

(12) 发明专利申请

(10) 申请公布号 CN 103250145 A

(43) 申请公布日 2013. 08. 14

(21) 申请号 201180058446. 4

代理人 邢德杰

(22) 申请日 2011. 12. 19

(51) Int. Cl.

(30) 优先权数据

G06F 17/00 (2006. 01)

12/975, 269 2010. 12. 21 US

(85) PCT申请进入国家阶段日

2013. 06. 04

(86) PCT申请的申请数据

PCT/US2011/065869 2011. 12. 19

(87) PCT申请的公布数据

W02012/087946 EN 2012. 06. 28

(71) 申请人 亚马逊技术股份有限公司

地址 美国内华达州

(72) 发明人 T · A · 瑟滕 S · 贾殷

J · R · 汉米尔顿 F · 卡塔诺 D · 魏

D · N · 桑德兰

(74) 专利代理机构 上海专利商标事务所有限公

司 31100

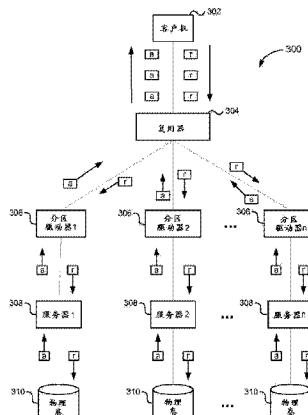
权利要求书2页 说明书18页 附图7页

(54) 发明名称

用于捕获数据集的技术

(57) 摘要

用于捕获数据集的技术（包括系统和方法）包括执行客户机端两阶段提交以确保一个或多个数据一致性条件。逻辑卷可以表示分布在多个物理存储装置中的数据集。指示一个或多个客户机装置阻止写入操作的至少认可。当一个或多个客户机装置已阻止写入操作的至少认可时，指示与物理存储装置进行通信的一个或多个服务器捕获数据集的对应部分。当已指示服务器捕获数据集的对应部分时，指示客户机装置恢复写入操作的至少认可。



1. 一种用于捕获数据集的计算机实施方法,其包括:

在配置有可执行指令的一个或多个计算机系统的控制下,

指示一个或多个应用程序延缓数据操作完成的至少认可,所述数据操作包括分布在多个分区中的数据集的操控,所述分区的每个是根据多个服务器的对应服务器的操作进行操作,所述一个或多个应用程序与所述服务器进行通信以用于至少写入到所述数据集;

响应于由所述应用程序延缓所述数据操作完成的认可,指示所述服务器捕获存储在所述数据集的对应分区中的所述数据集的对应部分;和

在指示所述服务器捕获所述数据集的所述对应部分之后,指示所述一个或多个应用程序恢复所述数据操作完成的至少认可。

2. 根据权利要求 1 所述的计算机实施方法,其中执行所述数据操作的请求源于计算装置且与所述一个或多个应用程序不同。

3. 根据权利要求 1 所述的计算机实施方法,其中所述多个分区存储在多个物理存储装置上。

4. 根据权利要求 1 所述的计算机实施方法,其中指示所述服务器捕获所述数据集的对应部分包括将令牌插入至所述服务器的请求流中。

5. 根据权利要求 1 所述的计算机实施方法,其还包括独立于执行所述数据操作的请求源于其的用户应用程序的操作生成捕获所述数据集的指令,且其中指示多个服务器的一个或多个客户机延缓数据操作完成的所述用户应用程序的至少认可是响应于捕获所述数据集的所述指令而执行。

6. 根据权利要求 1 所述的计算机实施方法,其中所述一个或多个存储界面的至少一个被配置来造成完成一个或多个数据操作,同时延缓对数据操作性能的至少认可的认可。

7. 根据权利要求 1 所述的计算机实施方法,其中指示一个或多个应用程序延缓数据操作完成的至少认可包括指示所述一个或多个应用程序延缓发布写入操作完成的认可。

8. 根据权利要求 1 所述的计算机实施方法,其中指示一个或多个应用程序延缓数据操作完成的至少认可包括指示所述一个或多个应用程序推迟写入请求的发布。

9. 一种可操作以引导数据集捕获的系统,其包括:

一个或多个处理器;和

存储器,其包括在由所述一个或多个处理器执行时造成所述系统进行以下步骤的可执行指令:

指示一个或多个应用程序延缓数据操作完成的至少认可,所述数据操作包括分布在多个分区中的数据集的操控,所述分区的每个是根据多个服务器的对应服务器的操作进行操作,所述一个或多个应用程序与所述服务器进行通信以用于至少写入到所述数据集;

响应于由所述应用程序延缓所述数据操作完成,指示所述服务器捕获存储在所述数据集的对应分区中的所述数据集的对应部分;和

在指示所述服务器捕获所述数据集的所述对应部分之后,指示所述一个或多个应用程序恢复所述数据操作完成的至少认可。

10. 根据权利要求 9 所述的系统,其中所述多个分区存储在多个物理存储装置上。

11. 根据权利要求 9 所述的系统,其中指示所述服务器捕获所述数据集的对应部分包括将令牌插入至所述服务器的请求流中。

12. 根据权利要求 9 所述的系统，其中执行所述数据操作的请求源于在计算装置上执行且与所述一个或多个应用程序不同的用户应用程序。

13. 根据权利要求 12 所述的系统，其中所述存储器包括在由所述一个或多个处理器执行时造成所述系统独立于所述用户应用程序的操作生成捕获所述数据集的指令的可执行指令，且其中指示多个服务器的一个或多个客户机延缓数据操作完成的所述用户应用程序的至少认可是响应于捕获所述数据集的所述指令而执行。

用于捕获数据集的技术

[0001] 发明背景

[0002] 联网的计算环境在计算组件的数量和类型以及计算环境中组件布置的复杂度的方面继续发展。某些这类计算环境提供虚拟化计算服务,其与主要实施计算服务的基础计算硬件不同程度地分离。将这种虚拟化用于虚拟化计算服务的用户和提供者存在许多优势。例如,虚拟化计算服务用户可以响应于需求增加而快速地(例如,以几分钟或几秒的级别)添加虚拟计算资源,并且如果需求下降那么也同样快速地释放虚拟计算资源用于其它用途。虚拟化计算服务用户的这种灵活性可使虚拟化计算服务提供者面临挑战和机会。

[0003] 虚拟化块装置是虚拟化计算服务的实例。虚拟化文件系统卷(“虚拟卷”)的用户可以创建虚拟卷、删除虚拟卷、重新调整虚拟卷大小和以其它方式重新配置虚拟卷而不用考虑如何分配基础计算资源的详情。用户还可以捕获存储在虚拟文件系统中的数据集,其中数据集捕获是某个时刻的数据集的表示。可以在不同时刻进行特定数据集的多次捕获,并且后期捕获可能取决于一个或多个早期捕获。例如,数据集的初次捕获可能涉及制作数据集的完整副本,然而数据集的后期捕获可能涉及复制从早期捕获以后已更改的数据。当出于不同原因而需要捕获时,捕获可以重建成卷。通常,为了使捕获在重建成卷时有用,必须满足关于输入/输出请求模式的特定性质。例如,捕获仅可能在其可保证或至少确保如果写入存在于捕获中,那么在提交所述写入之前确认的所有写入也在捕获中时有用。在某些情况下,这些条件可以相对容易实施。在其它情况下,诸如当跨多个服务器划分卷时,如果两个讨论中的写入前往不同服务器,那么必须注意维持这个性质。

附图说明

[0004] 将参考附图描述根据本公开内容的各个实施方案,其中:

[0005] 图1是示出了用于实施根据实施方案的方面的示例性环境的方面的示意图;

[0006] 图2是描绘了根据实施方案的示例性程序执行服务的方面的示意图;

[0007] 图3是可以用来实施本公开内容的实施方案的系统的构造的说明性实例的图示;

[0008] 图4是示出了分布式数据集的写入操作和捕获的时间线的图;

[0009] 图5是示出了分布式数据集的写入操作和捕获的另一时间线的图;

[0010] 图6是示出了分布式数据集的写入操作和捕获的另一时间线的图;

[0011] 图7是示出了分布式数据集的写入操作和捕获的另一时间线的图;和

[0012] 图8是可以用来实施本公开内容的各个实施方案的过程的说明性实例的流程图。

具体实施方式

[0013] 本文描述和提出的技术包括用于管理数据集捕获的方面的系统和方法。在实施方案中,逻辑卷划分在多个物理卷中,其中逻辑卷是以分布式方式物理地存储在物理卷中的数据集的表示。物理卷可以通过对应的物理存储装置(诸如硬盘驱动器或其它存储装置)进行存储,并且可以通过与多个服务器进行的通信进行访问。例如,可以由至少一个对应服务器服务于每个物理卷。在客户机装置上执行的客户机应用程序可以将用于访问数据的指

令发送到服务器并且服务器可以根据指令访问数据。例如，客户机可以发送执行写入操作的指令且接收指令的服务器可能造成要执行写入操作，并且发送对客户机操作执行所述操作的认可。客户机装置与每个服务器之间进行的通信可以经过一个或多个中间计算或联网装置。例如，客户机装置可以发送用于以下项的指令：对至少作为复用器操作的另一装置执行数据操作，从客户机装置接收指令和将指令传输到适当服务器。以此方式，客户机应用程序和客户机装置可以在无需保存数据集的特定部分存储在物理卷当中的记录的情况下进行操作。

[0014] 在实施方案中，进行数据集捕获，其中如所述，数据集捕获是某个时刻的数据集的表示。可以在不同时刻进行特定数据集的多次捕获，并且后期捕获可能取决于一个或多个早期捕获。例如，数据集的初次捕获可能涉及制作数据集的完整副本，然而数据集的后期捕获可能涉及复制从早期捕获以后已更改的数据。在本文描述的各个实施方案中，数据集的捕获方式保证如果写入存在于捕获中那么在提交所述写入之前确认或认可的所有写入也在捕获中。以此方式，捕获是以确保如果写入取决于早期写入并且所述写入是在捕获中那么所述写入和早期写入将在捕获中的方式进行。换句话说，如果应用程序进行两个写入（写入彼此依存），那么捕获将包括两个写入或不包含任何一个写入，从而避免逻辑不一致性，其中写入是在捕获中但是所述写入取决于其的另一写入不在捕获中。

[0015] 在实施方案中，使用客户机端两阶段提交提供上述保证。当捕获数据集时（诸如当接收执行捕获的指令时），指示一个或多个客户机装置阻止写入操作完成的认可。阻止写入操作完成的认可可以以任何合适方式（包括但不限于延缓写入操作的信息认可的发布和 / 或推迟发布写入请求）执行。阻止写入操作完成的认可还可以包括通过延缓认可的发布和 / 或推迟执行操作的新请求的发布来阻止其它操作（诸如读取操作）的认可。一个或多个客户机装置可以是与服务于逻辑卷分布在其中的物理卷的任何服务器进行通信的任何装置。例如，客户机装置可以是从在另一装置上执行的应用程序接收指令并且将指令分布到适当服务器的装置。客户机装置还可以是在其上执行应用程序的装置，或大体来说是涉及与服务于一个或多个物理卷的一个或多个服务器进行通信以参与对逻辑卷中数据的写入操作的处理和 / 或认可的任何装置。还可以指示一个或多个客户机装置延缓其它活动，诸如对数据集中数据的所有操作。

[0016] 当一个或多个客户机装置已如指示延缓活动时，指示服务于逻辑卷的物理卷的服务器进行数据集的对应部分的捕获。在实施方案中，指示服务器是通过将令牌插入至每个服务器的请求流中而完成，其中请求流是在数据集中执行操作的请求的序列，其通知对应服务器根据序列执行请求的操作。令牌是通知服务器执行由服务器服务的卷中数据集的一部分的捕获的任何信息。在执行通过应用程序执行生成的数据操作的请求在到达适当服务器之前经过多个装置的实施方案中，令牌可以插入从任何装置发送的请求的请求流中。在实施方案中，令牌插入从将指令发送到适当服务器的复用器发送的请求的请求流中。当每个服务器接收令牌（或应进行捕获的其它指示）时，服务器对服务器服务的物理卷进行捕获。服务器可以将捕获发送到另一数据存储区进行存储。而且，在实施方案中，当每个服务器接收令牌（或应进行捕获的其它指示）时，指示一个或多个客户机装置恢复延缓活动的处理。

[0017] 对于不同应用程序，可以在不同环境中实施不同方法。例如，图 1 示出了用于实施

根据各个实施方案的方面的示例性环境 100 的方面。如将明白,虽然基于 Web 的环境可以用于说明目的,但是可以适当利用不同环境实施各个实施方案。所示的环境 100 包括测试或开发部分(或端)和生产部分。生产部分包括电子客户机装置 102,其可以包括可操作以通过适当网络 104 发送和接收请求、消息或信息并且将信息传回到装置 102 的用户的任何适当装置。这些客户机装置的实例包括个人计算机、手机、手持消息传递装置、膝上型计算机、桌上型计算机、机顶盒、个人数据助理、电子书阅读器等等。

[0018] 网络 104 可以包括任何适当网络,包括内联网、互联网、蜂窝网、局域网、广域网、无线数据网或任何其它这类网络或其组合。用于这个系统的组件可以至少部分取决于选定网络和 / 环境的类型。用于经由这个网络进行通信的协议和组件是为人所熟知的并且本文将不再详细论述。可以通过有线或无线连接件和其组合启用网络通信。在这个实例中,网络 104 包括互联网,因为环境包括用于接收请求并响应于其而提供内容的 Web 服务器 106,但是如所属技术领域一般人员将明白对于其它网络,可利用作类似用途用的替代装置。

[0019] 说明性环境 100 包括至少一个应用程序服务器 108 和数据存储区 110。应了解可能存在多个应用程序服务器、层或者其它元件、过程或组件,其可以连接或以其它方式构造,可以进行交互以执行诸如获取来自适当数据存储区的数据的任务。如本文使用,术语“数据存储区”指的是能够存储、访问和 / 或检索数据的任何装置或装置组合,其可以包括任何标准的、分布式或群集式环境中的任何组合和数量的数据服务器、数据库、数据存储装置和数据存储介质。

[0020] 应用程序服务器 108 可以包括用于按需与数据存储区集成在一起以对客户机装置 102 执行一个或多个应用程序的方面的任何适当硬件和软件,并且甚至可以对应用程序处理大量数据访问和业务逻辑。应用程序服务器 108 与数据存储区 110 协作而提供访问控制服务,并且能够生成要传送给用户的内容(诸如文本、图形、音频和 / 或视频),在这个实例中所述内容可以以 HTML、XML 或另一适当结构化语言的形式通过 Web 服务器 106 提供给用户。

[0021] 可以由 Web 服务器 106 处理所有请求和响应的处理以及客户机装置 102 与应用程序服务器 108 之间进行的内容传递。应了解可无需 Web 服务器 106 和应用程序服务器 108 并且其仅是示例性组件,因为可以在如本文别处论述的任何适当装置或主机上执行本文论述的结构化代码。此外,可以使得测试自动化框架可以作为用户或应用程序可以订阅的服务提供的方式建构环境 100。测试自动化框架可以作为本文论述的不同测试模式的任何一个的实施方式提供,但是如本文论述或建议,还可以利用其它不同实施方式。

[0022] 环境 100 还可以包括开发端和 / 或测试端,其包括允许用户(诸如开发人员、数据管理员或测试员)访问系统的用户装置 118。用户装置 118 可以是(诸如)上文关于客户机装置 102 描述的任何适当装置或机器。环境 100 还可以包括开发服务器 120,例如在代码被部署和在生产端上执行并且外来用户可访问之前的开发和测试期间其与应用程序服务器 108 类似地起作用但是通常会运行代码。在某些实施方案中,应用程序服务器可以作开发服务器用,并且无法利用单独的生产和测试存储装置。

[0023] 数据存储区 110 可以包括用于存储涉及特定方面的数据的多个单独的数据表、数据库或其它数据存储机构和介质。例如,所示的数据存储区 110 包括用于存储生产数据 112 和用户信息 116 的机构,其可以用来为生产端提供内容。还示出了数据存储区 110 包括用

于存储可以与用于测试端的用户信息一起利用的测试数据 114 的机构。应了解可以存在存储在数据存储区 110 中的许多其它方面（诸如页面图像信息和访问权限信息），其可以适当存储在上述机构的任何一个中或在数据存储区 110 中的额外机构中。

[0024] 数据存储区 110 可通过与其相关的逻辑操作以从应用程序服务器 108 或开发服务器 120 接收指令，并且响应于其而获取、更新或以其它方式处理数据。在一个实例中，用户可以提交对特定类型项目的搜索请求。在这种情况下，数据存储区 110 可以访问用户信息 116 以验证用户识别，并且可以访问目录详情信息以获取关于所述类型项目的信息。接着，信息可以作为网页上列出且用户能够经由用户装置 102 上的浏览器查看的结果返回给用户。可以在专用浏览器网页或窗口中查看特定关注项目的信息。

[0025] 每个服务器通常将包括提供可执行程序指令用于服务器的一般管理和操作的操作系统，并且通常将包括存储在由服务器的处理器执行时允许服务器执行其预定功能的指令的计算机可读介质。尤其根据本公开内容，用于服务器的操作系统和一般功能的合适实施装置是已知的或可市售，并且容易由所属技术领域一般人员来实施。

[0026] 在一个实施方案中，环境 100 是分布式计算环境，其利用经由通信链路并使用一个或多个计算机网络或直接连接件互连的多个计算机系统和组件。然而，所属技术领域一般人员将明白这个系统同样可以组件数量小于或大于图 1 所示的组件数量的系统操作。因此，图 1 的系统 100 的描绘本质上应是说明性的，并且不限于本公开内容的范围。

[0027] 在至少一个实施方案中，环境 100 的一个或多个方面可以合并和 / 或并入分布式程序执行服务中。图 2 描绘了根据至少一个实施方案的示例性分布式程序执行服务 200 的方面。分布式程序执行服务 200 提供虚拟化计算服务，包括虚拟计算机系统服务 202 和虚拟数据存储区服务 204，其中通过相对高速的数据网使大量计算资源互连。这些计算资源可以包括处理器，诸如中央处理单元 (CPU)、易失性存储装置（诸如随机访问存储器 (RAM)）、非易失性存储装置（诸如快速存储器）、硬盘驱动器和光学驱动器、服务器（诸如上文参考图 1 描述的 Web 服务器 106 和应用程序服务器 108）、一个或多个数据存储区（诸如图 1 的数据存储区 110）以及互连网络中的通信带宽。图 2 未明确示出由分布式程序执行服务 200 管理的计算资源，因为其是用于着重示出虚拟化计算服务与实施其的计算资源的独立关系的分布式程序执行服务 200 的方面。

[0028] 分布式程序执行服务 200 可以利用计算资源以至少部分通过执行一个或多个程序、程序模块、程序组件和 / 或编程对象（统称“程序组件”）（包括和 / 或编译自以任何合适机器和 / 或编程语言指定的指令和 / 或代码）实施虚拟化计算服务。例如，计算资源可以在需要时进行分配或重新分配以使程序组件的执行变容易，和 / 或程序组件可以在需要时指派或重新指派给计算资源。这个指派可以包括（例如）用于增强执行效率的程序组件的物理重新定位。从虚拟化计算服务用户的角度来看，分布式程序执行服务 200 可以灵活地和 / 或按需供应例如与每个资源单元的商品款式定价规划相关的计算资源。

[0029] 分布式程序执行服务 200 还可以利用计算资源实施至少被配置来控制虚拟化计算服务的服务控制平面 206。服务控制平面 206 可以包括服务管理界面 208。服务管理界面 208 可以包括至少被配置来使虚拟化计算服务的用户和 / 或管理员能够提供、取消提供、配置和 / 或重新配置（统称“提供”）虚拟化计算服务的合适方面的基于 Web 的用户界面。例如，虚拟计算机系统服务 202 用户可以提供一个或多个虚拟计算机系统实例 210、212。用

户接着可以配置提供的虚拟计算机系统实例 210、212 以执行用户的应用程序。虚拟计算机系统实例 210 与 212 之间的省略号指示虚拟计算机系统服务 202 可以支持任何适量（例如，几千、几百万和更多）的虚拟计算机系统实例，但是为了清楚起见仅示出了两个。

[0030] 服务管理界面 208 还可以使用户和 / 或管理员能够指定和 / 或重新指定虚拟化计算服务政策。可以由服务控制平面 206 的服务政策强制执行组件 214 维持和强制执行这些政策。例如，服务管理界面 208 的存储管理界面 216 部分可以被虚拟数据存储区服务 204 的用户和 / 或管理员用来指定要由服务政策强制执行组件 214 的存储政策强制执行组件 218 维持和强制执行的虚拟数据存储区服务政策。虚拟计算机系统服务 202 和虚拟数据存储区服务 204 的不同方面和 / 或实用程序（包括虚拟计算机系统实例 210、212、低延时数据存储区 220、高耐用性数据存储区 222 和 / 或基础计算资源）可以受界面（诸如应用程序编程界面（API）和 / 或基于 Web 的服务界面）控制。在至少一个实施方案中，控制平面 206 还包括至少被配置来根据一个或多个工作流程与虚拟计算机系统服务 202 和虚拟数据存储区服务 204 的不同方面和 / 或实用程序的界面进行交互和 / 或引导与其进行交互的工作流程组件 246。

[0031] 在至少一个实施方案中，服务管理界面 208 和 / 或服务政策强制执行组件 214 可以创建接着由工作流程组件 246 维持的一个或多个工作流程和 / 或造成工作流程组件 246 创建其。工作流程（诸如提供工作流程和政策强制执行工作流程）可以包括被执行来执行工作（诸如提供或政策强制执行）的任务的一个或多个序列。如本文使用的术语工作流程不是任务本身，而是可以控制往返任务的信息的流动以及其控制的任务的执行次序的任务控制结构。例如，工作流程可以视作可管理和返回在执行期间任何时间的过程的状态的状态机。工作流程可以创建自工作流程模板。例如，提供工作流程可以创建自配置有服务管理界面 208 的参数的提供工作流程模板。举另一实例，政策强制执行工作流程可以创建自配置有服务政策强制执行组件 214 的参数的政策强制执行工作流程模板。

[0032] 工作流程组件 246 可以修改、进一步指定和 / 或进一步配置建立的工作流程。例如，工作流程组件 246 可以选择分布式程序执行服务 200 的特定计算资源来执行特定任务和 / 或指派给特定任务。这个选择可以至少部分基于如由工作流程组件 246 评估的特定任务的计算资源需要。举另一实例，工作流程组件 246 可以将额外和 / 或重复任务添加到建立的工作流程和 / 或重新配置建立的工作流程中的任务之间的信息流动。建立的工作流程的这个修改可以至少部分基于工作流程组件 246 进行的执行效率分析。例如，某些任务可以有效地并行执行，同时其它任务取决于先前任务的成功完成。

[0033] 虚拟数据存储区服务 204 可以包括多种类型的虚拟数据存储区，诸如低延时数据存储区 220 和高耐用性数据存储区 222。例如，低延时数据存储区 220 可以以相对低延时地保存可以由虚拟计算机系统实例 210、212 读取和 / 或写入（统称“访问”）的一个或多个数据集 224、226。数据集 224 与 226 之间的省略号指示低延时数据存储区 220 可以支持任何适量（例如，几千、几百万和更多）的数据集，但是为了清楚起见仅示出了两个。对于由低延时数据存储区 220 保存的每个数据集 224、226，高耐用性数据存储区 222 可以保存一组捕获 228、230。每组捕获 228、230 可以分别保存其相关数据集 224、226 的任何适量的捕获 232、234、236 和 238、240、242，如由省略号指示。每次捕获 232、234、236 和 238、240、242 可以提供在特定时刻各自数据集 224 和 226 的表示。这些捕获 232、234、236 和 238、240、242

可以用于后期检查,包括在捕获时刻各自数据集 224 和 226 到其状态的恢复。虽然分布式程序执行服务 200 的每个组件可以利用基础网络进行通信,但是图 2 着重示出低延时数据存储区 220 与高耐用性数据存储区 222 之间进行的数据传送 244,因为通过这个数据传送 224 对基础网络上的利用负载所做的贡献可是明显的。

[0034] 例如,低延时数据存储区 220 的数据集 224、226 可以是虚拟文件系统卷。低延时数据存储区 220 可以包括提供对基础数据存储硬件的访问的低开销虚拟化层。例如,低延时数据存储区 220 的虚拟化层可相对于高耐用性数据存储区 222 的等效层而呈低开销。用于根据至少一个实施方案建立和维护低延时数据存储区和高耐用性数据存储区的系统和方法是为所属技术领域熟练人员所知的,因此本文仅着重说明其特征的一些。在至少一个实施方案中,分别分配给低延时数据存储区 220 和高耐用性数据存储区 222 的基础计算资源的组实际上是断开的。

[0035] 低延时数据存储区 220 和 / 或高耐用性数据存储区 222 可以视作非局部的和 / 或相对于虚拟计算机系统实例 210、212 独立。例如,实施虚拟计算机系统服务 202 的物理服务器可以包括局部存储设施,诸如硬盘驱动器。这些局部存储设施可以是相对低延时的,但是其它方面受到限制,例如可靠度、耐用性、大小、吞吐量和 / 或可用性。此外,在局部存储装置中且分配给特定虚拟计算机系统实例 210、212 的数据可以具有对应虚拟计算机系统实例 210、212 的有效使用期,使得如果虚拟计算机系统实例 210、212 失效或取消提供,那么局部数据丢失和 / 或变得无效。在至少一个实施方案中,可以由多个虚拟计算机系统实例 210、212 有效地共享非局部存储装置中的数据集 224、226。例如,可以由虚拟计算机系统实例 210、212 将数据集 224、226 作为虚拟文件系统卷进行安装。

[0036] 可以至少部分通过块数据存储 (BDS) 服务 248 协助和 / 或运用其实施虚拟数据存储区服务 204 中的数据存储区,包括低延时数据存储区 220 和 / 或高耐用性数据存储区 222。BDS 服务 248 可以运用一组分配的计算资源 (包括多个块数据存储服务器) 协助一个或多个数据存储卷 (诸如文件系统卷) 的创建、读取、更新和 / 或删除。块数据存储卷和 / 或其数据块可以跨多个块数据存储服务器分布和 / 或复制以增强卷可靠度、延时、耐用性和 / 或可用性。举个例说,在某些实施方案中,存储块数据的多个服务器块数据存储系统可以组织成一个或多个池或其它组,每个池或组具有共同位于某个地理位置处 (诸如在一个或多个地理分布的数据中心的每个中) 的多个物理服务器存储系统,并且使用存储在数据中心的服务器块数据存储系统上的块数据卷的程序可以在数据中心处的一个或多个其它物理计算系统上执行。

[0037] BDS 服务 248 可以协助和 / 或实施在其传送通过分布式程序执行服务 200 的基础计算资源时数据块的本地缓存,包括实施低延时数据存储区 220 和 / 或高耐用性数据存储区 222 的数据存储区服务器处的本地缓存和实施虚拟计算机系统服务 202 的虚拟计算机系统服务器处的本地缓存。在至少一个实施方案中,高耐用性数据存储区 222 是独立于 BDS 服务 248 实施的档案质量数据存储区。高耐用性数据存储区 222 可以运用相对由 BDS 服务 248 操控的数据块来说较大的数据集进行工作。高耐用性数据存储区 222 可以独立于 BDS 服务 248 进行实施 (例如,运用不同界面、协议和 / 或存储格式)。

[0038] 每个数据集 224、226 可以具有随时间推移而不同的变化模式。例如,数据集 224 的变化速率可以高于数据集 226 的变化速率。然而,在至少一个实施方案中,整体的平均变化

速率的不足以表现出数据集变化的特征。例如,数据集 224、226 的变化速率可以本身具有随当前时间、星期几、季节(包括与假期和 / 或特殊事件相关的预期突发时间)和年度改变的模式。数据集 224、226 的不同部分可以与不同变化速率相关,并且每个变化速率“信号”本身可以由(例如)可运用傅里叶分析技术检测的独立信号源组成。任何合适的统计分析技术均可以用来对数据集变化模式进行建模,包括 Markov 建模和 Bayesian 建模。

[0039] 如上文描述,数据集 224 的初次捕获 232 可能涉及数据集 224 的基本完整复制和通过网络到高耐用性数据存储区 222 的传送 224(可以是“完全捕获”)。数据集 224 可能与不同类型的元数据相关。数据集 224 的捕获 232、234、236 中可以取决于数据集 224 的类型而包括所有这些元数据或不包括这些元数据的任何一个。例如,低延时数据存储区 220 可以取决于其在故障恢复情形下的重建成本来指定捕获中包括的元数据。初次捕获 232 外的捕获 234、236 可以是“增量的”,例如涉及复制由于一个或多个先前捕获引起的数据集 224 变化。捕获 232、234、236 可以布置在类层次中,使得特定捕获可以相对于捕获类的子层次增量(例如,每周调度的捕获相对于上周的每日捕获来说是多余的,但是相对于先前每周捕获来说是增量的)。取决于后续捕获 234、236 的频率,与完全捕获相比,对增量捕获来说基础计算资源上的利用负载可明显较小。

[0040] 例如,数据集 224 的捕获 232、234、236 可以包括实施低延时数据存储区 220 的一组服务器和 / 或存储装置的读取访问以及用于更新元数据例如以更新追踪数据集 224 的“脏”数据块的数据结构的写入访问。出于这种描述目的,如果数据集 224 的数据块已从(相同类别和 / 或类型的)最近捕获以后有所改变,那么其是脏的(相对于特定类别和 / 或类型的捕获)。在从低延时数据存储区 220 传送 224 到高耐用性数据存储区 222 之前,可以由所述组服务器压缩捕获 232、234、236 数据和 / 或对其进行加密。在高耐用性数据存储区 222 处,接收的捕获 232、234、236 可以再次写入到基础组的服务器和 / 或存储装置。因此,每次捕获 232、234、236 涉及对有限的基础计算资源的负载,包括服务器负载和网络负载。

[0041] 例如,可以利用存储管理界面 216 手动请求数据集 224 的捕获 232、234、236。在至少一个实施方案中,可以根据数据集捕获政策自动调度捕获 232、234、236。根据至少一个实施方案的数据集捕获政策可以用存储管理界面 216 指定,并且与一个或多个特定数据集 224、226 相关。数据集捕获政策可以指定针对数据集捕获的固定或灵活调度。固定数据集捕获调度可以指定某天的特定时间、一周中某天、某月和 / 或任何合适时间和日期的捕获。固定数据集捕获调度可以包括循环捕获(例如,每周末午夜、每周五上午 2 点、每月初上午 4 点)以及开关捕获。

[0042] 灵活的数据集捕获政策可以指定捕获会在特定时间窗(例如,每天上午 2 点至上午 6 点、周日某刻、月末停业后)内或以特定频率(例如,每小时一次、每天两次、每周一次、每月一次)发生。在至少一个实施方案中,灵活的数据集捕获政策可以指定捕获被调度来达成合适目的、目标和 / 或条件(统称“捕获条件”)。例如,每次捕获 232、234、236 可以具有相关财政和 / 或计算资源成本,并且灵活的数据集捕获政策可以指定捕获 232、234、236 或捕获组 228 的成本目标和 / 或成本差异,包括每个时段的预算和 / 或每次捕获的平均成本。举另一实例,在至少一个实施方案中,数据集 224 的一部分的数据丢失的概率至少是在给定时间数据集 224 中未捕获数据的数量的函数。因此,灵活的数据集捕获政策可以指定数据集 224 的一部分的数据丢失的概率,并且存储政策强制执行组件 218 可以调度数据集

224 的捕获以通过使数据集 224 中未捕获数据的数量保持低于相关的未捕获数据目标和 / 或差异来达成目标。

[0043] 数据集捕获政策可以指定固定调度、灵活调度和捕获条件的任何合适组合。数据集捕获政策还可以指定捕获有效期限和 / 或捕获保留目的、目标和 / 或条件。例如，可以对每日捕获指定七天有效期限，可以对每周捕获指定四周有效期限，和 / 或可以对每月捕获指定一年有效期限。捕获可以具有未指定和 / 或无限的有效期限，因此需要手动删除。此外，特定捕获可以受保护，例如可能需要通过指定组的授权用户进行手动删除。捕获 232、234、236 和 / 或捕获组 228、230 可以与成本（例如，每千兆字节用于存储的定期费用）相关，并且数据集捕获政策可以指定要自动删除捕获 232、234、236 以达成成本目标和 / 或差异。数据捕获保留政策的强制执行可以分析相关捕获组 228、230 以优选删除多余捕获和 / 或禁止删除可能防止在对应最近捕获 232 的时间数据集 224 恢复到其状态的捕获。

[0044] 图 3 示出了根据实施方案的捕获可以记录在其中的环境的示意图 300。在这个实例中，客户机 302 根据一组可执行指令操作。客户机可以是计算装置（诸如上文描述的计算装置）或在计算装置上操作的模块。在实施方案中，客户机 302 在其操作时利用数据、创建数据和并且以其它方式使用数据。在实施方案中，当客户机 302 操作时，客户机生成用于结合存储在逻辑卷中的数据执行数据操作（诸如创建操作、读取操作、更新操作和删除操作）的多个请求，其中逻辑卷是物理地存储在一个或多个物理存储装置中的数据集的表示。在这个实例中，如下文更多论述，由多个物理卷服务于由客户机装置 302 使用的本地卷。物理存储装置可以是块存储装置或其它存储装置。而且，当数据集存储在多个物理存储装置中时，物理存储装置可以分布在多个位置（诸如共同数据中心中的位置和 / 或不同地理位置）中。因此，从客户机 302 的角度来看，无论逻辑卷是什么或有多少物理卷用来存储数据集均进行执行数据操作的请求。客户机 302 可能能够或无法识别和 / 或指定识别物理存储装置本身的信息。

[0045] 如下文描述，当客户机 302 进行执行数据操作的不同请求时，客户机将多个请求发送到复用器 304，其中复用器是可操作以适当地分布来自客户机 302 的请求使得可以履行请求的计算装置或模块。在这个实例中，由包含字母“r”的方框表示请求。请求可以是对复用器 304 或复用器 304 与客户机 302 之间的中间系统提出的网页服务请求，但是一般来说，可以根据任何合适协议以任何合适方式提出请求。

[0046] 在实施方案中，当复用器 304 从客户机 302 接收请求时，复用器 304 将请求分布到适当的分区驱动器 306，其中分区驱动器可以是可操作以将请求传达到可访问物理数据存储区 310 的服务器 308 的模块。复用器 304 和分区驱动器 306 可以作为共同装置的部分执行使得复用器 304 与分区驱动器 306 之间进行的通信是在存储器中执行且因此相对较快。然而，复用器 304 和一个或多个分区驱动器 306 可以分布在网络上。此外，虽然出于说明目的，图 300 中每个阶段的请求示为相同，但是请求可以转译成不同格式以容纳接收请求的不同组件。例如，从客户机 302 到复用器 304 的请求可以呈一种格式，同时对应从复用器 304 到分区驱动器 306 的请求的信息可以呈另一格式。

[0047] 在实施方案中，当执行由客户机 302 请求的数据操作时，从物理卷 310 发送操作完成的认可，其中对客户机 302 执行所述操作，如图 300 中由包含字母“a”的方框所示。正如请求，认可可以在不同组件之间传达时呈不同格式。而且，虽然出于说明目的，图 300 示出

了每个服务器 308 与单个物理卷 310 进行通信,但是服务器可以与多个物理卷和由客户机 302 访问的数据集的部分进行通信并且可以存储在与服务器进行通信的一个或多个物理卷中。

[0048] 如论述,可以出于不同原因进行数据集捕获。在许多情况下,可以根据关于哪次捕获表示某个时段的数据集的准确度的一个或多个保证进行捕获。保证可能涉及在接近开始捕获数据集的过程之时执行数据操作的请求。例如,捕获可以以下述方式进行:保证捕获将包括在开始捕获之前(诸如在发送和/或接收进行捕获的请求之前)认可的所有写入和捕获将不包括在开始捕获之后(即,在返回成功之后)提交的任何写入。这些保证可以脱离关于在开始之前或期间提交和在开始期间或之后认可的写入的不确定性。存在用于处理这些写入的多个选项,包括提供某种形式的强一致性保证、较弱保证或根本不提供保证。

[0049] 在强一致性保证的情况下,对于某个时间 t (其可以介于客户发送请求的时间与客户接收响应的时间之间),可以保证捕获包括在 t 之间提交的所有写入(包括在 t 之后认可的某些写入)或不包括在 t 之后认可的写入(包括在 t 之前提交的某些写入)。在由多个服务器服务于逻辑卷的情况下,强一致性保证通常需要默许和激励逻辑卷的任何客户机,从而导致可能的明显性能损失。此外,强一致性保证可能不足以满足某些客户使用情况,其需要甚至更强的保证,其中保证捕获包括在 t 之前提交和认可的所有写入。这个额外保证可能需要遵循客户应用程序。例如,应用程序可以选择检查点,停止写入,等待要认可的所有写入,且接着恢复写入。一般来说,所有这些步骤需要强加客户可能不需要并且可能以恢复写入会高延时的代价的限制。

[0050] 然而,即使强一致性保证改善延时,但是不提供额外捕获一致性保证可能导致不合希望的结果,如图 4 所示。图 4 示出了包括写入和捕获有关事件的时间线的说明性实例。特定来说,图 4 示出了由两个服务器(图中称作服务器 A 和服务器 B)服务于逻辑卷的情况。在图 4 所示的实例中,几乎同时开始对应每个服务器的分区的捕获而不用在服务器中进行捕获的其它协调。在这个时间线中(其中时间从左前往右),客户机相继发布两个写入,写入 0 和写入 1。而且在这个实例中,客户机等待在提交写入 1 之前写入 0 的认可,从而导致三种可能的一致情况:捕获不包含写入 0 也不包含写入 1,捕获包含写入 0,或捕获包含写入 0 或写入 1。然而,在图 4 所示的情况下,可能进行其中仅存在写入 1 的不一致捕获,因为关于服务器 A 的捕获是在认可写入 0 之前开始。如果在客户机上执行的应用程序取决于写入 0 和写入 1 的次序,那么如果逻辑卷(或其一部分)从捕获恢复,那么逻辑卷中的数据可能是无意义的(损坏的),从而可能造成客户机和任何应用程序发生故障,这取决于逻辑卷中的数据。

[0051] 为了避免这些后果,可以给定关于捕获一致性的保证弱于上文描述的强一致性,但是其提供严格的排序保证,其中对于某个时间 t (在捕获初始化期间):捕获将包括在 t 之前认可的所有写入;和捕获将不包括在 t 之后提交的写入。换句话说,对于所有写入(写入 0 和写入 1),如果在认可写入 0 之后提交写入 1,那么如果写入 1 在捕获中,那么写入 0 也必须在捕获中。这仅是原定保证的更严格版本,其足以提供不会破坏两个写入之间的因果关系的保证。在实施方案中,这是通过推迟写入请求的认可直到捕获初始化之后为止而实施;即,当开始分区捕获时,分区停止认可写入请求直到指示其恢复认可为止。在此期间,服务器可以继续处理输入的写入请求,但是服务器推迟(阻止)请求的认可直到适当时间

为止,诸如当已认可已运用另一服务器(或所有其它服务器)开始的捕获时。如果每个服务器进行阻止直到最后一个服务器已开始阻止为止,那么可达成上述保证。

[0052] 图5示出了与图4所示的时间线类似的时间线。在这个实例中,实施用于维持上文描述的较弱条件的算法。在这种情形下,客户机阻止等待写入0的认可。同时,服务器A和服务器B继续进行对逻辑卷的其各自分区开始捕获。一旦已开始所有捕获,客户机立即恢复等待写入0的认可并且服务器释放写入认可。如所示,服务器A认可至客户机的写入0,并且客户机接着继续发布接着由服务器B认可的写入1。应注意与强保证不同的是客户机无需在捕获可以开始之前等待对所有未决请求的认可,从而造成所得延时明显较少。

[0053] 然而,这个较弱保证导致可以视作不寻常的某些行为。例如,图6示出了无序数据包递送造成在写入0之后认可的写入1包括在不包括写入0的捕获中的情况。在这种情况下,写入1在捕获中,但是写入0不在捕获中。应注意这不违反上文描述的较弱一致性保证,因为写入1是在认可写入0之前提交,且因此两个写入不具有任何因果关系。事实上,在这种情况下,提交两个写入的次序是无关紧要的;这种情况也出现于可适当自由地对写入进行重新排序的现代输入/输出(I/O)调度器。这个原理可以采取下述否定形式:对于任何两个写入(写入0和写入1),如果写入1是在认可写入0之后提交并且如果写入1在捕获中,那么写入0也必须在捕获中。

[0054] 一种满足这组条件的保证方式是在服务于逻辑卷的分区存储在其上的物理存储装置的服务器上执行两阶段提交(服务器端两阶段提交)。在服务器端两阶段提交的情况下,当接收执行捕获的请求时,指示每个服务器停止接受新写入。在所有服务器已停止之后,请求服务器进行捕获且接着开始再次接受新写入。这保证了在最后一个写入进入捕获中之前确认的所有写入也将在捕获中,从而通过容纳取决于其它写入的写入但不容纳其它写入取决于其的写入避免捕获中的逻辑不一致性。

[0055] 虽然可用于上述原因,但是服务器端两阶段提交的性能可产生操作问题。例如,如果在一个服务器中处理开始提交阶段存在困难(例如,服务器无法认可服务器已停止接受新写入),那么捕获过程无法继续进行直到克服困难为止。换句话说,在等待一个或多个其它服务器的响应时,一个或多个服务器可以是闲置的。因此,一个服务器的问题可以造成许多服务器发生延迟。此外,由于卷划分到其中的物理存储装置的数量增加,所以快照将花较长时间才开始的机会增加。这种服务器延迟可能造成不良的应用程序性能和/或机能障碍。

[0056] 图7示出了实施根据实施方案的过程的说明性实例,其确保可保证上述较弱条件。在这个示例性图700中,执行客户机端两阶段提交。特定来说,如图所示,客户机将执行数据操作的请求传达到图中识别为服务器A和服务器B的两个服务器。例如,客户机可以请求对存储在分布在由服务器A和服务器B服务的两个物理卷中的虚拟卷中的数据进行的操作的性能。客户机请求可以根据在客户机上或在与客户机进行通信的另一计算装置中执行的应用程序。在图中提供的实例中,客户机通过驱动层702连接到服务器。驱动层702可以是一个或多个计算装置或可以在一个或多个计算装置上实施。驱动层可以包括多个驱动器,或一般来说包括其它存储界面,诸如上文结合图3论述的驱动器。虽然未示出,但是复用器可以从客户机接收请求并将请求转发到驱动层的适当驱动器。在这个实例中,驱动层可以具有两个驱动器,每个服务器A和服务器B具有一个驱动器。

[0057] 在图 7 所示的说明性实施方案中,当客户机发送请求时,请求被驱动层 702 接收并转发到适当服务器。例如,在实施方案中,如果客户机请求在一个或多个数据块上执行操作,那么驱动层确定服务器可访问适用的数据块并将请求发送到每个适用服务器。举个例说,如图 7 所示,客户机发送被驱动层 702 接收且示为写入 0 的写入请求。驱动层 702 确定写入 0 应用于存储在由服务器 A 服务的物理存储装置上的数据块且因此将请求转发到服务器 A,服务器 A 接着执行请求的操作并将执行操作的认可发送到将认可转发到客户机的驱动层 702。

[0058] 如实例中所示,客户机将另一写入请求(写入 1)提交到驱动层 702,其确定写入 1 应用于可由服务器 B 访问的数据且因此将请求转发到服务器 B。当执行请求的操作时,服务器 B 将操作性能的认可发送到将认可转发到客户机的驱动层 702。在这个实例中,在客户机接收执行写入 1 的认可之前,由客户机发送写入 2,从而指示写入 2 不取决于写入 1。因此,根据上文论述的较弱保证,如果写入 2 在捕获中,那么写入 1 未必需要在捕获中,但是其可以在捕获中。如所述,来自客户机的请求可以指定对存储在多个物理存储装置中的数据进行的操作。举个例说,请求可以指定对多个数据块进行的操作,一些数据块存储在由服务器 A 服务的物理存储装置中并且其它数据块存储在由服务器 B 服务的物理存储装置中。在这个实施方案中,驱动层 702 可以将适用于由服务器 A 服务的物理存储装置的请求发送到服务器 A 并且将适用于由服务器 B 服务的物理存储装置的另一请求发送到服务器 B。换句话说,驱动层 702 可以将客户机请求分成多个请求并将多个请求的每个发送到适当服务器。

[0059] 又如图 7 所示,客户机提交执行图中识别为写入 2 的操作的另一请求。驱动层 702 接收请求,确定写入 2 应用于存储在由服务器 B 服务的物理存储装置中的数据将写入 2 转发到服务器 B。然而,在所述实例中,在驱动层接收已接收写入 2 的认可之前,发送和接收捕获分布在由服务器 A 和服务器 B 服务的物理存储装置中的逻辑卷的请求。可以由客户机或另一装置(诸如根据调度、用户提交的捕获请求或以任何其它方式的发送捕获请求的装置)发送捕获请求。在实施方案中,当驱动层 702 接收捕获请求时,驱动层 702 停止处理认可和/或请求。驱动层 702 接着将捕获令牌插入从驱动层 702 到每个服务器的请求流中。例如,参考图 3,每个驱动器可以将捕获令牌插入从所述驱动器到对应服务器的请求流中。由驱动层 702 接收的认可和/或请求可以保存在存储器中直到令牌已插入请求流中为止。此外,指示驱动层 702 的一个或多个驱动器停止处理请求可以是存储器中进行的操作且因此相对较快。

[0060] 在实施方案中,捕获令牌是在由服务器接收时向服务器指示服务器应在处理源于客户机的额外请求之前执行存储在对应物理存储装置中的逻辑卷的一部分的捕获的任何信息。令牌可以是对服务器的明确指令或可以是其它信息。举个例说,从驱动层到服务器的请求流中的请求可以包括每当进行逻辑卷捕获时改变的号码或其它识别符。服务器可以检测识别符的变化,且作为响应而在处理源于客户机的额外请求(一般来说是指示服务器进行捕获的任何信息)之前捕获逻辑卷的一部分。此外,插入请求流中的信息可以诸如取决于对特定捕获确保哪些保证而改变。

[0061] 如图所示,在已转发对服务器执行写入 1 和写入 2 的请求之后并且在从服务器 B 接收执行写入 1 和写入 2 的认可之前,由驱动层 702 接收捕获请求。此外,如所示,在将捕获令牌插入从驱动层 702 到服务器 A 的请求流中之后但在将捕获令牌插入从驱动层 702 到

服务器 B 的请求流中之前, 驱动层接收写入 1 和写入 2 的认可。为了遵循上文论述的条件组, 驱动器层 702 延迟将写入 1 和写入 2 的认可转发到客户机 702。驱动层 702 可以将写入 1 和写入 2 的认可保存在存储器中直到捕获令牌已插入所有适当请求流中且接着将认可转发到客户机为止。可以由驱动层 702 延迟其它操作, 诸如请求到适当服务器的转发。以此方式, 与服务器端两阶段提交不同, 服务器 A 和服务器 B 能够处理有用请求, 而非必须等待其它服务器。此外, 由于指示驱动层 702 停止处理请求直到将捕获令牌插入至服务器的请求流中为止可以是快于指示服务器停止处理请求的存储器中进行的请求, 所以一个或多个服务器能够早于服务器端两阶段提交再次开始处理请求和 / 或认可。

[0062] 应注意图 7 提供了客户机端两阶段提交的特定发生事件的说明性实例并且预期不同变化是在本公开内容的范围内。例如, 图 7 示出了划分在两个物理卷中的逻辑卷, 由对应服务器服务于每个物理卷。然而, 出于执行数据操作的目的, 逻辑卷可以划分在两个以上物理卷中并且驱动律师可以因此与两个以上服务器进行通信。此外, 图 7 示出了客户机装置通过驱动层与服务器进行通信的特定配置。然而, 如所示, 本文描述的原理适用于这个配置和其它配置 (通常包括客户机与服务器直接或间接进行通信以对以分布式方式存储的数据集中的数据执行操作的任何配置) 的变化。

[0063] 图 8 示出了用于记录信息的过程 800 的说明性实例的流程图。例如, 过程 800 可以用来实施上文图 7 中所示的原理。过程 800 (或本文描述的任何其它过程, 或其变化和 / 或组合) 的部分可以在配置有可执行指令的一个或多个计算机系统的控制下执行并且可以作为共同在硬件的一个或多个处理器或其组合上执行的代码 (例如, 可执行指令、一个或多个计算机程序、或一个或多个应用程序) 实施。代码可以 (例如) 以包括可由一个或多个处理器执行的多个指令的计算机程序的形式存储在计算机可读存储介质上。

[0064] 在实施方案中, 接收 802 捕获请求。如上文论述, 捕获请求是向接收者指示应捕获数据集的任何信息。捕获请求可以是 (例如) 用于捕获数据集的可执行指令。捕获请求还可以是被接收者用来应用处理逻辑以确定是否应捕获数据集的信息。捕获请求可以接收自任何合适源 (诸如在其上执行使用数据集的应用程序的计算装置) 或另一计算装置 (诸如调度数据集捕获的计算装置)。捕获请求还可以生成在执行过程 800 的至少一部分或其变化的计算装置内部。例如, 管理存储至少一个数据集的存储装置的计算装置可以基于调度或其它因素确定应进行捕获并且其可以据此操作。换句话说, 捕获请求可以取决于或独立于使用存储在数据集中的信息的应用程序。

[0065] 在实施方案中, 当接收捕获请求时, 请求第一客户机装置延缓一个或多个数据处理活动。延缓一个或多个数据处理活动的指令可以是或包括延缓执行数据集操作的认可的指令和 / 或将来自应用程序的请求转发到一个或多个服务器的指令。参考作为说明性实例的图 3, 客户机装置可以是驱动器或在其上执行多个驱动器的计算装置。客户机装置还可以是图中描绘的客户机 302 或另一装置。一般来说, 被指示延缓一个或多个数据处理活动的客户机装置可以是与结合数据存储区操作的服务器进行通信的任何装置。当已请求第一客户机装置延缓一个或多个数据处理活动时, 接着可以确定 806 是否存在需要所述指令的额外客户机装置, 并且如果存在, 那么指示 808 下一个客户机装置在数据集中延缓一个或多个数据处理活动, 且再次确定 806 是否需要被指示延缓请求处理的额外客户机装置。

[0066] 因此, 在实施方案中, 当已指示所有客户机装置延缓一个或多个数据处理活动并

且确定不存在应被指示延缓一个或多个数据处理活动的额外客户机装置时,指示服务于数据集分区的一组服务器捕获数据集的对应部分。指示服务器捕获数据集的对应部分可以包括指示 810 第一服务器捕获数据集的对应部分并且确定 812 是否存在应被指示捕获数据集的额外服务器。在实施方案中,指示第一服务器是响应于从客户机装置接收客户机装置已延缓一个或多个数据处理活动的处理的认可而执行。如果存在应被指示捕获数据集的额外服务器,那么可以指示 814 下一个服务器捕获数据集的对应部分,并且可以再次确定 812 是否存在应被指示捕获数据集的对应部分的额外服务器。

[0067] 当已指示所有服务器捕获数据集的对应部分时,可以确定 812 不存在应被指示捕获数据集的对应部分的额外服务器,并且可以指示客户机装置恢复一个或多个数据处理活动的处理。指示客户机装置恢复处理一个或多个数据处理活动可以包括指示 816 第一客户机装置恢复一个或多个数据处理活动,确定 818 是否应指示额外客户机装置恢复一个或多个数据处理活动的处理,并且如果存在应被指示恢复一个或多个数据处理活动的处理的额外客户机装置,那么指示 820 客户机装置恢复一个或多个数据处理活动的处理直到已指示所有客户机装置为止。

[0068] 条款 :

[0069] 条款 1. 一种用于捕获数据集的计算机实施方法,数据集分布在多个物理存储装置中,可由至少一个对应服务器访问每个物理存储装置,至少一个对应服务器可操作以从一个或多个存储界面接收造成对数据集中的数据执行对应操作的指令,所述方法包括:

[0070] 在配置有可执行指令的一个或多个计算机系统的控制下,

[0071] 接收捕获数据集的指令;

[0072] 响应于捕获数据集的指令;

[0073] 将延缓对应用程序进行的写入操作完成的至少认可的第一命令发到一个或多个存储界面;

[0074] 对于一个或多个存储界面的每个,接收存储界面已延缓写入操作完成的至少认可的第一通知;

[0075] 在对于一个或多个存储界面的每个接收第一通知之后,指示每个服务器存储界面使捕获存储在对应物理存储装置中的数据集的一部分的指令入队且随后恢复认可写入操作的完成;和

[0076] 存储数据集的表示,所述表示包括数据集的捕获部分。

[0077] 条款 2. 根据条款 1 所述的计算机实施方法,其还包括:

[0078] 从应用程序接收执行写入操作的请求;

[0079] 对于每个请求,选择在其中执行对应操作的物理存储装置并且在选定物理存储装置中执行对应操作。

[0080] 条款 3. 根据条款 1 所述的计算机实施方法,其中应用程序是在与一个或多个存储界面不同的计算机系统上执行。

[0081] 条款 4. 根据条款 1 所述的计算机实施方法,其中指示每个存储界面使捕获存储在对应物理存储装置中的数据集的一部分的指令入队且随后恢复认可写入操作的完成包括对于所述服务器的每个服务器,将令牌插入从一个或多个存储界面到服务器的请求流中。

[0082] 条款 5. 根据条款 1 所述的计算机实施方法,其中一个或多个存储界面包括多个存

储界面,传达执行数据操作的请求包括 :

[0083] 从应用程序接收请求;

[0084] 对于所述请求的每个请求,选择多个存储界面的存储界面,其中一个或多个存储界面的每个被配置来指示对应服务器造成在对应物理存储装置中完成对应请求的操作。

[0085] 条款 6. 根据条款 1 所述的计算机实施方法,其中一个或多个存储界面的至少一个被配置来造成完成一个或多个数据操作,同时延缓对数据操作性能的至少认可的认可。

[0086] 条款 7. 一种用于捕获数据集的计算机实施方法,其包括:

[0087] 在配置有可执行指令的一个或多个计算机系统的控制下,

[0088] 指示一个或多个应用程序延缓数据操作完成的至少认可,数据操作包括分布在多个分区中的数据集的操控,每个分区是根据多个服务器的对应服务器的操作进行操作,一个或多个应用程序与服务器进行通信以用于至少写入到数据集;

[0089] 响应于由应用程序延缓数据操作完成的认可,指示服务器捕获存储在数据集的对应分区中的数据集的对应部分;和

[0090] 在指示服务器捕获数据集的对应部分之后,指示一个或多个应用程序恢复数据操作完成的至少认可。

[0091] 条款 8. 根据条款 7 所述的计算机实施方法,其中执行数据操作的请求源于计算装置且与所述一个或多个应用程序不同。

[0092] 条款 9. 根据条款 7 所述的计算机实施方法,其中多个分区存储在多个物理存储装置上。

[0093] 条款 10. 根据条款 7 所述的计算机实施方法,其中指示服务器捕获数据集的对应部分包括将令牌插入至服务器的请求流中。

[0094] 条款 11. 根据条款 7 所述的计算机实施方法,其还包括独立于执行数据操作的请求源于其的用户应用程序的操作生成捕获数据集的指令,且其中指示多个服务器的一个或多个客户机延缓数据操作完成的所述用户应用程序的至少认可是响应于捕获所述数据集的所述指令而执行。

[0095] 条款 12. 根据条款 7 所述的计算机实施方法,其中一个或多个存储界面的至少一个被配置来造成完成一个或多个数据操作,同时延缓对数据操作性能的至少认可的认可。

[0096] 条款 13. 根据条款 7 所述的计算机实施方法,其中指示一个或多个应用程序延缓数据操作完成的至少认可包括指示一个或多个应用程序延缓发布写入操作完成的认可。

[0097] 条款 14. 根据条款 7 所述的计算机实施方法,其中指示一个或多个应用程序延缓数据操作完成的至少认可包括指示一个或多个应用程序推迟写入请求的发布。

[0098] 条款 15. 一种可操作以引导数据集捕获的系统,其包括:

[0099] 一个或多个处理器;和

[0100] 存储器,其包括在由一个或多个处理器执行时造成系统进行以下步骤的可执行指令:

[0101] 指示一个或多个应用程序延缓数据操作完成的至少认可,数据操作包括分布在多个分区中的数据集的操控,每个分区的是根据多个服务器的对应服务器的操作进行操作,一个或多个应用程序与服务器进行通信以用于至少写入到数据集;

[0102] 响应于由应用程序延缓数据操作完成,指示服务器捕获存储在数据集的对应分区

中的数据集的对应部分 ; 和

[0103] 在指示服务器捕获数据集的对应部分之后, 指示一个或多个应用程序恢复数据操作完成的至少认可。

[0104] 条款 16. 根据条款 15 所述的系统, 其中多个分区存储在多个物理存储装置上。

[0105] 条款 17. 根据条款 15 所述的系统, 其中指示服务器捕获数据集的对应部分包括将令牌插入至服务器的请求流中。

[0106] 条款 18. 根据条款 15 所述的系统, 其中执行数据操作的请求源于在计算装置上执行且与一个或多个应用程序不同的用户应用程序。

[0107] 条款 19. 根据条款 18 所述的系统, 其中存储器包括在由一个或多个处理器执行时造成系统独立于应用程序的操作生成捕获数据集的指令的可执行指令, 且其中指示多个服务器的一个或多个客户机延缓数据操作完成的所述用户应用程序的至少认可是响应于捕获所述数据集的所述指令而执行。

[0108] 条款 20. 具有指令共同存储在其上的一个或多个计算机可读存储介质, 所述指令在由一个或多个处理器执行时造成一个或多个处理器进行以下步骤 :

[0109] 指示一个或多个存储界面延缓一个或多个活动结合数据集的处理 ;

[0110] 在一个或多个存储界面延缓一个或多个活动的处理之后, 将令牌插入多个请求流中, 多个请求流的每个是从一个或多个存储界面到多个服务器的服务器, 多个服务器的每个可操作以造成根据对应请求流中的请求完成数据集分区中的数据操作, 每个令牌是执行多个数据集分区的数据集分区的捕获的指示 ; 和

[0111] 在将令牌插入一个或多个请求流中之后, 指示一个或多个存储界面恢复一个或多个活动的处理。

[0112] 条款 21. 根据条款 20 所述的一个或多个计算机可读存储介质, 其中数据集分区分布在多个物理存储装置中。

[0113] 条款 22. 根据条款 20 所述的一个或多个计算机可读存储介质, 其中每个请求流包括执行从一个或多个存储界面的对应计算装置到多个服务器的对应服务器的数据操作的请求的集合。

[0114] 条款 23. 根据条款 20 所述的一个或多个计算机可读存储介质, 其中对于每个服务器, 请求流包括执行来自一个或多个存储界面的数据操作的多个请求。

[0115] 条款 24. 根据条款 20 所述的一个或多个计算机可读存储介质, 其还包括在由一个或多个处理器执行时造成一个或多个处理器进行以下步骤的指令 :

[0116] 从应用程序接收执行数据操作的请求 ; 和

[0117] 对于来自应用程序的每个请求 :

[0118] 选择适当服务器用于执行请求的操作 ; 和

[0119] 指示选定服务器造成完成请求的操作。

[0120] 条款 25. 一种用于捕获数据集的计算机实施方法, 其包括 :

[0121] 在配置有可执行指令的一个或多个计算机系统的控制下,

[0122] 使用多个服务器在多个分布式数据集分区中执行数据集操作 ;

[0123] 指示与多个服务器进行通信的一个或多个存储界面阻止一个或多个活动结合数据集的至少完成的认可 ; 和

[0124] 在一个或多个存储界面阻止至少一个或多个活动的认可之后,将令牌插入从一个或多个存储界面到服务器的一个或多个请求流中,每个令牌是服务器接收在处理令牌时在处理令牌已插入其中的对应请求流期间执行多个数据集分区的数据集分区的捕获的令牌的指示。

[0125] 条款 26. 根据条款 25 所述的计算机实施方法,其中数据集分区分布在多个物理存储装置中。

[0126] 条款 27. 根据条款 25 所述的计算机实施方法,其中每个请求流包括执行从一个或多个存储界面的对应计算装置到多个服务器的对应服务器的数据操作的请求的集合。

[0127] 条款 28. 根据条款 25 所述的计算机实施方法,其中对于每个服务器,请求流包括来自一个或多个存储界面的执行数据操作的多个请求。

[0128] 条款 29. 根据条款 25 所述的计算机实施方法,其还包括 :

[0129] 从应用程序接收执行数据操作的请求 ;和

[0130] 对于来自应用程序的每个请求 :

[0131] 选择与适当服务器进行通信的适当存储界面以造成完成请求的操作 ;和

[0132] 造成选定存储界面与适当服务器进行通信以完成请求的操作。

[0133] 各个实施方案还可在多种操作环境中实施,在某些情况下其可包括可用来操作任何多种应用程序的一个或多个用户计算机、计算装置或处理装置。用户或客户机装置可包括任何多种通用个人计算机,诸如运行标准操作系统的桌上型或膝上型计算机以及运行移动软件并且能够支持多种联网和消息传递协议的蜂窝、无线和手持装置。这个系统还可包括运行多种市售操作系统和其它已知应用程序用于(诸如)开发和数据库管理目的的多个工作站。这些装置还可包括其它电子装置,诸如能够经由网络进行通信的虚拟终端、瘦客户机、游戏系统和其它装置。

[0134] 多数实施方案利用所属技术领域熟练人员将熟悉是用于支持使用任何多种市售协议(诸如 TCP/IP、OSI、FTP、UPnP、NFS、CIFS 和 AppleTalk)进行的通信的至少一个网络。网络可是例如局域网、广域网、虚拟专用网、互联网、内联网、外联网、公共电话交换网、红外线网、无线网和其任何组合。

[0135] 在利用 Web 服务器的实施方案中,Web 服务器可运行任何多种服务器或中间层应用程序,包括 HTTP 服务器、FTP 服务器、CGI 服务器、数据服务器、Java 服务器和业务应用程序服务器。服务器还可能够响应于来自用户装置的请求而诸如通过执行一个或多个 Web 应用程序(其可以作为以任何编程语言(诸如 Java®、C、C# 或 C++)或任何脚本语言(诸如 Perl、Python 或 TCL)以及其组合写入的一个或多个脚本或程序实施)执行程序或脚本。服务器还可以包括数据库服务器,包括不限于 Oracle®、Microsoft®、Sybase® 和 IBM® 市售的服务器。

[0136] 环境可包括如上文论述的多种数据存储区以及其它存储器和存储装置。这些装置可以常驻在多个位置中,诸如跨网络位于一个或多个计算机本地或远离任何或所有计算机(和 / 或常驻在其中)的存储介质上。在特定组的实施方案中,信息可以常驻在所属技术领域熟练人员熟悉的存储区网络(“SAN”)中。类似地,用于执行归属于计算机、服务器或其它网络装置的功能的任何必要文件可以适当存储在本地和 / 或远端。在系统包括计算机化装置的情况下,每个这类装置可包括可以经由总线电连接的硬件元件,所述元件包括例

如至少一个中央处理单元 (CPU)、至少一个输入装置（例如，鼠标、键盘、控制器、触屏或按键）和至少一个输出装置（例如，显示装置、打印机或扬声器）。这个系统还可以包括一个或多个存储装置，诸如硬盘驱动器、光学存储装置和固态存储装置（诸如随机访问存储器（“RAM”）或只读存储器（“ROM”））以及可移式介质装置、存储卡、闪存卡等。

[0137] 这些装置还可包括如上文描述的计算机可读存储介质读取器、通信装置（例如，调制解调器、网卡（无线或有线）、红外线通信装置等）和工作存储器。计算机可读存储介质读取器可与计算机可读存储介质（表示用于暂时和 / 或更持久容纳、存储、传输和检索计算机可读信息的远程、本地、固定和 / 或可移式存储装置以及存储介质）连接或被构造来收纳其。系统和不同装置通常还将包括位于至少一个工作存储器装置内的多种软件应用程序、模块、服务或其它元件，包括操作系统和应用程序，诸如客户机应用程序或 Web 浏览器。应明白替代实施方案可以具有上文描述的众多变化。例如，还可使用自定义硬件和 / 或特定元件可以以硬件、软件（包括便携式软件，诸如 applet）或两者进行实施。此外，可以采用至其它计算装置的连接件，诸如网络输入 / 输出装置。

[0138] 用于容纳代码或代码的部分的存储介质和计算机可读介质可包括所属技术领域中已知或使用的任何适当介质，包括以任何方法或技术实施用于存储和 / 或传输信息（诸如计算机可读指令、数据结构、程序模块或其它数据）的存储介质和通信介质，诸如但不限于易失性和非易失性、可移式和非可移式介质，包括可用来存储所希望信息并且可通过系统装置访问的 RAM、ROM、EEPROM、快速存储器或其它存储器技术、CD-ROM、数字多用盘 (DVD) 或其它光学存储装置、磁带盒、磁带、磁盘存储装置或其它磁存储装置、或任何其它介质。基于本文提供的本公开内容和教学，所属技术领域一般人员将明白用于实施各个实施方案的其它方式和 / 或方法。

[0139] 因此，说明书和图应视作以说明之义而非限制之义。然而，将明白在不背离如权利要求书中提出的本发明的更广精神和范围的情况下可以进行本文未描述的各种修改和变化。

[0140] 其它变化是在本公开内容的精神内。因此，虽然所公开的技术具有各种修改和替代结构，但是其特定说明的实施方案已在图中示出并且上文已进行详细描述。然而，应了解并非旨在将本发明限于所公开的特定形式，但相反，本发明涵盖落于如随附权利要求书中定义的本发明的精神和范围内的所有修改、替代结构和等效物。

[0141] 除非本文另有指示或上下文中明确反驳，否则在描述所公开的实施方案的上下文中（尤其在下文权利要求书的上下文中）术语“一”、“一个”和“所述”以及类似参考的使用应解释成涵盖单数和复数。除非另有说明，否则术语“包括”、“具有”、“包含”和“含有”应解释成开放式术语（即，意味着“包括但不限于”）。术语“连接的”应解释成部分或全部包括在内、彼此连接或连接在一起，即使其间存在某些中间物。除非本文另有指示，否则本文的值范围的列举仅旨在用作个别指落于所述范围内的每个单独值的速记方法，并且如果本文个别地列举所述范围，那么每个单独值并入说明书中。除非本文另有指示或上下文中另有明确反驳，否则本文描述的所有方法可以任何合适次序执行。除非另有要求，否则本文提供的任何和所有实例或示例性语言（例如，“诸如”）的使用仅旨在更好地说明本发明的实施方案并且不会对本发明的范围强加限制。说明书中的语言不应解释成指示如实行本发明必需的任何非要求元件。

[0142] 本文描述本公开内容的优选实施方案，包括发明者已知用于实行本发明的最佳模式。所属技术领域一般人员可以在读取前文描述时明白优选实施方案的变化。发明者预期熟练技术人员会适当采用这些变化，并且发明者希望以本文具体描述的方式外的其它方式实行本发明。因此，本发明包括如适用法律允许随附其的权利要求书中列举的标的的所有修改和等效物。此外，除非本文另有指示或上下文中另有明确反驳，否则本发明涵盖上述元件的所有可能变化的任何组合。

[0143] 本文引用的所有参考文献（包括公开、专利申请和专利）是与好像个别和具体指示每个参考文献是以引用的方式并入本文并且在本文提出其全部内容相同的程度以引用的方式并入本文。

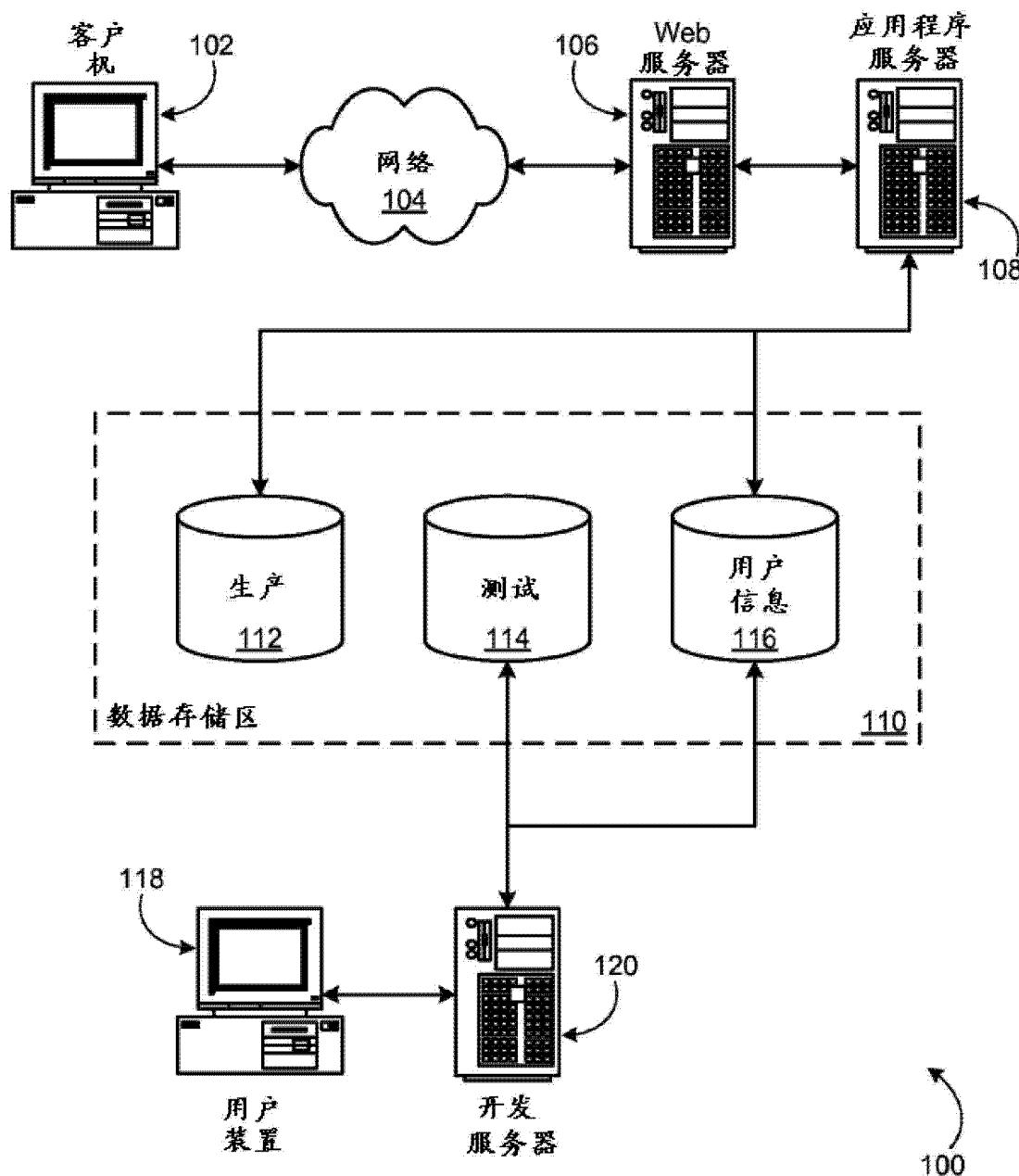


图 1

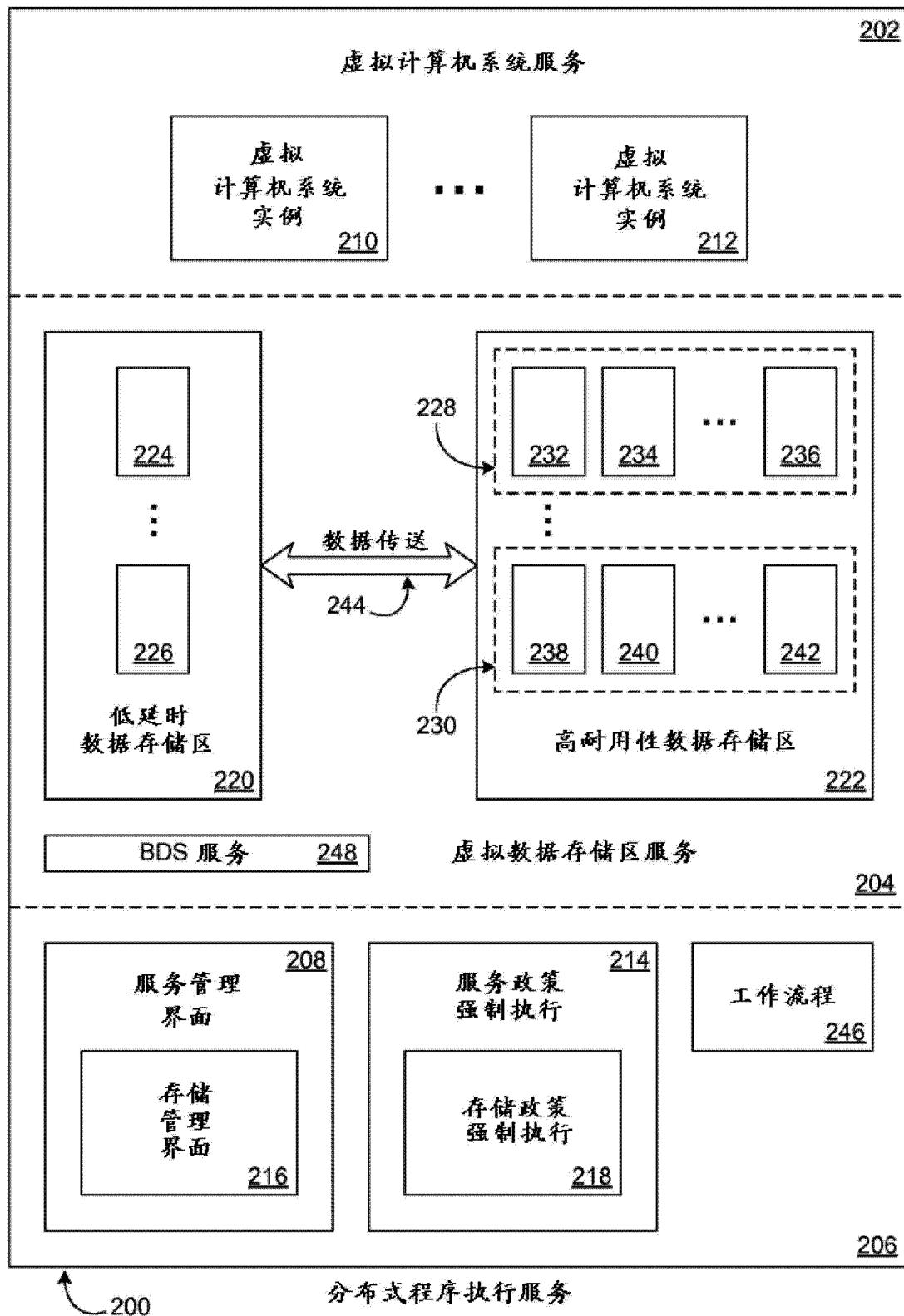


图 2

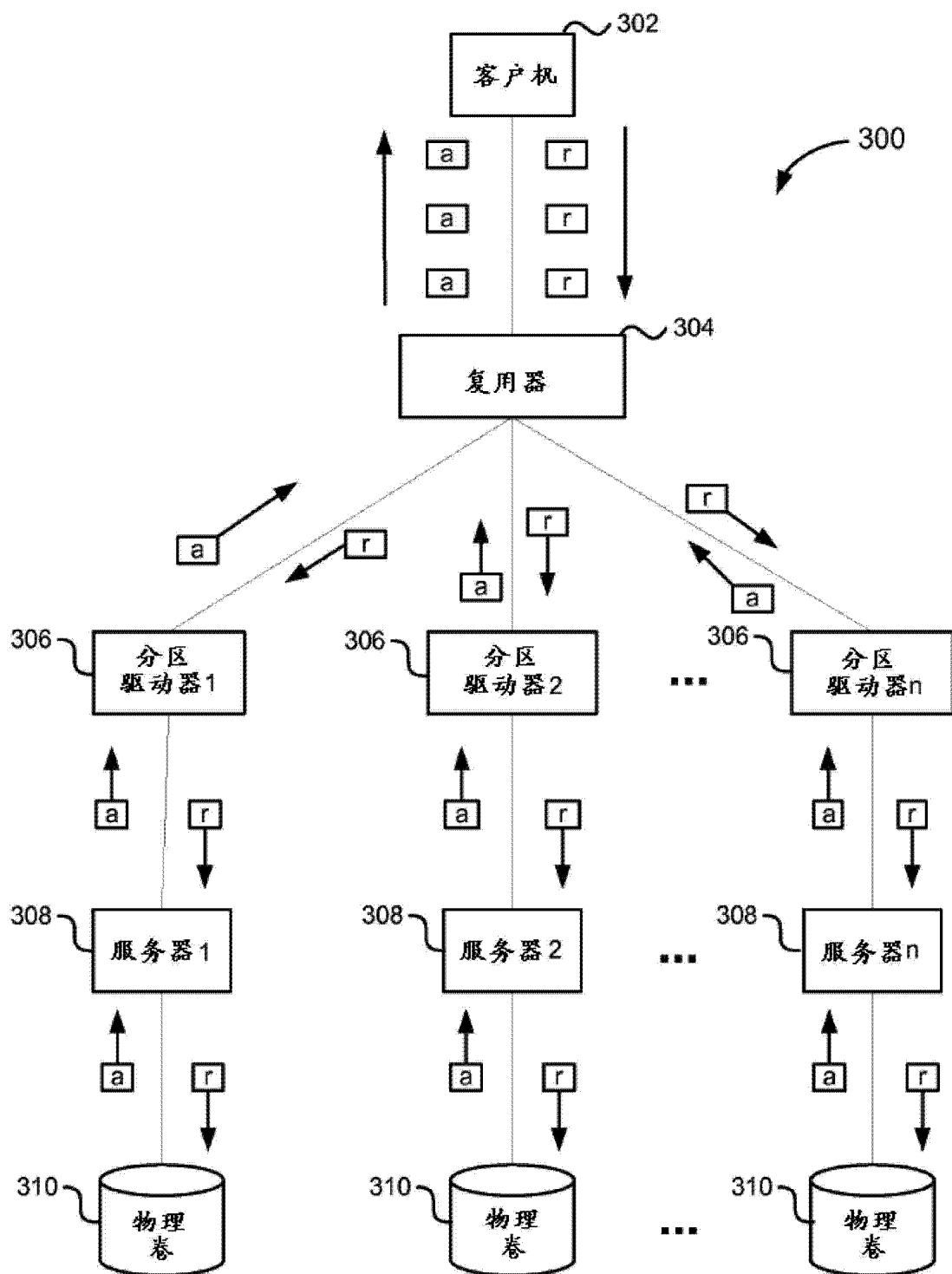


图 3

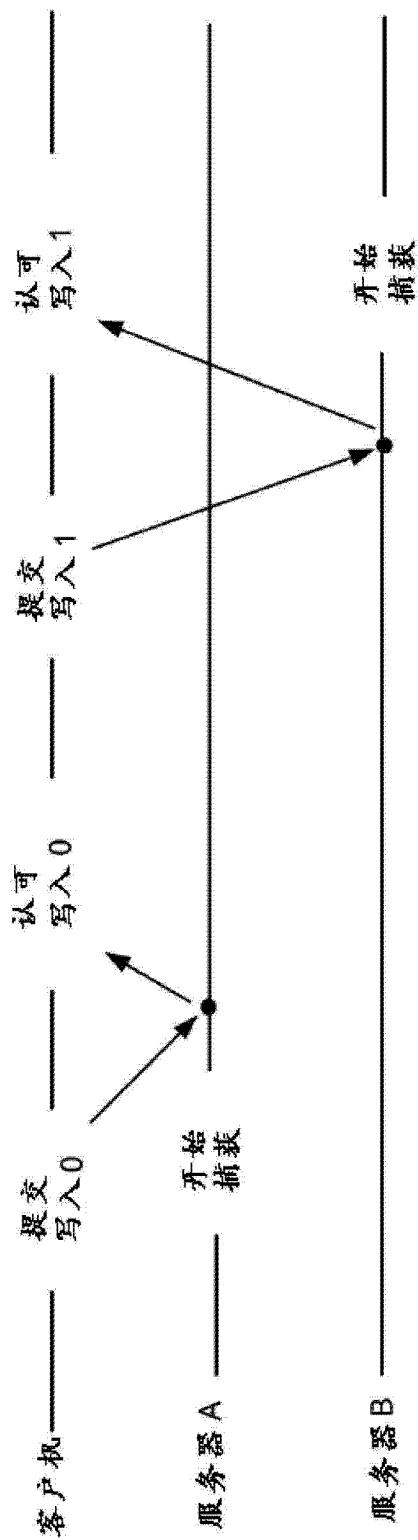


图 4

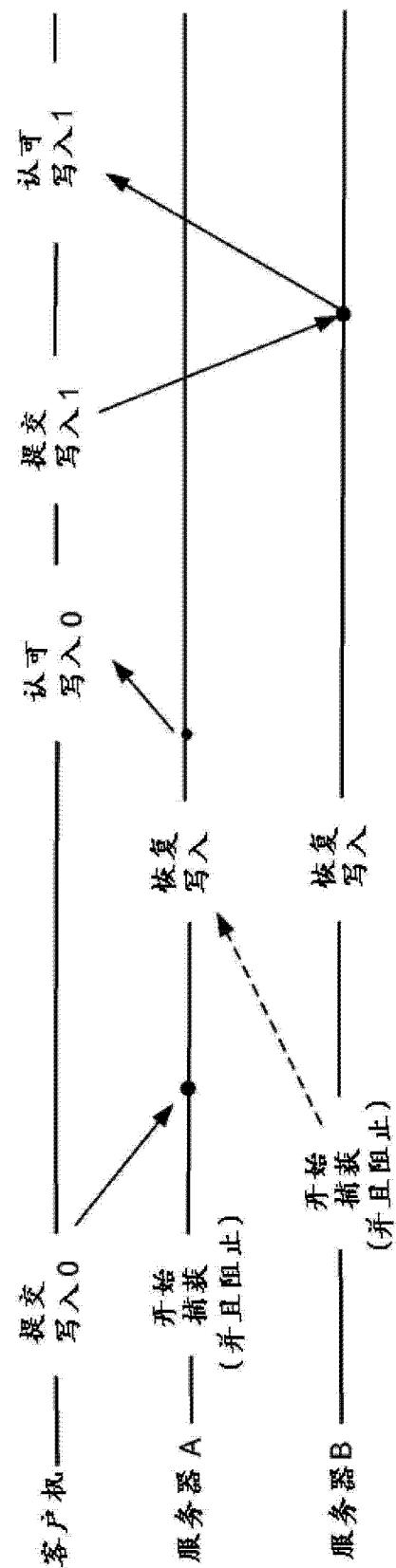


图 5

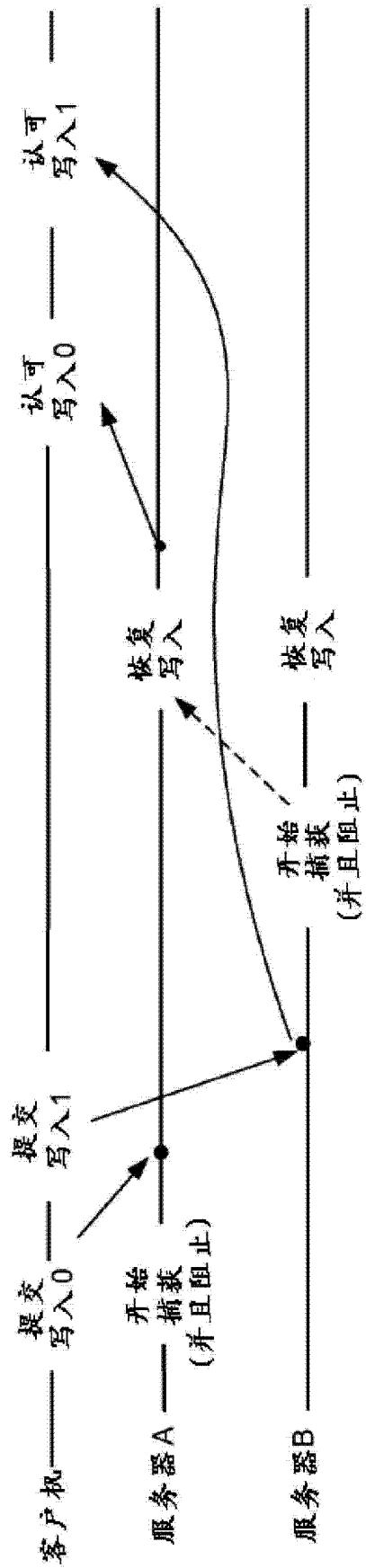


图 6

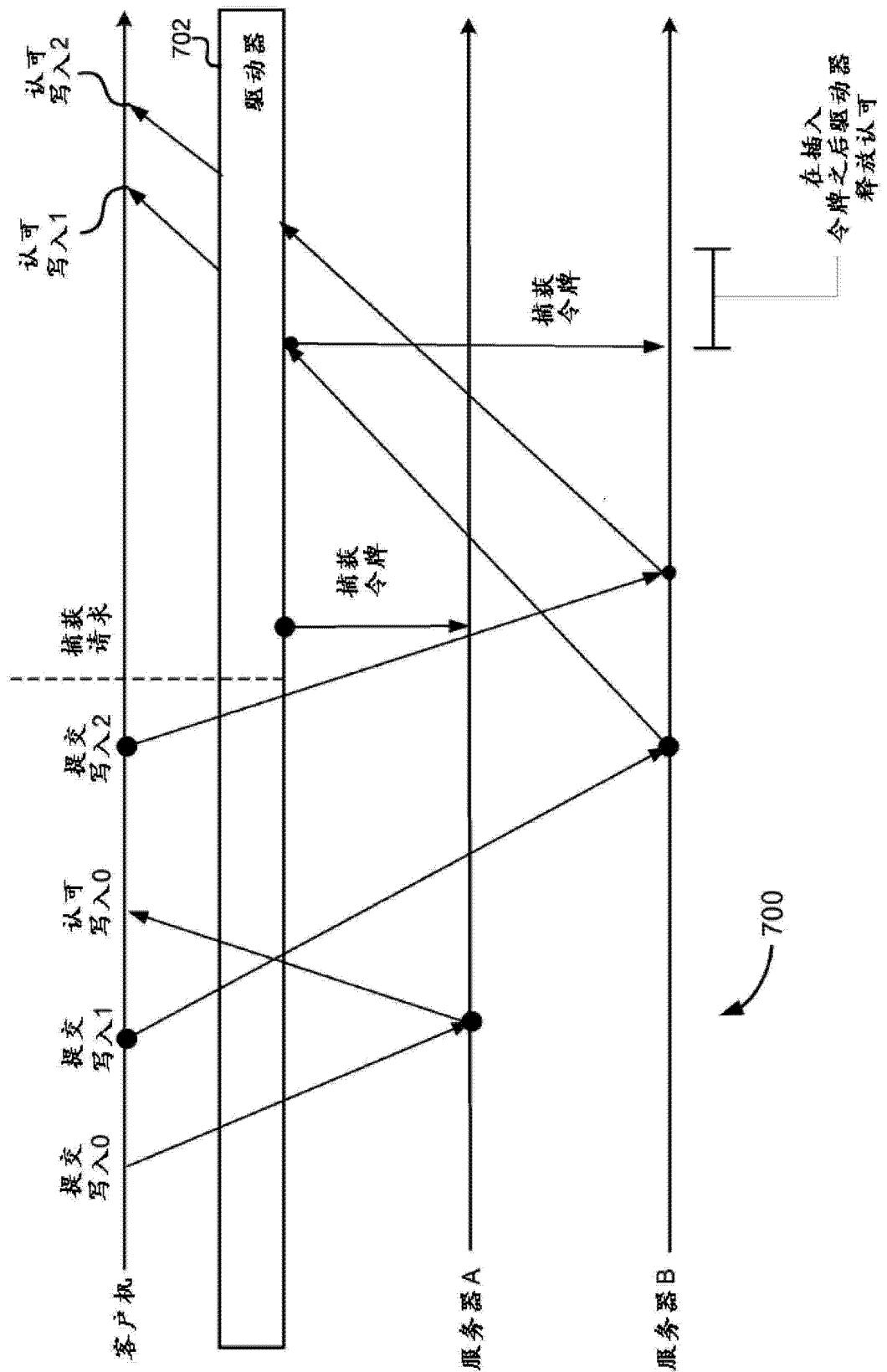


图 7

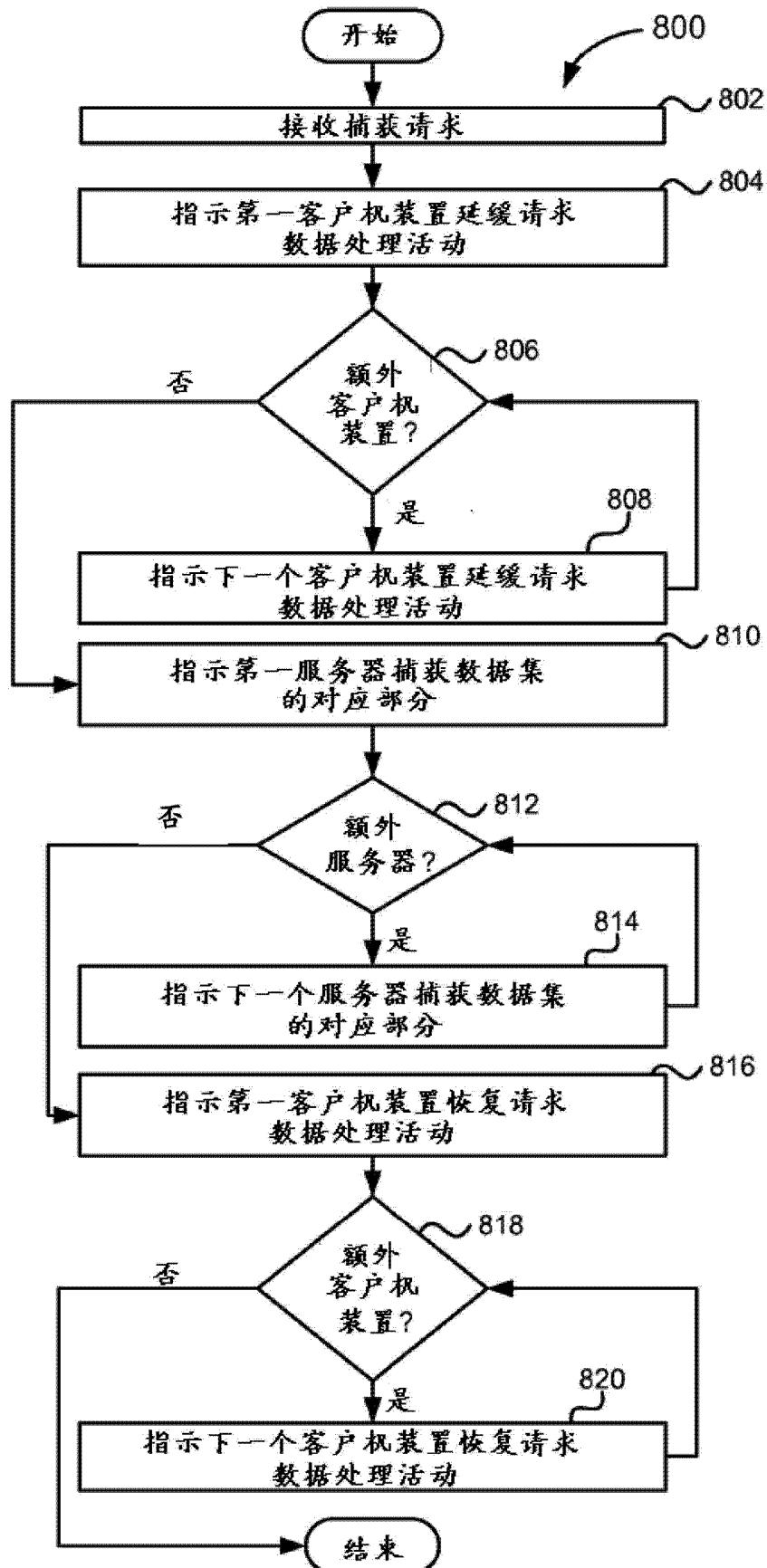


图 8