(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 105283831 B
(45)授权公告日 2018.11.16

(21)申请号 201480030836.4
(22)申请日 2014.05.27
(65)同一申请的已公布的文献号
申请公布号 CN 105283831 A
(43)申请公布日 2016.01.27
(30)优先权数据
13/904,989 2013.05.29 US
(85)PCT国际申请进入国家阶段日
2015.11.27
(86)PCT国际申请的申请数据
PCT/US2014/039480 2014.05.27
(87)PCT国际申请的公布数据
WO2014/193770 EN 2014.12.04
(73)专利权人 微软技术许可有限责任公司
地址 美国华盛顿州
(72)发明人 V·库兹耐特索夫 V·R·尚卡尔 A·达玛托 D·A·迪昂
(74)专利代理机构 上海专利商标事务所有限公司 31100
代理人 顾嘉运
(51)Int.Cl.
G06F 3/06(2006.01)
(56)对比文件
CN 102160047 A,2011.08.17,权利要求1,说明书第[0020],[0025],[0030]段.
CN 102077193 A,2011.05.25,全文.
US 2003/0065782 A1,2003.04.03,全文.
审查员 曾鹏飞

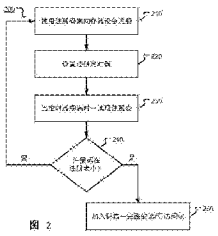
权利要求书2页 说明书12页 附图9页

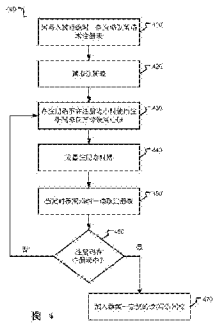
(54)发明名称

群集中的分布式存储防御

(57)摘要

各实施例提供了用于使得能够访问存储设备的方法和系统。具体而言，节点可请求对具有对存储设备的读和写访问权的群集的准入。寻求访问存储设备的节点首先必须由该群集中的其它节点批准。作为请求的一部分，寻求访问存储设备的节点向存储设备发送注册码。在注册定时器期满之际，寻求访问存储设备的节点从该存储设备接收注册表并确定其注册码是否被存储在该注册表中。如果注册码被存储在注册表中，则该节点已在该群集中被接受并因此已被授予对存储设备的读和写访问权。





CN 105283831 B

1.一种用于使得能够访问存储设备的方法,所述方法包括:

向存储设备发送注册码,其中所述存储设备连接到节点群集中的至少一个节点;

设置第一注册定时器,其中所述第一注册定时器提供了供所述节点群集中的每一节点确定寻求访问所述存储设备的节点是否应被授予所请求的访问权的时间段;

在所述第一注册定时器期满之际,从所述存储设备接收注册表;

确定所述注册码是否被存储在所述注册表中;以及

当所述注册码被存储在所述注册表中时,加入所述群集,其中加入所述群集使得能够对所述存储设备进行写访问。

2.如权利要求1所述的方法,其特征在于,还包括在加入所述群集后擦洗所述注册表。

3.如权利要求2所述的方法,其特征在于,擦洗所述注册表包括:

从所述存储设备接收所述注册表;

读取所述注册表;

将一个或多个注册码从所述注册表中清除,其中所述一个或多个注册码中的每一个与所述群集中的节点相关联。

4.如权利要求1所述的方法,其特征在于,所述注册码包括群集标识符和节点标识符。

5.一种用于使得能够访问存储设备的系统,所述系统包括:

用于向存储设备发送注册码的装置,其中所述存储设备连接到节点群集中的至少一个节点;

用于设置第一注册定时器的装置,其中所述第一注册定时器提供了供所述节点群集中的每一节点确定寻求访问所述存储设备的节点是否应被授予所请求的访问权的时间段;

用于在所述第一注册定时器期满之际,从所述存储设备接收注册表的装置;

用于确定所述注册码是否被存储在所述注册表中的装置;以及

用于当所述注册码被存储在所述注册表中时,加入所述群集的装置,其中加入所述群集使得能够对所述存储设备进行写访问。

6.如权利要求5所述的系统,其特征在于,还包括用于在加入所述群集后擦洗所述注册表的装置。

7.如权利要求6所述的系统,其特征在于,用于擦洗所述注册表的装置包括:

用于从所述存储设备接收所述注册表的装置;

用于读取所述注册表的装置;

用于将一个或多个注册码从所述注册表中清除的装置,其中所述一个或多个注册码中的每一个与所述群集中的一个节点相关联。

8.如权利要求5所述的系统,其特征在于,所述注册码包括群集标识符和节点标识符。

9.一种用于使得能够访问存储设备的方法,所述方法包括:

将写入命令从群集中的节点发送到所述存储设备,其中所述节点具有相关联的注册码;以及

在接收到所述写入命令已被拒绝的通知之际:

向所述存储设备请求注册表;

确定与所述节点相关联的注册码是否存在于所述注册表中;以及

当与所述节点相关联的注册码不存在于所述注册表中时:

向所述存储设备发送新注册码；

设置第一注册定时器，其中所述第一注册定时器提供了供所述群集中的每一节点确定寻求访问所述存储设备的节点是否应被授予所请求的访问权的时间段；

在所述第一注册定时器期满之际，从所述存储设备接收所述注册表；

确定所述新注册码是否被存储在所述注册表中；以及

当所述新注册码被存储在所述注册表中时，加入所述群集，其中加入所述群集使得所述节点能够对所述存储设备进行写入。

10.如权利要求9所述的方法，其特征在于，还包括在加入所述群集后擦洗所述注册表。

11.一种包括用于执行如权利要求9-10中的任一项所述的方法的装置的计算机系统。

12.一种具有指令的计算机可读存储介质，所述指令在被执行时使机器执行如权利要求1-4，9-10中的任一项所述的方法。

群集中的分布式存储防御

[0001] 背景

[0002] 在其中节点群集具有对存储设备的访问权的典型的共享存储情形中，该群集中的至少一个节点连接到存储设备。因此，连接到存储设备的节点负责处理该存储设备的防御。然而，在其中群集具有对多个存储设备的访问权的情形中，该群集中的单个节点可能未连接到每一存储设备。因此，一些存储设备可能是不受保护的。

[0003] 各实施例正是对于这些和其它一般考虑事项而做出的。而且，尽管讨论了相对具体的问题，但是应当理解，各实施例不应被限于解决本背景技术中所标识的具体问题。

[0004] 概述

[0005] 提供本概述来以简化形式介绍将在以下详细描述部分中进一步描述的一些概念。本概述并不旨在标识出所要求保护的主题的关键特征或必要特征，也不旨在用于帮助确定所要求保护的主题的范围。

[0006] 本公开的各实施例提供了用于使得能够访问存储设备以及用于保护可由群集中的各个节点访问的一个或多个存储设备的方法和系统。具体而言，一个或多个实施例描述了节点可如何被准入群集并由此获取对连接到节点群集中的至少一个节点的存储设备的读和写访问权。另外，一个或多个实施例提供了各节点可监视与群集相关联的注册表并移除来自未经识别的节点的条目。对于那些被移除的节点，该节点可尝试向注册表重新注册以寻求对群集的重新准入。

[0007] 如将在以下解释的，寻求访问存储设备的节点使用群集通信协议来进入群集。一旦被准入该群集，节点就有资格获取对该群集所利用的一个或多个存储设备的访问权。为了获取对存储设备的访问权，寻求访问存储设备的节点向存储设备发送注册码。在向存储设备注册后，节点设置注册定时器。在各实施例中，注册定时器等同于期间群集中的每一节点有机会确定寻求访问存储设备的节点是否应被授予所请求的访问权的时间段。在注册定时器期满之际，寻求访问存储设备的节点从存储设备接收注册表。一旦接收到注册表，该节点就确定其注册码是否被存储在该注册表中。如果注册码被存储在注册表中，则该节点被准许访问存储设备。更具体而言，该节点被授予对存储设备的写访问权。

[0008] 各实施例可被实现为计算机进程、计算系统、或者诸如计算机程序产品或计算机可读介质等制品。计算机程序产品可以是计算机系统可读的并编码了用于执行计算机进程的指令的计算机程序的计算机存储介质。计算机程序产品还可以是计算系统可读并编码了用于执行计算机过程的指令的计算机程序的载体上的传播信号。

[0009] 附图简述

[0010] 参考以下附图描述非限制性和非穷尽的实施例，其中：

[0011] 图1示出了根据本公开的一个或多个实施例的其中群集中的多个节点连接到相应存储设备的系统；

[0012] 图2示出了根据本公开的一个或多个实施例的用于请求群集中的成员资格的方法。

[0013] 图3示出了根据本公开的一个或多个实施例的用于确定对与节点群集相关联的存

储设备的访问权的方法；

[0014] 图4示出了根据本公开的一个或多个实施例的用于请求群集中的重新准入的方法；

[0015] 图5是示出根据本公开的一个或多个实施例的群集中的各个节点可如何连接到物理存储设备的框图；

[0016] 图6是解说可以与本公开的一个或多个实施例一起使用的计算设备的示例物理组件的框图；

[0017] 图7A和7B是可以与本公开的一个或多个实施例一起使用的移动计算设备的简化框图；以及

[0018] 图8是解说可以与本公开的一个或多个实施例一起使用的分布式计算系统的简化框图；

[0019] 详细描述

[0020] 以下将参考形成本发明一部分并示出各具体示例性实施例的附图更详尽地描述各个实施例。然而，各实施例可以以许多不同的形式来实现，并且不应将其解释为限制此处所阐述的各实施例；相反地，提供这些实施例以使得本公开变得透彻和完整，并且将这些实施例的范围完全传达给本领域普通技术人员。各实施例可按照方法、系统或设备来实施。因此，这些实施例可采用硬件实现形式、全软件实现形式或者结合软件和硬件方面的实现形式。因此，以下详细描述并非是局限性的。

[0021] 图1示出了根据本公开的一个或多个实施例的其中群集102中的多个节点连接到相应存储设备的系统100。如图1所示，群集102可包括多个节点102A-102D。尽管示出了四个节点，但构想群集102可具有多于四个节点或少于四个节点。在特定实施例中，节点可以是计算设备，诸如例如个人计算机、平板、膝上型计算机、智能电话、个人数字助理等。在其它实施例中，节点可以是服务器计算设备。

[0022] 图1还示出群集102中的每一节点连接到一个或多个存储设备。在特定实施例中，存储设备可以是直接附连的存储设备(即，直接连接到主机系统或设备的存储设备)。还构想存储设备可由群集中的多个节点使用一条或多条路径来访问。例如，一个或多个节点可以物理地连接到存储设备，而群集中的其它节点可以使用远程路径来连接到存储设备。另外，单个节点可具有到各个存储设备的多个物理连接以及到各个存储设备的一个或多个远程连接。还构想群集中的每一节点能够查看该群集中的每一其它节点的活动和连接。综上，系统100可以是非对称的，这表现在一些存储设备对一些节点可用，而其它存储设备对这些节点不可用。

[0023] 例如，如图1所示，节点102A和102B连接到存储设备104，节点102C连接到存储设备104和存储设备106，而节点102D连接到存储设备106和存储设备108。在特定实施例中，存储设备104－108包括存储池。由于在群集102中不存在可访问存储池中的每一存储设备的单个节点，因此群集102中的每一节点负责运行防御算法以确保只有作为该群集的一部分的节点才可具有对存储设备的读和写访问权。由此，群集102中的每一节点并发地保护存储池中的它们连接到的存储设备。

[0024] 回头参考图1，节点102A、102B和102C中的每一节点连接到存储设备104。如所讨论的，这些节点中的每一节点可具有到存储设备104的物理连接或远程连接(即，经由具有到

存储设备104的物理连接的节点的到存储设备104的连接)。因为节点102A、102B和102C连接到存储设备104,所以每一节点都具有对存储设备104的读和写访问权。此外,群集102中的节点102A、102B和102C中的每一节点都能检测到群集102中的其它节点的存在并确定该群集中的每一其它节点的活动。

[0025] 在各实施例中,对特定存储设备的权限通过持久保留来确定。即,甚至当存储设备(诸如例如存储设备104)离线或已被重启时,该存储设备也维持特定节点的保留。出于讨论目的,特定节点的保留在节点保留特定存储设备并防止另一未经授权的节点访问该存储设备时发生。

[0026] 回头参考上述示例,节点102A、102B和102C中的每一节点具有对存储设备104的读和写访问权,因为节点102A、102B和102C中的每一节点都是群集102的一部分。如将在下文中详细解释的,群集102中的每一节点在时间t运行一个防御算法以确定群集102中的任一其它节点是否已丢失到以下各项任一个的连接:(i)群集中的其它节点或者(ii)存储设备104。

[0027] 例如,如果节点102A丢失到节点102B和102C的连接或者到存储设备104的连接,则节点102B和102C独立地确定节点102A应不再具有对存储设备104的(至少)写访问权并由此应被禁止访问该存储设备。当连接已被丢失时,节点102B和102C承担节点102A的工作负载并且还采取各个步骤来确保节点102A不再能够对存储设备104进行写入,因为准许节点102A对该存储设备进行写入可能损坏存储设备104上的数据。构想尽管节点102A可能已经丢失到节点102B和102C的连接,但节点102A仍然可能具有到存储设备104的连接。同样,构想如果节点102A丢失到存储设备104的连接,则节点102A仍然可能连接到节点102B和/或节点102C。

[0028] 返回到上述示例,为了禁止节点102A对存储设备104进行写入,节点102B或102C向存储设备104发送将节点102A从节点注册表中清除的请求。更具体而言,向存储设备发送将与节点102A相关联的注册码从节点注册表中清除的请求。结果,存储设备104将不再接受来自与节点102A相关联的物理路径或远程路径的写入命令。在特定实施例中,尽管节点102A将不具有对存储设备104的写访问权,但节点102A仍将具有对存储设备104的读访问权。

[0029] 在特定实施例中,节点102B或102C中的任一节点可以彼此独立地发送清除请求。在另一实施例中,群集中的各个节点可被要求就特定节点是否应被清除达成一致。在又一实施例中,节点自己可确定它是否应从群集中移除。例如,如果节点102A确定它已经丢失到其它节点之一或到存储设备104的连接,则节点102A可移除从其自身到存储设备的一条或多条路径或者指示存储设备104将其注册码从节点注册表中移除。

[0030] 参考节点注册表,在特定实施例中,节点注册表由存储设备104维护并且列出具有对存储设备104的写访问权的节点。在各实施例中,节点注册表包括具有对该存储设备的写访问权的每一节点的注册码。在特定实施例中,注册码包括具有以下格式的64位整数:(i)群集全局唯一标识符(对于群集中的所有节点是相同的)的32位散列;(ii)8位码修订;(iii)8位节点号;以及(iv)16位签名。尽管阐述了注册码的特定大小和配置,但构想注册码可具有任何数量的位并且还具有各种配置,只要注册码对于每一节点是唯一的。

[0031] 如将在下文中更详细地解释的,一旦节点被清除,该节点就可请求重新准入群集。为了请求重新准入,已被清除的节点可以向存储设备发送经更新的注册码。一旦该节点已

经向存储设备重新注册，群集中的每一其它节点做出关于该节点是否应被重新准入群集的判定。基于群集中的各个节点的判定，寻求重新准入的节点可被授予重新准入或被拒绝重新准入。在各实施例中，群集中的各个节点可将其判定基于任何数量的因素，包括寻求重新准入的节点的连接速度、寻求重新准入的节点的可靠性、寻求重新准入的节点具有的对存储池中的其它存储设备的访问权等。

[0032] 图2示出了根据本公开的一个或多个实施例的用于请求群集中的成员资格的方法200。在特定实施例中，方法200可由节点用来请求诸如图1的群集102等群集中的成员资格。如上文所讨论的，一旦节点已被准入群集，该节点就可具有对该特定存储设备的读和写访问权。

[0033] 具体而言，加入方节点可尝试使用群集通信协议来与其它节点通信以获取对活动群集的准入。在这种情况下，一旦加入方节点进入活动状态，它就将执行方法200以获取对存储的访问权。如果加入方节点无法经由群集协议与其它加入或活动节点通信并且加入方节点相信可能不存在活动群集，则加入方节点可执行方法200以获取对存储的访问并由此变成第一活动节点。出于讨论的目的，当节点请求对群集的访问权并获取对该群集的准入时，该节点被视作活动节点或进入活动状态。例如，正在运行群集通信协议并且当前正在参与群集成员资格的所有节点都被认为是活动节点。另外，具有对与群集相关联的一个或多个存储设备的访问权的节点被认为是存储节点。在各实施例中，存储节点是活动节点集合的子集。

[0034] 如图2所示，方法200在节点已使用群集通信协议来被准入群集后开始。一旦被准入群集，一个或多个实施例提供节点通过使用注册码来向与该群集相关联的一个或多个存储设备(诸如例如存储设备104(图1))注册210来寻求访问所述存储设备。如所讨论的，注册码可包括具有以上讨论的各种分量的64位整数。

[0035] 一旦注册码已被发送到存储设备，流程继续至操作220，其中设置注册定时器。在特定实施例中，该注册定时器可由正在请求访问存储设备或存储池的节点来维护。然而，构想存储设备或群集中的另一节点也可维护该注册定时器。在各实施例中，注册定时器的长度等同于时间段t。在特定实施例中，时间段t是3秒。更具体而言，时间段t等同于群集中的任一其它节点执行应每3秒(考虑可能由于CPU负载、I/O等待时间等而出现的任何延迟)进行一次的清除所花费的时间。

[0036] 在定时器期满之际，流程继续至操作230，其中从存储设备读取节点注册表。如上讨论的，节点注册表由存储设备(或者由存储池中的至少一个存储设备)维护，并且包含与群集中的每一节点相关联的每一注册码的列表。

[0037] 一旦注册表被接收到并读取，流程继续至操作240，其中确定节点的注册码是否被包含在节点注册表中。如果该节点的注册码被包含的节点注册表中，则群集中的每一节点已经运行防御算法并且存储设备已经接受该节点访问该存储设备的请求。结果，流程继续至操作250，其中准许该节点访问存储设备，更具体而言具有对存储设备的写访问权。

[0038] 然而，如果在操作240中确定请求方节点的注册码不在节点注册表中，则流程继续回到操作210并且该节点尝试第二次向存储设备注册。该方法重复并且请求访问存储设备的节点再次请求并读取注册表以确定其注册码是否被存储在节点注册表中。

[0039] 图3示出了根据本公开的一个或多个实施例的用于确定对与节点群集相关联的存

储设备的访问权的方法300。在特定实施例中，方法300由群集中的被视作活动节点(即，正在运行群集通信协议并且当前正在参与群集成员资格的所有节点)的每一节点执行。方法300还可由被认为是存储节点(即，可访问与群集相关联的一个或多个存储设备的任何节点)的节点来执行。

[0040] 方法300开始于一节点从节点注册表中“擦除”310注册码。具体而言，一节点查找不是活动群集的一部分的其它节点。当注册码从盘注册表中擦除时，与被擦除的注册码相关联的节点不再具有对与群集相关联的特定存储设备或存储池的(至少)写访问权。在特定实施例中，当群集中的一个节点相信该群集中的另一节点应不再具有对特定存储设备或存储池的写访问权时，擦除注册码。这可能是由于一个节点丢失到群集中的另一节点的连接、节点将其自身从群集中移除、或者节点丢失到存储设备的连接中的一个情形所导致的结果。在其中节点已将自身从群集中移除的情形中，该节点可以向存储设备发送指示其注册码应从节点注册表中移除的请求。在另一实施例中，群集中的一个其它节点可请求在擦除过程期间将该节点从群集中清除(即，移除)。在特定实施例中，当一节点被从群集中清除时，其它节点可被配置成防止来自被移除的节点的命令到达存储设备。

[0041] 如图3所示，擦除过程具有各种子选项。擦除过程开始于子操作311，其中一节点读取由存储设备维护的节点注册表。如上文所讨论的，节点注册表包含与群集中的每一节点相关联的所有注册码的列表。

[0042] 流程然后继续至子操作312，其中不具有群集中的活动成员资格的一个或多个节点被从群集中清除。在各实施例中，群集中的每一节点可被该群集中的每一其它节点查看到并且可以通过远程连接或物理连接来连接到一个或多个存储设备。由于群集中的每一节点都可具有该群集中的每一其它节点的视图，因此正在读取节点注册表的节点能确定该群集中的哪些节点在该节点注册表中具有相关联的注册码。由此，一节点擦洗从存储设备接收到的注册表。如果注册码在表中但节点不是活动的，则该节点被清除。

[0043] 在特定实施例中，一节点可以不被清除，直到群集中的多个节点达成相似判定(即，确定将被清除的节点在节点注册表中不具有注册码)。在其它实施例中，当单个节点达成将被清除的节点在节点注册表中不具有相关联的注册码的判定时，可清除该节点。

[0044] 一旦已经发送清除节点的请求，运行算法的节点就确定313该节点是否已被成功清除。如果该节点未被成功清除，则运行算法的节点执行自检以确定其自己的注册码是否存在于节点注册表中。如果其自己的注册码不存在于节点注册表中，则流程继续至子操作314并且该节点进入“访客状态”并使用一个或多个群集协议来寻求重新准入该群集。如果节点已被成功清除，则流程继续至操作315，并且该节点在存储设备尚未被保留的情况下保留该存储设备。即，该节点取得存储设备的所有权并且然后与在同一群集中的其它节点共享对该存储设备的访问权。

[0045] 当节点注册表中的注册码已被擦除时，流程继续至操作320，其中设置第二定时器。在各实施例中，第二定时器与图2中讨论的注册定时其成比例。例如，擦除算法在所有节点上并行地运行(例如，每一节点独立于所有其它节点地每3秒擦除码一次)。在特定实施例中，擦除算法并行地运行，因为一个节点可能无法可靠地告知另一节点是否也正在擦洗注册表或者一个节点是否已经丢失其到存储的连接并因此无法擦洗注册表。

[0046] 当一节点加入群集时，该节点向该群集注册并等待获得准入。如果该节点被准入，

一个实施例提供该节点尽其所能快地擦洗并显现盘。设置三个第二擦洗定时器并且该节点能确认其自身在存储上的注册。

[0047] 在定时器期满之际，流程继续至操作330，其中该节点显现到存储设备的一条或多条路径。即，该节点可确定哪些路径(物理或远程)需要被连接到群集中的其它节点和/或存储设备。例如，参考图1，如果节点102B已加入群集102并且节点102A已建立或具有到存储设备104的物理连接，则该物理路径被通告给群集中的其它节点，并且该群集中的其它节点(例如102D)可建立到节点102A的远程路径并利用节点102A与存储设备104之间的物理路径。

[0048] 图4示出了根据本公开的一个或多个实施例的用于请求对群集的重新准入的方法400。在特定实施例中，寻求重新准入群集的节点可能已经将其自身从群集中移除或者可能已经被群集中的另一节点移除，如以上参考图3讨论的。在各实施例中，方法400开始于群集中的一节点向存储设备发送写入命令并被通知该写入不成功。如果写入不成功，则该节点可请求沿不同路径发送该命令。另外地或另选地，该节点可请求取消进行中的所有其它待决命令。在接收到通知之际，该节点向存储设备请求410节点注册表。如上文讨论的，节点注册表由存储设备维护并且包括与群集中的各个节点相关联的各个注册码。

[0049] 当节点注册表已被请求方节点接收到时，该节点读取420注册表以确定其自己的注册码是否被包含在该注册表中。当确定该节点的注册码不在节点注册表中时，该节点使用另一注册码来向存储设备注册430。如上文所讨论的，注册码可以是具有群集全局唯一标识符的32位散列、8位码修订、8位节点号以及16位签名的64位整数。在特定实施例中，由于节点必须重新注册，因此该节点可生成新注册码，其中该注册码的至少一部分被递增或改变。由此，该节点、存储设备或群集中的其它节点可以跟踪必须为该节点建立新路径的次数。在一些实施例中，该节点已经请求重新准入的次数可影响该节点获取对群集的重新准入。另外，对注册码的改变帮助确保来自该节点和/或来自与该节点相关联的路径的写入命令(在该节点被重新准入群集时)能与来自该节点的可能仍然正在等待被写入到存储设备的旧写入命令区分开来。

[0050] 一旦该节点已向存储设备注册其新生成的注册码，就设置440注册定时器。如先前所讨论的，注册定时器的长度等同于时间段t的约2.5倍。即，时间段t等同于群集中的每一节点运行本文公开的防御算法所花费的时间。

[0051] 在定时器期满之际，流程继续至操作450并且寻求对存储设备的写访问权的节点读取由存储设备维护的注册表。如果确定460新生成的注册码在节点注册表中，则该节点被授予对存储设备的写访问权。

[0052] 然而，如果在操作460中确定该节点的新生成的注册码不在节点注册表中，则流程继续回到操作430，该节点重新注册并重设注册定时器。在特定实施例中，如果该节点已被拒绝所确定的次数，则该节点将不再寻求对存储设备的写访问权。在特定实施例中，该节点可以在所确定的时间段已经过去后再次寻求对群集的准入或对存储设备的访问权。

[0053] 图5是示出根据本公开的一个或多个实施例的群集中的两个节点可用来使用远程路径和/或物理路径来访问物理盘的系统500的框图。如以上参考图1到4所讨论的，群集中的各个节点可以连接到存储池中的一个或多个存储设备。这些连接(或轴)可以是物理连接或远程连接。如将在以下解释的，节点可利用各种路径来连接到一个或多个存储设备。

[0054] 在图5所示的示例性实施例中，群集可具有两个节点，即节点A 510和节点B 520。每一节点都可具有到物理盘530的物理连接。尽管只示出了两个节点，但构想群集可包括不止两个节点。另外，构想每一节点都可具有到一个或多个存储设备的物理连接。如图5所示，节点可经由多条不同路径来访问物理盘。例如，节点A 510具有到物理盘530的物理路径并且还具有经由节点B 520的目标524的到物理盘530的远程路径。在特定实施例中，单个节点可具有到同一盘的多条物理路径。在这些实施例中，节点将向群集中的所有其它节点展示所有这些物理路径。

[0055] 也如图5所示，节点A可访问虚拟盘511，聚集到物理盘530的多条物理路径和远程路径的多路径对象512、实例化经由另一节点(诸如例如节点B 520)的到物理盘530的一条或多条远程路径的远程路径对象513、用于通告节点A510与物理盘530之间的一条或多条物理连接并且允许群集中的其它节点经由节点A 510获得到物理盘530的连接(例如，经由远程路径)的目标对象514、以及实例化到物理盘530的一条或多条物理连接或路径的物理路径对象515。

[0056] 同样，节点B 520可访问虚拟盘521、聚集从节点B 520到物理盘530的多条物理路径和远程路径的多路径对象522、实例化经由另一节点(诸如例如节点A 510)的到物理盘530的一条或多条远程路径的远程路径对象523、向群集中的其它节点通告到物理盘530的物理路径的目标对象524、以及实例化到物理盘530的一条或多条物理连接或路径的物理路径对象525。尽管对于节点A 510和节点B 520两者都示出了一条远程路径，但构想单个节点可具有多条远程路径。还构想每一节点可具有多条物理路径。

[0057] 在各实施例中，经由其向物理盘530发送各种命令的优选路径是物理路径。例如，当检测到新盘时，群集中的一个或多个节点将对盘进行注册或保留。如上文所讨论的，该过程包括运行以上参考图2讨论的防御算法并且随后创建从该节点到物理盘的物理路径。在特定实施例中，每一物理路径或物理路径每一新实例具有注册码，该注册码包括群集标识符、节点标识符和再生标识符(在每一次物理路径被实例化时递增的该物理路径的唯一号码)。在各实施例中，路径的注册码可以等同于相关联的节点的注册码。一旦物理连接被建立并且节点使用注册码来向盘注册，该节点的多路径对象和目标对象就被通知该新建立的物理路径。该信息然后被传送到群集中的其它节点，以使得其它节点可经由具有到物理盘的物理连接的节点的目标来建立远程路径。

[0058] 如上文所讨论的，构想一个或多个节点可能丢失到群集中的一个或多个其它节点或到物理盘的连接。在这种情况下，群集中的一个连接的节点将请求移除来自断开连接的节点的一条或多条路径，并且还请求存储设备停止获取来自与断开连接的节点相关联的一条或多条路径(例如，物理路径或远程路径)的写入请求。同样，与具有到断开路径的节点的远程连接的每一节点相关联的目标也可停止接收来自断开节点的命令。这些动作防止断开连接的节点发送可能在线上但尚未完成的对存储设备的附加和/或重复写入。即，将注册码从存储设备中移除并通过目标阻塞写入命令有助于确保断开连接的节点无法使用物理路径或远程路径来对盘进行写入。

[0059] 例如，参考图5，节点A 510可能丢失其经由其物理路径对象515的到物理盘530的物理连接。然而，如图所示，节点A 510还具有经由节点B 520的目标对象524的到物理盘530的远程路径513。另外，在丢失到物理盘530的连接之前，节点A 510可能已经向物理盘530发

送尚未完成的写入命令。由于节点A 510已经丢失到物理盘530的连接，因此节点A 510可能不知道其写入命令已被执行还是拒绝。

[0060] 然而，如果节点A 510被允许立即与物理盘530重新连接并且要么重新提交可能已被或未被执行的命令，要么如果节点A 510被准许向物理盘530发送附加命令(这些命令可能由于节点A 510丢失其连接而失序)，则这些动作可导致物理盘530中的数据被损坏。为了防止此类损坏，节点B 520抢占与节点A 510相关联的物理路径和/或所有远程路径。

[0061] 一旦与节点A 510相关联的物理和/或远程路径被抢占，物理盘530将不会接受来自与节点A 510相关联的路径的命令。由于针对每一节点的每一路径都具有相关联的标识符，因此物理盘530可基于每一路径的标识符来确定哪些命令与该节点相关联。在特定实施例中，物理盘530在物理路径之间进行区分。由此，如果从物理盘的角度来看I/O经由远程路径到来，则该I/O将看上去就像它来自主存该远程路径连接到的目标的节点。简而言之，在目标处执行远程路径I/O栏选，同时在物理盘530层面执行物理路径I/O栏选。

[0062] 继续该示例，群集中的每一节点可看到每一其它节点的每一轴(spindle)或路径。由此，节点B 520可看到节点A 510已丢失到物理盘530的连接。结果，节点B 520将丢失远程路径523。然而，如果节点A 510无法与群集中的其它节点通信，则节点B 520可指示物理盘530拒绝来自节点A 510的写入命令。

[0063] 在特定实施例中，一旦物理盘530开始拒绝来自节点A 510的物理路径的命令，节点A 510的多路径对象512就检测到对命令的拒绝。结果多路径对象512可查询所有其它现存物理路径以确定任何路径是否有效。如果一条物理路径仍然有效，则将该有效物理路径添加到多路径对象512。然而，如果不存在有效物理路径对象，则创建新的多路径对象，并且物理路径对象515用新注册码来实例化新物理路径。当被生成时，新物理路径及其相关联的注册码将具有使其远离与旧物理路径相关联的现在不存在的标识符的新再生标识符。

[0064] 另外，当节点(诸如例如节点A 510)使用新标识符来请求重新准入群集时，该新标识符被通告给群集中的其它节点。由此，其它节点的远程路径对象可使用节点A 510的物理路径的新标识符来连接到物理盘530。如上文所讨论的，由于物理盘530知道不接受来自旧物理路径的命令，因此当节点A 510通过以上参考图2－4描述的方法来寻求重新准入群集时物理盘接受来自新物理路径及其相关联的标识符的命令。

[0065] 回头参考图5，如果节点A 510和节点B 520丢失到彼此的连接，则在应用的高速缓存中可能存在尚未被写入到虚拟盘的数据或者在虚拟盘中可能存在尚未被写入到物理盘的数据。因此，各实施例提供了处理完来自断开连接的节点的路径上的所有剩余命令，并且不再接受来自与断开路径的节点相关联的路径的命令。

[0066] 本文描述的实施例和功能可通过多种计算系统来操作，包括但不限于台式计算机系统、有线和无线计算系统、移动计算系统(如移动电话、上网本、图形输入板或平板型计算机、笔记本计算机、和膝上型计算机)、手持设备、多处理器系统、基于微处理器或可编程消费电子产品、小型计算机、以及大型计算机。

[0067] 此外，本文所述的实施例和功能性可在分布式系统(如基于云的计算系统)上操作，其中应用功能性、存储器、数据存储和检索、以及各种处理功能可在诸如因特网或内联网之类的分布式计算网络上彼此远程地操作。各种类型的用户界面和信息可经板载计算设备显示器或经与一个或多个计算设备相关联的远程显示单元被显示。例如，各种类型的用

户界面和信息可在墙壁表面上被显示和交互，各种类型的用户界面和信息被投射在墙壁表面上。与可用于实施本发明的各实施例的许多计算系统的交互包括：键击输入、触摸屏输入、语音或其他音频输入、手势输入(其中相关联的计算设备配备有用于捕捉和解释用于控制计算设备的功能性的用户手势的检测(如相机)功能性)等。

[0068] 图6-8及相关联的描述提供了其中可实施本发明的各实施例的各种操作环境的讨论。然而，关于图6-8所示出和讨论的设备和系统是用于示例和说明的目的，而非对可被用于实施本文所述的本发明的各实施例的大量计算设备配置的限制。

[0069] 图6是示出可用来实施本发明的各实施例的计算设备105的物理组件(即硬件)的框图。下面描述的计算设备组件可适用于上述节点或计算设备。在基本配置中，计算设备105可包括至少一个处理单元602和系统存储器604。取决于计算设备的配置和类型，系统存储器604可包括，但不限于，易失性存储(例如，随机存取存储器)、非易失性存储(例如，只读存储器)、闪存、或这些存储器的任何组合。系统存储器604可包括操作系统605和适合于运行各种软件应用620的一个或多个程序模块606。操作系统605例如可适合于控制计算设备105的操作。此外，本发明的实施例可结合图形库、其他操作系统、或任何其他应用程序来实践，并且不限于任何特定应用或系统。该基本配置在图6中由虚线608内的那些组件示出。计算设备105可具有附加特征或功能。例如，计算设备105还可以包括诸如例如磁盘、光盘或磁带之类的附加数据存储设备(可移动和/或不可移动)。这些附加存储在图6中由可移动存储设备609和不可移动存储设备610示出。

[0070] 如上所述，可在系统存储器604中存储多个程序模块和数据文件。尽管在处理单元602上执行，但是程序模块606可以执行包括但不限于下列过程：图1-4中所示的方法的各阶段中的一个或多个。根据本发明的实施例可使用的其他程序模块可包括电子邮件和联系人应用、字处理应用、电子表格应用、数据库应用、幻灯片演示应用、绘图或计算机辅助应用等。

[0071] 此外，本发明的实施例可在包括分立电子元件的电子电路、包含逻辑门的封装或集成电子芯片、利用微处理器的电路、或在包含电子元件或微处理器的单个芯片上实践。例如，可以通过片上系统(SOC)来实施本发明的各实施例，其中，可以将图6中示出的每个或许多组件集成到单个集成电路上。这样的SOC设备可包括一个或多个处理单元、图形单元、通信单元、系统虚拟化单元以及各种应用功能性，所有这些都作为单个集成电路被集成到(或“烧录到”)芯片基板上。当通过SOC操作时，在此所述的功能可以通过与计算设备105的其他组件一起集成在单个集成电路(芯片)上的应用专用逻辑来操作。本发明的实施例还可使用能够执行诸如例如，AND(与)、OR(或)和NOT(非)的逻辑运算的其他技术来实践，包括但不限于，机械、光学、流体和量子技术。另外，本发明的实施例可在通用计算机或任何其他电路或系统中实践。

[0072] 计算设备105也可具有一个或多个输入设备612，如键盘、鼠标、笔、语音输入设备、触摸输入设备等等。也可包括输出设备614，如显示器、扬声器、打印机等等。上述设备是示例，并且可使用其他设备。计算设备104可包括允许与其他计算设备618进行通信的一个或多个通信连接616。合适的通信连接616的示例包括但不限于RF发射机、接收机和/或收发机电路；通用串行总线(USB)、并行和/或串行端口。

[0073] 本文所使用的术语计算机可读介质可包括计算机存储介质。计算机存储介质可包

括以任何方法或技术实现的用于存储诸如计算机可读指令、数据结构、或程序模块等信息的易失性和非易失性、可移动和不可移动介质。系统存储器604、可移动存储设备609和不可移动存储设备610都是计算机存储介质(即,存储器存储)的示例。计算机存储介质可以包括RAM、ROM、电可擦除只读存储器(EEPROM)、闪存或其他存储器技术、CD-ROM、数字多功能盘(DVD)或其他光存储、磁带盒、磁带、磁盘存储或其他磁性存储设备、或可用于存储信息且可以由计算机设备105访问的任何其他制造品。任何这样的计算机存储介质都可以是计算设备105的一部分。计算机存储介质不包括载波或其他传播或已调数据信号。

[0074] 通信介质由诸如载波或其他传输机制等已调制数据信号中的计算机可读指令、数据结构、程序模块或其他数据来体现,并包括任何信息传递介质。术语“已调制数据信号”可以描述以对该信号中的信息进行编码的方式设定或者改变其一个或多个特征的信号。作为示例而非限制,通信介质包括诸如有线网络或直接线连接等有线介质,以及诸如声学、射频(RF)、红外线和其他无线介质等无线介质。

[0075] 图7A和7B示出可用来实施本发明的各实施例的移动计算环境700,例如移动电话、智能电话、平板个人计算机、膝上型计算机等。参考图7A,示出了用于实现各实施例的移动计算设备700的一个实施例。在一基本配置中,移动计算设备700是具有输入元件和输出元件两者的手持式计算机。移动计算设备700通常包括显示器705以及允许用户将信息输入到移动计算设备700中的一个或多个输入按钮710。移动计算设备700的显示器705也可充当输入设备(如触摸屏显示器)。如果包括,则可任选的侧输入元件715允许进一步的用户输入。侧输入元件715可以是旋转开关、按钮、或任何其他类型的手动输入元件。在替代实施例中,移动计算设备700可结合更多或更少的输入元件。例如,在某些实施例中,显示器705可以不是触摸屏。在又一替代实施例中,移动计算设备700是诸如蜂窝电话之类的便携式电话系统。移动计算设备700还可包括可选的小键盘735。可选的小键盘735可以是物理小键盘或者在触摸屏显示器上生成的“软”小键盘。在各个实施例中,输出元件包括用于显示图形用户界面(GUI)的显示器705、可视指示器720(如发光二极管)、和/或音频换能器725(如扬声器)。在某些实施例中,移动计算设备700结合振动换能器来向用户提供触觉反馈。在又一实施例中,移动计算设备700结合诸如音频输入(如传声器插孔)、音频输出(如耳机插孔)、以及视频输出(如HDMI端口)之类的输入和/或输出端口,用于将信号发送到外部设备或从外部设备接收信号。

[0076] 图7B是示出移动计算设备的一个实施例的架构的框图。即,移动计算设备700可结合系统(即架构)702以实现某些实施例。在一个实施例中,系统702被实现为能够运行一个或多个应用(如浏览器、电子邮件、日历、联系人管理器、消息收发客户端、游戏、以及媒体客户端/播放器)的“智能手机”。在某些实施例中,系统702被集成为计算设备,诸如集成的个人数字助理(PDA)和无线电话。

[0077] 一个或多个应用程序766可被加载到存储器762中并在操作系统764上或与操作系统764相关联地运行。应用程序的示例包括电话拨号程序、电子邮件程序、个人信息管理(PIM)程序、文字处理程序、电子表格程序、因特网浏览器程序、消息通信程序等等。系统702还包括存储器762内的非易失性存储区768。非易失性存储区768可被用于存储在系统702断电的情况下不会丢失的持久信息。应用程序766可使用信息并将信息存储在非易失性存储区768中,如电子邮件应用使用的电子邮件或其他消息等。同步应用(未示出)也驻留于系统

702上且被编程为与驻留在主机计算机上的对应的同步应用交互，以保持非易失性存储区768中存储的信息与主机计算机处存储的相应信息同步。应当理解的，其他应用也可被加载到存储器762并在移动计算设备700上运行。

[0078] 系统702具有可被实现为一个或多个电池的电源770。电源770还可包括外部功率源，如补充电池或对电池充电的AC适配器或加电对接托架。

[0079] 系统702还可包括执行发射和接收无线电频率通信的功能的无线电772。无线电772通过通信运营商或服务供应商方便了系统702与“外部世界”之间的无线连接。来往无线电772的传输是在操作系统764的控制下进行的。换言之，无线电772接收的通信可通过操作系统764传播到应用程序766，反之亦然。

[0080] 可以使用可视指示器720来提供可视通知，和/或可以使用音频接口774来经由音频换能器725产生可听通知。在所示实施例中，可视指示符720是发光二极管(LED)，而音频换能器725是扬声器。这些设备可直接耦合到电源770，使得当被激活时，即使为了节省电池功率而可能关闭处理器760和其它组件，它们也保留一段由通知机制指示的保持通电时间。LED可被编程为无限地保持通电，直到用户采取动作指示该设备的通电状态。音频接口774用于向用户提供听觉信号并从用户接收听觉信号。例如，除了被耦合到音频换能器725之外，音频接口774还可被耦合到话筒来接收可听输入，例如便于电话通话。根据各本发明的各实施例，话筒也可充当音频传感器来便于对通知的控制，如下文将描述的。系统702可进一步包括允许板载相机730的操作来记录静止图像、视频流等的视频接口776。

[0081] 实现系统702的移动计算设备700可具有附加特征或功能。例如，移动计算设备700还可包括附加数据存储设备(可移动和/或不可移动)，例如磁盘、光盘或磁带。这种附加存储设备在图7B中用非易失性存储区768示出。

[0082] 由移动计算设备700生成或捕捉的且经系统702存储的数据/信息可如上所述被本地存储在移动计算设备700上，或数据可被存储在可由设备通过无线电772或通过移动计算设备700和与移动计算设备700相关联的一分开的计算设备之间的有线连接访问的任何数量的存储介质上，该计算设备如例如因特网之类的分布式计算网络中的服务器计算机。如应理解的，此类数据/信息可经移动计算设备700、经无线电772或经分布式计算网络来被访问。类似地，这些数据/信息可根据已知的数据/信息传送和存储手段来容易地在计算设备之间传送以存储和使用，这些手段包括电子邮件和协作数据/信息共享系统。

[0083] 图8示出了如上所述的用于提供和维护群集中的成员资格的系统的体系结构的一个实施例。例如，节点注册表、标识符以及节点之间和节点与物理盘之间的各种路径可被存储在不同的通信信道信道或其它存储类型中。例如，各种标识符可使用目录服务822、web门户824、邮箱服务826、即时消息收发存储828或社交联网网站830来存储。服务器820可以向群集中的一个或多个其它服务器或节点提供数据和/或连接类型。作为一个示例，服务器820可以是在web上通过网络815向客户机提供数据的web服务器。作为示例，客户端计算设备可被实现为计算设备105并体现为个人计算机、平板计算设备610和/或移动计算设备700(如智能电话)中。客户端计算设备105、610、700的这些实施例中的任一个可从存储816获得内容。

[0084] 以上参考例如根据本发明的各实施方式的方法、系统和计算机程序产品的框图和/或操作图示描述了本发明的各实施方式。框中所注明的各功能/动作可以按不同于任何

流程图所示的次序出现。例如，取决于所涉及的功能/动作，连续示出的两个框实际上可以基本上同时执行，或者这些框有时可以按相反的次序来执行。

[0085] 本申请中提供的一个或多个实施例的描述和说明不旨在以任何方式限制或约束如权利要求所要求保护的发明范围。本申请中提供的实施例、示例和细节被认为是足以传达所有权，且使得他人能够制作并使用所要求保护的发明的最佳模式。所要求保护的发明不应被理解为限制于本申请中所提供的任何实施例、示例或细节。不管是以组合的方式还是分开的方式示出和描述，各种特征(结构上的和方法逻辑上的)旨在被选择性地包括或忽略，以产生具有特定的特征集的实施例。在被提供本申请的描述和说明的情况下，本领域的技术人员能够想象到落在所要求保护的发明的更宽泛方面以及本申请中所具体化的一般发明概念的精神内的替代实施例并不背离该更宽泛的范围。

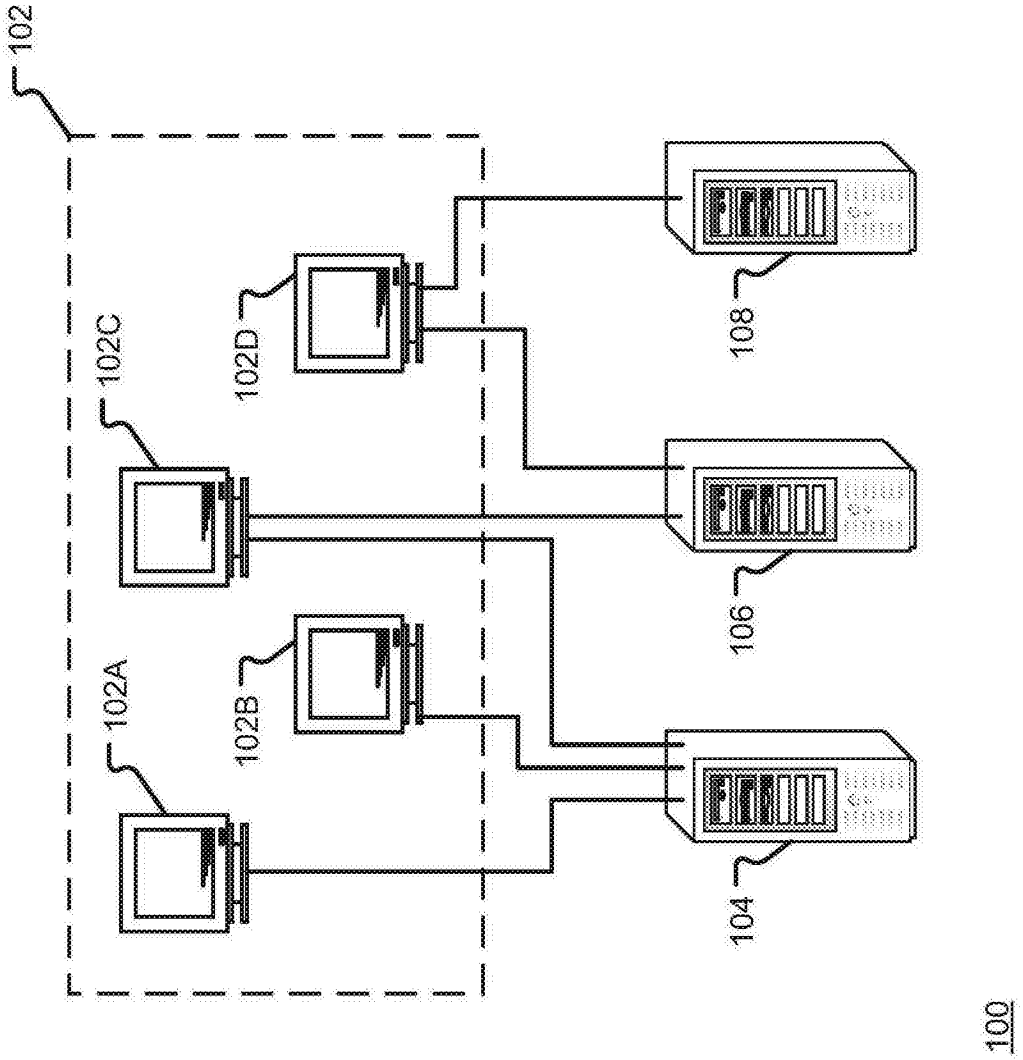


图1

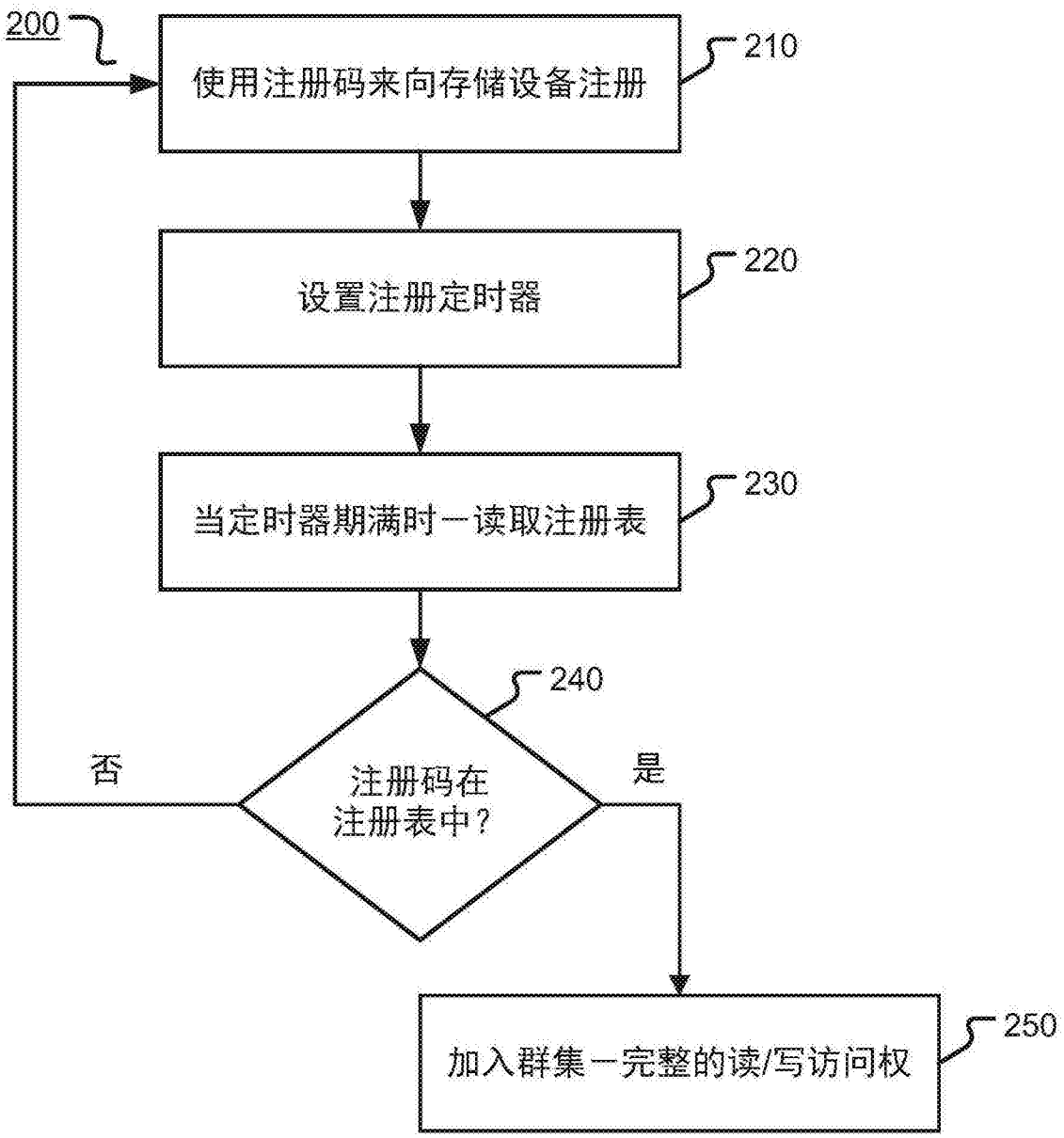

200
使用注册码来向存储设备注册
210
设置注册定时器
220
当定时器期满时一读取注册表
230
240
注册码在
注册表中？
否
是
加入群集一完整的读/写访问权
250

图2

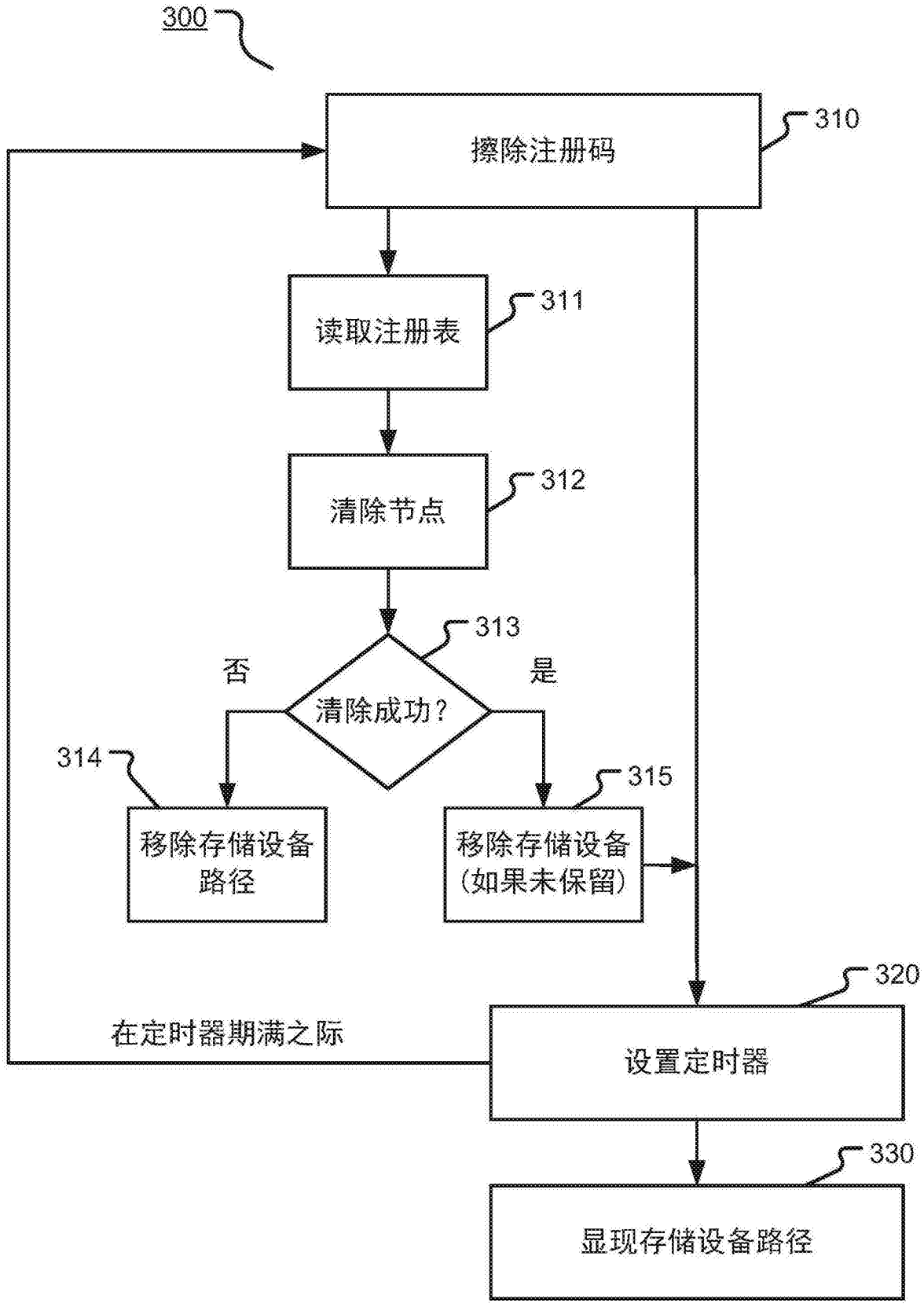

300
擦除注册码
310
读取注册表
311
清除节点
312
313
清除成功？
否
是
314
移除存储设备路径
315
移除存储设备（如果未保留）
320
在定时器期满之际
设置定时器
330
显现存储设备路径

图3

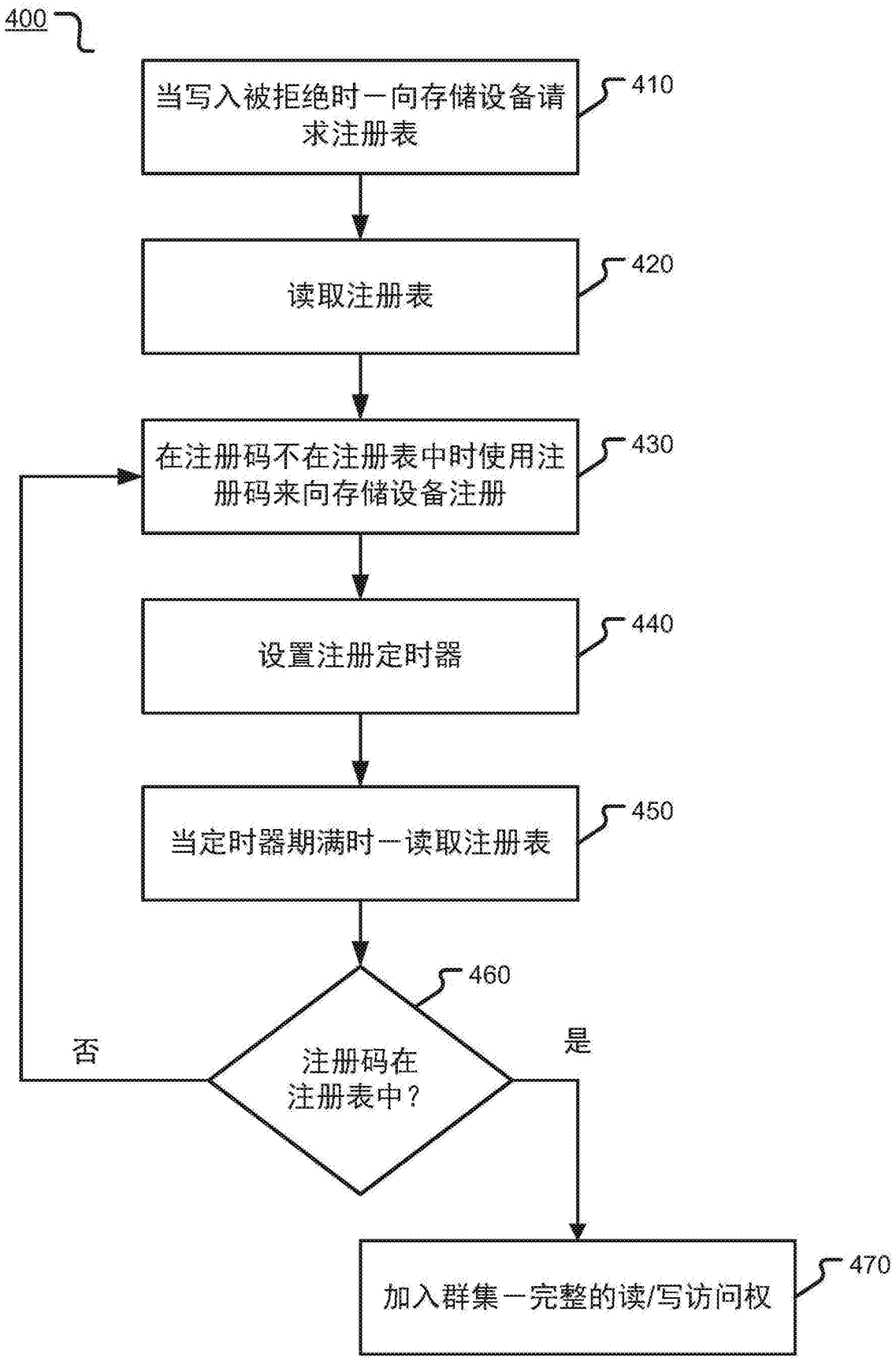

400
当写入被拒绝时一向存储设备请求注册表
410
读取注册表
420
在注册码不在注册表中时使用注册码来向存储设备注册
430
设置注册定时器
440
当定时器期满时一读取注册表
450
460
注册码在注册表中？
否
是
加入群集一完整的读/写访问权
470

图4

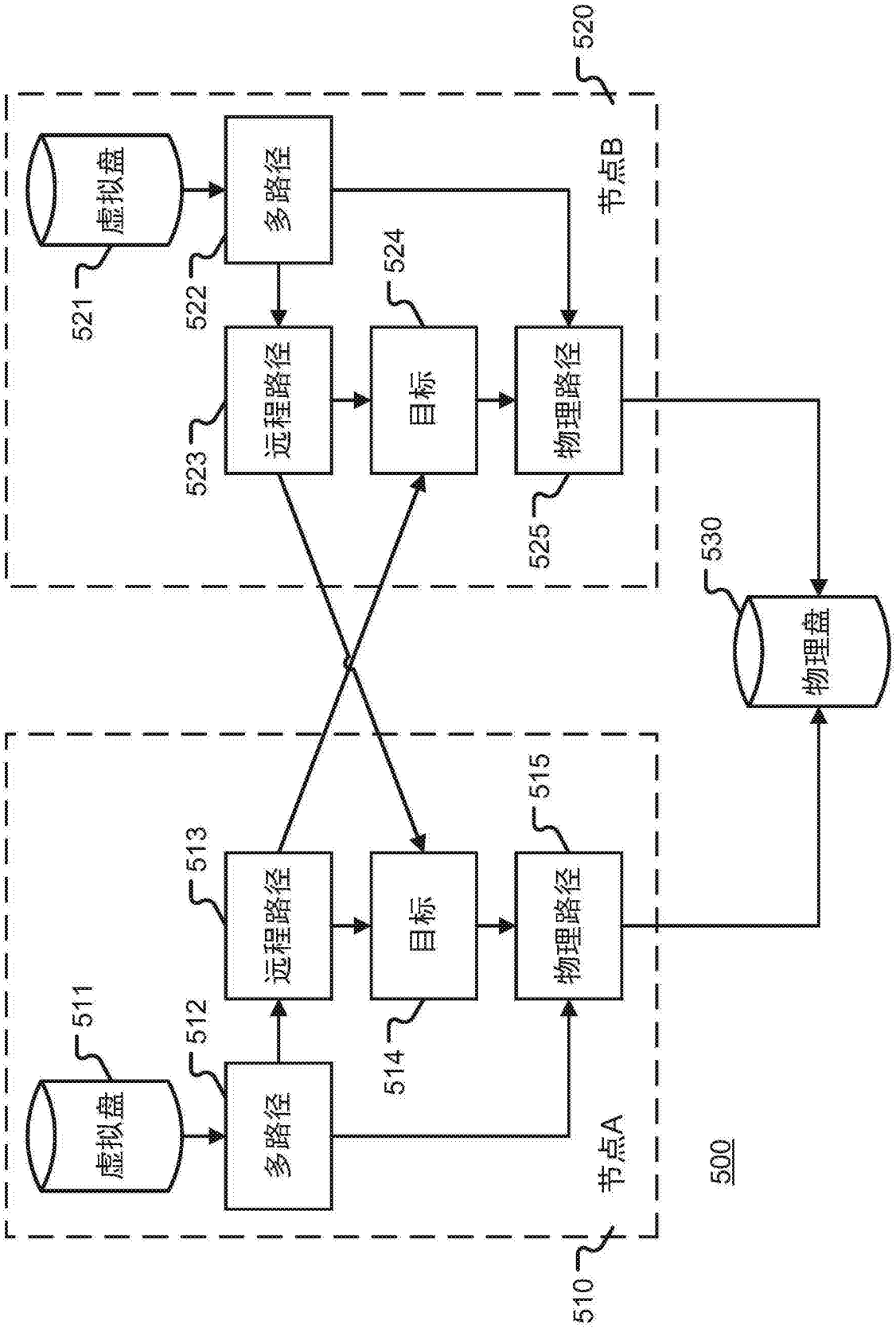

520
节点B
虚拟盘
521
多路径
522
远程路径
523
目标
524
物理路径
525
530
物理盘
515
513
远程路径
目标
物理路径
511
虚拟盘
512
多路径
514
节点A
500
510

图5

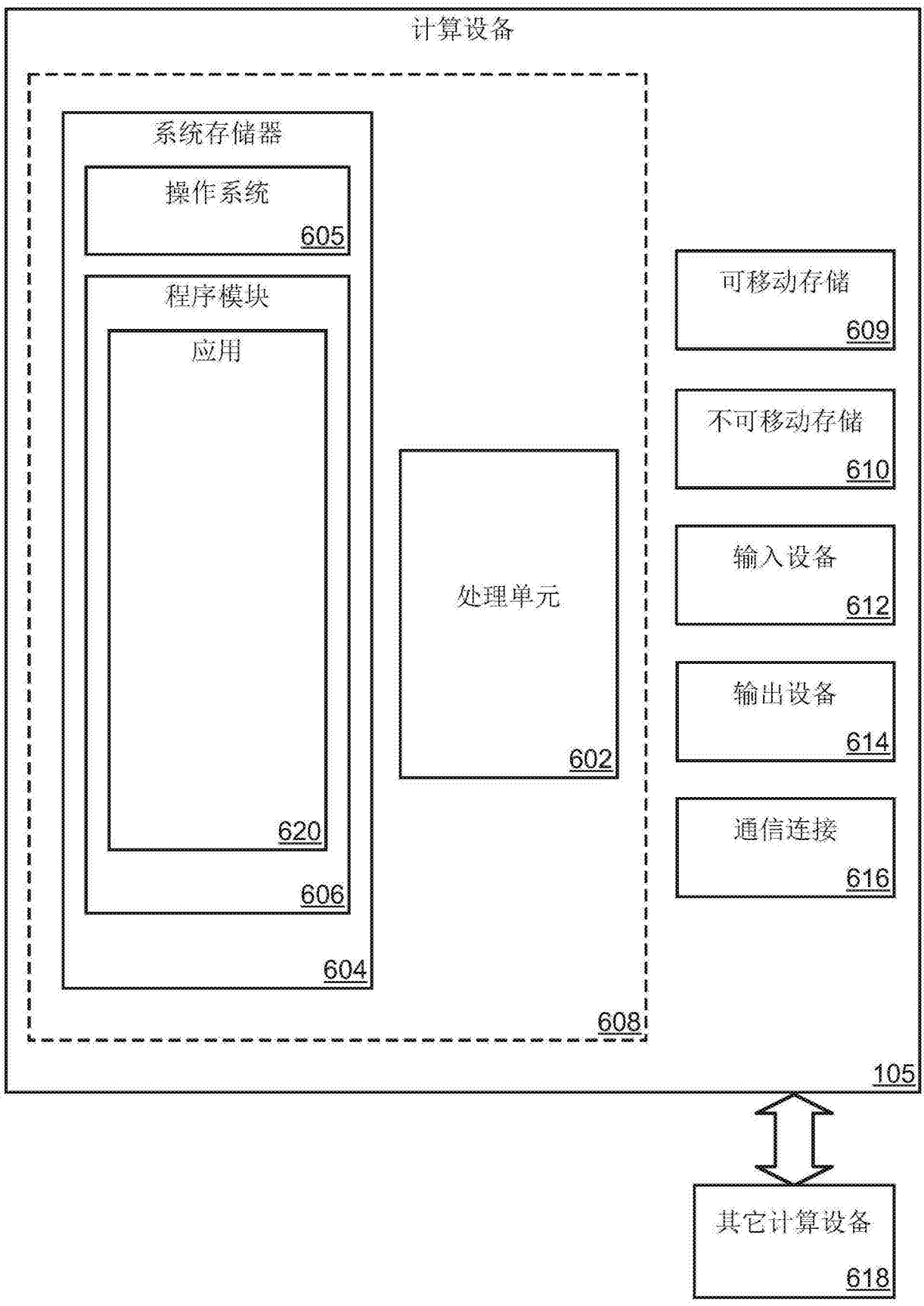

计算设备
系统存储器
操作系统
605
程序模块
应用
620
606
604
处理单元
602
608
可移动存储
609
不可移动存储
610
输入设备
612
输出设备
614
通信连接
616
105
其它计算设备
618

图6

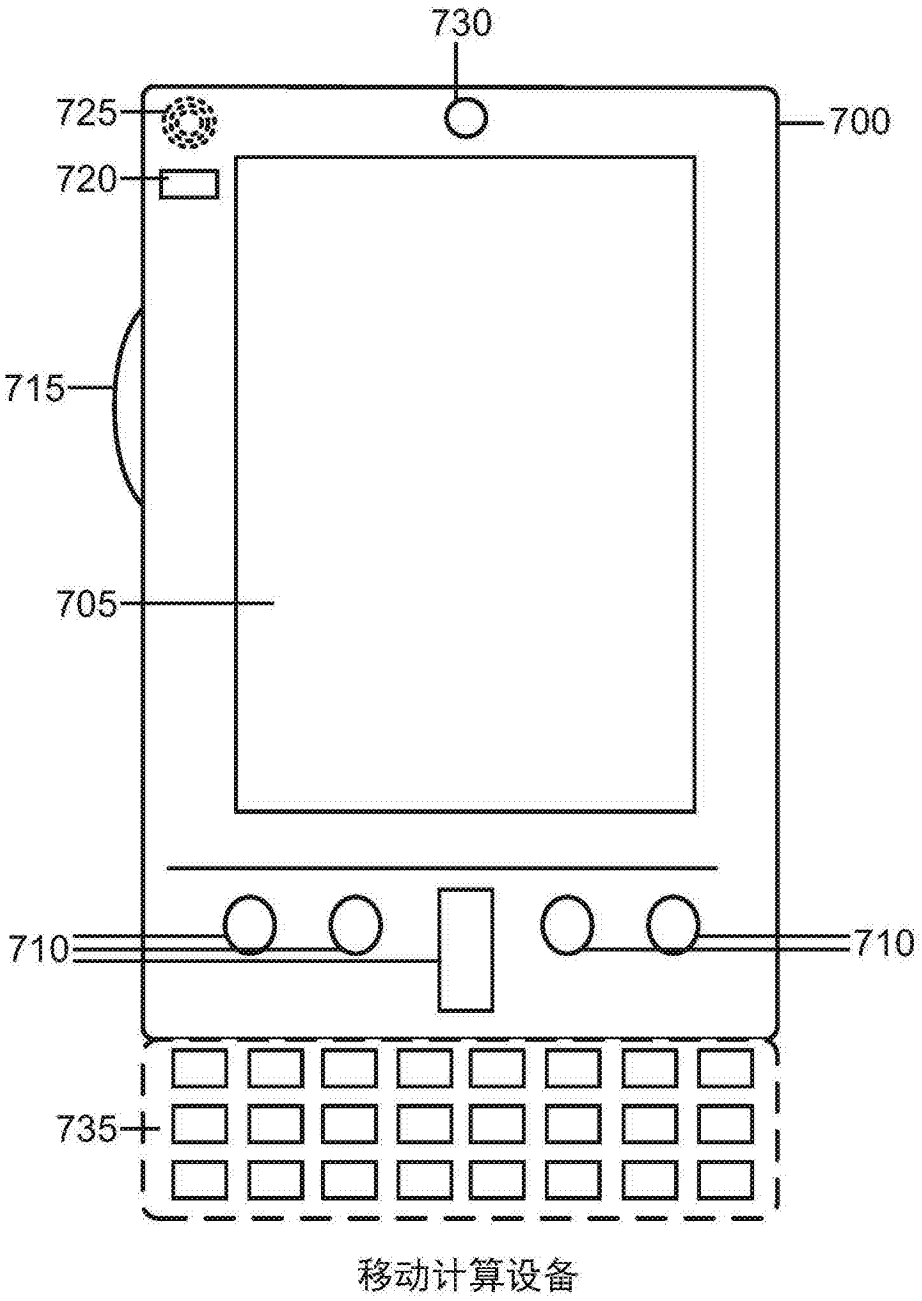


图7A

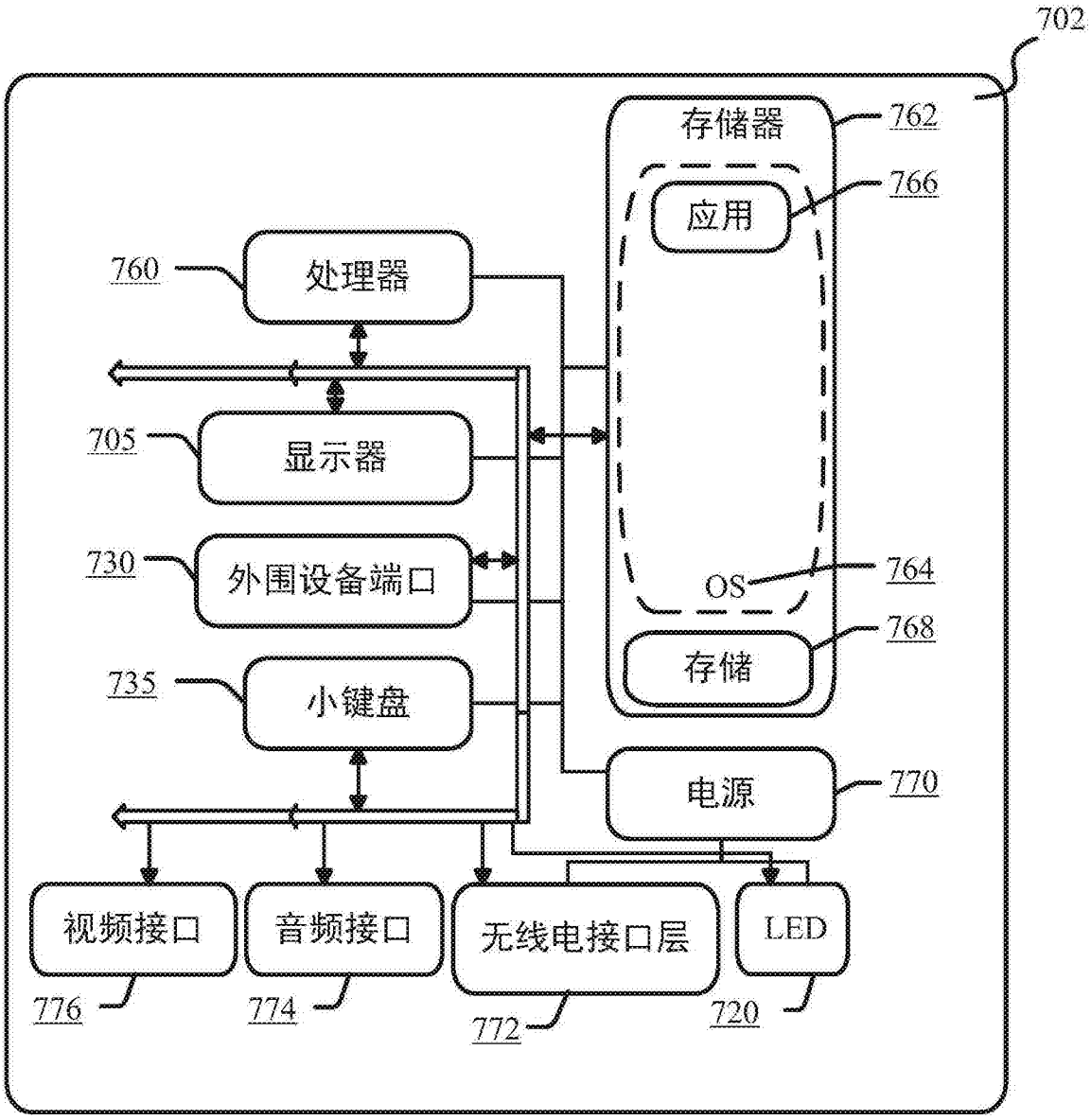

702
存储器
762
766
应用
760
处理器
705
显示器
730
外围设备端口
764
OS
768
存储
735
小键盘
770
电源
视频接口
音频接口
无线电接口层
LED
776
774
772
720

图7B

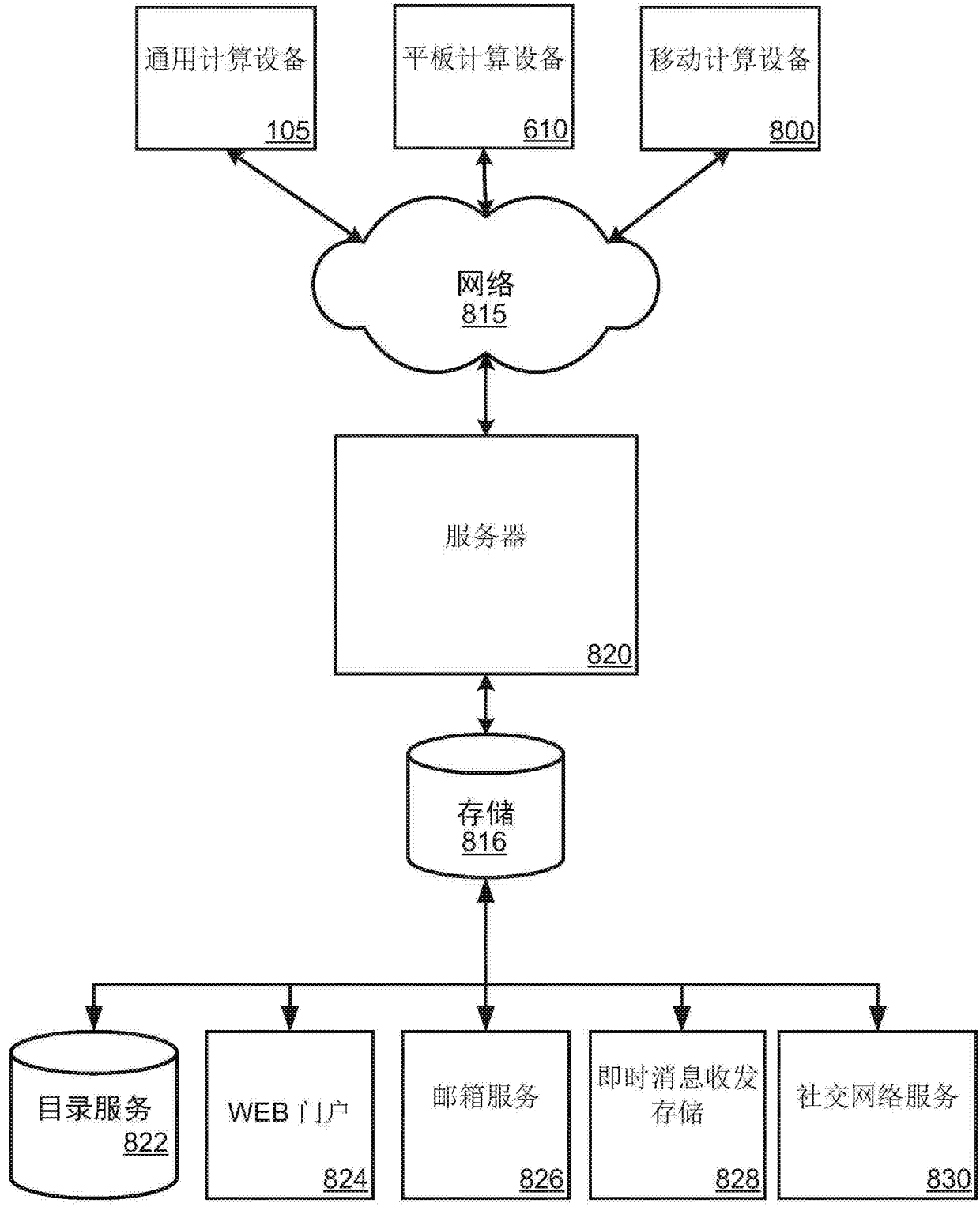

通用计算设备
105
平板计算设备
610
移动计算设备
800
网络
815
服务器
820
存储
816
目录服务
822
WEB 门户
824
邮箱服务
826
即时消息收发
存储
828
社交网络服务
830

图8