x1000, R-密钥: D
122,122c
250,250b
RDMA读取（成功）
VA: 0x3000, R-密钥: F
122,122d
存储器主机
（主机进程）
110n
保护域1
270,270a
存储器区域1
VA: 0x1000, R密钥: D
114a
存储器区域2
VA: 0x2000, R密钥: E
114b
（一个或多个）
相关联的RDMA连接
250
250a
保护域2
270,270b
存储器区域3
VA: 0x3000, R密钥: F
114c
（一个或多个）
相关联的RDMA连接
250
250b
250b

图3B

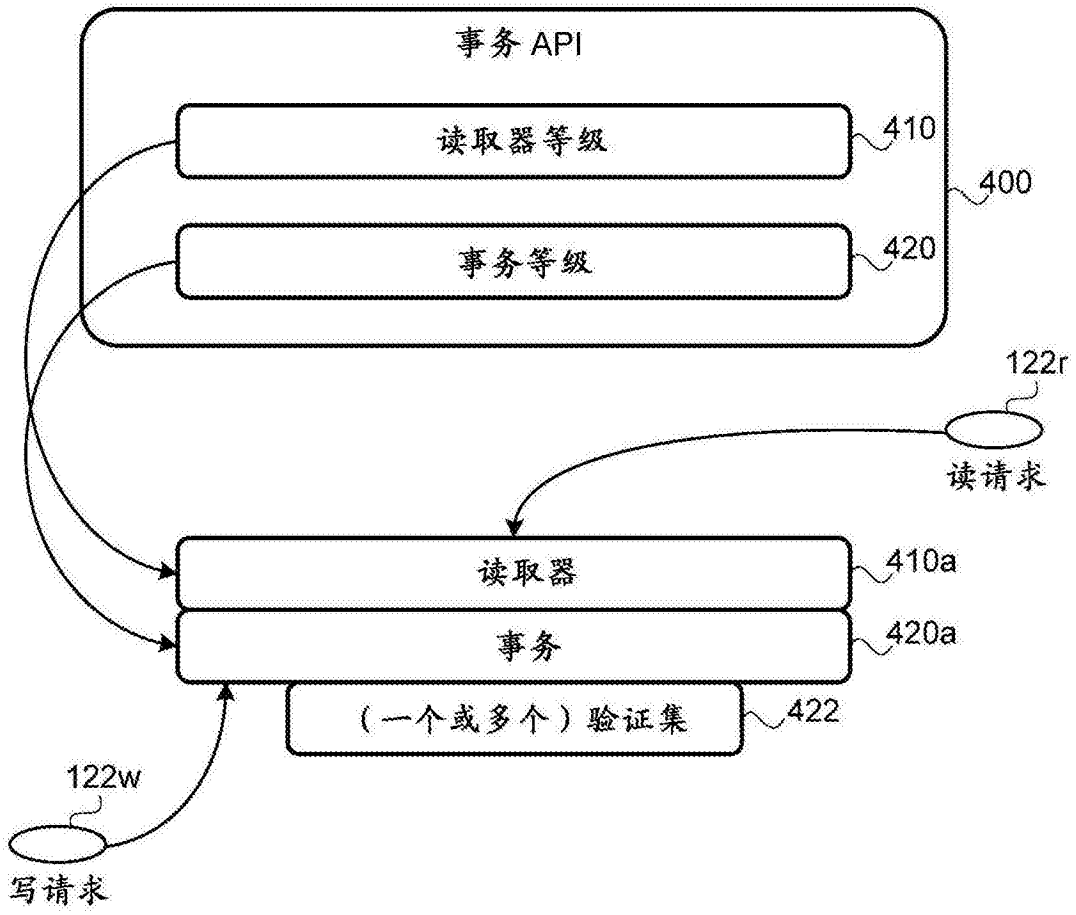

事务 API
读取器等级
410
400
事务等级
420
122r
读请求
读取器
410a
事务
420a
（一个或多个）验证集
422
122w
写请求

图4A

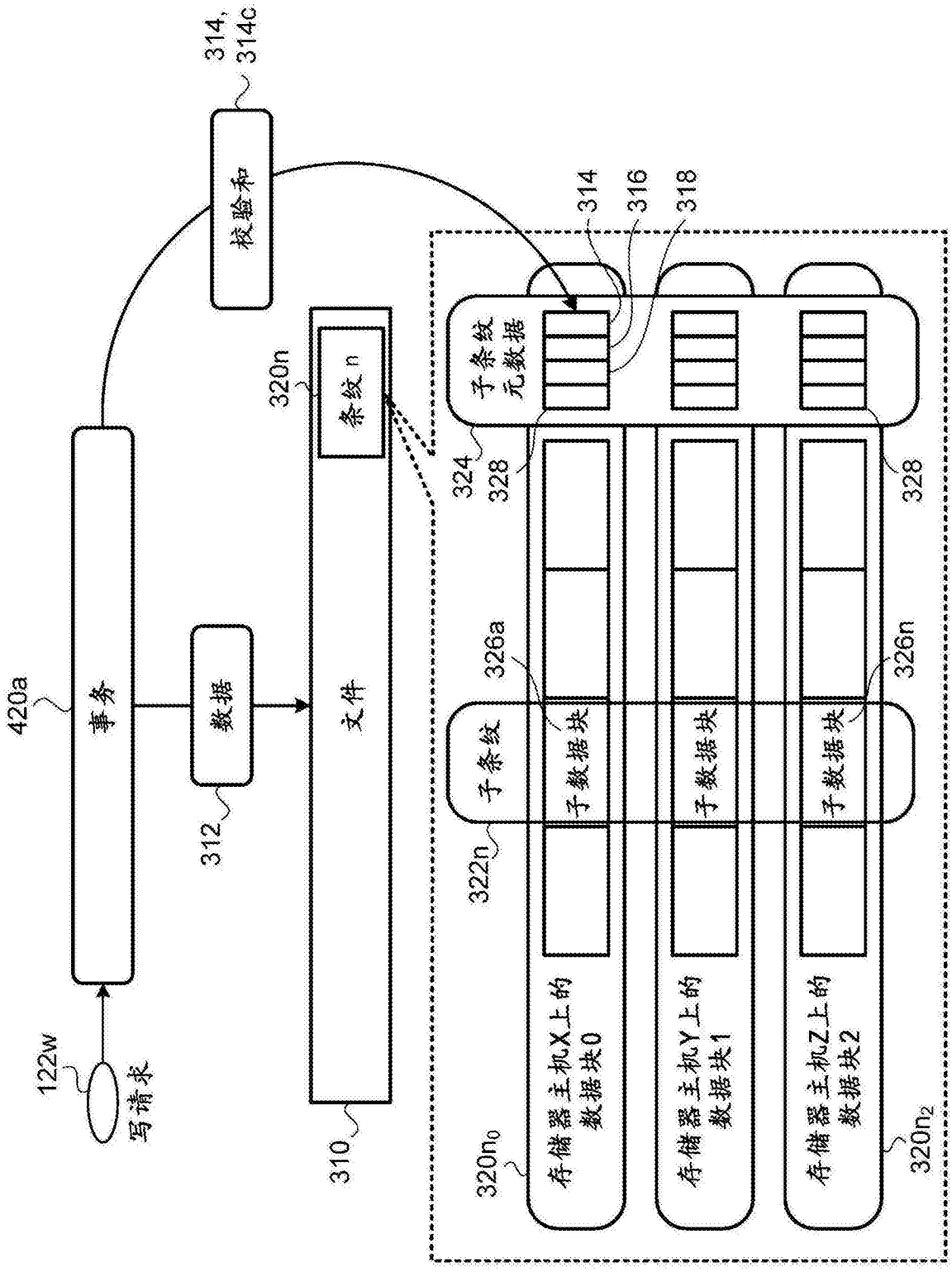

122w
写请求
420a
事务
数据
312
文件
310
320n
条纹 n
314, 314c
校验和
314
316
318
子条纹元数据
324
328
328
326a
326n
子条纹
子数据块
子数据块
子数据块
322n
320n0
存储器主机X上的数据块0
存储器主机Y上的数据块1
存储器主机Z上的数据块2
320n2

图4B

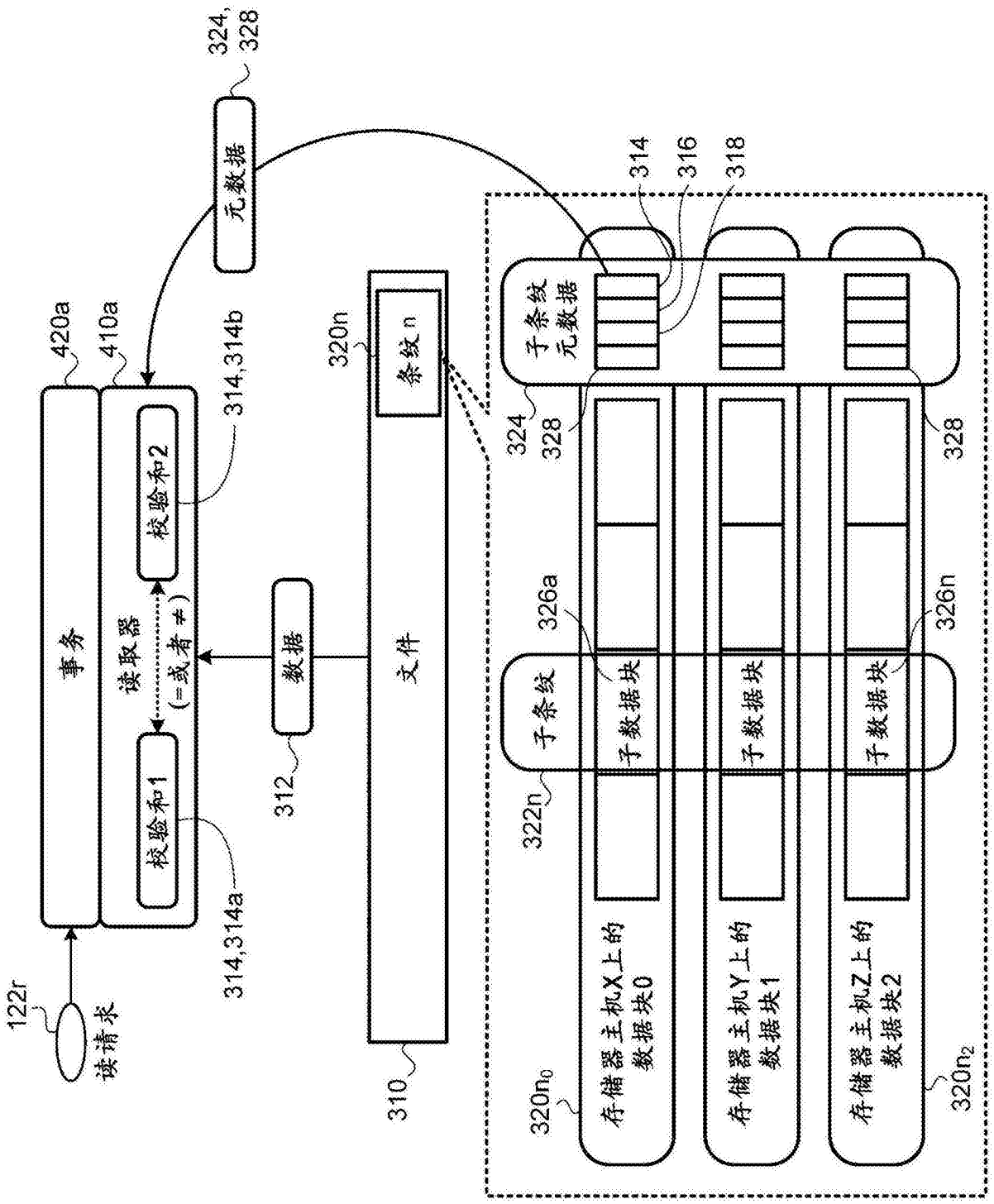

324, 328
元数据
420a
410a
314,314b
320n
条纹 n
314
316
318
子条纹 元数据
事务
读取器 (=或者≠)
校验和2
数据
文件
324
328
328
326a
326n
子条纹
子数据块
子数据块
子数据块
322n
312
校验和1
122r
读请求
314,314a
310
320n0
存储器主机X上的数据块0
存储器主机Y上的数据块1
存储器主机Z上的数据块2
320n2

图4C

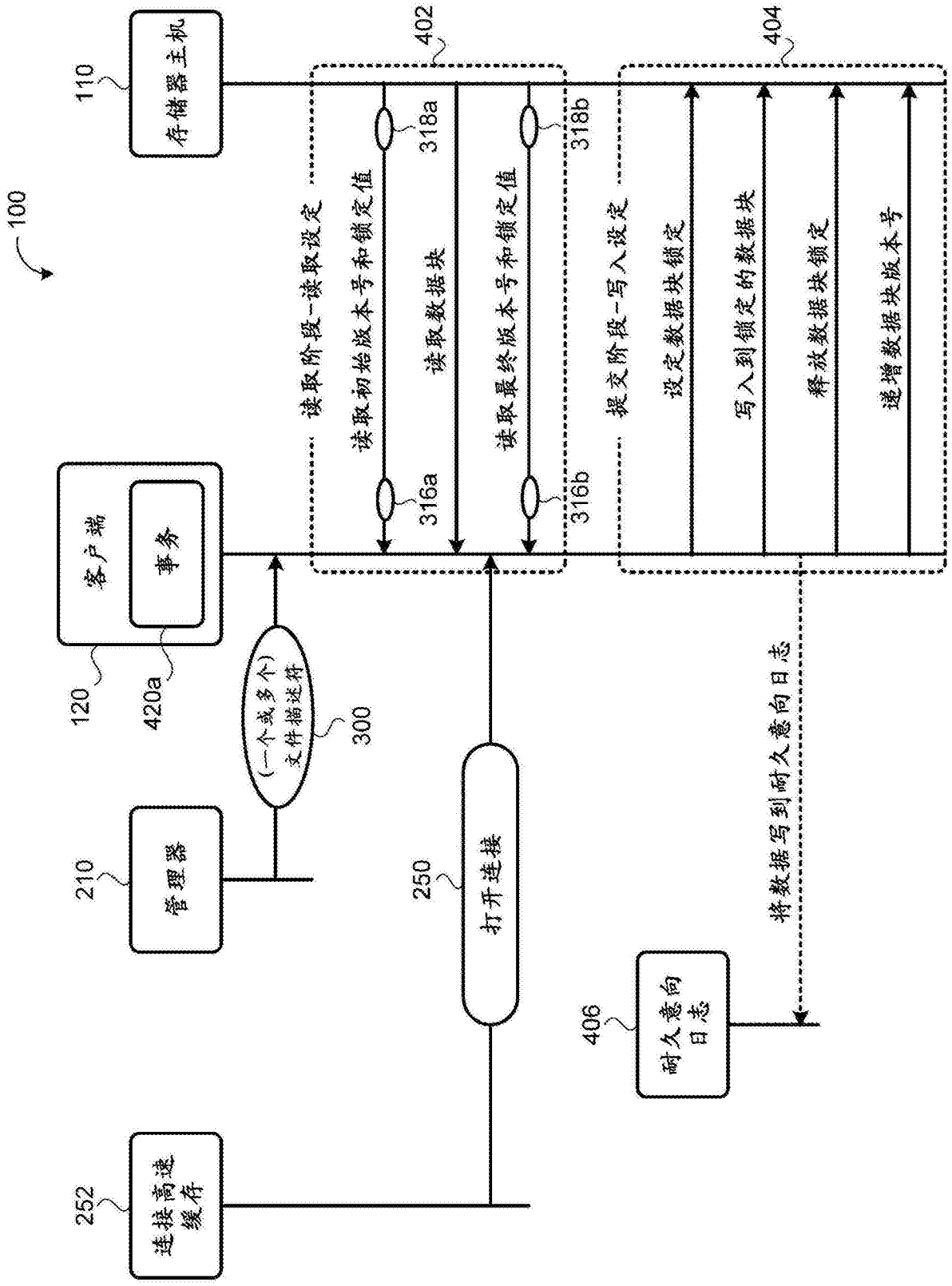

100
110
存储器主机
402
404
读取阶段-读取设定
读取初始版本号和锁定值
318a
316a
读取数据块
读取最终版本号和锁定值
318b
316b
提交阶段-写入设定
设定数据块锁定
写入到锁定的数据块
释放数据块锁定
递增数据块版本号
客户端
事务
120
420a
(一个或多个)
文件描述符
300
210
管理器
250
打开连接
将数据写到耐久意向日志
406
耐久意向
日志
252
连接高速
缓存

图4D

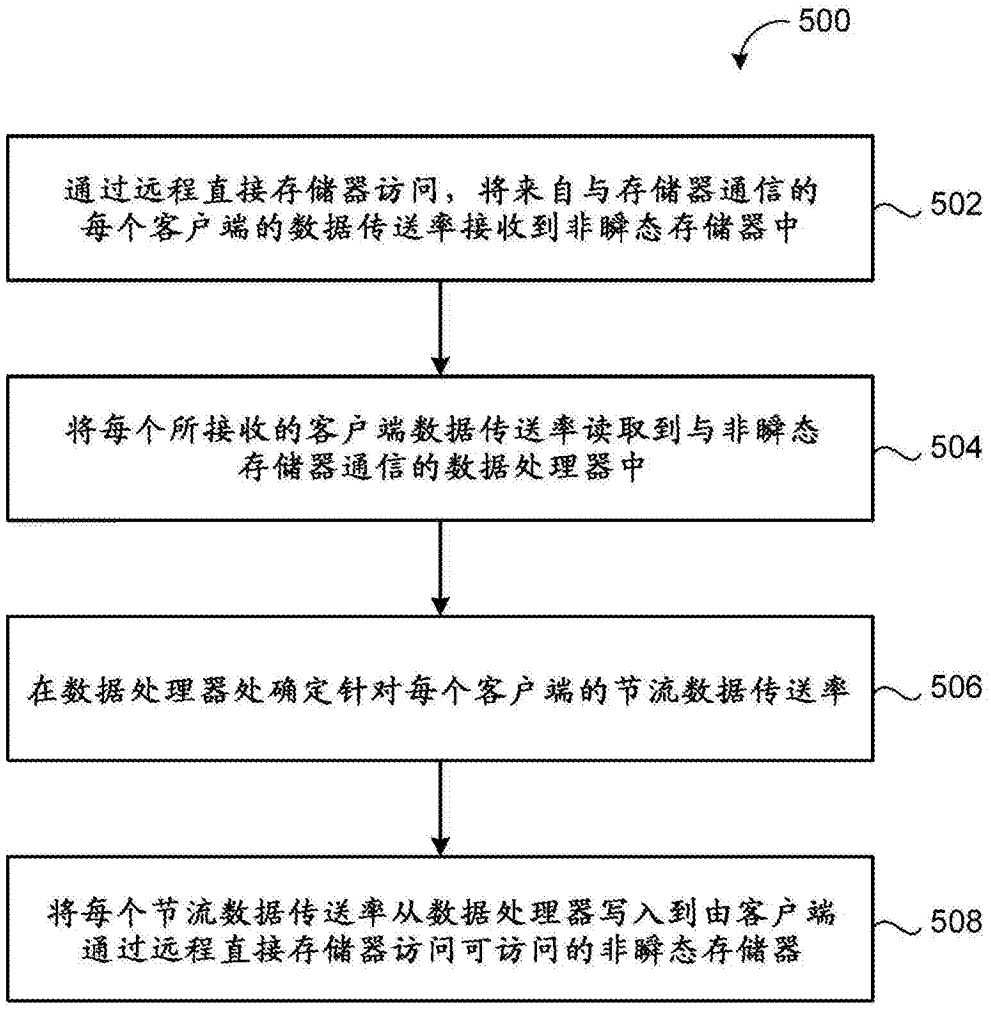

500
通过远程直接存储器访问，将来自与存储器通信的每个客户端的数据传送率接收到非瞬态存储器中
502
将每个所接收的客户端数据传送率读取到与非瞬态存储器通信的数据处理器中
504
在数据处理器处确定针对每个客户端的节流数据传送率
506
将每个节流数据传送率从数据处理器写入到由客户端通过远程直接存储器访问可访问的非瞬态存储器
508

图5