



(12)发明专利

(10)授权公告号 CN 104335159 B

(45)授权公告日 2018.01.16

(21)申请号 201380019653.8

(22)申请日 2013.02.27

(65)同一申请的已公布的文献号
申请公布号 CN 104335159 A

(43)申请公布日 2015.02.04

(30)优先权数据
13/408892 2012.02.29 US

(85)PCT国际申请进入国家阶段日
2014.10.13

(86)PCT国际申请的申请数据
PCT/US2013/027884 2013.02.27

(87)PCT国际申请的公布数据
W02013/130507 EN 2013.09.06

(73)专利权人 赛门铁克公司
地址 美国加利福尼亚州

(72)发明人 P·R·达什

(74)专利代理机构 北京纪凯知识产权代理有限公司 11245

代理人 赵蓉民

(51)Int.Cl.
G06F 3/06(2006.01)
G06F 11/20(2006.01)

(56)对比文件
CN 1834933 A, 2006.09.20,
CN 1549974 A, 2004.11.24,
CN 1900914 A, 2007.01.24,
US 2005198452 A1, 2005.09.08,
US 6976186 B1, 2005.12.13,
US 7885923 B1, 2011.02.08,
CN 100527092 C, 2009.08.12,

审查员 李艳军

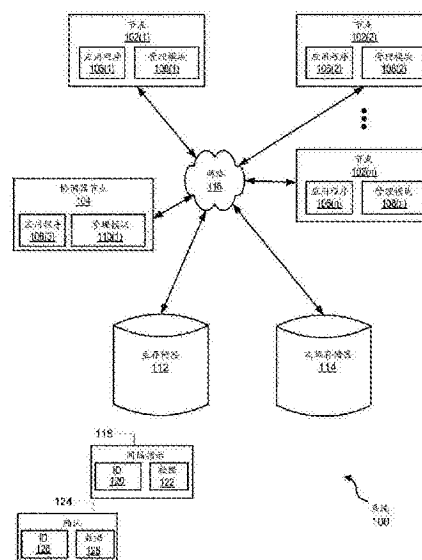
权利要求书6页 说明书21页 附图17页

(54)发明名称

用于间隔控制复制的方法、系统和设备

(57)摘要

本发明描述了用于间隔控制复制的各种系统和方法。例如,一种方法涉及一节点,所述节点接收指示第一间隔开始的第一间隔指示。在接收到所述第一间隔指示之后,所述节点上的应用程序启动第一写入,这包括所述节点启动第一本地写入和第一远程写入。所述第一本地写入访问本地存储器。所述第一远程写入访问远程存储器并且复制所述第一本地写入。所述节点在将来自所述应用程序的所述第一写入视为完成之前,等待指示所述本地和远程写入已完成以及针对第二间隔指示的确认。



1. 一种用于数据复制的方法,包括:
在本地节点处,接收来自应用程序的第一写入和第二写入,以及
第一间隔指示,其中
所述第一间隔指示指示第一间隔开始,并且
所述第一间隔指示是多个间隔指示中的一者;
从所述本地节点发送第一间隔确认至协调器节点,其中,
所述第一间隔确认指示接收到所述第一间隔指示;
经由所述本地节点,启动第一本地写入、第一远程写入和第二远程写入,其中
响应于接收到所述第一间隔指示,启动所述第一本地写入、所述第一远程写入和所述
第二远程写入,
所述第一本地写入访问本地存储器,
所述第一远程写入访问第一远程存储器,
所述第一远程写入复制所述第一本地写入,
所述第二远程写入访问第二远程存储器,以及
所述第二远程写入复制所述第一本地写入;
在所述本地节点处,接收第一远程写入确认,其中
所述第一远程写入确认指示所述第一远程存储器完成了所述第一远程写入;
在所述本地节点处,接收第二远程写入确认,其中
所述第二远程写入确认指示所述第二远程存储器完成了所述第二远程写入;
在所述本地节点处,接收第二间隔指示,其中
响应于发送所述第一间隔确认,接收所述第二间隔指示;以及
经由所述本地节点,启动第二本地写入、第三远程写入和第四远程写入,其中,
响应于所述应用程序发出所述第二写入、接收到所述第一远程写入确认,和接收到所
述第二间隔指示,启动所述第二本地写入、所述第三远程写入和所述第四远程写入。
2. 根据权利要求1所述的方法,还包括
确认所述第一写入的完成,其中
响应于接收到所述第一远程写入确认、所述第二远程写入确认以及所述第二间隔指
示,执行所述确认;以及
响应于所述应用程序从管理模块接收所述第一写入的所述完成的确认,由所述应用程
序发出所述第二写入。
3. 根据权利要求1所述的方法,其中
在协调器节点处,确定多个节点中的每个节点均已接收到所述第二间隔指示,其中
所述确定包括接收确认,确认所述多个节点中的每一者均已接收到所述第二间隔指
示,并且所述协调器节点通信地耦合到所述多个节点。
4. 根据权利要求1所述的方法,其中
所述第一远程写入具有第一关联时间标记,并且
所述第一关联时间标记指示所述第一间隔。
5. 根据权利要求1所述的方法,还包括:
如果在接收到所述第二间隔指示之前所述第一本地写入和所述第一远程写入已完成,

则确定新闻隔。

6. 根据权利要求1所述的方法,还包括:

应用程序发出第一写入,其中

进一步响应于所述应用程序发出所述第一写入,启动所述第一本地写入、所述第一远程写入和所述第二远程写入,并且

所述第一写入访问第一数据,

所述应用程序发出所述第二写入,其中

响应于接收到所述第一远程写入确认和所述第二间隔指示,所述应用程序发出所述第二写入,

所述第二写入访问第二数据,

所述第二数据依赖于所述第一数据。

7. 根据权利要求1所述的方法,其中

所述第一远程写入访问第一数据,

所述第二远程写入访问第二数据,并且

所述第二数据依赖于所述第一数据。

8. 根据权利要求1所述的方法,还包括:

经由所述本地节点,启动第三本地写入、第五远程写入和第六远程写入,其中

响应于接收到第三间隔指示,启动所述第三本地写入、所述第五远程写入和所述第六远程写入,

所述第三本地写入访问所述本地存储器,

所述第五远程写入访问所述第一远程存储器,

所述第五远程写入复制所述第三本地写入,

所述第六远程写入访问所述第二远程存储器,

所述第一远程写入访问第一数据,

所述第三远程写入访问第二数据,

所述第一远程写入访问第三数据,并且

所述第三数据、所述第二数据和所述第一数据彼此无关。

9. 根据权利要求1所述的方法,还包括:

确定第一节点是否还未对第一量的所述多个间隔指示作出响应;

响应于确定所述第一节点还未对所述第一量的所述多个间隔指示作出响应,在间隔指示中指示所述第一节点未响应。

10. 根据权利要求9所述的方法,还包括:

确定所述第一节点是否在延迟之后对间隔指示作出响应;并且

响应于确定所述第一节点对所述间隔指示作出响应,处理由所述第一节点在所述延迟期间执行的任何远程写入。

11. 根据权利要求1所述的方法,还包括:

确定第一节点是否访问特定数据,其中

所述特定数据包括第一数据相关性特性;并且

响应于确定所述第一节点正在访问所述特定数据,所述第一节点执行所述特定数据的

批处理。

12. 一种用于数据复制的系统,所述系统包括:

应用程序,其中

所述应用程序被配置成发出第一写入和第二写入;以及

管理模块,其中

所述管理模块被配置成

在本地节点处,接收来自所述应用程序的所述第一写入和所述第二写入,以及第一间隔指示,其中

所述第一间隔指示指示第一间隔开始,

从所述本地节点发送第一间隔确认至协调器节点,其中

所述第一间隔确认指示接收到所述第一间隔指示,

经由所述本地节点,启动第一本地写入、第一远程写入和第二远程写入,其中

响应于以下各项,启动所述第一本地写入、所述第一远程写入和所述第二远程写入:

所述应用程序发出所述第一写入,和

所述管理模块接收所述第一间隔指示,

所述第一本地写入访问本地存储器,

所述第一远程写入访问第一远程存储器,

所述第一远程写入复制所述第一本地写入,

所述第二远程写入访问第二远程存储器,以及

所述第二远程写入复制所述第一本地写入;

在所述本地节点处,接收第一远程写入确认,其中

所述第一远程写入确认指示所述第一远程存储器完成了所述第一远程写入,

在所述本地节点处,接收第二远程写入确认,其中

所述第二远程写入确认指示所述第二远程存储器完成了所述第二远程写入,

在所述本地节点处,接收第二间隔指示,其中

响应于发送所述第一间隔确认,接收所述第二间隔指示;以及

经由所述本地节点,启动第二本地写入、第三远程写入和第四远程写入,其中

响应于所述应用程序发出所述第二写入、接收到所述第一远程写入确认,和接收到所述第二间隔指示,启动所述第二本地写入、所述第三远程写入和所述第四远程写入。

13. 根据权利要求12所述的系统,其中

所述管理模块进一步被配置成

确认所述第一写入的完成,其中

所述管理模块响应于以下项确认所述第一写入的所述完成:

接收到所述第一远程写入确认,

接收到所述第二远程写入确认,以及

接收到所述第二间隔指示;以及

所述应用程序被配置成响应于从所述管理模块接收到所述第一写入的所述完成的确认,发出所述第二写入。

14. 根据权利要求12所述的系统,其中

多个节点包括所述系统，

所述管理模块进一步响应于以下项来确认所述第一写入

确定所述多个节点中的每一者均已接收到所述第二间隔指示。

15. 一种用于数据复制的设备，包括：

用于在本地节点处，接收来自应用程序的第一写入和第二写入以及第一间隔指示的装置，其中

所述第一间隔指示指示第一间隔开始，

所述第一间隔指示为多个间隔指示中的一者；

用于从所述本地节点发送第一间隔确认至协调器节点的装置，其中

所述第一间隔确认指示所述第一间隔指示的接收；

用于经由所述本地节点，启动第一本地写入、第一远程写入和第二远程写入的装置，其中

响应于接收到所述第一间隔指示启动所述第一本地写入、

所述第一远程写入和所述第二远程写入，

所述第一本地写入访问本地存储器，

所述第一远程写入访问第一远程存储器，

所述第一远程写入复制所述第一本地写入，

所述第二远程写入访问第二远程存储器，以及

所述第二远程写入复制所述第一本地写入；

用于在所述本地节点处，接收第一远程写入确认的装置，

其中

所述第一远程写入确认指示所述第一远程存储器完成了所述第一远程写入，

用于在所述本地节点处，接收第二远程写入确认的装置，其中

所述第二远程写入确认指示所述第二远程存储器完成了所述第二远程写入；

用于在所述本地节点处接收第二间隔指示的装置，其中响应于发送所述第一间隔确认，接收所述第二间隔指示；以及

用于经由所述本地节点启动第二本地写入、第三远程写入和第四远程写入的装置，其中

所述第二本地写入、所述第三远程写入和所述第四远程写入响应于以下各项而被启动：

所述应用程序发出所述第二写入，

接收到所述第一远程写入确认，以及接收到所述第二间隔指示。

16. 根据权利要求15所述的设备，其中

所述第一远程写入和所述第一本地写入中的每一者均具有第一关联时间标记，并且

所述第一关联时间标记指示所述第一间隔。

17. 根据权利要求15所述的设备，还包括

用于确认所述第一写入的完成的装置，其中

响应于接收到所述第一远程写入确认、所述第二远程写入确认和所述第二间隔指示，执行所述确认，并且

响应于所述应用程序从管理模块接收到所述第一写入的所述完成的确认,所述应用程序发出所述第二写入。

18.一种计算机可读存储介质,其包含用于引起处理器执行一种用于数据复制的方法的指令,所述方法包含:

在本地节点处,接收来自应用程序的第一写入和第二写入以及第一间隔指示,其中

所述第一间隔指示指示第一间隔开始,并且

所述第一间隔指示为多个间隔指示中的一者;

从所述本地节点发送第一间隔确认至协调器节点,其中所述第一间隔确认指示所述第一间隔指示的接收;

经由所述本地节点,启动第一本地写入、第一远程写入和第二远程写入,其中

响应于接收到所述第一间隔指示,启动所述第一本地写入、所述第一远程写入和所述第二远程写入,

所述第一本地写入访问本地存储器,

所述第一远程写入访问第一远程存储器,

所述第一远程写入复制所述第一本地写入,

所述第二远程写入访问第二远程存储器,以及

所述第二远程写入复制所述第一本地写入;

在所述本地节点处,接收第一远程写入确认,其中

所述第一远程写入确认指示所述第一远程存储器完成了所述第一远程写入,

在所述本地节点处,接收第二远程写入确认,其中

所述第二远程写入确认指示所述第二远程存储器完成了所述第二远程写入,

在所述本地节点处,接收第二间隔指示,其中

响应于发送所述第一间隔确认,接收所述第二间隔指示;以及

经由所述本地节点,启动第二本地写入、第三远程写入和第四远程写入,其中

响应于所述应用程序发出所述第二写入、接收到所述第一远程写入确认,和接收到所述第二间隔指示,启动所述第二本地写入、所述第三远程写入和所述第四远程写入。

19.根据权利要求18所述的计算机可读存储介质,其中

所述第一远程写入和所述第一本地写入中的每一者均具有第一关联时间标记,并且

所述第一关联时间标记指示所述第一间隔。

20.根据权利要求18所述的计算机可读存储介质,还包括

确认所述第一写入的完成,其中

响应于接收到所述第一远程写入确认、所述第二远程写入确认和所述第二间隔指示,执行所述确认,并且

响应于所述应用程序从管理模块接收到所述第一写入的所述完成的确认,所述应用程序发出所述第二写入。

21.根据权利要求18所述的计算机可读存储介质,其中

在协调器节点处,确定多个节点中的每个节点均已接收到所述第二间隔指示,其中

所述确定包括接收确认,确认所述多个节点中的每一者均已接收到所述第二间隔指示,并且所述协调器节点通信地耦合到所述多个节点。

22. 根据权利要求18所述的计算机可读存储介质,其中
所述第一远程写入具有第一关联时间标记,并且
所述第一关联时间标记指示所述第一间隔。

用于间隔控制复制的方法、系统和设备

技术领域

[0001] 本申请涉及数据复制。具体地讲,本申请涉及管理分布式环境中的复制。

背景技术

[0002] 复制是一种用于将数据损失降到最低并且提高数据可用性的技术,其中复制的数据副本被分布并且存储在分布式计算环境中的一个或多个远程站点或节点。例如,当执行站点迁移时,存储数据的物理磁盘或与此类磁盘关联的节点可能出现故障。在此类故障事件中,为了确保数据完整性和可用性,可利用远程复制数据副本。复制经常与其他高可用性技术(诸如群集)结合以提供强大的数据存储解决方案,在现代计算系统中出现需要时,其能够被用于消除数据损失或将数据损失降到最低,并且利用可用数据快速恢复。

发明内容

[0003] 基于主机的复制随着群集中的节点增加而不能扩展良好。在这种环境中实施写入次序的保真度需要的消息传递减慢了应用程序I/O和复制事件。为了改善这种复制性能,本文公开的内容提供了基本上降低节点之间需要的消息传递。贯穿本文公开的内容更详细地提供关于发明内容的公开。

[0004] 本文公开了一种用于数据复制的方法,包括:在本地节点处,接收来自应用程序的第一写入和第二写入,以及第一间隔指示,其中所述第一间隔指示指示第一间隔开始,并且所述第一间隔指示是多个间隔指示中的一者;从所述本地节点发送第一间隔确认至协调器节点,其中,所述第一间隔确认指示接收到所述第一间隔指示;经由所述本地节点处,启动第一本地写入、第一远程写入和第二远程写入,其中响应于接收到所述第一间隔指示,启动所述第一本地写入、所述第一远程写入和所述第二远程写入,所述第一本地写入访问本地存储器,所述第一远程写入访问第一远程存储器,所述第一远程写入复制所述第一本地写入,所述第二远程写入访问第二远程存储器,以及所述第二远程写入复制所述第一本地写入;在所述本地节点处,接收第一远程写入确认,其中所述第一远程写入确认指示所述第一远程存储器完成了所述第一远程写入;在所述本地节点处,接收第二远程写入确认,其中所述第二远程写入确认指示所述第二远程存储器完成了所述第二远程写入;在所述本地节点处,接收第二间隔指示,其中响应于发送所述第一间隔确认,接收所述第二间隔指示;以及经由所述本地节点,启动第二本地写入、第三远程写入和第四远程写入,其中,响应于所述应用程序发出所述第二写入、接收到所述第一远程写入确认,和接收到所述第二间隔指示,启动所述第二本地写入、所述第三远程写入和所述第四远程写入。

[0005] 本文公开了一种用于数据复制的系统,所述系统包括:应用程序,其中所述应用程序被配置成发出第一写入和第二写入;以及管理模块,其中所述管理模块被配置成在本地节点处,接收来自所述应用程序的所述第一写入和所述第二写入,以及第一间隔指示,其中所述第一间隔指示指示第一间隔开始,从所述本地节点发送第一间隔确认至协调器节点,其中所述第一间隔确认指示接收到所述第一间隔指示,经由所述本地节点,启动第一本地

写入、第一远程写入和第二远程写入,其中响应于以下各项,启动所述第一本地写入、所述第一远程写入和所述第二远程写入:所述应用程序发出所述第一写入,和所述管理模块接收所述第一间隔指示,所述第一本地写入访问本地存储器,所述第一远程写入访问第一远程存储器,所述第一远程写入复制所述第一本地写入,所述第二远程写入访问第二远程存储器,以及所述第二远程写入复制所述第一本地写入;在所述本地节点处,接收第一远程写入确认,其中所述第一远程写入确认指示所述第一远程存储器完成了所述第一远程写入,在所述本地节点处,接收第二远程写入确认,其中所述第二远程写入确认指示所述第二远程存储器完成了所述第二远程写入,在所述本地节点处,接收第二间隔指示,其中响应于发送所述第一间隔确认,接收所述第二间隔指示;以及经由所述本地节点,启动第二本地写入、第三远程写入和第四远程写入,其中响应于所述应用程序发出所述第二写入、接收到所述第一远程写入确认,和接收到所述第二间隔指示,启动所述第二本地写入、所述第三远程写入和所述第四远程写入。

[0006] 本文公开了一种用于数据复制的设备,包括:用于在本地节点处,接收来自应用程序的第一写入和第二写入以及第一间隔指示的装置,其中所述第一间隔指示指示第一间隔开始,所述第一间隔指示为多个间隔指示中的一者;用于从所述本地节点发送第一间隔确认至协调器节点的装置,其中所述第一间隔确认指示所述第一间隔指示的接收;用于经由所述本地节点,启动第一本地写入、第一远程写入和第二远程写入的装置,其中响应于接收到所述第一间隔指示启动所述第一本地写入、所述第一远程写入和所述第二远程写入,所述第一本地写入访问本地存储器,所述第一远程写入访问第一远程存储器,所述第一远程写入复制所述第一本地写入,所述第二远程写入访问第二远程存储器,以及所述第二远程写入复制所述第一本地写入;用于在所述本地节点处,接收第一远程写入确认的装置,其中所述第一远程写入确认指示所述远程存储器完成了所述第一远程写入,用于在所述本地节点处,接收第二远程写入确认的装置,其中所述第二远程写入确认指示所述远程存储器完成了所述第二远程写入;用于在所述本地节点处接收第二间隔指示的装置,其中响应于发送所述第一间隔确认,接收所述第二间隔指示;以及用于经由所述本地节点启动第二本地写入、第三远程写入和第四远程写入的装置,其中所述第二本地写入、所述第三远程写入和所述第四远程写入响应于以下各项而启动:所述应用程序发出所述第二写入,接收到所述第一远程写入确认,以及接收到所述第二间隔指示。

附图说明

[0007] 通过参考附图,本申请的实施例可能更便于理解,并且其众多对象、特征和优点对于本领域的技术人员来说变得显而易见。

[0008] 图1是根据一个实施例示出的分布式存储系统的框图。

[0009] 图2A-2D是根据一个或多个实施例示出的用于间隔控制复制中节点操作的方法的流程图。

[0010] 图3是根据一个实施例示出的用于间隔控制复制中协调器节点操作的方法的流程图。

[0011] 图4-7是根据一个实施例示出的在间隔控制复制期间通过分布式存储系统的节点进行的访问的图。

[0012] 图8-10是根据一个实施例示出的在间隔控制复制期间各个节点与存储设备之间进行通信的图。

[0013] 图11是根据一个实施例示出的用于间隔控制复制中协调器节点操作的方法的流程图。

[0014] 图12是根据一个实施例示出的节点的各种组件的框图。

[0015] 图13是根据一个实施例示出的协调器节点的各种组件的框图。

[0016] 图14是示出其中可以实施本申请的实施例的网络体系结构的框图。

[0017] 图15是示出适用于实施本申请的实施例的计算机系统的示例的框图。

[0018] 由于本申请的实施例容易受到各种修改形式和替代形式的影响,因此提供具体实施例作为附图和具体实施方式中的示例。应当理解,附图和具体实施方式并非旨在将实施例限制为所公开的具体形式。相反,其目的在于涵盖所附权利要求所限定的本发明精神和范围内的所有修改形式、等同形式和替代形式。

具体实施方式

[0019] 尽管已结合多个实施例描述了本发明,但本发明并非旨在限制为本文中示出的具体形式。相反,本发明旨在涵盖可以合理地包含在所附权利要求所限定的本发明范围内的替代形式、修改形式和等同形式。

[0020] 现代分布式存储环境可包括经由一个或多个互连网络连接的多个存储对象。互连网络提供基础结构以连接分布式共享存储环境的各种元素。存储系统经常使用数据冗余机制确保数据完整性、一致性和可用性。数据冗余的其他用途可包括备份数据、分布式负载共享、灾难恢复或时间点分析和报告。获得数据冗余的一种方法是将数据从主存储系统拷贝或复制到第二或复制存储系统。换句话说,存储系统可将写入到数据块主副本的数据复制到冗余,或将所述数据块的副本复制到其他次级存储系统。在一些设计中,当执行数据I/O时同步完成该拷贝。在其他设计中,该复制可与第二存储系统的数据状态异步执行,该数据状态滞后主存储状态某一时间间隔,该时间间隔可为几分之一秒到多个小时,具体取决于设计目的和所用的技术。还可使用周期性复制,其中写入到第二存储系统的写入(例如)在每个周期得以累积并且分批写入。

[0021] 图1是示出包括节点和存储器的集合的分布式系统100的框图。分布式系统100(例如集群)包括多个节点,例如节点102(1)、102(2)-102(n)和协调器节点104。每个节点可使用网络(例如网络116)与主存储器(例如,存储器112)通信。另外,每个节点可使用网络(例如网络116)与次级存储器(例如,存储器114)通信。网络116可包括局域网(LAN)、广域网(WAN)、存储区域网(SAN)、以太网和/或其任何组合。应当注意,分布式系统100可包括不同数量的节点。

[0022] 每个节点102(1)-102(n)可包括应用程序106(1)-106(n)中的一者或多者以及管理模块108(1)-108(n)中的一者或多者。类似地,协调器节点104可包括应用程序106(3)和管理模块110。每个此类应用程序106(1)-106(n)可发起到存储器的读写以及从存储器进行的读写。例如,应用程序106(1)可发起写入到主存储器112。节点102(1)随后可执行该写入到主存储器112。该写入也可复制到次级存储器114,诸如通过使用管理模块106(1)。管理模块106(1)可针对分布式系统100中的节点102(1)以及其他节点执行复制和其他服务。在一

个实施例中,假定当每个应用程序发出写入到特定数据时,该应用程序直到完成第一时间写入才写入到该数据(或到从属数据(dependent data))。在一个实施例中,与应用程序关联的管理模块向该应用程序确认此类第一时间写入完成。在从其关联管理模块接收到该确认时,应用程序随后可发出相关写入。在一个实施例中,管理模块可与驻存在相同节点上的应用程序关联,例如,管理模块108(1)与节点102(1)上的应用程序108(1)关联。然而,在其他实施例中,可设想管理模块与应用程序之间的其他关联。

[0023] 如本文所述的间隔控制复制可在分布式或集群计算环境上实施,诸如图1中所示的环境。如图1所示,源节点集(诸如节点102(1)-102(n))可将数据写入到一个或多个主存储设备,诸如主存储器112,并且也可将数据写入复制到复制目标,诸如次级存储器114。图1中所示的系统可实施如本文所述的间隔控制复制。在一些实施例中,间隔控制复制涉及节点,诸如协调器节点104,从而将间隔指示发送到分布式系统100中的其他节点。这些其他节点(例如节点102(1)-102(n))以本文所述的方式执行写入到主存储器112和次级存储器114。来自节点(诸如来自节点102(1))的每一数据写入均可包括当前间隔的标识符。间隔控制复制可与至少三种复制模式一起使用,即同步复制、异步复制和周期性复制,如本文所述。

[0024] 同步复制通常确保在更新主副本时创建数据的远程副本。远程副本的创建可与主副本基本上类似。在同步复制中,直到在主存储器112和次级存储器114(例如,镜像站点)两者处确认完成才将I/O操作视为完成。在一个实施例中,可通过应用程序和/或管理模块在两个位置处重试和/或回滚未完成操作,以便确保远程副本(基本上)为主副本的镜像。在复制期间,在更新主副本时创建数据的远程副本。

[0025] 例如,如果节点102(1)使用主存储器112和次级存储器114写入到第一数据对象,则使用间隔控制复制可确保在某间隔中开始的写入不在相同间隔中完成。此外,直到节点102(1)完成其(第一)写入到对象主存储器112和次级存储器114,节点102(1)或节点102(2)才发出写入到该第一数据对象或依赖于第一写入的第二数据对象。这样确保在使用间隔控制复制时,在与原始I/O写入相同的间隔内,没有同一节点或其它节点写入相关I/O。间隔控制复制可应用于将数据写入到主存储器112和次级存储器114的节点。直到节点(例如,特定子集的所有节点)移至间隔(m+1),实施在间隔m中接收I/O的间隔控制复制的节点才将完成该I/O。在一些实施例中,当节点确认该下一个间隔时其已“移至”下一个间隔。以下附图中描述了此类确认的各种实施方式。根据实施方式,假定当节点启动间隔(m+2)时节点移至间隔(m+1)。在另一个实施方式中,假定当节点发回确认间隔(m+1)的消息时,该节点已移至间隔(m+1)。换言之,对于给定节点,在间隔m中接收的任何节点上的I/O彼此无关,因为在间隔m+1之前,没有其他节点可以完成这些I/O中的任何一者。

[0026] 在一个实施例中,对于同步复制,节点在发起与前一个I/O写入相关的下一个I/O写入之前,等待来自主存储器的确认、来自第二存储器的确认和用于下一个间隔的间隔指示。例如,用于节点102(1)的管理模块108(1)在向应用程序106(1)确认第一I/O写入(即,第一本地写入和第一远程写入)完成之前,等待来自主存储器112的有关第一本地写入(例如,具有用于间隔m的关联ID)的确认、来自第二存储器的有关第一远程写入(例如,具有用于间隔m的关联ID)的确认、用于下一个间隔(例如,m+1)的间隔指示以及分布式系统100中的其他节点已移至下一个间隔(m+1)的确认。一旦应用程序106(1)从其关联的管理模块108(1)

接收到此类确认,应用程序106 (1) 就可发起与第一I/O写入相关的下一个I/O写入。

[0027] 在一个实施例中,对于异步复制,节点在发起与前一个I/O写入相关的下一个I/O写入之前,等待来自主存储器的确认和用于下一个间隔的间隔指示。换句话说讲,与同步复制相比之下,异步复制中的节点在向应用程序发出确认之前不等待来自第二存储器的确认。例如,用于节点102 (1) 的管理模块108 (1) 在向应用程序106 (1) 确认第一I/O写入(即,第一本地写入)完成之前,等待来自主存储器112的有关第一本地写入(例如,具有用于间隔m的关联ID)的确认和用于下一个间隔(例如,m+1)的间隔指示以及分布式系统100中的其他节点已移至该下一个间隔(m+1)的确认。一旦应用程序106 (1) 从其关联的管理模块108 (1) 接收到此类确认,应用程序106 (1) 就可发起依赖于第一I/O写入的下一个I/O写入。然而,在某一时刻,次级存储器确认远程写入。次级存储器114在为下一个间隔(m+1)应用任何I/O写入之前针对间隔(m)应用所有I/O写入。

[0028] 在一个实施例中,针对周期性复制,节点以特定周期将数据从主存储器拷贝到次级存储器。例如,可通过节点收集要复制的数据,诸如通过使用快照和/或使用另一种技术。在一个实施例中,节点可使用本地存储器,诸如日志和/或暂存区域,其中可累积远程写入。在特定时间周期,可将累积的该数据写入到次级存储器,诸如通过使用多个远程写入(例如,批量写入)。周期性复制中的节点在向应用程序发送确认之前,不等待来自次级存储器的确认。例如,用于节点102 (1) 的管理模块108 (1) 在向应用程序106 (1) 确认第一I/O写入(即,第一本地写入)完成之前,等待来自主存储器112的有关第一本地写入(例如,具有用于间隔m的关联ID)的确认和用于下一个间隔(例如,m+1)的间隔指示以及分布式系统100中的其他节点已移至该下一个间隔(m+1)的确认。一旦应用程序106 (1) 从其关联的管理模块108 (1) 接收到此类确认,应用程序106 (1) 就可发起依赖于第一I/O写入的下一个I/O写入。如上所述,每个此类I/O写入被累积并且随后按特定周期写入到次级存储器。

[0029] 对于所有三种复制类型,在一个实施例中,每个此类间隔指示可以实施为图1的间隔指示118。间隔指示118可包括间隔ID 120和数据122,但也可设想其他实施方式。每个节点随后可由每个所接收的间隔指示118的间隔ID 120确定当前间隔。例如,节点(诸如协调器节点)发送间隔指示,诸如以间隔m开始。接收到间隔指示的每个节点针对该间隔执行I/O操作。协调器节点可以特定时间表(即,按特定间隔)发送这些间隔。间隔指示118的数据122还可包括其他信息,诸如以对先前间隔指示的确认作为响应的其他节点的列表。

[0030] 这些间隔的长度可针对每个分布式存储系统动态确定。在许多情况下,远程I/O写入延迟是限制因素,即,向远程存储器的节点写入与来自该远程存储器的该写入确认之间的实耗时间。在一个实施方式中,协调器节点104可例如通过使用管理模块110 (1) 确定间隔长度,诸如基于历史数据和/或分析。例如,如果完成针对节点110 (1) -110 (n) 的本地I/O需要不到1ms,并且完成针对这些节点的远程I/O平均需要3ms至10ms,则可将间隔设定为5ms。在该例子中,针对需要7ms的I/O(例如,通过节点102 (1) 进行的远程写入),如果该I/O写入在间隔m中开始,则其在间隔(m+1)中完成。当节点102 (1) 启动该远程I/O写入时,节点102 (1) -102 (n) 处于间隔m中。在该间隔m期间,没有其他节点写入到相同数据。接着,节点随后接收用于间隔(m+1)的间隔指示。另一个节点(例如节点102 (n))将直到节点102 (1) 完成远程I/O才尝试写入到相同或从属数据。

[0031] 在一个实施例中,协调器节点104向节点102 (1) -102 (n) 发送间隔通知。响应于接

收到这些间隔通知中的一者,每个节点102(1)-102(n)还可完成在前一个间隔中开始的I/O操作。在一个实施例中,间隔通知可表示间隔开始。在另一个实施例中,间隔通知可表示其他时间点,诸如间隔的结束、间隔的中间点等。例如,间隔指示118可在ID 120中包括关联的间隔编号。间隔编号可与间隔通知关联,诸如在时间0处的间隔通知可具有间隔编号1,下一个间隔通知可具有间隔编号2等。然而,在其他实施方式中,间隔指示可使用不同方法指示从一个间隔指示至下一个间隔指示的进展。

[0032] 在一个实施例中,节点102(1)接收第一间隔指示(例如从协调器节点104)。第一间隔指示指示第一间隔开始。在接收到第一间隔指示时,节点102(1)可启动本地写入到主存储器112。此时,节点102(1)还可启动远程写入到次级存储器114。在一个实施例中,远程写入可作为复制过程(例如复制写入到主存储器的数据的一些或全部)的一部分执行。远程写入和本地写入均可包括关联的时间标记。该时间标记可表示间隔通知(即,指示第一间隔通知)的间隔编号。应当注意,每个节点在每个间隔期间发起多个写入,并且这些写入与当前间隔关联。然而,复制模式控制用于节点的应用程序如何发起相关写入。

[0033] 在一个实施例中,当节点102(1)-102(n)接收并且确认下一个间隔通知(例如,第二间隔通知),假定本地写入和远程写入均已完成时,节点102(1)可确认完成来自第一间隔的I/O写入。在一个实施例中,可确定此类间隔的频率(例如,通过协调器节点104),使得在大多数情况下,在远程存储器114确认完成前一个间隔中的远程写入之前,将第二间隔指示发送至节点102(1)-102(n)(并且也被该节点接收和确认)。这会确保在等待下一个间隔指示之际不延迟应用程序I/O完成;换言之,即使引入间隔控制,应用程序I/O延迟也保持不变。在一个实施例中,远程存储器114可确认已完成远程写入到节点102(1)-102(n)中的任何一者和/或协调器节点104。在一个实施例中,在接收第二间隔指示之后,节点102(1)-102(n)中的任何一者将继续发起额外的I/O写入到本地存储器112和远程存储器114。与上面所述类似,直到节点102(1)-102(n)接收并且处理第三间隔指示并且这些额外的本地和远程写入已完成,节点102(1)-102(n)才(例如,分别向应用程序106(1)-106(n))确认这些额外的I/O写入完成。

[0034] 在一个实施例中,协调器节点104从节点102(1)-102(n)接收确认。通过节点102(1)-102(n)发送这些确认以确认接收到间隔指示。在一个实施例中,每个此类确认可以实施为图1的确认124。确认124可包括间隔ID 126和数据128,但也可设想其他实施方式。例如,每个节点102(1)-102(n)可向协调器节点104发送确认(例如,指示第一间隔的该确认124的ID 126),确认接收到第一间隔指示。协调器节点104随后可响应于从节点102(1)-102(n)接收到确认而发送下一个间隔指示(例如,指示第二间隔的该间隔指示118的ID 120),其中这些确认指示通过节点102(1)-102(n)接收到指示第一间隔的间隔指示。

[0035] 在一个实施例中,如果节点102(1)-102(n)中的一者不发回此类确认,则协调器节点104可延迟特定时间发送下一个间隔指示,此后协调器节点104将向节点102(1)-102(n)重新开始发送下一个间隔指示,同时保持跟踪此类过失节点错过的间隔。在下一个间隔中发送的数据可包括活动节点和任何过失节点的列表。可将该数据记录在主存储器和次级存储器中。一旦过失节点开始对间隔指示作出响应,过失节点就可处理任何累积和/或进行中的写入。在一个实施方式中,当过失节点重新开始对间隔指示作出响应时,此类累积写入可以使用指示时间的间隔进行全部处理。

[0036] 虽然本文主要参照从应用程序、文件系统、卷管理器和类似过程到存储设备的写入请求来描述间隔控制复制,但一般来讲,间隔控制复制可用于为任何类型的数据流提供一致的点或快照。例如,间隔控制复制可在分布式或群集环境中横跨日志集合、横跨存储对象集合或横跨消息流集提供一致性校验点或快照。另外,术语“I/O”、“访问”和/或“写入”在本文可互换使用以指诸如在间隔控制复制期间由源节点写入到复制目标的数据。术语“应用程序”用于指在源节点上执行并且将数据复制到复制目标的任何过程。因此,如本文所用的“应用程序”可指卷管理器、文件系统、数据库应用程序、存储复制器或最终用户应用程序以及复制数据的任何其他过程。

[0037] 图2A为根据一个实施例示出的用于间隔控制复制的方法200的流程图。如根据本发明将会理解,可修改该方法以便得到替代实施例。此外,此实施例中的步骤按照顺序显示。然而,特定步骤发生的顺序可与显示的顺序不同,特定步骤可同时执行,特定步骤可与其他步骤组合,并且特定步骤在另一个实施例中可能不存在。参照结合图1所述的元素的变型形式描述方法200。图2A示出了在同步或异步复制模式下操作时节点(诸如节点102(1)-102(N))的操作。如果适用,对这两种模式之间的操作差异进行说明。

[0038] 在元素202中,节点在第一间隔期间开始本地和远程I/O写入。例如,节点102(1)可接收第一间隔的指示(诸如从协调器节点104)。一旦节点102(1)接收到该指示,节点102(1)就开始本地写入到主存储器112以及远程写入到次级存储器114。远程写入可作为复制的一部分执行,并且可由节点102(1)的管理模块启动。可由管理模块108(1)启动本地和远程写入,作为应用程序106(1)执行I/O写入的一部分。

[0039] 在元素204中,在同步复制模式下操作时,节点等待下一个间隔、来自次级存储器的远程写入确认以及来自本地存储器的本地写入确认。例如,节点102(1)等待来自主存储器112和次级存储器114的本地和远程写入(在元素202中启动)确认以及等待下一个间隔指示。在一个实施方式中,节点102(1)还可等待分布式系统100中的其他节点102(n)-102(n)已移至该下一个间隔的确认。

[0040] 在元素204中,在异步复制模式下操作时,节点等待下一个间隔和来自本地存储器的本地写入确认。例如,节点102(1)等待来自主存储器112的本地写入(在元素202中启动)确认,以及等待下一个间隔指示。在一个实施方式中,节点102(1)还可等待分布式系统100中的其他节点102(n)-102(n)已移至该下一个间隔的确认。然而,在异步模式下操作时,节点不会等待来自次级存储器的远程写入确认。

[0041] 在元素206中,节点向应用程序确认I/O写入完成。例如,节点102(1)的管理模块108(1)向应用程序106(1)发送在元素202中启动的I/O写入已完成的确认。一旦应用程序106(1)接收到该确认,应用程序106(1)就可发起I/O写入,该写入依赖于元素202的I/O写入。

[0042] 图2B为根据一个实施例的示出用于使用间隔控制复制的周期性复制模式的方法225的流程图。如根据本发明将会理解,可修改该方法以便导出替代实施例。此外,此实施例中的步骤按照顺序显示。然而,特定步骤发生的顺序可与显示的顺序不同,特定步骤可同时执行,特定步骤可与其他步骤组合,并且特定步骤在另一个实施例中可能不存在。参照结合图1所述的元素的变型形式描述方法225。图2A示出了在周期性复制模式下操作时节点(诸如节点102(1)-102(N))的操作。

[0043] 在元素226中,节点在第一间隔期间开始本地写入并且累积远程I/O写入。例如,节点102(1)可接收第一间隔的指示(诸如从协调器节点104)。一旦节点102(1)接收到该指示,节点102(1)就均开始本地写入到主存储器112以及累积远程写入到次级存储器114。远程写入可作为复制的一部分执行,并且可由节点102(1)的管理模块启动。可由管理模块108(1)启动本地和远程写入作为应用程序106(1)执行I/O写入的一部分。因此,在一个实施例中,节点可执行前端记录,其中当执行本地I/O写入时将远程I/O写入收集到日志。在另一个实施例中,节点可执行后端记录,其中从主存储器读取数据并且随后将数据写入到日志。

[0044] 在元素228中,节点确定最近周期是否已结束。换句话说,节点确定其中启动最近远程写入(元素226)的周期是否已完成。在一个实施例中,每个节点可保持跟踪其自己的周期。在另一个实施例中,协调器节点可保持跟踪用于分布式存储系统的周期。作为间隔指示的一部分(例如,数据字段的部分),协调器节点还可向节点发送每个周期和/或周期完成的指示。例如,可将间隔设定为5ms,并且可将周期设定为15秒,使得每个周期可包括约3,000个间隔。

[0045] 在元素230中,节点等待下一个间隔和来自本地存储器的本地写入确认。例如,节点102(1)等待来自主存储器112的本地写入(在元素226中启动)的确认,以及等待下一个间隔指示。在一个实施方式中,节点102(1)还可等待分布式系统100中的其他节点102(n)-102(n)已移至该下一个间隔的确认。然而,在周期性模式下操作时,节点不会等待来自次级存储器的远程写入确认。

[0046] 在元素232中,节点启动累积的远程I/O写入。例如,节点102(1)的管理模块108(1)启动累积的远程I/O写入(例如,在上一个周期中累积)。在一个实施例中,节点获取主存储器的至少一部分的快照以生成累积的远程I/O写入。在一个实施例中,可由日志而非使用应用程序数据的快照累积I/O写入。累积的I/O写入随后被写入到次级存储器。在一个实施例中,如果在周期中包括用于间隔(m)的任何I/O写入,则也包括用于前一个间隔(m-1)的所有I/O写入。在元素232之后,节点执行元素228。

[0047] 应当注意,在周期性和异步复制模式下操作时,次级存储器仍确认远程写入。然而,在上述两种情况下,节点的管理模块在向其关联应用程序确认I/O写入完成之前不会等待这些远程写入确认。由于远程I/O写入具有关联的间隔信息(例如,指示其中启动了用于该远程I/O写入的应用程序写入的间隔),因此次级存储器可按正确顺序应用这些远程I/O写入,即在处理间隔为(m+1)的远程写入之前处理具有间隔(m)的远程写入。在一个实施例中,次级存储器可根据远程I/O写入的间隔排队和处理这些写入。通过根据间隔处理远程写入,次级存储器是一致的,即,不会针对从属数据形成写入冲突。

[0048] 在一个实施例中,当远程写入被接收但尚未完成时,诸如当远程写入在次级存储器处排队时,次级存储器可发送确认。在一个实施例中,当远程写入被实际写入到存储器时,该次级存储器发送确认。次级存储器还可针对正在存储但未完成的写入发送单独确认,并且当完成次级写入时发送另一个确认。

[0049] 在元素234中,节点向应用程序确认I/O写入完成。例如,节点102(1)的管理模块108(1)向应用程序106(1)发送在元素202中启动的I/O写入完成的确认。一旦应用程序106(1)接收到该确认,应用程序106(1)就可发起I/O写入,该写入依赖于元素202的I/O写入。

[0050] 图2C为根据一个实施例示出的用于间隔控制复制的方法250的另一个流程图。如

根据本发明将会理解,可修改该方法以便导出替代实施例。此外,此实施例中的步骤按照顺序显示。然而,特定步骤发生的顺序可与显示的顺序不同,特定步骤可同时执行,特定步骤可与其他步骤组合,并且特定步骤在另一个实施例中可能不存在。参照结合图1所述的元素的变型形式描述方法250。方法250是上面参照图2A所述的方法200的变型形式。图2C示出了在同步或异步复制模式下操作时节点(诸如节点102(1)-102(N))的操作。如果适用,对这两种模式操作之间的差异进行说明。

[0051] 在元素252中,确定是否接收到间隔指示。例如,节点102(1)确定是否从协调器节点104接收到间隔指示。所接收的间隔指示可包括间隔ID和数据。间隔ID可指示当前间隔。例如,节点102(1)接收用于间隔m的间隔指示。

[0052] 在元素254中,节点可将I/O写入与当前间隔关联。例如,一旦节点102(1)接收到间隔指示,节点102(1)就可将本地和远程I/O写入与间隔指示的间隔ID关联。如上所述,远程写入可作为复制的一部分执行,并且可由节点102(1)的管理模块启动。可由管理模块108(1)启动本地和远程写入作为应用程序106(1)执行I/O写入的一部分。根据需要,这两种写入可基本上同时执行。

[0053] 在元素256中,确定是否可执行批处理。例如,节点102(1)可以是复制大量数据(例如,从主存储器112至远程存储器114)。此类复制的一部分可具有延迟敏感性,例如,当复制事务日志时。在这种情况下,对于选择设备,节点在发回完成确认(例如,向协调器节点104)之前可能不等待间隔完成。

[0054] 如果确定(即在元素256中)执行批处理,则执行元素258。在一个实施方式中,节点102(1)可在未接收到针对所有未完成写入的确认的情况下执行许多I/O写入。例如,如果要复制100个数据元素,则可使用相同间隔ID(例如,100)写入这100个数据元素。在一个实施例中,在次级存储器中的任何其他I/O写入之前处理被分批的所有此类I/O,因为这些可为相同间隔中的相关I/O。管理模块还可检查正由节点发出的任何I/O是否不与正被分批的I/O重叠。

[0055] 在元素260中,在同步复制模式下操作时,如果确定节点102(1)未正在执行批处理,则节点等待下一个间隔以及来自次级存储器的远程写入确认。节点还可等待来自本地存储器的本地写入确认。在一个实施例中,元素260类似于方法200的元素204。例如,节点102(1)接收用于间隔(m+1)的间隔指示。应当注意,节点102(1)可接收下一个间隔指示并且按照任何顺序接收来自本地存储器112和来自远程存储器114的确认。

[0056] 在元素260中,在异步复制模式下操作时,如果确定节点102(1)未执行批处理,则节点等待下一个间隔以及来自本地存储器的本地写入确认。在一个实施例中,元素260类似于方法200的元素204。例如,节点102(1)接收用于间隔(m+1)的间隔指示。应当注意,节点102(1)可接收下一个间隔指示并且按照任何顺序接收来自本地存储器112的确认。然而,在异步模式下操作时,节点不会等待来自次级存储器的远程写入确认。

[0057] 在元素262中,节点102(1)确定是否存在网络和/或协调器节点错误。例如,如果自节点102(1)接收到前一个间隔指示起经过了特定时间周期,则节点102(1)可确定此类网络和/或协调器节点错误。在元素264中,节点102(1)可启动此类网络/协调器节点错误的处理。例如,节点102(1)可启动一过程,在该过程中另一个节点(例如,节点102(2))成为用于分布式存储系统100的协调器节点。一旦对网络/协调器节点的错误进行了处理,方法250就

返回到元素256。

[0058] 在元素266中,节点102(1)确定分布式网络中的其他节点是否处于下一个间隔中。在一个实施例中,如果节点102(1)接收到用于间隔(m+2)的间隔指示,则节点102(1)确定网络中的其他节点已移至下一个间隔(即,(m+1))。在另一个实施例中,如果节点102(1)接收到来自协调器节点的间隔确认指示,指示网络中的其他节点已移至下一个间隔(即,(m+1)),则节点102(1)确定网络中的其他节点已移至下一个间隔(即,(m+1))。在又一个实施例中,如果节点102(1)接收到下一个子间隔指示,则节点102(1)确定网络中的其他节点已移至下一个间隔(即,间隔(m+1))。

[0059] 在元素268中,节点向应用程序确认I/O写入完成。例如,节点102(1)的管理模块108(1)向应用程序106(1)发送在元素254中启动的I/O写入完成的确认。一旦应用程序106(1)接收到该确认,应用程序106(1)就可发起I/O写入,该写入依赖于元素254的I/O写入。换句话说,对于同步复制,直到满足元素260的条件并且其他节点处于下一个间隔中,节点102(1)才确认I/O写入(在元素254中开始)完成。

[0060] 图2D是根据一个实施例示出的用于间隔控制复制的方法275的另一个流程图。如根据本发明将会理解,可修改该方法以便导出替代实施例。此外,此实施例中的步骤按照顺序显示。然而,特定步骤发生的顺序可与显示的顺序不同,特定步骤可同时执行,特定步骤可与其他步骤组合,并且特定步骤在另一个实施例中可能不存在。参照结合图1所述的元素的变型形式描述方法275。方法275可由分布式存储系统中的每个节点执行以选择复制类型。

[0061] 在元素276中,根据一个实施例,节点可接收复制类型的选择。例如,节点或协调器节点中的一者或多者可接收该选择。该选择可在开始复制过程之前接收,或可在复制过程期间接收。在一个实施例中,用于分布式系统中的所有节点的管理模块(包括协调器节点)使用相同复制模式。因此,任何复制模式变化均在节点之间被达成一致。例如,协调器节点可向其他节点的管理模块发送复制模式通知。在一个实施例中,协调器节点在切换至另一个复制模式之前可等待来自其他节点的复制模式确认消息。在一个实施例中,元素276在开始复制之前执行。在另一个实施例中,元素276在复制期间执行,使得可动态更改复制模式(例如,通过用户输入和/或来自另一个节点/模块的输入)。

[0062] 在元素278中,确定所述选择是否指示同步复制。如果所述选择指示同步复制,则节点在同步模式(元素280)中操作。如果所述选择不指示同步复制,则在元素282中,确定所述选择是否指示异步复制。如果所述选择指示异步复制,则节点在异步(元素290)模式下操作。如果所述选择不指示异步复制,则节点在周期性模式(元素284)下操作。

[0063] 当在同步模式(元素280)下操作时,节点(和/或协调器节点)可计算(元素286)用于该节点的指标。在元素288中,基于这些指标,确定是否切换至异步复制。在一个实施例中,这些指标包括次级存储器在预定时间周期内是否可发送远程写入的确认。在一个实施例中,这些指标包括网络统计值、节点的响应和/或分布式存储系统的次级存储器的响应。如果确定指示应当切换至异步复制,则该节点可将其操作切换至异步复制(元素290)。例如,如果次级存储器的延迟太大,则分布式存储系统将在异步复制模式而非同步复制模式下更有效地操作。

[0064] 在异步模式(元素290)下操作时,节点(和/或协调器节点)可计算(元素292)用于

该节点的指标。在元素294中,基于这些指标,以及原始的同步模式(元素276),确定是否切换至同步复制。在一个实施例中,这些指标包括次级存储器在预定时间周期内是否可发送远程写入的确认。在一个实施例中,这些指标包括网络统计值、节点的响应和/或分布式存储系统的次级存储器的响应。例如,指标可指示足够多个(例如,全部)未完成远程写入可写入到次级存储器;可动态确定被写入的所述足够多个未完成远程I/O写入的阈值。如果确定指示应当切换至同步复制,则该节点可将其操作切换至同步复制(元素280)。例如,如果使用指标进行的确定指示这样将会更有效,并且如果原始选择针对同步模式,则可将模式切换至同步模式。

[0065] 图3是根据一个实施例的用于间隔控制复制中的协调器节点操作的方法的流程图。如根据本发明将会理解,可修改该方法以便导出替代实施例。此外,此实施例中的步骤按照顺序显示。然而,特定步骤发生的顺序可与显示的顺序不同,特定步骤可同时执行,特定步骤可与其他步骤组合,并且特定步骤在另一个实施例中可能不存在。参照结合图1所述的元素的变型形式描述方法300。图3示出了同步、异步或周期性复制中的协调器节点的操作。在一个实施例中,除非另外指明,否则协调器节点在这三种复制模式下以基本类似方式操作。另外,在一个实施例中,与协调器模块关联的应用程序(例如,应用程序106(3))还可发起使用本文所述的三种复制模式中的一者所复制的I/O写入。换句话讲,除了作为协调器节点操作之外,协调器节点还可作为其他节点中的一者操作。在一个实施例中,对于包括次级站点的较大存储系统,协调器节点也可向次级站点中的一个或多个节点和/或向次级站点中的对等协调器节点发送间隔信息。

[0066] 在元素302中,协调器节点根据时间间隔发送间隔指示。协调器节点104可使用管理模块110(1)保持跟踪时间间隔。例如,协调器节点104可在第一间隔期间向节点102(1)-102(n)发送指示时间间隔m的间隔指示。类似地,例如,协调器节点104可在第二间隔期间向节点102(1)-102(n)发送指示时间间隔(m+1)的间隔指示。

[0067] 在一个实施例中,当分布式存储系统在周期性复制模式下操作时,作为间隔指示的一部分(例如,数据字段的部分),或以单独消息的形式,协调器节点还可向节点发送每个周期的指示和/或周期完成的指示。例如,协调器节点可发送指示当前周期的周期指示符。协调器节点还可发送指示下一个周期已开始的周期变化指示符。

[0068] 在元素304中,协调器节点确定是否应当调整时间间隔。当时间间隔太大或太小时可作出此类确定。元素304的确定可为动态的,即,管理模块110(1)可例如基于历史数据分析确定时间间隔太短(或太长)。例如,如果均在相同间隔中开始和完成远程I/O(例如,远程I/O向启动该远程I/O的节点发回确认),则间隔可能太长。此类分析可使用统计值执行,并且管理模块110(1)可使用阈值确定是否有足够的远程I/O写入在相同间隔中开始和完成。例如,如果仅1%的此类远程I/O写入在相同间隔中开始和完成,则管理模块110(1)可确定间隔不需要调整。另一方面,如果此类远程I/O写入的90%在相同间隔中开始和完成,则管理模块110(1)可确定间隔确实需要调整。

[0069] 在元素306中,协调器节点调整时间间隔。例如,如果管理模块110(1)确定时间间隔太短,则管理模块110(1)根据所确定将时间间隔调整至更长间隔。类似地,如果管理模块110(1)确定时间间隔太长,则管理模块110(1)根据所确定将时间间隔调整至较短间隔。

[0070] 图4-7示出了间隔控制复制的各种实施例的总体时序图。图4-7参照结合图1所述

的元素的变型形式以及图2A-2D和3等等中所示的流程图的方法加以描述。节点1可由节点102(1)实施,并且节点2可由节点102(2)实施。为了简化说明,假定两个节点可基本上同时处理来自协调器节点的消息,并且因此基本上同时开始新的间隔。然而,在接收此类消息的节点之间可存在小而有限的延迟和/或偏差。

[0071] 图4A为根据一个实施例的通过实施间隔控制复制的分布式存储系统的两个节点进行的I/O写入的总体时序图400。协调器节点(例如,协调器节点104)可向节点1和2发送指示间隔1-5 402(1)-402(5)的间隔指示。具体地讲,间隔1在402(1)开始并且在402(2)结束,间隔2在402(2)开始并且在402(3)结束等。图4A示出了在同步、异步或周期性复制中的节点的操作。

[0072] 在一个实施例中,在接收到间隔指示1时,节点1和2均可更新其内部数据结构(例如,通过使用管理模块),指示间隔1已开始。节点1随后可开始第一I/O 404(1)。当I/O 404(1)对数据进行访问时,没有其他节点可执行I/O写入到重叠或从属数据。虽然每个间隔仅示出一个I/O,但两个节点均可在每个间隔中执行数百个独立的I/O。一旦节点1接收到下一个间隔的指示,例如,指示间隔2 404(2)已开始,节点1就可完成I/O 404(1)。在接收到该下一个间隔的指示时,节点2可开始I/O写入406(1),该写入406(1)可依赖于I/O写入404(1)。

[0073] 然而,节点1在第二间隔中完成I/O写入404(1)。在接收到间隔3的间隔指示时,节点2直到间隔3才完成其潜在的数据相关I/O写入406(1)。类似地,通过节点1进行的潜在数据相关I/O写入404(2)直到第三间隔才开始,并且直到第四间隔才完成。为了将节点处理间隔消息时的微小时间差考虑在内,直到所有节点已接收到下一个间隔指示才发生I/O完成。这可能需要节点和/或协调器节点之间的额外消息。

[0074] 在图4B中所示的一个实施例中,接收间隔指示可能被视为确认所有节点均接收到前一个间隔。图4B是根据一个实施例的通过实施间隔控制复制的分布式存储系统的两个节点进行的I/O写入的总体时序图450。协调器节点(例如,协调器节点104)可向节点1和2发送指示间隔1-5 502(1)-502(5)的间隔指示。图4B示出了在同步、异步或周期性复制中的节点的操作。

[0075] 在图4B所示的实施方式中,用于节点1的管理模块将确认在间隔‘m’中开始的任何I/O写入将在间隔‘m+2’中完成,因为此时所有节点均已确认接收到间隔‘m+1’。例如,节点1可在间隔1(即,间隔m)中启动I/O写入454(1)。直到节点1接收到用于间隔3(即,间隔(m+2))的间隔指示,才确认该I/O写入(即,节点1的管理模块不向启动该I/O写入的应用程序确认该写入)。在一个实施例中,当节点接收到新的间隔指示(诸如‘m+1’)时,该节点将针对未来的I/O写入使用这个新的间隔指示,即使其他节点尚未接收到这个新的间隔指示。

[0076] 图5是根据一个实施例的通过分布式存储系统的两个节点进行的I/O写入的总体时序图500。协调器节点(例如,协调器节点104)可向节点1和2发送指示间隔1-5 502(1)-502(5)的间隔指示。图5示出了延迟可如何产生针对节点1和2的I/O写入问题。图5示出了在同步、异步或周期性复制中的节点的操作。

[0077] 在一个实施例中,在接收到间隔指示1时,节点1和2均可更新其内部数据结构(例如,通过使用管理模块),指示间隔1已开始。节点1随后可开始第一I/O 504(1)。当I/O 504(1)对数据进行访问时,没有其他节点可执行I/O写入到相同数据。一旦节点1接收到下一个间隔的指示,例如,指示间隔2 404(2)已开始,节点1就可完成I/O 404(1)。然而,如果存在

网络或延迟问题,则节点2不会接收到间隔2已开始的间隔指示。因此,节点2随后可开始I/O 506 (1),该I/O 506 (1)依赖于节点1的I/O 504 (1)。这样可能存在问题,例如,因为I/O写入506 (1)依赖于I/O 504 (1),然而两者均在相同间隔内执行。

[0078] 图6是根据一个实施例的通过分布式存储系统的两个节点进行的I/O写入的总体时序图600。协调器节点(例如,协调器节点104)可向节点1和2发送指示间隔1-4 602 (1)-602 (4)的间隔指示。图6示出了间隔控制复制可如何用于克服I/O写入的可能问题的一个实施例(图5所示)。图6示出了在同步、异步或周期性复制中的节点的操作。

[0079] 在一个实施例中,在接收到间隔指示1时,节点1和2均可更新其内部数据结构(例如,通过使用管理模块),指示间隔1已开始。节点1随后可开始第一I/O 604 (1)。当I/O 604 (1)对数据进行访问时,没有其他节点可执行I/O写入到相同数据。一旦节点1接收到下一个间隔的指示,例如,指示间隔2 604 (2)已开始,节点1就可完成I/O 604 (1)。然而,如果存在网络或延迟问题,则节点2不会接收到间隔2已开始的间隔指示。延迟608示出了在节点1接收到第二间隔602 (2)的间隔指示之后节点2多久接收到此类指示。

[0080] 在一个实施例中,节点1和2均可向协调器节点发送确认,指示接收到最近的间隔指示。例如,节点1和2均可发送间隔确认,指示接收到用于间隔1 602 (1)的间隔指示。在接收到用于间隔2 602 (2)的间隔指示时,节点1可发送间隔确认,指示接收到该间隔指示。然而,节点2直到延迟608之后才接收到该第二间隔指示(例如,由于网络延迟或其他问题)。因此,节点2直到延迟608之后才发送接收到第二间隔指示的确认。在一个实施例中,协调器节点可发送额外的指示符以指示剩余节点(包括节点2)何时确认接收到第二间隔指示。节点1在接收到此类额外指示之后完成I/O 604 (1),其中I/O 604 (1)被延迟额外量610。因此,由节点2进行的相关I/O写入606 (1)不在与I/O写入604 (1)相同的间隔中开始。另外,协调器节点随后可偏移并且调整下一个间隔指示符的时间,使得节点1和2均(例如针对间隔4 602 (4))基本上同时接收到下一个间隔指示符。

[0081] 图7是根据一个实施例的由分布式存储系统的两个节点进行的I/O写入的总体时序图700。协调器节点(例如,协调器节点104)可向节点1和2发送指示间隔1-5 702 (1)-702 (5)的间隔指示。图7示出了当节点2未确认间隔指示时处理I/O写入的一个实施例。例如,如果存在网络/延迟问题,则节点2不会接收到间隔2已开始的间隔指示。节点2还可能太忙而不能确认接收到所述第二间隔指示。因此,协调器节点可针对延迟708延迟发送下一个间隔指示。图7示出了在同步复制中的节点的操作。

[0082] 在延迟708之后,协调器节点可向分布式系统中的节点恢复发送间隔指示。在一个实施例中,协调器节点可向节点发送额外的指示消息,指示哪些节点已确认接收到间隔指示和/或哪些节点(即,过失节点)还未确认接收到间隔指示。在另一个实施例中,协调器节点可向节点发送部分间隔指示消息,指示哪些节点已确认接收到间隔指示和/或过失节点。在延迟710之后节点1可完成I/O写入704 (1),然而,间隔可变化。

[0083] 在一个实施例中,对于同步复制,如果过失节点尝试启动写入到次级存储器,则此时无法正确识别间隔/将间隔与来自过失节点的任何此类进行中I/O写入关联。因此,此类I/O写入将直到过失节点恢复确认来自协调器节点的间隔指示才会包括在次级存储器中。在过失节点(例如,第二节点)没有对间隔指示作出响应期间,第二节点可向主存储器和/或次级存储器发送I/O写入,并且将这些I/O写入与最后为第二节点所知的间隔关联。

[0084] 在一个实施例中,一旦第二节点重新开始对间隔通知(例如,对新闻隔)作出响应,协调器节点就可向第二节点通知错过的间隔范围。第二节点随后可将任何未完成和进行中的I/O写入关联为属于从协调器节点所接收的新闻隔(例如,间隔100)。在一个实施例中,这可包括将已完成的到主存储器和/或次级存储器的I/O写入关联为属于该新闻隔。换句话说讲,自上一个确认间隔直至现在所处理的写入被处理为好像其在上一个所接收间隔中接收一样。这是可接受的,因为第二节点在不向协调器节点确认新闻隔指示的情况下可能无法确认应用程序。第二节点还可恢复I/O写入706(1)。

[0085] 图8-10是根据一个实施例示出的在间隔控制复制期间各个节点与存储设备之间进行通信的图。图8-10参照结合图1所述的元素的变型形式以及图2A-2D和3等等中所示的流程图的方法有所描述。在一些实施例中,节点向协调器节点发送消息以确认接收到间隔指示。在一个实施例中,当节点接收到指示间隔(m+2)的间隔指示时,假定在间隔(m+1)中活动(即,无过失)的所有节点确实向协调器节点确认接收到前一个间隔。在一个实施例中,每当活动成员资格改变时(例如,当一些节点未确认间隔时,或早前有过失但现在活动的节点),间隔指示(例如,数据122)便可指示。在一个实施例中,当间隔指示不包括有关其他节点的信息时,这可指示所有活动节点均确认了前一个间隔。除非另外指明,否则图9-10示出了在同步复制模式中的节点的操作。应当注意,图8-10示出了在相同单个复制模式中的节点的操作(例如,所有节点在同步复制模式中操作)。

[0086] 图8是根据一个实施例的在实施间隔控制复制的分布式存储系统的各种元素之间进行通信的时序图800。元素802(1)示出了第一节点(例如,节点1 102(1)),元素802(2)示出了第二节点(例如,节点2 102(2)),元素804(1)示出了协调器节点(例如,节点104),元素806示出了本地存储器(例如,主存储器112),并且元素808示出了远程存储器(例如,次级存储器114)。元素826示出了可启动各种通信的时间点。图8示出了在同步、异步或周期性复制中的节点的操作。

[0087] 在时间T1 826(1),协调器节点804向第一节点802(1)和第二节点802(2)发送指示第一间隔的间隔指示810(1)。在同步和异步复制模式中,在接收到间隔指示810(1)时,以及在第一间隔期间,第一节点802(1)开始第一I/O写入,即第一本地写入812(1)(到本地存储器806)和第一远程写入814(1)到远程存储器808。第一远程写入814(1)可包括指示第一间隔的间隔ID。在周期性复制模式中,第一节点802(1)相反可累积第一远程写入。在所有三种复制模式中,并且仍在第一间隔期间,本地存储器806还可向第一节点802(1)发送本地写入确认消息816(1)(确认完成第一本地写入)。第一节点802(1)向协调器节点804发送间隔确认消息816(1)以确认接收到第一间隔指示810(1)。类似地,第二节点802(2)向协调器节点804发送间隔确认消息818(1)以确认接收到第一间隔指示810(1)。

[0088] 在时间T2 826(2),协调器节点804向第一节点802(1)和第二节点802(2)发送指示第二间隔的间隔指示810(2)。在同步和异步复制模式中,并且在接收到间隔指示810(2)时,并且在第二间隔期间,第二节点802(2)开始第二本地写入812(2)(到本地存储器806)和第二远程写入814(2)到远程存储器808。在周期性复制模式中,第二节点802(2)可累积第二远程写入。在所有三种复制模式中,并且仍在第二间隔期间,本地存储器806可向第一节点802(1)发送本地写入确认消息820(2)(确认完成第二本地写入)。在同步复制中,在第二间隔期间,远程存储器808可向第一节点802(1)发送远程写入确认消息822(1)(确认完成第一远程

写入)。应当注意,在异步复制中,远程存储器808还可发送远程写入确认消息822(1),然而,第一节点802(1)的管理模块在确认完成第一写入到其关联应用程序之前不会等待接收该远程写入确认消息822(1)。

[0089] 应当注意,第二I/O写入(即第二本地写入812(2)和第二远程写入814(2))与第一I/O写入无关。第二远程写入814(2)可包括指示第二间隔的间隔ID。第一节点802(1)和第二节点802(2)分别向协调器节点804发送间隔确认消息816(2)和818(2)以确认接收到第二间隔指示810(2)。一旦写入一批累积的远程写入(即在周期结束时),远程存储器808就可发送确认。在周期性复制中,针对一个周期累积写入(在数据卷和/或日志中)并且在周期结束之后(即,已发送针对之前周期的数据之后)将所述写入发送至远程存储器。

[0090] 在时间T3 826(3),协调器节点804向第一节点802(1)和第二节点802(2)发送指示第三间隔的间隔指示810(3)。在同步复制中操作时,在第三间隔期间,第一节点802(1)可基于接收到来自本地和远程存储器的确认(确认完成第一本地和远程I/O写入并且所有节点均处于第二间隔中)来确认第一I/O写入完成。在异步和周期性复制中操作时,在第三间隔期间,第一节点802(1)可基于接收到来自本地存储器的确认(确认完成第一本地I/O写入并且所有节点均处于第二间隔中)来确认第一I/O写入完成。例如,在所有三种复制模式中,第一节点的管理节点可向第一节点的应用程序确认第一I/O写入完成。该应用程序可请求发起第三I/O写入。然而,由于第三I/O写入依赖于第一I/O写入(即,在由这些写入访问的数据中存在数据相关性),因此应用程序等待发起第三I/O写入,直至其接收到第一I/O写入被完成的确认为止。在接收到间隔指示810(3)时,并且在第三间隔期间,第一节点802(1)开始第三本地写入812(2)(到本地存储器806)和第三远程写入814(2)到远程存储器808。第三远程写入814(3)可包括指示第三间隔的间隔ID。

[0091] 图9是根据一个实施例的在实施间隔控制复制的分布式存储系统的各种元素之间进行通信的时序图900。元素902(1)示出第一节点(例如,节点1 102(1)),元素902(2)示出第二节点(例如,节点2 102(2)),元素904(1)示出协调器节点(例如,节点104),元素906示出本地存储器(例如,主存储器112),并且元素908示出远程存储器(例如,次级存储器114)。元素926示出可启动各种通信的时间点。图9示出了使用额外指示消息的间隔控制复制的实施例。节点(例如,节点902(1)和902(2))向协调器节点904发送消息以确认接收到间隔指示。作为响应,协调器节点904向节点(例如,节点902(1)和902(2))发送额外指示消息,指示哪些节点已确认接收到间隔指示。

[0092] 在时间T1 926(1),协调器节点904向第一节点902(1)和第二节点902(2)发送指示第一间隔的间隔指示910(1)。在接收到间隔指示910(1)时,并且在第一间隔期间,第一节点902(1)开始第一I/O写入,即,第一本地写入912(1)(到本地存储器906)和第一远程写入914(1)到远程存储器908。第一远程写入914(1)可包括指示第一间隔的间隔ID。第一节点902(1)和第二节点902(2)还分别向协调器节点904发送间隔确认916(1)和917(1),确认接收到间隔指示910(1)。作为响应,协调器节点904向节点902(1)和902(2)发送间隔确认指示消息918(1)。间隔确认指示消息918(1)可指示哪些节点已确认接收到间隔指示。仍在第一间隔期间,本地存储器906可向第一节点902(1)发送本地写入确认消息920(1),该消息确认完成第一本地写入。

[0093] 在时间T2 926(2),协调器节点904向第一节点902(1)和第二节点902(2)发送指示

第二间隔的间隔指示910(2)。第一节点902(1)向协调器节点发送间隔确认916(2),确认第一节点的902(1)接收到第二间隔指示910(2)。然而,第二节点902(2)直到延迟930之后,才向协调器节点904发送间隔确认,确认第二节点的902(2)接收到第二间隔指示910(2)。因此,协调器节点904等待(例如,至少延迟930的持续时间)直到在第二节点902(2)将间隔确认917(2)发送至协调器节点904之后才向节点902(1)和902(2)发送间隔确认指示消息918(2)。在该例子中,由于第二I/O写入与第一I/O写入相关(即,在由这些写入访问的数据中存在数据相关性),因此第二节点902(2)等待直到从协调器节点904接收到间隔确认指示消息918(2)才启动第二本地写入912(2)和第二远程写入914(2)。第二远程写入914(2)可包括指示第二间隔的间隔ID。

[0094] 应当注意,在元素916(2)与917(2)之间可进行多个I/O写入,但为了简单起见在图9中未示出。例如,在元素916(2)之后,来自第一节点902(1)的这些I/O写入使用第二间隔,但在第二节点902(2)上这些I/O写入可使用第一或第二间隔(根据第二节点是否已处理910(2))。在一个实施例中,在元素917(2)之后,第二节点902(2)即使在接收到918(2)之前也将使用第二间隔。

[0095] 仍在第二间隔期间,本地存储器906可将本地写入确认消息920(2)发送至第二节点902(2)。在第二间隔期间,远程存储器908可向第一节点902(1)发送远程写入确认消息924(1)(确认完成第一远程写入)。在时间T3 926(3),协调器节点904将间隔指示910(3)发送至节点1 902(1)和节点2902(2)。

[0096] 图10是根据一个实施例的在实施间隔控制复制的分布式存储系统的各种元素之间进行通信的时序图1000。元素1002(1)示出第一节点(例如,节点1 102(1)),元素1002(2)示出第二节点(例如,节点2 102(2)),元素1004(1)示出协调器节点(例如,节点104),元素1006示出本地存储器(例如,主存储器112),并且元素1008示出远程存储器(例如,次级存储器114)。元素1026示出可启动各种通信的时间点。

[0097] 图10示出了使用子间隔指示和确认消息的间隔控制复制的实施例。在该具体实施中,不发送两条消息(间隔指示和间隔确认指示),而是将间隔分成两半。在原本为间隔的中间处,可发送新的部分间隔指示消息,该消息还用作前一个间隔的间隔确认指示。利用这种方法,即使写入完成等待两个间隔,有效消逝持续时间仍与在图9的实施方式中相同。

[0098] 节点(例如,节点1002(1)和1002(2))向协调器节点1004发送消息以确认接收到间隔指示。作为响应,协调器节点1004在部分间隔时间(例如,原始间隔时间的中点)向节点发送部分间隔指示消息以指示哪些节点已确认接收到前一个间隔指示。在一个实施例中,节点1002(1)和1002(2)确认接收到所有间隔指示。

[0099] 在时间T1 1026(1),协调器节点1004向第一节点1002(1)和第二节点1002(2)发送指示第一间隔的间隔指示1010(1)。在接收到间隔指示1010(1)时,并且在第一间隔期间,第一节点1002(1)开始第一本地写入1012(1)(到本地存储器1006)和第一远程写入1014(1)到远程存储器1008。第一节点1002(1)和第二节点1002(2)还分别向协调器节点1004发送间隔确认1016(1)和1017(1),确认接收到间隔指示1010(1)。仍在第一间隔期间,本地存储器1006可向第一节点1002(1)发送本地写入确认消息1020(1),该消息确认完成第一本地写入。响应于接收到间隔确认1016(1)和1017(1),协调器节点1004在时间T1.5 1026(2)向节点1002(1)和1002(2)发送部分间隔确认指示消息1010(2),该时间可基本上为正使用的间

隔的约0.5。每个部分间隔确认指示消息还用作间隔确认指示消息并且可指示哪些节点已确认接收到前一个间隔指示。在一个实施例中,第一节点1002(1)和第二节点1002(2)还分别向协调器节点1004发送间隔确认1016(2)和1017(2),确认接收到间隔指示1010(2)。

[0100] 在时间T2 1026(3),协调器节点1004向第一节点1002(1)和第二节点1002(2)发送指示第二间隔的间隔指示1010(2)。第一节点1002(1)向协调器节点1004发送间隔确认1016(3),确认第一节点1002(1)接收到第二间隔指示1010(3)。然而,第二节点1002(2)直到在延迟1030之后才向协调器节点1004发送间隔确认,确认第二节点的1002(2)接收到第二间隔指示1010(2)。因此,协调器节点1004等待(例如,至少延迟1030的持续时间)直到在第二节点1002(2)将间隔确认1017(4)发送至协调器节点1004之后才向节点1002(1)和1002(2)发送下一个间隔指示1018(4)。

[0101] 在该例子中,由于第二I/O写入与第一I/O写入相关(即,在由这些写入访问的数据中存在数据相关性),第二节点1002(2)等待直到从协调器节点1004接收到下一个间隔指示1018(4)才启动第二本地写入1012(2)和第二远程写入1014(2)。第二远程写入1014(2)可包括指示第二间隔的间隔ID。仍在第二间隔期间,本地存储器1006可向第二节点1002(2)发送本地写入确认消息1020(2)。在第二间隔期间,远程存储器1008可向第一节点1002(1)发送远程写入确认消息1024(1)(确认完成第一远程写入)。在时间T3 1026(5),协调器节点1004向第一节点1002(1)和第二节点1002(2)发送间隔指示1010(5)。在一个实施例中,可将部分间隔视为定期间隔,正因如此,将在第三间隔中确认在第一间隔中开始的应用程序写入(确保所有节点已移至第二间隔)。将部分间隔视为定期间隔可增加间隔粒度(即,使得间隔较小)而不添加额外消息。

[0102] 图11是根据一个实施例的示出用于间隔控制复制中协调器节点的操作的方法的流程图。如根据本发明将会理解,可修改该方法以便导出替代实施例。此外,此实施例中的步骤按照顺序显示。然而,特定步骤发生的顺序可与显示的顺序不同,特定步骤可同时执行,特定步骤可与其他步骤组合,并且特定步骤在另一个实施例中可能不存在。参照结合图1所述的元素的变型形式描述方法1100。方法1100是上文参照图3所述的方法300的变型形式。图11示出在同步、异步或周期性复制中的协调器节点的操作。在一个实施例中,除非另外指明,否则协调器节点在这三种复制模式中以基本上类似方式操作。

[0103] 在元素1102中,协调器节点根据时间间隔发送间隔指示。协调器节点104可使用管理模块110(1)保持跟踪时间间隔。例如,管理模块可在第一间隔期间向节点102(1)-102(n)发送指示时间间隔m的间隔指示。类似地,例如,管理模块可在第二间隔期间向节点102(1)-102(n)发送指示时间间隔(m+1)的间隔指示。

[0104] 在元素1104中,协调器节点确定是否从基本上所有节点接收到(接收到前一个间隔指示的)确认。在一个实施例中,协调器节点确定是否从分布式存储系统中的所有节点接收到确认。在另一个实施例中,协调器节点确定是否仅从分布式存储系统中的预期节点子集接收到确认。协调器节点可确定预期提供哪些节点通过确认回应。如果从分布式I/O系统的所有预期节点(即所有节点或预期节点子集)接收确认,则方法1100执行元素1105。

[0105] 在元素1105中,执行时间间隔检查。在一个实施例中,元素1105包括方法300的元素304和306。同样,如果需要时间间隔调整,则对时间间隔进行调整。如果不需要时间间隔调整,则不对时间间隔进行调整。在一个实施例中,过失节点(即,预期提供确认但未提供确

认的节点)用在元素1106中指定的延迟阈值内的此类确认作出响应,随后元素1105可因此偏移和/或调整时间间隔(例如,上面结合图6所述的情况)。方法1100随后执行元素1102。

[0106] 在元素1106中,如果协调器节点未从预期节点接收确认,则协调器节点确定在接收这些确认方面的延迟是否长于阈值。在一个实施例中,可在执行方法1100之前确定该阈值的值。在另一个实施例中,可例如在执行方法1100期间动态确定该阈值的值。例如,可通过分析节点之间通信速度的管理模块来计算该阈值的值。如果延迟不大于阈值,则方法1100执行元素1108,诸如其中协调器模块在再次执行元素1104之前等待一定量的时间。接着,方法1100进行到元素1114,其中协调器节点使用下一个时间间隔发送下一个间隔指示。

[0107] 在一个实施例中,在元素1102和/或1114中,当分布式存储系统在周期性复制模式中操作时,作为间隔指示的一部分(例如,数据字段的部分)或以单独消息的形式,协调器节点还可向节点发送每个周期和/或周期完成的指示。例如,协调器节点可发送指示当前周期的周期指示符。协调器节点还可发送指示下一个周期已开始的周期变化指示符。

[0108] 在元素1116中,协调器节点确定是否从基本上所有节点(包括过失节点)接收到(接收到元素1114的前一个间隔指示的)确认。在一个实施例中,协调器节点确定是否从分布式存储系统中的所有节点接收到确认。在另一个实施例中,协调器节点确定是否仅从分布式存储系统中的预期节点子集接收到确认。协调器节点可确定哪些节点需要提供确认回应。如果从分布式I/O系统中的所有预期节点(即所有节点或预期节点子集,并且包括过失节点)接收到确认,则方法1100执行元素1102。如果未从分布式I/O系统中的所有预期节点中接收到确认,则方法1100再次执行元素1114。协调器节点还可与其他节点进行信息通信,该信息指示过失节点可处理I/O写入。在一个实施例中,由于任何过失节点可能未接收到早前的消息,因此协调器节点保持向过失节点发送新消息,该消息指示这些过失节点需要特殊处理。

[0109] 在一个实施例中,一旦过失节点开始从协调器接收间隔指示,过失节点就将在确认这些间隔指示之前采取纠正措施。如果该纠正措施过程较长,则在过失节点能够确认间隔指示之前可有更多一些间隔。一旦过失节点确认间隔指示,就可将该节点视为具有良好记录的节点。在另一个实施例中,过失节点可向协调器节点确认其处于纠正措施过程中,并且在后台进行纠正措施处理。在一个实施例中,如上面参照图7所述,该纠正措施过程包括过失节点可将任何未完成和进行中的I/O写入关联为属于从协调器节点接收的新闻隔。

[0110] 图12为节点1200(诸如图1中所述的节点102(1)-102(n))的框图。节点1200包括应用程序1202、管理模块1204、文件系统1206、内存1210和/或一个或多个处理器1212。应当注意,在一些实施例中,可将这些元素中的一者或多者结合。例如,内存1210可包括应用程序1202、管理模块1204和/或文件系统1206中的一者或多者。另外应当注意,管理模块1204可实施为软件和/或硬件模块。另外应当注意,在一些实施例中,可不使用节点1200的元素中的一者或多者。处理器1212可执行应用程序1202、管理模块1204和/或文件系统1206中的一者或多者。管理模块1204可实施图1的管理模块108(1)-108(n)。管理模块1204可实施方法200和/或250的至少部分。

[0111] 图13为协调器节点1300(诸如图1中所述的协调器节点104)的框图。协调器节点1300包括应用程序1302、管理模块1304、文件系统1306、内存1310和/或一个或多个处理器1312。应当注意,在一些实施例中,可将这些元素中的一者或多者结合。例如,内存1310可包

括应用程序1302、管理模块1304和/或文件系统1306中的一者或多者。另外应当注意,管理模块1304可实施为软件和/或硬件模块。另外应当注意,在一些实施例中,可不使用节点1300的元素中的一者或多者。处理器1312可执行应用程序1302、管理模块1304和/或文件系统1306中的一者或多者。管理模块1304可实施图1的管理模块110(1)。管理模块1304可实施方法300和/或1100的至少部分。

[0112] 可以使用不同的计算机系统和网络来实施网络体系结构的元素。下面参照图14描述了一个此类网络环境的例子。图14是示出网络体系结构1400的简化框图,网络体系结构1400中提供一个或多个客户端,其通过各种网络连接来访问服务器。如图14中所示,客户端1402(1)-(N)耦合到网络1410,因此可以通过网络1410访问服务器1406(此服务器可以用于实现图1、12和/或13的节点)。而其他服务器(未示出)则可用于实施图1、12和/或13的系统节点。可以使用,例如,台式计算机、膝上型计算机、工作站、服务器、蜂窝电话、智能手机、有网络功能的个人数字助理(PDA)或类似设备来实施客户端。可供客户端1402(1)-(N)用于访问服务器1406的网络1410的例子是互联网。或者,可以通过采用以太网、IEEE 802.11x或一些其他通信协议的局域网(LAN)提供对服务器1406的访问。如将会理解的,可通过将客户端直接耦合到服务器1406来对其进行访问(未示出)。

[0113] 另外如图14所示,服务器1406耦合至服务器存储设备1408,该存储设备1408包括数据卷,诸如主存储器112和/或次级存储器114。服务器存储设备1408可实施为单个存储设备或存储设备的集合。服务器存储设备1408还可实施为存储区域网络,存储区域网络将会使远程存储设备耦合到服务器(例如服务器1406),例如,使得远程存储设备对于服务器操作系统呈现为本地连接的存储设备。

[0114] 根据本发明,本领域中的技术人员将会知道,服务器存储设备1408可通过任何类型的计算机可读存储介质实施,所述计算机可读存储介质包括但不限于内部或外部硬盘驱动器(HDD)、光驱(例如CD-R、CD-RW、DVD-R、DVD-RW等)、闪存驱动器(例如USB记忆棒等)、磁带驱动器等。或者,本领域中的技术人员也将会知道,根据本发明,网络体系结构1400可以包括其他组件,例如路由器、防火墙等与当前网络的讨论没有密切关系并且不会在本文中进一步讨论的组件。本领域中的技术人员还将了解,其他配置也是可行的。例如,客户端1402(1)-(N)可以不使用服务器或互联网而直接耦合到服务器存储设备1408;服务器1406可用于实施客户端和服务;可以在不使用客户端1402(1)-(N)的情况下实施网络体系结构1400;等等。

[0115] 作为网络体系结构1400、服务器1406的示例性具体实施,对由客户端1402(1)-(N)生成的数据的服务请求将会指向存储在服务器存储设备1408中的数据。可用图1、12和/或13中所示的方式使用其他服务器中的一者实施节点、代理和/或管理模块的功能中的任何一者。

[0116] 图15示出了适用于实施本发明的计算机系统1510的框图。计算机系统1510可示例性说明图1的联网系统中的各种计算机系统,诸如节点和/或协调器节点等等。计算机系统1510包括用于互连计算机系统1510的主要子系统(诸如中央处理器1514)的总线1512、系统内存1517(通常为RAM,但也可以包括ROM、闪存RAM等)、输入/输出控制器1518、外部音频设备(诸如经由音频输出接口1522的扬声器系统1520)、外部设备(诸如经由显示适配器1526的显示屏1524)、串行端口1528和1530、键盘1532(与键盘控制器1533连接)、存储接口1534、

可操作用于接收软盘1538的软盘驱动器1537、可操作用于与光纤通道网络1590连接的主机总线适配器(HBA)接口卡1535A、操作用于与SCSI总线1539连接的主机总线适配器(HBA)接口卡1535B以及可操作用于接收光盘1542的光盘驱动器1540。此外,还包括鼠标1546(或其他经由串行端口1528耦合到总线1512的点击设备)、调制解调器1547(经由串行端口1530耦合到总线1512)以及网络接口1548(直接耦合到总线1512)。

[0117] 总线1512允许在中央处理器1514和系统内存1517之间进行数据通信,如前文提及,系统内存1517可能包括只读存储器(ROM)或闪存存储器(均未示出)以及随机存取存储器(未示出)。RAM是将操作系统和应用程序装载到其中的主内存。除了其他代码,ROM或闪存存储器可以包括控制基本硬件操作(例如与外围组件的交互)的基本输入输出系统(BIOS)。与计算机系统1510驻存在一起的应用程序通常存储在计算机可读介质上并经由该计算机可读介质进行访问,计算机可读介质例如硬盘驱动器(例如固定磁盘1544)、光盘驱动器(例如光盘驱动器1540)、软盘单元1537或其他存储介质。此外,通过网络调制解调器1547或接口1548进行访问时,可以根据应用程序和数据通信技术以电子信号的形式对应用程序进行调制。

[0118] 存储接口1534与计算机系统1510的其他存储接口一样,可以连接到标准的计算机可读介质以存储和/或检索信息,例如固定磁盘驱动器1544。固定磁盘驱动器1544可以是计算机系统1510的一部分,或者可以是独立的,可以通过其他接口系统进行访问。调制解调器1547可提供经由电话链路至远程服务器的直接连接,或经由互联网服务提供方(ISP)至互联网的直接连接。网络接口1548可提供经由直接网络链路至远程服务器的直接连接,或经由POP(入网点)至互联网的直接连接。网络接口1548可提供使用无线技术的此类连接,包括数字蜂窝电话连接、蜂窝数字分组数据(CDPD)连接、数字卫星数据连接等。

[0119] 很多其他设备或子系统(未示出)可以通过相似的方式进行连接(例如文档扫描仪、数码相机等)。相反,不需要提供图15中显示的所有设备即可操作本发明。可以使用与图15中所示方法不同的方法来互连设备和子系统。计算机系统的操作(例如图15中所示的操作)是本领域中所熟知的,在本申请中不进行详细讨论。为了实施本发明,用于基于此类应用程序对其他应用程序的依赖性而对其自动执行操作的代码(例如上面参照图2A-2D、3和11的方法所述)可以存储在计算机可读存储介质中,例如系统内存1517、固定磁盘1544、光盘1542或软盘1538中的一者或多者。在处理器1510执行指令期间,内存1520还可以用于存储临时变量或其他中间信息。计算机系统1510上提供的操作系统可以是MS-DOS[®]、MS-WINDOWS[®]、OS/2[®]、UNIX[®]、Linux[®]或其他已知的操作系统。

[0120] 此外,关于本文所述的信号,本领域中的技术人员将会认识到,可以将信号从第一块直接传输到第二块,或者可以在块之间修改信号(例如,放大、衰减、延迟、锁存、缓冲、反转、过滤或其他修改方式)。尽管上述实施例的信号被特性化为从一个块传输到下一个块,但本发明的其他实施例可以包括经过修改的信号来代替这些直接传输的信号,只要信号的信息和/或功能方面在块之间传输。在某种程度上,由于所涉及的电路的物理限制(例如,不可避免地存在一些衰减和延迟),第二块上的信号输入可以概念化为根据来自第一块的第一信号输出派生得到的第二信号。因此,正如本文所用,根据第一信号派生的第二信号包括第一信号或对第一信号所做的任何修改,无论是由于电路限制还是由于通过不会改变第一

信号的信息和/或最终功能方面的其他电路元件。

[0121] 尽管已结合多个实施例描述了本发明,但并非旨在将本发明限制为本文中示出的具体形式。相反,本发明旨在涵盖可以合理地包含在所附权利要求所限定的本发明范围内的替代形式、修改形式和等同形式。

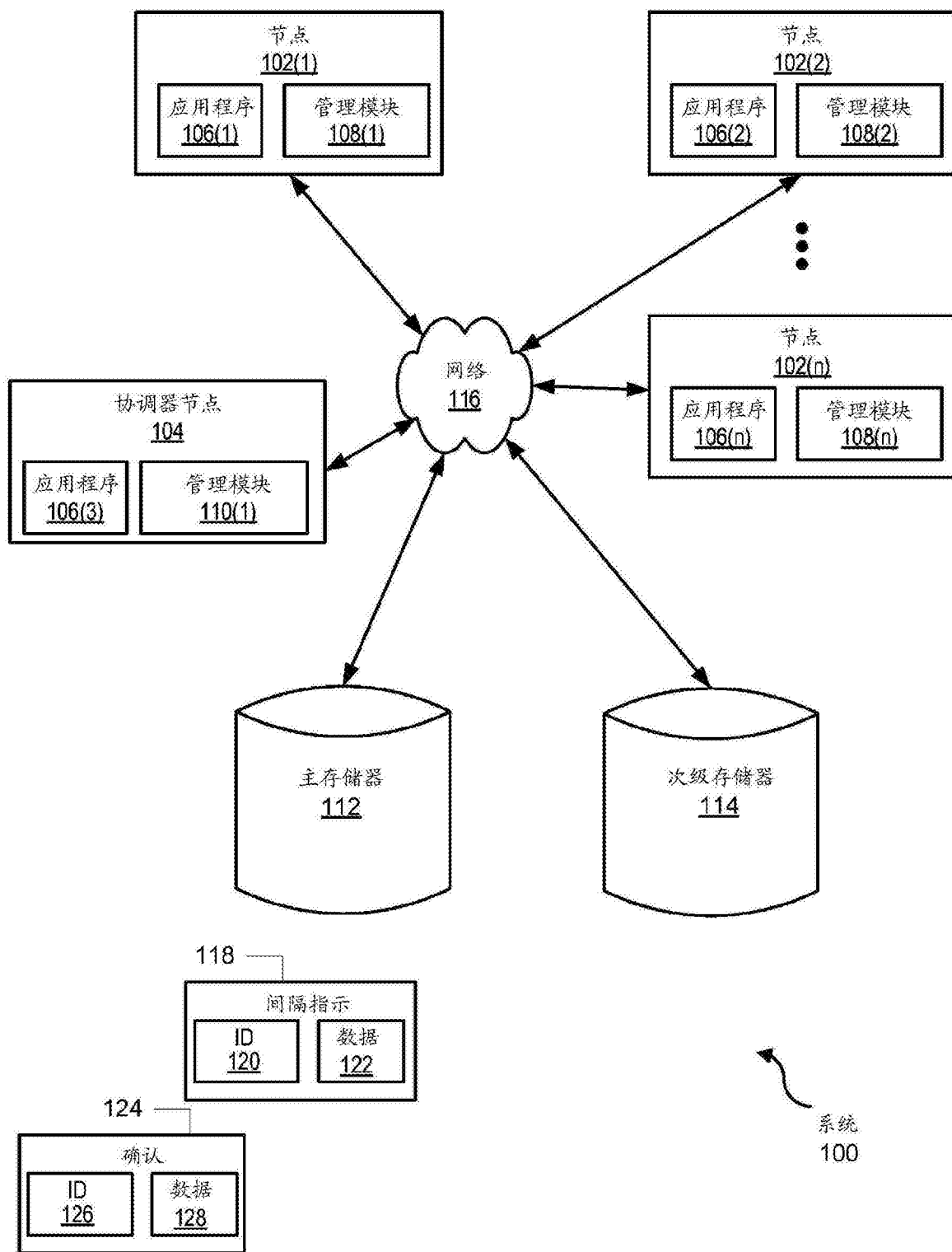


图1

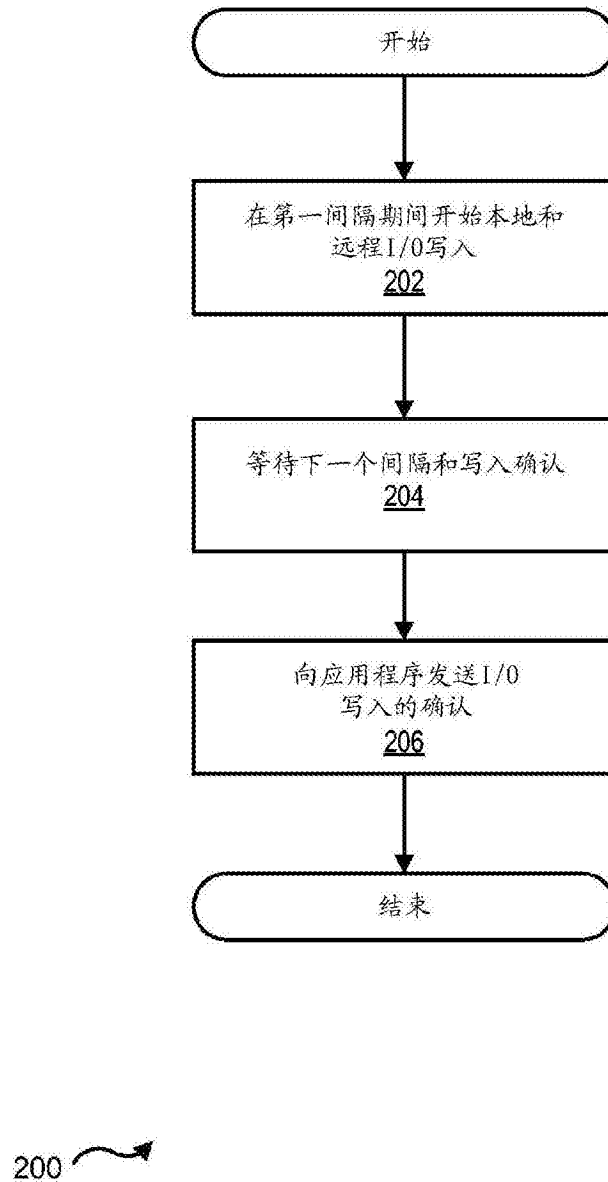
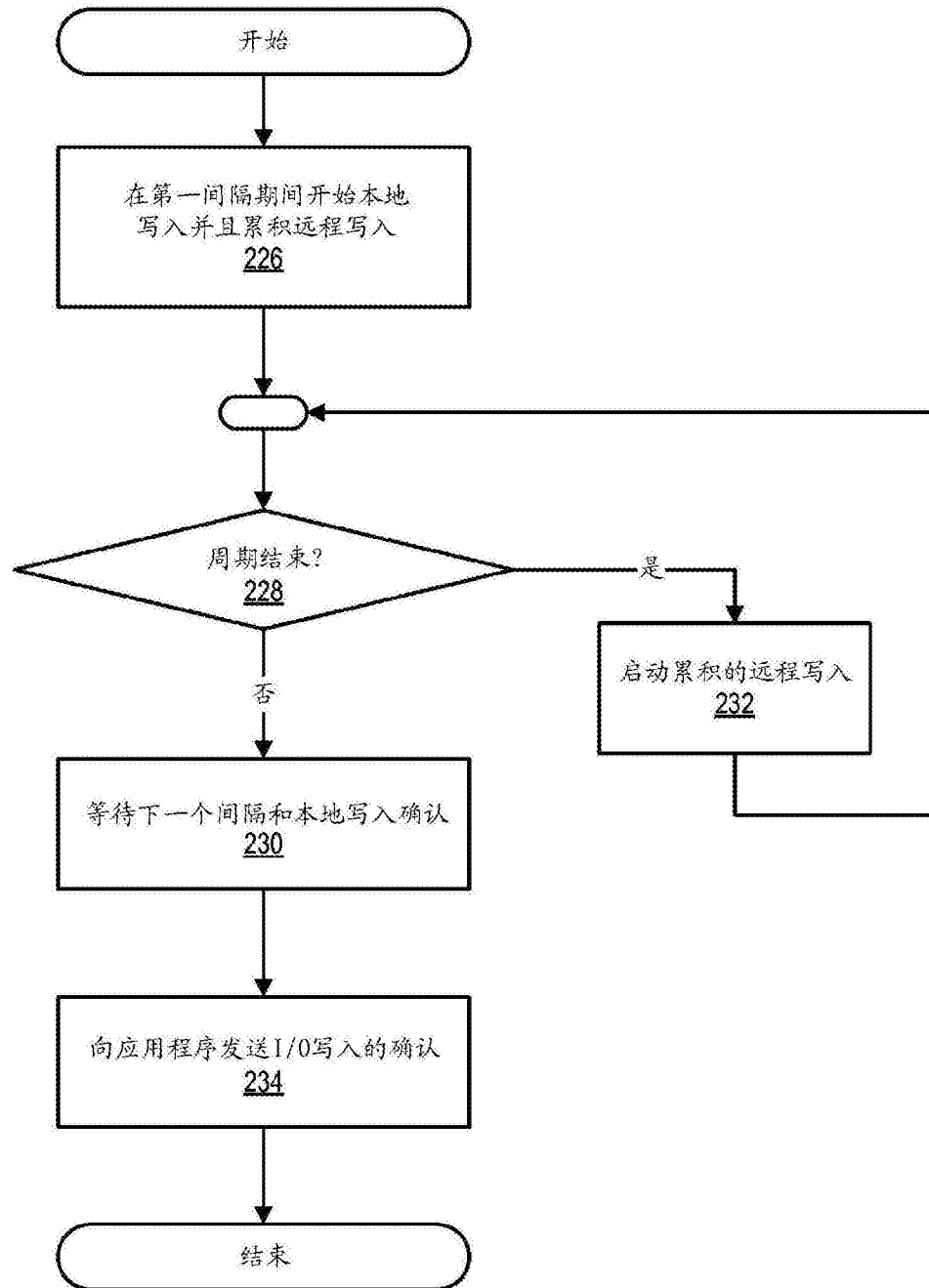


图2A



225

图2B

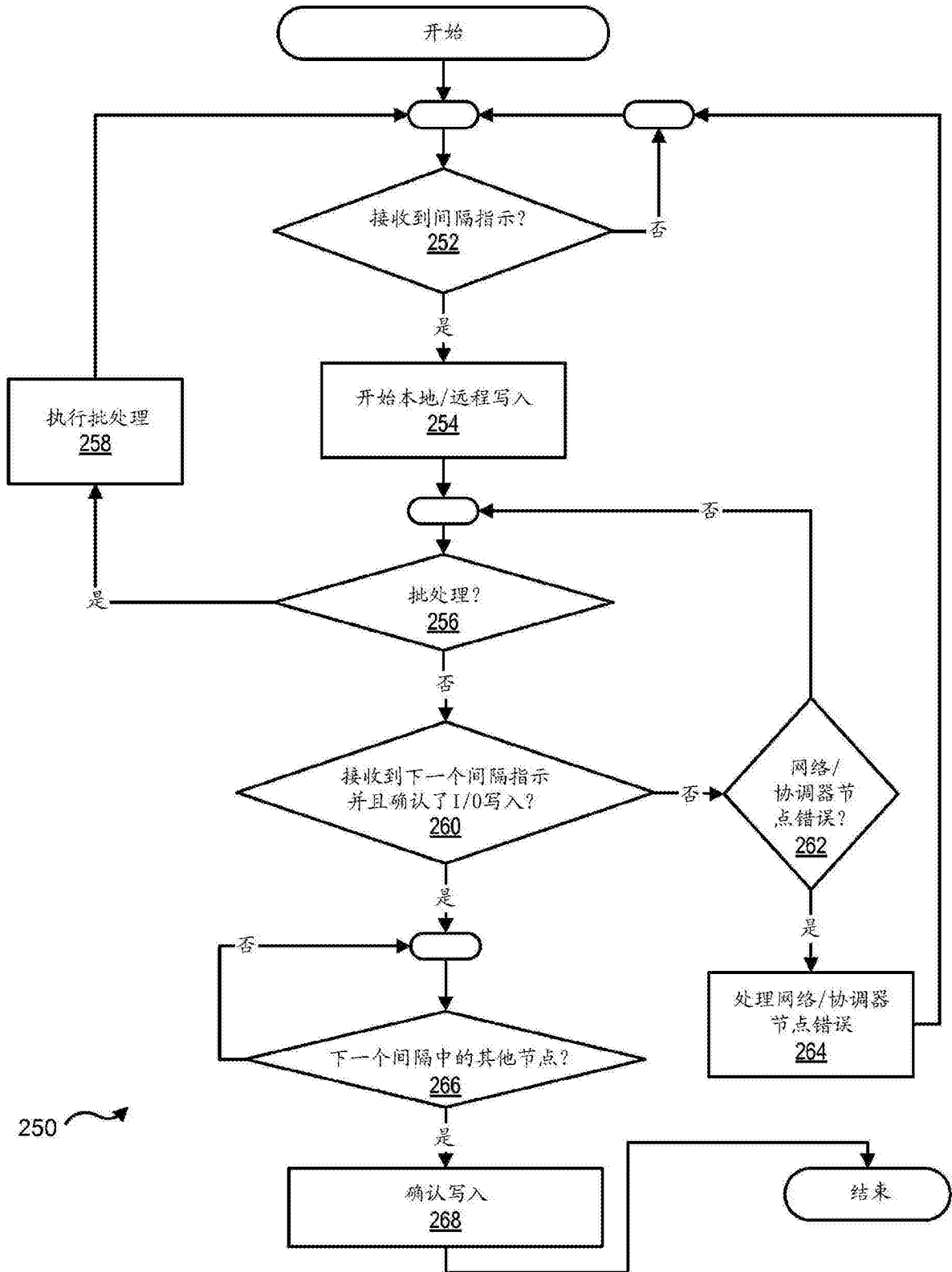


图2C

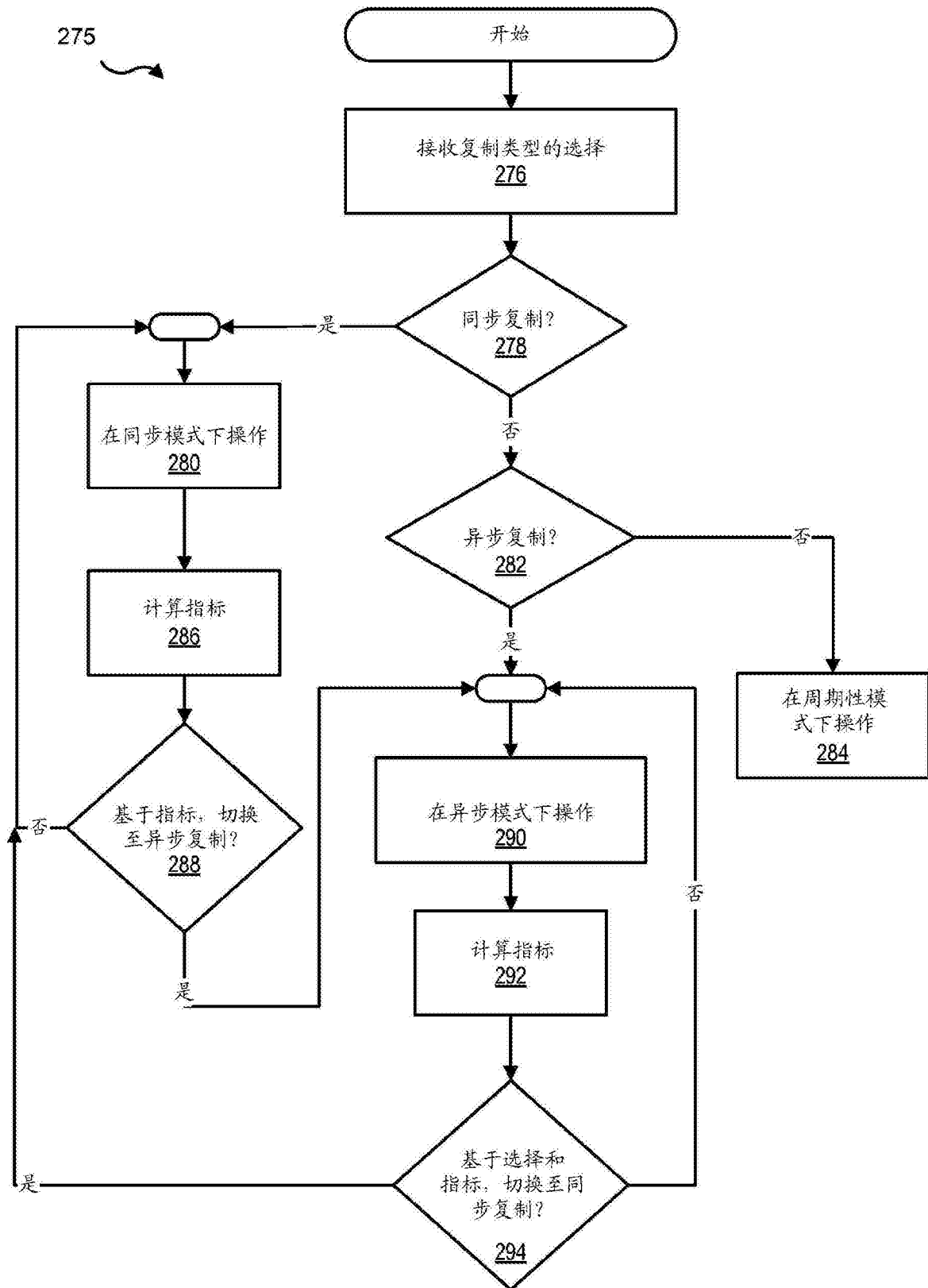
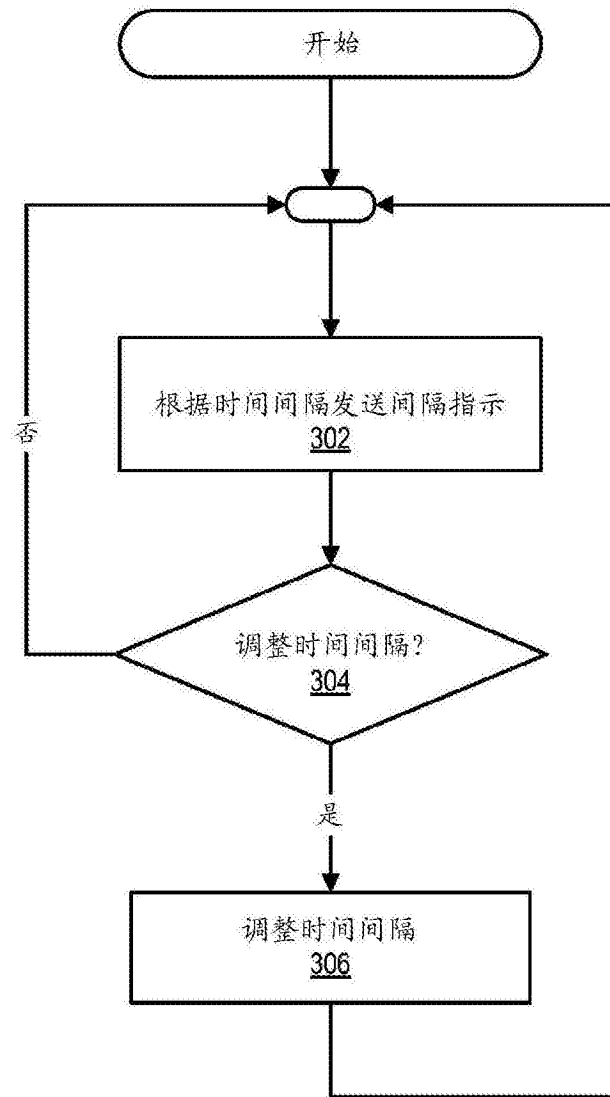


图2D



300 ↗

图3

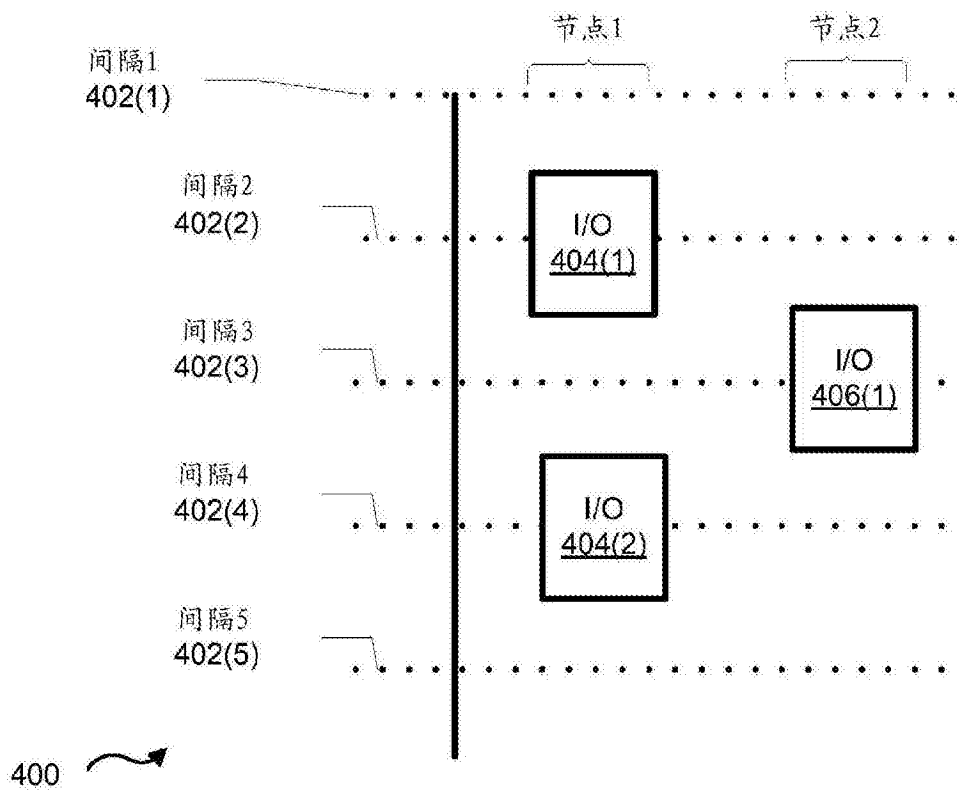


图4A

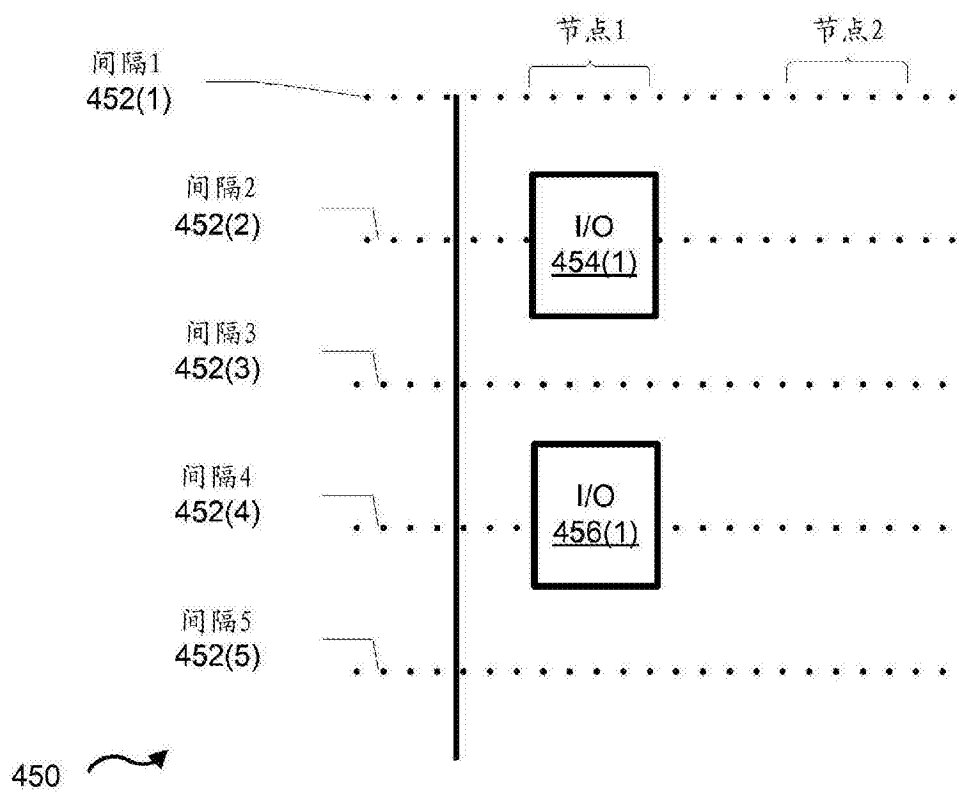


图4B

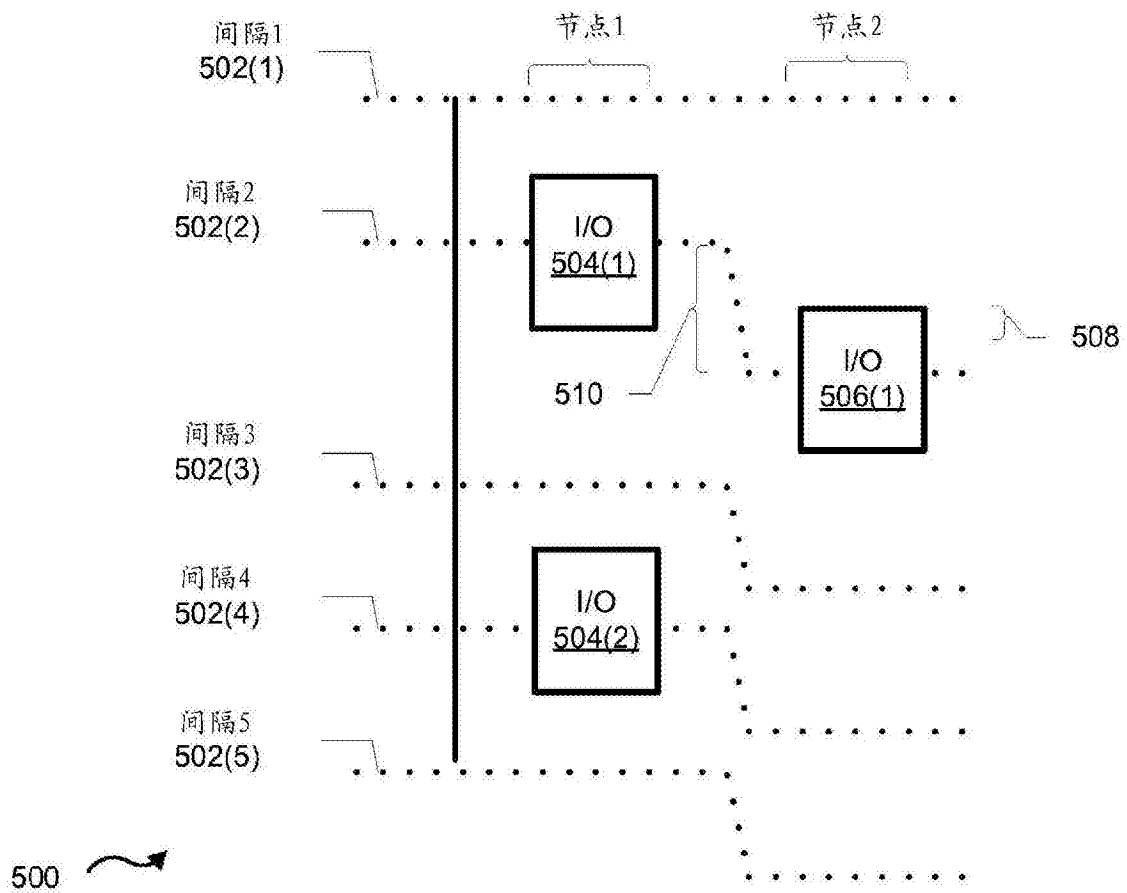


图5

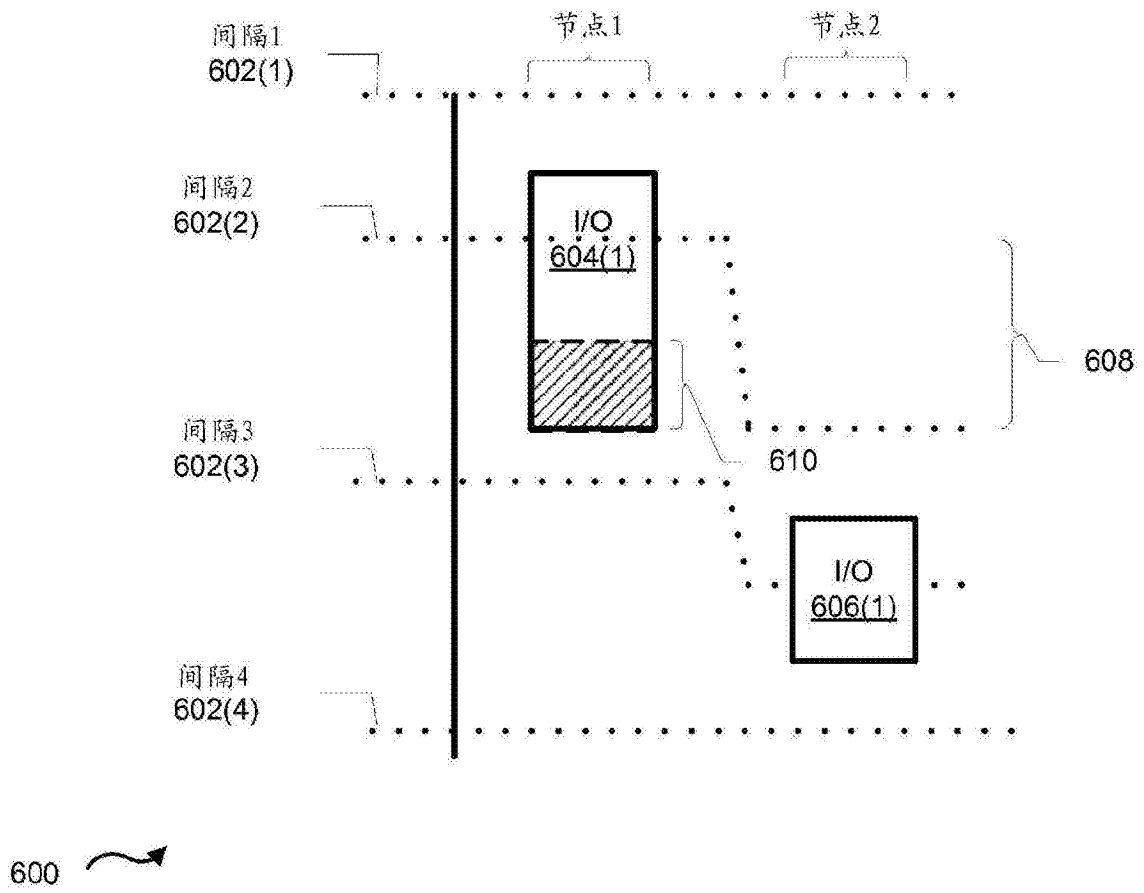


图6

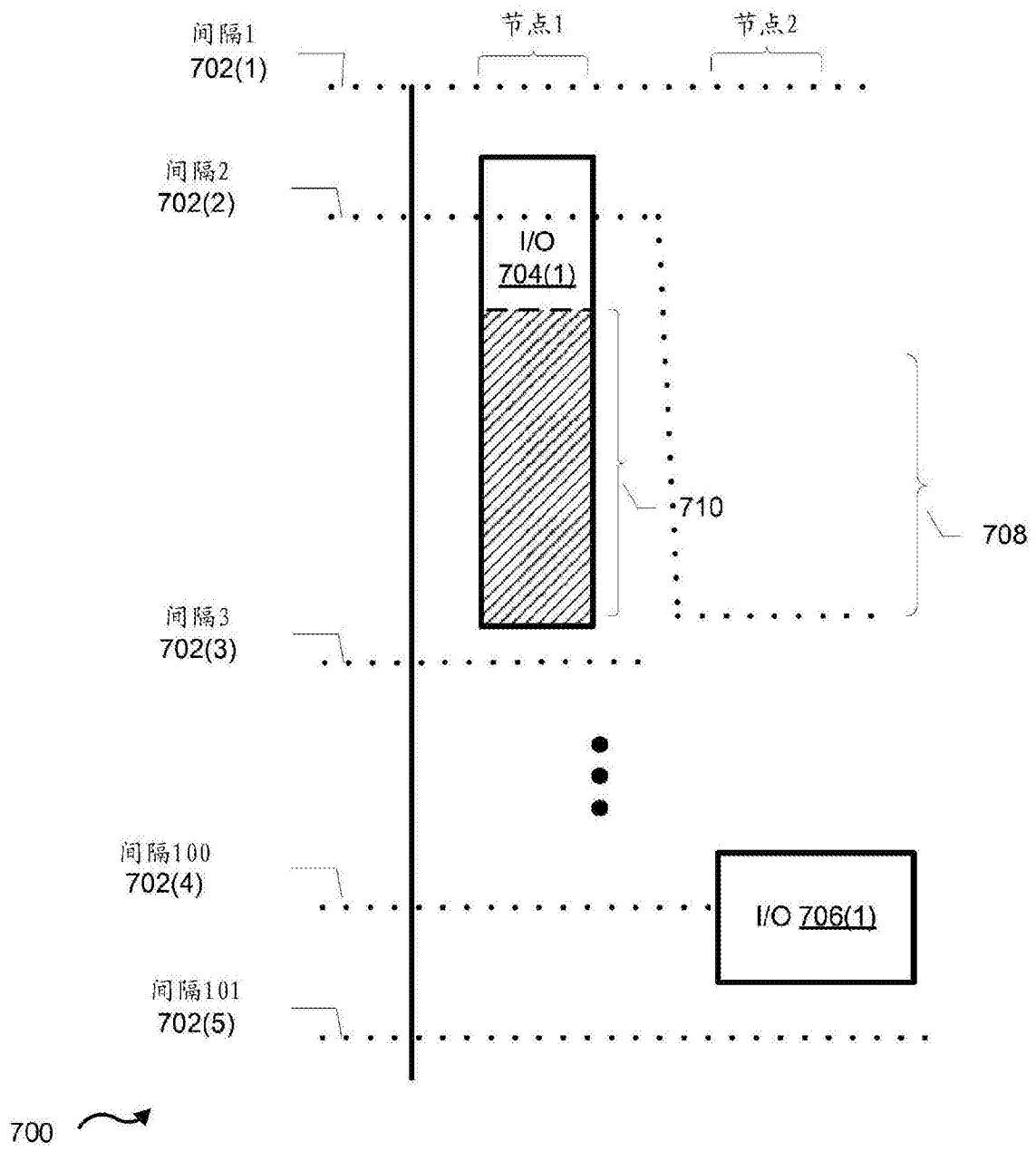


图7

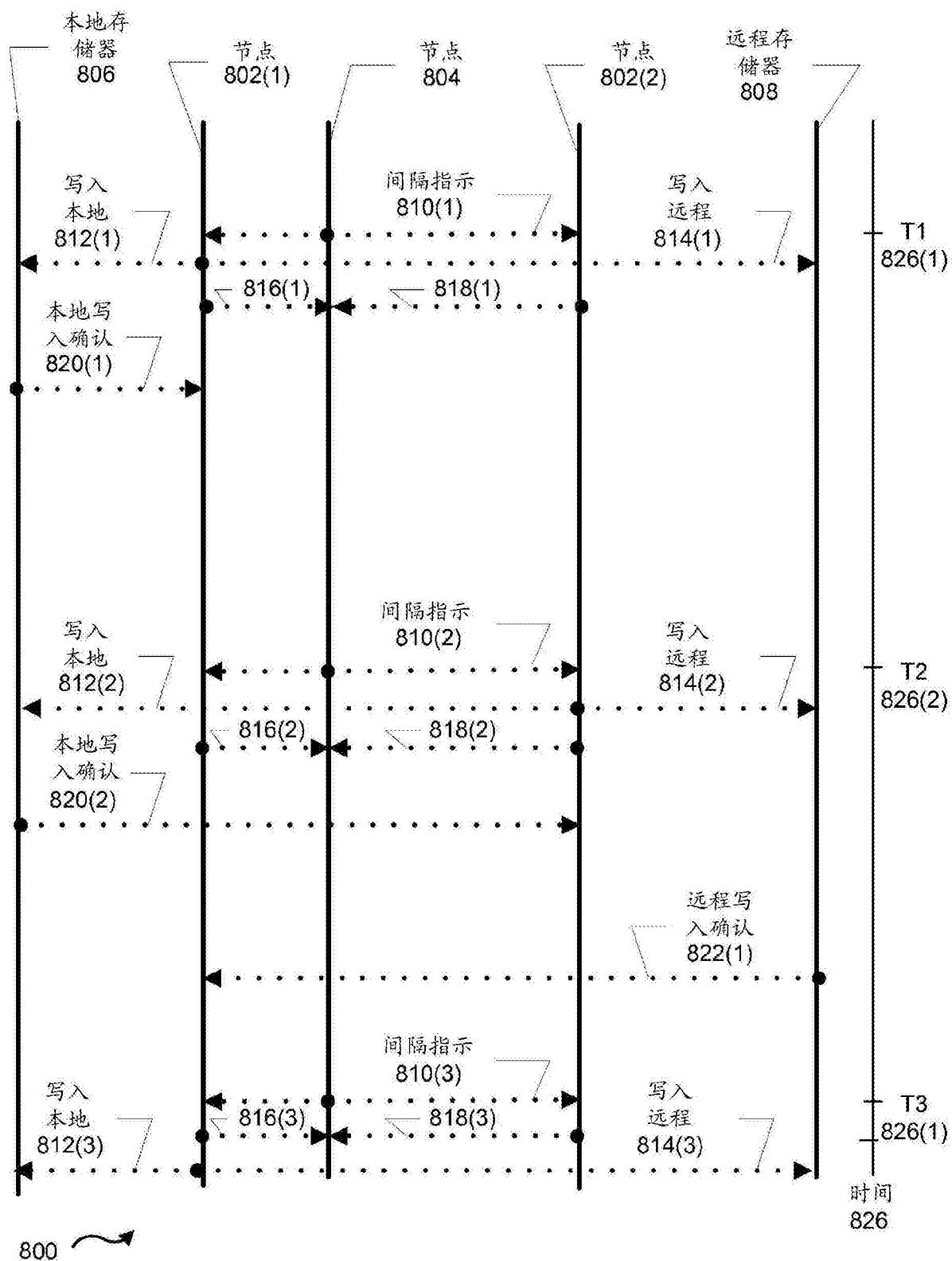


图8

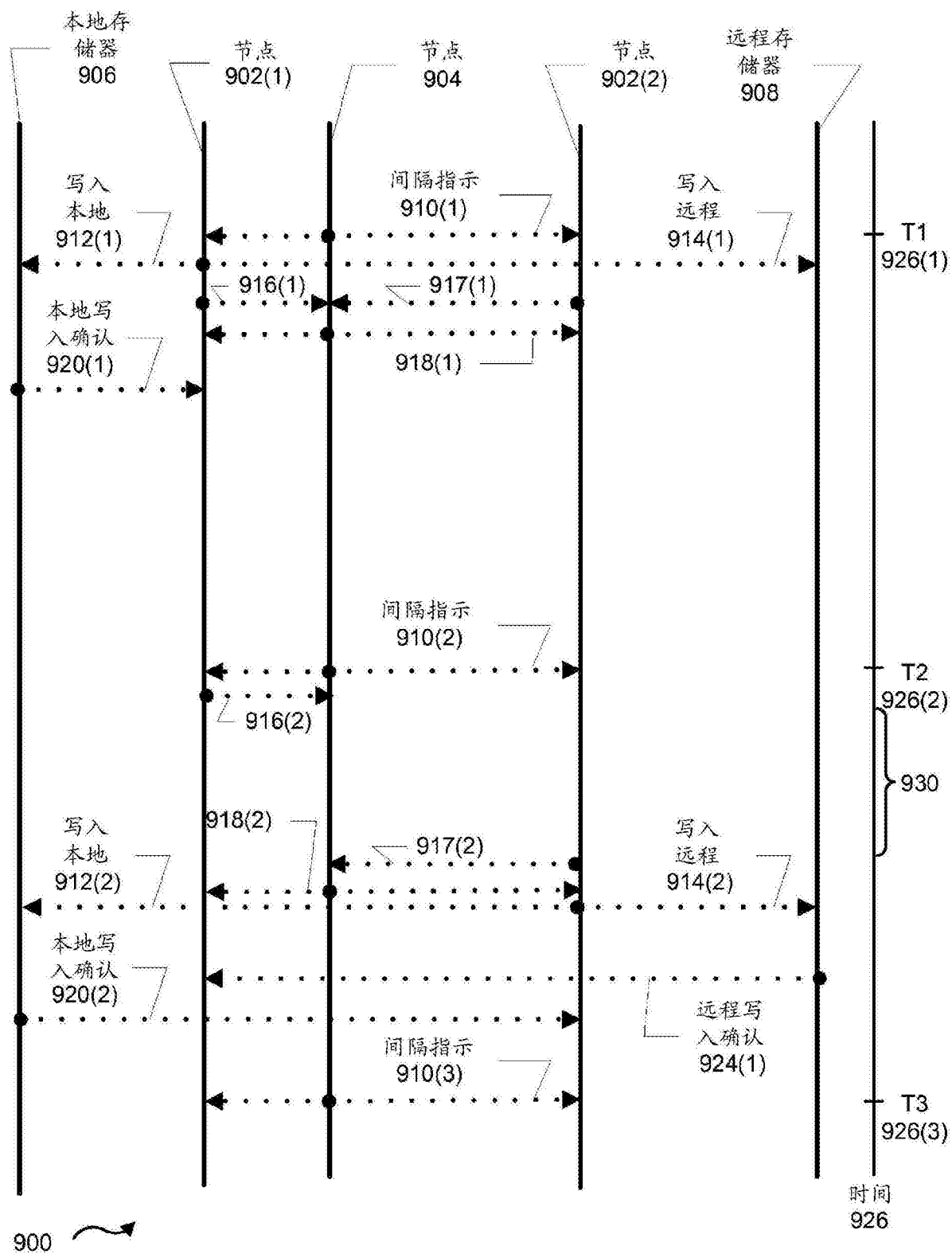


图9

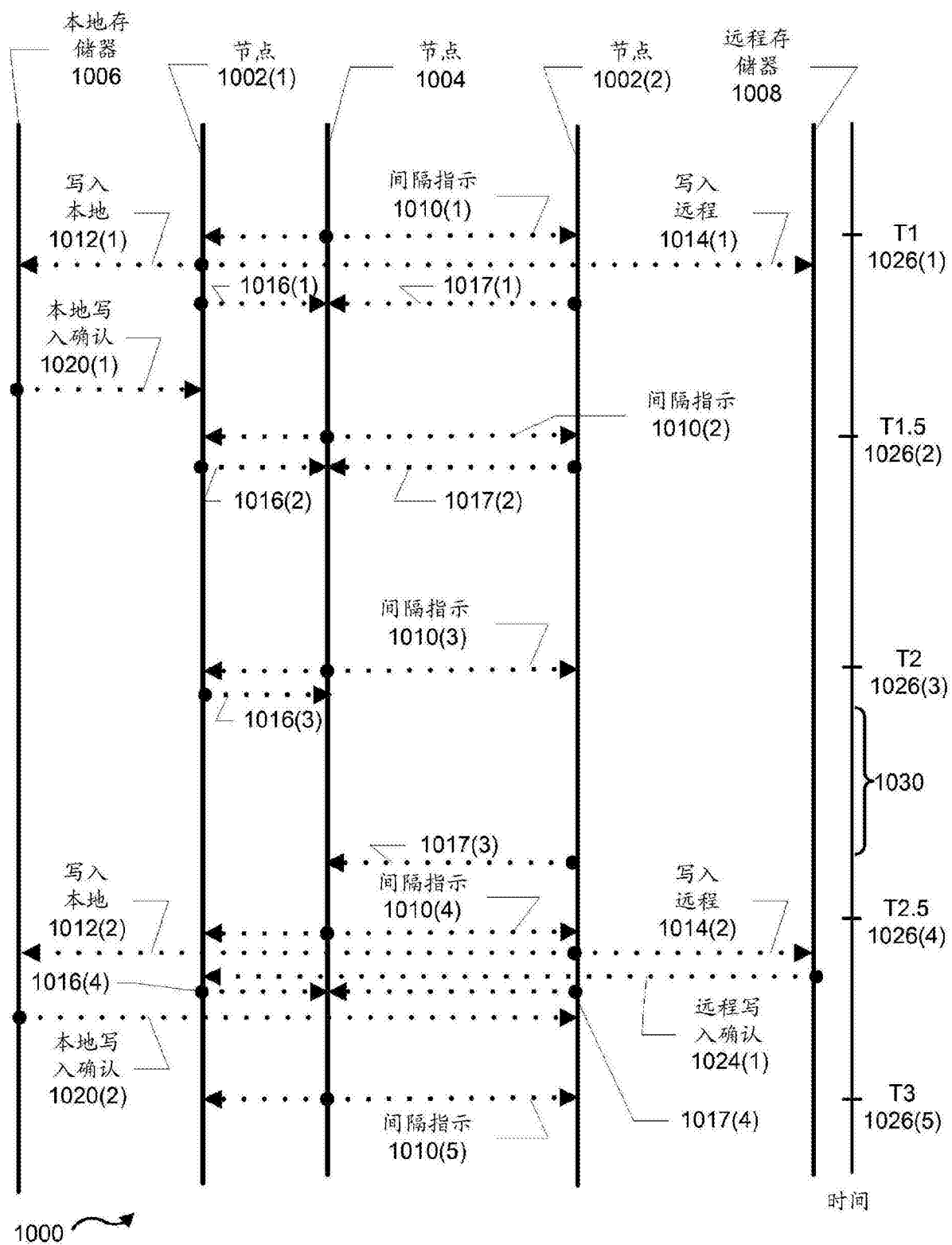


图10

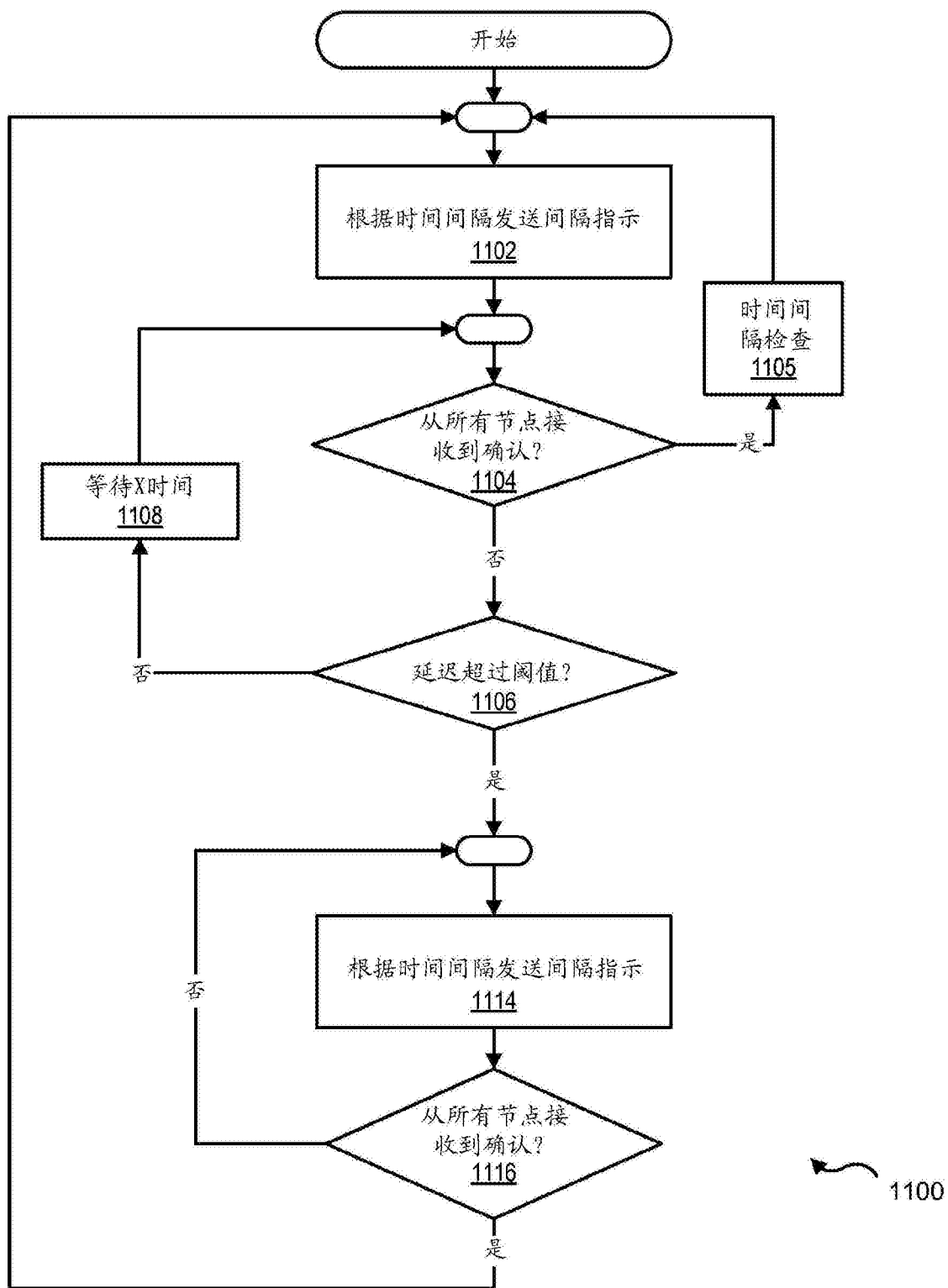


图11

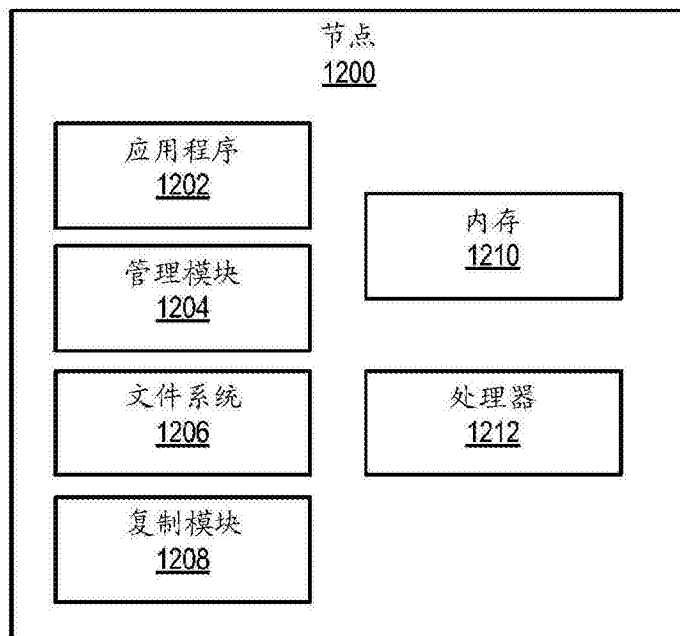


图12

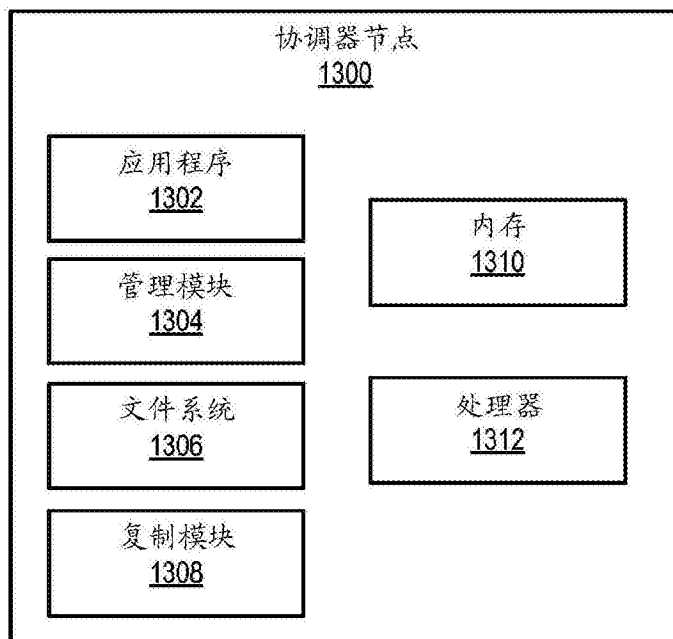


图13

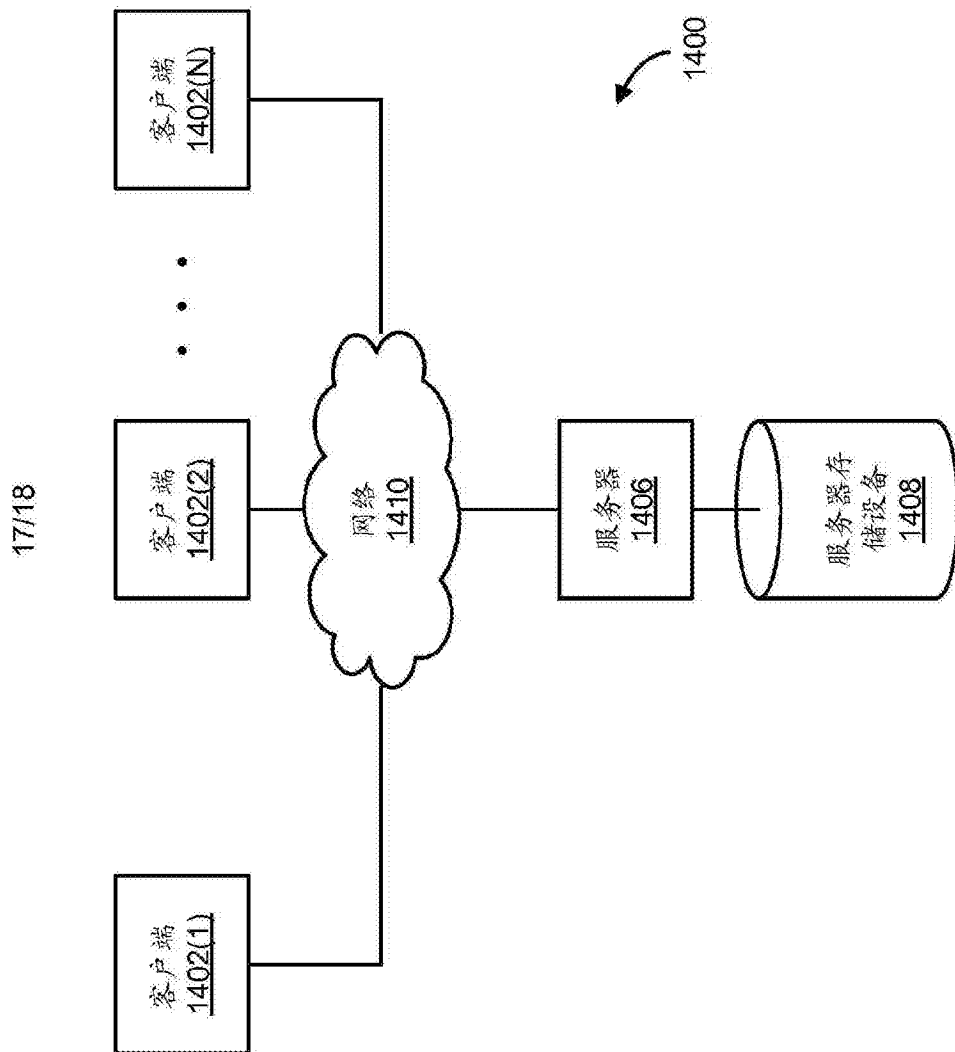


图14

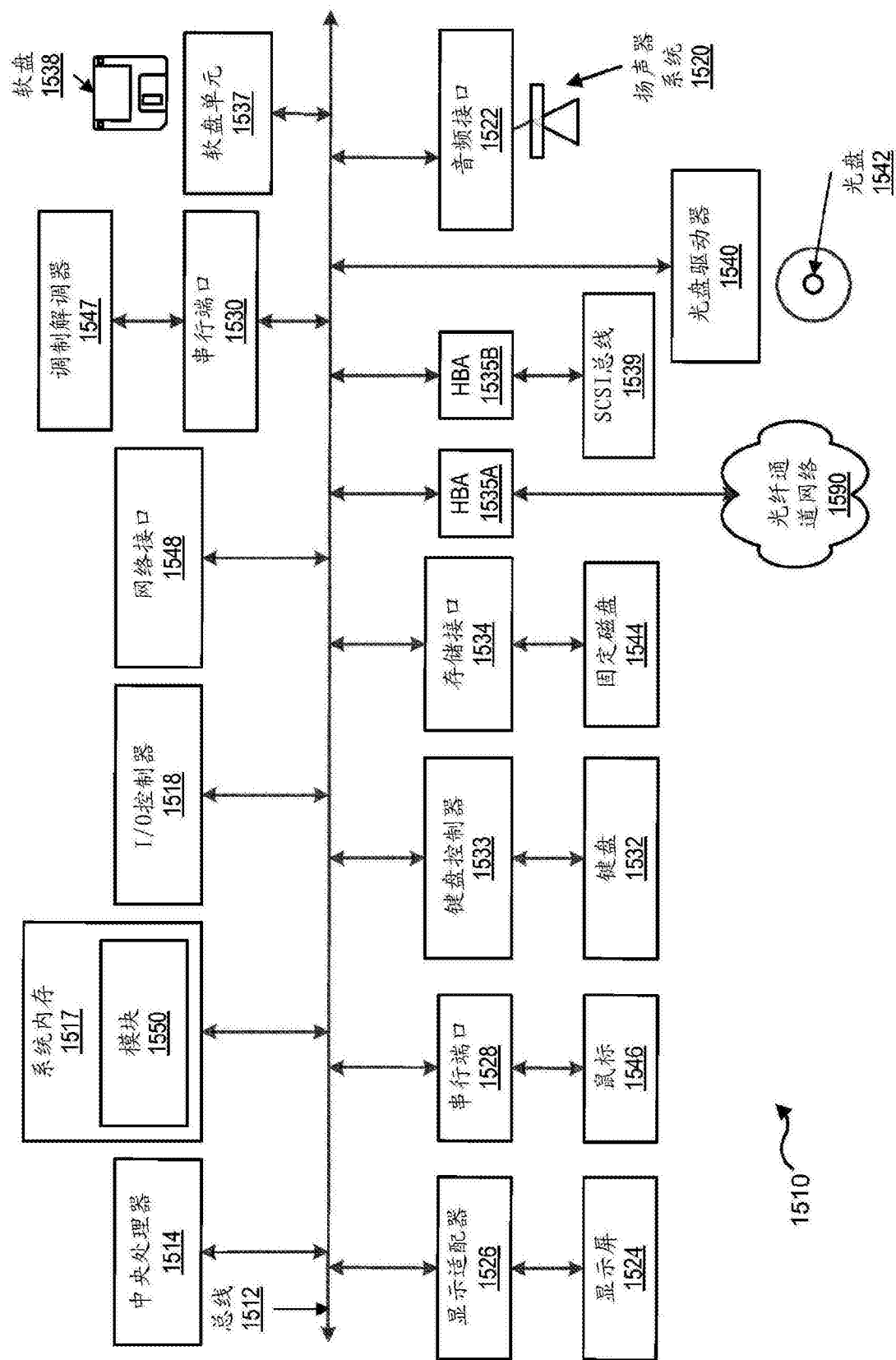


图15