



(12)发明专利

(10)授权公告号 CN 105683929 B

(45)授权公告日 2018.10.09

(21)申请号 201480060023.X

(22)申请日 2014.05.06

(65)同一申请的已公布的文献号

申请公布号 CN 105683929 A

(43)申请公布日 2016.06.15

(30)优先权数据

IN5044/CHE/2013 2013.11.08 IN

(85)PCT国际申请进入国家阶段日

2016.04.29

(86)PCT国际申请的申请数据

PCT/CN2014/076844 2014.05.06

(87)PCT国际申请的公布数据

W02015/067013 EN 2015.05.14

(73)专利权人 华为技术有限公司

地址 518129 广东省深圳市龙岗区坂田华为总部办公楼

(72)发明人 普拉森纳·文卡特斯·拉玛穆尔蒂

(51)Int.Cl.

G06F 12/16(2006.01)

G06F 12/00(2006.01)

H04L 12/18(2006.01)

(56)对比文件

CN 101729365 A, 2010.06.09,

CN 103384988 A, 2013.11.06,

US 7792987 B1, 2010.09.07,

US 2012054533 A1, 2012.03.01,

审查员 李艳军

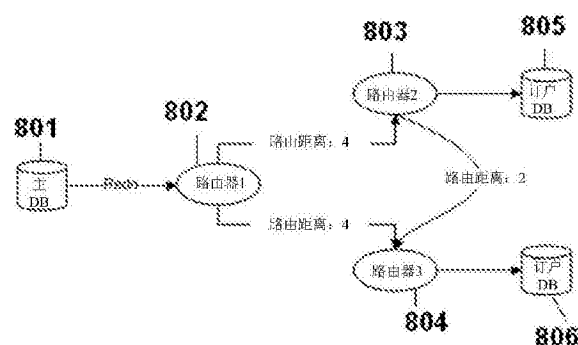
权利要求书3页 说明书10页 附图10页

(54)发明名称

用于数据库和存储器感知路由器的方法和设备

(57)摘要

描述了一种用于优化分布式网络中的复制的方法和系统。本发明允许所述网络中运行的一个或多个路由器确定所述网络的当前集群拓扑；所述路由器识别并优化网络中正使用的数据复制流/服务；所述路由器基于所述集群拓扑确定路由方案；基于所述路由方案对所述网络内的数据包进行路由；所述路由器将预定义的策略应用到对应于所述识别的数据复制流的预定义的路由器组。



基于缓存和拓扑感知的最优路由方法

1. 一种用于优化分布式网络中的复制的方法,所述分布式网络用于承载多个数据包和数据库服务器之间的通信,其特征在于,所述方法包括:

所述网络中运行的路由器确定所述网络的当前集群拓扑;

所述路由器基于所述多个数据包识别并优化所述网络中正使用的数据复制流;

所述路由器基于所述集群拓扑确定路由方案;

基于所述路由方案对所述网络内的数据包进行路由;

所述路由器将预定义的策略应用到预定义的路由器组,所述预定义的路由器组对应于识别的数据复制流,其中,所述预定义的路由器组包括至少一个边缘路由器,最靠近所述数据库服务器运行的边缘路由器为主数据库节点边缘路由器MDER;

所述MDER通过在所述至少一个边缘路由器上开启缓存并将指纹附着到与所述缓存关联的净荷上来优化所述数据复制流。

2. 根据权利要求1所述的方法,其特征在于,所述主数据库节点边缘路由器包含配置的数据,所述配置的数据包括:

数据库节点边缘路由器的互联网协议地址,

数据库集群的唯一标识符,以及

信道和相关的策略配置。

3. 根据权利要求1所述的方法,其特征在于,数据包的路由包括:

第一路由器创建路径及路径标签,以用于通过使用第二路由器的缓存数据将第一数据包路由到第三路由器;

所述第一路由器确定待传输到所述第三路由器的数据包在所述第二路由器中已缓存并使用已缓存的指纹和所述路径标签;

所述第二路由器接收所述指纹和对应的路径标签,处理已缓存的数据包,将所述数据包转换为实际数据,并将所述实际数据转发到所述第三路由器。

4. 根据权利要求1所述的方法,其特征在于,还包括:

所述至少一个边缘路由器在所述网络中通告/发布所述识别的数据复制流;

数据库感知路由器接收并存储所述通告/发布的信息;

所述数据库感知路由器订阅所述主数据库节点边缘路由器MDER;

在所述订阅后将集群中的所述数据库感知路由器添加为订阅数据库边缘路由器SDER;

为每个添加的数据库感知路由器配置对应的SDER;

其中所述主数据库节点边缘路由器MDER在接收订阅请求时用于:

检查是否已订阅所请求的数据库感知路由器或SDER;

在确定已订阅所请求的数据库感知路由器或SDER后在MDER-SDER链路上开启缓存。

5. 根据权利要求1所述的方法,其特征在于,还包括:

订阅者节点对应的主DB节点将REDO日志信息作为数据包发送到该订阅者节点;

所述主数据库节点边缘路由器MDER拦截所述数据包用于确定REDO日志信息;

在确定所述数据包包含REDO日志信息并且未开启缓存后缓存所述数据包。

6. 根据权利要求5所述的方法,其特征在于,包括:

所述主数据库节点边缘路由器通过查找边缘路由器信息提取数据包的目的地址;

为开启缓存的边缘路由器生成所述数据包的指纹并存储和转发所述指纹;

通过标签标记所述数据包并转发所述指纹;以及

当确定已存在指纹时,所述主数据库节点边缘路由器生成所述数据包的标签并仅转发指纹。

7.一种用于优化分布式网络中的复制的系统,所述分布式网络用于承载多个数据包和数据库服务器之间的通信,其特征在于,所述系统包括:

在所述网络中运行的一个或多个路由器,用于:

确定所述网络的当前集群拓扑;

基于所述多个数据包识别并优化所述网络中正使用的数据复制流;

基于所述集群拓扑确定路由方案,以及

将预定义的策略应用到预定义的路由器组,所述预定义的路由器组对应于识别的数据复制流,其中,所述预定义的路由器组包括至少一个边缘路由器,最靠近所述数据库服务器运行的边缘路由器为主数据库节点边缘路由器MDER;

所述MDER通过在所述至少一个边缘路由器上开启缓存并将指纹附着到与所述缓存关联的净荷上来优化所述数据复制流。

8.根据权利要求7所述的系统,其特征在于,所述主数据库节点边缘路由器包含配置的数据,所述配置的数据包括:

数据库节点边缘路由器的互联网协议地址,

数据库集群的唯一标识符,以及

信道和相关的策略配置。

9.根据权利要求7所述的系统,其特征在于,当路由数据包时,所述系统用于:

通过第一路由器创建路径及路径标签,以用于通过使用第二路由器的缓存数据将第一数据包路由到第三路由器;

通过所述第一路由器确定待传输到所述第三路由器的数据包在所述第二路由器中缓存并使用已缓存的指纹和所述路径标签;

通过所述第二路由器接收所述指纹和对应的路径标签,处理已缓存的数据包,将所述数据包转换为实际数据,并将所述实际数据转发到所述第三路由器。

10.根据权利要求7所述的系统,其特征在于,其中:

所述至少一个边缘路由器通告/发布所述识别的数据复制流;

数据库感知路由器用于接收并存储所述通告/发布的信息并订阅所述主数据库节点边缘路由器;

在所述订阅后将集群中的所述数据库感知路由器添加为订阅数据库边缘路由器SDER;

为每个添加的数据库感知路由器配置对应的订阅数据库边缘路由器SDER;

其中所述主数据库节点边缘路由器MDER在接收订阅请求时用于检查是否已订阅所请求的数据库感知路由器或SDER,以及在确定已订阅所请求的数据库感知路由器或SDER后在MDER-SDER链路上开启缓存。

11.根据权利要求7所述的系统,其特征在于,相应的主数据库节点将REDO日志信息作为数据包发送到订阅者节点;所述主数据库节点边缘路由器MDER拦截所述数据包用于确定REDO日志信息;在确定所述数据包包含REDO日志信息并且未开启缓存后缓存所述数据包。

12.根据权利要求11所述的系统,其特征在于,所述系统还用于:

所述主数据库节点边缘路由器通过查找边缘路由器信息提取数据包的目的地址；
为开启缓存的边缘路由器生成所述数据包的指纹并存储和转发所述指纹；
通过标签标记所述数据包并转发指纹；以及
当确定已存在指纹时，所述主数据库节点边缘路由器生成所述数据包的标签并仅转发指纹。

用于数据库和存储器感知路由器的方法和设备

技术领域

[0001] 本发明涉及用于维护网络环境中的分布式存储器的设备和相关方法,具体而言涉及但不限于一种用于优化分布式存储器的网络环境的系统和方法。

背景技术

[0002] 分布式存储器为当今企业最重要的关键需求之一。数据跨地域保存并且维护专用系统,以确保了跨地域的冗余备份。

[0003] 具有分布式结构最重要的原因之一包括但不限于:数据量及其庞大、为了使程序可扩展及利用多系统和多核CPU架构的优势。另一方面,网站服务器需要在全球分布以保证低延迟和故障倒换等。

[0004] 在典型场景中,系统管理员可以跨越多个物理位置分发(例如,数据库中的)数据集。分布式数据库可以位于因特网、公司内网或其他公司网络的网络服务器上。由于分布式数据库存储了多台计算机的数据,分布式数据库通过允许事务在许多机器(而非限制为一台机器)上处理改善了终端用户感知的性能。

[0005] 本发明范围内的分布式存储器可表示信息(例如,以复制方式)存储在多个网络节点上的计算机网络。因此,分布式存储还可表示用户在多个网络节点上存储信息的分布式数据库,或者用户在多个对等网络节点上存储信息的计算机网络。

[0006] 所述领域的普通技术人员可能很了解,在通信网络中,一个节点是一个连接点,表示再分点(redistribution point)或通信端点(例如,一些终端设备)。物理网络节点是连接到网络的有源电子设备,能够通过通信信道发送、接收或转发信息(例如,数据包)。

[0007] 通常复制涉及利用查找分布式数据库中的变化的专门软件模块。一旦识别了变化,复制过程使得所有数据库看起来相同,即包含相似的数据。复制过程取决于分布式数据库的大小和数目,可能繁琐、复杂且耗时。该过程还需要大量计算机资源。

[0008] 因此,长久以来一直存在优化数据库复制以节省内存、CPU周期等宝贵的网络资源的需求。

[0009] 已提出了许多方案来降低由于在网络的多个节点上复制数据而导致的网络资源的损耗。然而,现有方案在具有集群拓扑感知的应用级别或不具有数据库集群拓扑感知的路由器级别(L2/L3)上操作。

[0010] 仅基于“路由器级别”方法论的方案存在很多缺点,包括但不限于由于主节点侧的过多订阅而导致的网络中的冗余。冗余消除技术本身很昂贵,其被统一应用而不考虑信息的性质(即,不是上下文特定的)。

[0011] 如果不了解底层数据库的拓扑,多余的内容会被缓存或者缓存未被最优地使用。

[0012] 此外,传统方案无法基于数据库内容为各种内容提供策略。例如,不可以使用这些方案来实施例如特殊路径,该特殊路径尝试为初始镜像信息最大化MTU(最大传输单元)。

[0013] 因此,长久以来一直存在提供一种用于分布式存储器的优化方法和系统的需求,该方法和系统避免了传统方法呈现的上述缺点及许多其他缺点。确切地说,需要一种用于

分布式存储器的优化方法和系统,该方法和系统可以在任意数据存储方案或任意数据库中
使用并且无论使用何种复制机制都能最优地运行。这种方案应当具有成本效益并且优先使
用现有的路由器基础结构(例如,MPLS、MPLS-TE、CSPF)来实现所需的改进。

发明内容

[0014] 出于总结的目的,本文已描述了一些方面、优点以及新颖的特征。应理解的是,并
不是按照任何一个特定实施例就可以实现所有此类优点。因此,本发明可以以如下方式体
现或执行:实现或优化本文教导的一个优点或一组优点,而无需实现本文可能教导或建议
的其他优点。

[0015] 本发明的一个目的在于消除上述提到的和许多其他缺点,本发明提供了一种通过
使存储器/数据库路由器感知底层的数据存储器/数据库集群拓扑和复制流来优化存储器
和数据库复制的方法和系统。

[0016] 本发明的典型实施例中的系统和对应方法提供了一种路由器(用户对其进行最低
配置)确定存储器/数据库复制拓扑,以及利用拓扑感知装备路由器以优化数据存储器/数
据库复制流的机制。

[0017] 本发明的实施例提供了路由器对底层网络的集群拓扑的确定。

[0018] 一种用于优化分布式网络中的复制的方法和系统,该网络用于承载多个数据包和
数据库【服务器】以及用于多个数据包和数据库【服务器】之间的通信,所述方法和系统包括
所述网络中运行的一个或多个路由器确定所述网络的当前集群拓扑;所述路由器识别并优
化网络中正使用的数据复制流/服务;所述路由器基于所述集群拓扑确定路由方案;基于所
述路由方案对所述网络内的数据包进行路由;所述路由器将预定义的策略应用到对应于所
述识别的数据复制流的预定义的路由器组。

[0019] 在本发明的另一实施例中,路由器使用一组边缘路由器确定所述网络的当前集群
拓扑,最靠近所述数据库服务器运行的边缘路由器为数据库节点边缘路由器。

[0020] 在本发明的另一实施例中,数据库节点边缘路由器包含配置的数据,包括数据库
节点边缘路由器的互联网协议地址、数据库集群的唯一标识符、信道和相关的策略配置。

[0021] 在本发明的又一实施例中,主边缘路由器通过在边缘路由器上开启缓存并将指纹
附着到与所述缓存关联的净荷上来优化数据复制流。

[0022] 在本发明的又一实施例中,本发明上下文内的数据包的路由包括第一路由器创建
路径和标签用于通过使用第二路由器的缓存数据将第一数据包路由到第三路由器,第一路
由器确定待传输到第三路由器的数据包在所述第二路由器中缓存并使用已缓存的指纹和
所述路径标签;第二路由器接收所述指纹和对应的标签,处理所述接收的数据包,将所述数
据包转换为实际数据,并将所述实际数据转发到第三路由器。

[0023] 在本发明的又一实施例中,所述方法和系统用于通过所述边缘路由器在所述网络
中通告/发布所述识别的数据复制流/服务;通过数据库感知路由器接收并存储所述通告/
发布的信息;通过所述数据库感知路由器订阅主数据库边缘路由器;在所述订阅后将集群
中的所述数据库感知路由器添加为SDER(订阅数据库边缘路由器);为每个添加的数据库感
知路由器配置对应的SDER(订阅数据库边缘路由器),其中主数据库边缘路由器(MDER)在接
收订阅请求时用于检查是否已订阅所请求的数据库感知路由器或SDER;以及在确定已订阅

所请求的数据库感知路由器或SDER后在MDER-SDER链路上开启缓存。

[0024] 在本发明的又一实施例中,所述方法和系统用于通过订阅者节点对应的主DB节点将REDO日志信息作为数据包发送到订阅者节点;通过MDER(主数据库边缘路由器)拦截所述数据包用于确定REDO日志信息;在确定所述数据包包含REDO日志信息并且未开启缓存后缓存所述数据包。

[0025] 在本发明的又一实施例中,所述方法和系统用于通过主数据库边缘路由器使用查找边缘路由器信息提取数据包的目的地址;为开启缓存的边缘路由器生成所述数据包的指纹并存储和转发指纹数据和包;给所述数据包贴上标签并转发指纹;以及当确定已存在指纹时通过主数据库边缘路由器生成所述数据包的标签并仅转发指纹。

[0026] 在本发明的另一实施例中,基于DB集群确定的拓扑的主数据库边缘路由器呈现出对缓存的最优利用。

[0027] 在本发明的另一实施例中,本发明提供了多个路由器对复制流或重做流的优化。

[0028] 在本发明的另一实施例中,描述了一种用于对包进行路由和缓存的方法。

[0029] 本发明的实施例中公开的内容提供了一种系统(和方法),为从DB(数据库操作)角度表达的不同性质的流应用特定策略。

[0030] 参照本发明的本文中所述的所有实施例适用于所述方法和所述对应系统。

[0031] 所述领域的技术人员将参照附图,从实施例的详细描述中易明了本发明的这些和其他实施例,本发明不限于所公开的任意特定实施例。

附图说明

[0032] 为了更好地理解本文所述的系统和方法的实施例,并为了更清楚地示出如何实现这些实施例,将通过示例对附图进行参照,其中:

[0033] 图1示出了现有的单主复制机制。

[0034] 图2示出了现有的多主复制。

[0035] 图3示出了传统复制中的冗余。

[0036] 图4示出了用于解决分布式网络中的过多订阅的现有组播方案的图解表示。

[0037] 图5示出了用于冗余消除的现有方案的图解表示。

[0038] 图6示出了用于包级冗余检测的现有方案的图解表示。

[0039] 图7示出了示例性分布式/复制数据库。

[0040] 图8示出了基于缓存和拓扑感知的最优路由方法。

[0041] 图9示出了示例性MDER(主DB节点边缘路由器)中配置的数据。

[0042] 图10示出了MDER(主DB节点边缘路由器)在分布式网络中进行通告和订阅的流程图。

[0043] 图11示出了REDO缓存优化的流程图。

[0044] 图12示出了用于REDO缓存的序列图。

[0045] 图13示出了基于缓存的最优路由的示例性网络图。

[0046] 图14示出了基于缓存的最优路由的路径创建。

[0047] 图15示出了地理分布式多主DB集群。

具体实施方式

[0048] 现将参考附图描述示例性实施例。但是,本发明可以具体体现为许多不同的形式且不应被解释为仅限于本文所阐述的实施例;提供这些实施例是为使本发明详尽完整并向本领域技术人员充分传达本发明的范围。在附图中所示的特定的示例性实施例的详细描述中使用的术语并非意在限制。在附图中,相同的数字指相同的元件。

[0049] 说明书在若干位置可能涉及“一种”、“一个”或“一些”实施例。这并不一定意味着每个这样的引用都是相同的实施例,或者该特征仅应用于单个实施例。也可以组合不同实施例的单独的特征以提供其他实施例。

[0050] 除非另外明确规定,否则本文所用的单数形式“一”和“所述”包括其复数形式。应进一步了解,本说明书中所用的术语“包括”用于说明存在所述特征、整数、步骤、操作、元件和/或部件,但并不排除存在或添加一个或多个其他特征、整数、步骤、操作、元件、部件和/或它们的组合。应理解,当某个元件被称作“连接到”或“耦合到”另一个元件时,它可以直接连接或耦合到另外的元件,或者可能存在中介元件。此外,本文所用的“连接到”或者“耦合到”可包括可操作地连接或耦合。本文所用的术语“和/或”包括相关联的所列项目中的一者或多者的任何和所有组合和安排。

[0051] 除非另有定义,否则本文所用的所有术语(包括技术和科学术语)都具有与本领域普通技术人员公知的含义相同的含义。应进一步理解,术语例如常用的字典中定义的那些术语应被解释为具有与它们在相关领域和本发明中的含义一致的含义,并且不应以理想化或过于正式的方式被解释,除非在本文中明确地如此定义。

[0052] 附图描绘了仅示出一些元件和功能实体的简化结构,所有这些元件和功能实体都是逻辑单元,其实施方式可以不同于所示出的。所示出的连接是逻辑连接;实际的物理连接可以不同。对本领域的技术人员显而易见的是该结构还包括其他功能和结构。应当理解,在通信中使用的功能、结构、元件和协议与本发明无关。因此,这里不需要更详细地讨论它们。

[0053] 另外,附图中所述及所绘的所有逻辑单元包括单元发挥其功能所需的软件和/或硬件部件。进一步地,每个单元自身包括隐含理解的一个或多个部件。这些部件可操作地互相耦合并用于互相通信以执行所述单元的功能。

[0054] 本发明所公开的系统中提供的特征可在一个或多个实施例中通过远程无线,和/或通过无线网络访问。此类无线网络服务提供商操作并维护计算系统和环境,例如服务器系统和架构。通常,服务器架构包括提供无线网络服务的基础机构(例如,硬件、软件和通信线路)。在某些实施例中,本发明的实施例中的操作可以通过空中接口(on-air interface)或空口无线执行。

[0055] 基本上,本文所述的操作为网络、设备、计算机或机器执行的操作,或者在一些实施例中结合与计算机和机器进行交互的操作员或用户执行的操作。本文所述的程序、模块、过程、方法、数据等仅仅是示例性实施方式,不涉及或限于任意特定计算机、装置或计算机语言。相反,各种类型的通用计算机或设备可根据本文所述的启示与所构建的程序一起使用。

[0056] 应理解本发明的实施例可包括在各种类型的通信网络中,这些通信网络在本发明的范围内,但不限于实施例。术语“网络”和“系统”经常互换使用。

[0057] 当前,现有机制有利地为给定的一组边缘路由器提供了一种机制以根据一些最低配置确定数据库集群拓扑。最靠近主DB(数据库)节点的边缘路由器为“主DB节点边缘路由器”或简言之“MDER”。创建由多个网络节点(例如,路由器)组成的集群时配置MDER(“主DB节点边缘路由器”)。路由器可提供CLI(命令行视图)或SNMP(简单网络管理协议)或其他配置机制来启动该配置。MDER(“主DB节点边缘路由器”)中配置的数据包括:

[0058] a.主DB(数据库)节点的IP地址

[0059] b.数据库集群的唯一标识符

[0060] c.信道和相关的策略配置

[0061] 因此,本发明范围内的边缘路由器用于检测数据存储器网络、复制数据库集群、分布式数据库和分布式数据存储器的拓扑。众所周知,路由器为一种转发计算机网络之间的数据包的设备,创建了叠加互连网络。路由器通常可连接到来自不同网络的两个或多个数据线路。当数据包来自其中一个线路时,路由器读取包中的地址信息来确定其目的地。随后,路由器使用其路由表或路由策略中的信息将包引导至其行程中的下一网络。因此,路由器在因特网上进行“流量引导”功能。

[0062] 本发明范围内的所属领域的普通技术人员注意到本发明范围内的集群由一组松散连接或紧密连接的计算机组成,这些计算机一起工作并且通常被视为单个系统。

[0063] 一旦配置了集群,MDER(“主DB节点边缘路由器”)使用合适的通告/发布机制在整个网络通告/发布该服务。通告方式之一为使用IETF(工程任务组)草案“draft-keyupdate-bgp-services-02”中引入的BGP(边界网关协议)通告/发布机制。

[0064] 本发明还提供了一种用于服务通告/发布的框架,该框架在后面的段落中详述。

[0065] 为了清晰起见并有效地理解本发明的技术贡献,本文提供了少量提出的现有方案,参照附图简要地描述了这些方案。

[0066] 图1示出了现有的单主复制机制。本文中,对主数据库(101)进行的事务(104)分别通过事务(105)和(106)传播至订户数据库(102)和订户数据库(103)。因此,冗余副本在订户数据库(102)和(103)中保存。

[0067] 图2示出了现有的多主复制机制。此处,可对多个主数据库(201、202、203)进行维护(同步且独立更新)。对特定主数据库进行的任意更新需要传播至剩余主数据库以保持一致性。例如,当主数据库(201)随事务(204)更新时,主数据库(201)将更新传播至剩余主数据库(202、203)。剩余主数据库也会如此这样将其更新传播至所有其他主数据库。

[0068] 图3示出了现有/传统复制中的冗余。此处,主数据库(301)具有两个订户数据库(302、303)。对主数据库(301)进行的任意更改/更新需要传播至所有主数据库的订户(302、303)。因此,无论何时对主数据库(301)进行更新,主数据库(301)向订户数据库(302)和订户数据库(303)提供两份更新副本。这两份副本通过多个中间路由器在分布式网络中传播。因此,发往主数据库的一份更新副本在每个路由器处复制用于向订户数据库(302和303)递送,该复制过程使用了宝贵的网络资源并增加了多种类型的冗余。

[0069] 因此,又一次维护了分布式网络中的冗余。

[0070] 图4示出了用于解决分布式网络中的过多订阅的现有组播方案的图解表示。

[0071] 到目前为止,尝试避免现有分布式网络中的冗余包括解决数据路由问题并集中于数据库复制。只提出了部分方案。这些方案可分类为

- [0072] 尝试减少网络中的过多订阅和冗余
- [0073] 尝试针对特定内容的路由
- [0074] 图5示出了用于冗余消除的现有方案的图解表示。
- [0075] 类似地,图6示出了用于包级冗余检测的现有方案。
- [0076] 然而,当这些方案(如图5和图6中所示)应用到一般的数据库复制优化时,这些方案存在以下但不限于以下的缺点:
- [0077] 高计算成本,因为它们尝试减少“字节”级的冗余,因此会大大增加边缘路由器的负载;
- [0078] 这些方案未利用传输数据的性质,因为这些方案对所有数据应用统一优化,统一优化在实际的分布式网络中是次优的;
- [0079] 由于数据库集群的拓扑未知,这些方案无法提供数据库感知的路由方案。
- [0080] 避免过度订阅的基本方案为使用组播方案。然而,组播方案仅可用于UDP(用户数据报协议)。大多数复制实施方案基于TCP(传输控制协议),这些复制实施方案不含最优的组播方案。UDP(用户数据报协议)组播方案还缺乏可靠性。另外,使用组播方案需要在数据库实施中进行大量更改,这使得方案变得尤为复杂。
- [0081] 图7示出了本发明的示例性分布式/复制数据库。
- [0082] 本发明的一个主要方面涉及由在本发明的分布式网络中运行的路由器确定集群拓扑。这种确定可利用IETF(因特网工程任务组)进行标准化的现有机制,例如包括但不限于“使用BGP(边界网关协议)的服务通告,draft-keyupate-bgp-services-02”
- [0083] 本发明通过使用上述提及的机制,允许路由器使用BGP协议通告(即,发布)服务。因此,典型实施例中的本发明使用该机制来发布数据库复制服务。上述提及的机制仅仅是与本发明的目标一致的用于通告的示例性机制。
- [0084] 之后,最靠近“主DB节点”的边缘路由器配置有数据库复制信息。基于该配置,路由器使用BGP(或其他合适的机制)在整个分布式网络上公告/发布该服务。
- [0085] 类似地,当订户数据库节点增至该集群时,与该节点最靠近的边缘路由器还配置为将该角色标记为订户。基于该配置,路由器订阅主数据库路由器。通过该机制路由器能够了解底层的数据库集群拓扑。
- [0086] 本发明的另一方面涉及由路由器对复制流(例如在数据库复制情况下需要对重做流进行优化)进行优化。
- [0087] 如本文档的背景技术部分所提及的,多订户DB(数据库)集群中存在大量冗余数据。为每个订户数据库(DB)复制相同的数据。主数据库(DB)节点的边缘路由器通过在边缘路由器(其附着的订户数据库(DB)多于一个)上开启缓存来优化这些相同的数据。随后,对于相同数据的进一步传输,使用缓存中的一个副本。本发明关注的焦点问题在于,通常在重做流中,由于网络中存在订户,所以所有包需要重传多次。主DB(数据库)节点的边缘路由器通过使用该机制不需要使用“滑动窗口”缓存技术(类似拉宾指纹),但可以使用类似MD5的固定净荷缓存。随后,该指纹附着到净荷,并在所有路由器上进行同步。
- [0088] 本质上,各种路由器-路由器链路上存在选择性地、自动开启的缓存。
- [0089] 参照图7,主DB节点(701)保存了数据库的主副本。通过主DB路由器(702)可访问主DB节点(701)。主DB路由器(702)具体配置有主DB信息。另外,主DB路由器(702)进一步用于

通过使用BGP通告发现DB集群拓扑。

[0090] 实际上,主DB路由器(702)首先缓存发送的REDO但是有利地发送第二REDO的指纹,从而节省珍贵的网络资源,

[0091] 进一步地,订户DB路由器(703)和(707)配置有订户DB信息。订户DB路由器(703)和(707)订阅主DB路由器服务。

[0092] 图8示出了基于缓存和拓扑感知的最优路由方法。

[0093] 本发明的该方面涉及不同拓扑的路由概念。典型拓扑之一为多主数据库集群。作为示例性实施例,数据库集群中存在3个节点,所有节点均允许写入。这时候,每个节点中的每个变化将同步到剩余两个节点以保持底层的分布式网络的一致性。分布式网络可采用现有“网状”或“星型”方案中的任意一者。

[0094] 在任一方案中,提出了与本发明的目标一致的多个缓存以及基于这些缓存的路由重新计算。

[0095] 在不考虑底层拓扑的情况下,当存在多个订户但每个订户附着到不同的边缘路由器时,简单的缓存技术是不够的。

[0096] 在如图8适当所示的场景中,以及通过缓存技术,还需要对数据包进行合适地路由。因此,当主DB节点路由器识别DB集群的拓扑为上述性质时,主DB节点路由器计算到每个节点的路径。基于该计算,DB集群在可以使用缓存的网络中建立路径。在示例性实施例中,这些路径可以通过RSVP-TE(针对流量工程扩展的资源预留协议)或MPLS-TE机制建立。

[0097] 图8示出了本发明的该实施例,该实施例涉及一些特定条件下的包的实际路由和缓存。通常,路由器1使用路由器2中的缓存数据创建路径以及用于向路由器3进行路由的标签,反之亦然。

[0098] 当路由器1确定包在路由器2中缓存并且包需要立刻传输到路由器3时,路由器1使用已缓存的指纹以及该路径标签。路由器2在接收指纹和特定标签时处理包并将数据包转换为实际数据并随后将实际数据转发到路由器3。

[0099] 该技术特征本质上节省了“2个单元距离”,否则这些就是冗余数据。

[0100] 本发明的另一实施例涉及为从DB操作角度表达的不同性质的流(例如,“重做流(REDO stream)”、“全镜像流(Full Mirror stream)”、“优先级流(Priority Stream)”)应用特定策略。

[0101] 数据库感知路由器配置有各种连接参数以及策略。一个示例性实施例包括具有最大MTU(最大传输单元)的初始镜像流的连接以及压缩。数据库感知路由器使用提及的策略创建标签交换路径。这种标签交换路径通过CSPF和LDP实施。主数据库节点边缘路由器在流传输中确定流的性质并应用合适的策略。

[0102] 本质上,创建基于路由器配置的最大MTU(最大传输单元)路径以允许路由该路径上的初始镜像(完全同步)包。

[0103] 图9示出了示例性MDER(主DB节点边缘路由器)中配置的数据。

[0104] 图10示出了MDER(主DB节点边缘路由器)在分布式网络中进行通告和订阅的流程图。

[0105] 与本发明一致的由MDER(主DB节点边缘路由器)进行的通告和订阅包括但不限于以下步骤。

[0106] 起初MDER(主DB节点边缘路由器)在配置有复制服务时将该服务通告给整个分布式网络。任意DB感知路由器接收该通告并存储通告信息。当DB感知路由器配置有订户信息及相应的集群信息时,DB感知路由器向MDER订阅服务。此时,该节点成为SDER(订户数据库边缘路由器)。示例性实施例中的订阅可以通过XMPP、Jabber或SIP呈现(RFC-3265)等订阅协议之一来处理。

[0107] 对于每个添加到集群中的订户DB节点,配置相应的SDER(订户数据库边缘路由器)并且该SDER订阅服务。因此,在实施例中,如果SDER(订户数据库边缘路由器)配置有2个订户DB节点,那么该SDER将订阅两次服务。MDER(主DB节点边缘路由器)在接收订阅时检查该SDER(订户数据库边缘路由器)是否已订阅。如果该SDER(订户数据库边缘路由器)已订阅该服务,那么MDER在MDER-SDER链路上开启缓存。

[0108] 在另一实施例中,主数据库将REDO日志信息发送给其订户。MDER(主DB节点边缘路由器)知道REDO日志的概念。MDER(主DB节点边缘路由器)拦截来自主DB节点的包并尝试确定该包是否为REDO日志包。在典型实施例中的所述确定可通过配置IP地址和端口号实现。

[0109] 此处,特定的(源IP+目的IP+目的端口)还可与信道关联并且该信道可与REDO日志策略关联。或者,MDRR可配备有REDO消息的报文格式并且通过使用DPI(深度报文检测)技术可以确定包的性质。因此,路由器对数据复制流量流性质的确定以及策略的选择性应用是本发明的关键技术特征。

[0110] 参照图10解释通告和订阅步骤。主DB边缘路由器(1001)配置了复制(主)服务。之后,复制服务由主DB边缘路由器通告/发布(1002)。相应的订户DB边缘路由器存储复制服务通告(1003)并配置复制(订户)服务(1004)。

[0111] 之后,订户DB边缘路由器确定是否已存在这种服务的通告/发布。如果存在这种服务的通告,那么由订户DB边缘路由器发送订阅并且由主DB边缘路由器接收该订阅(1007)。

[0112] 主DB边缘路由器在接收所述订阅(1007)时确定路由器是否已订阅(1008),如果已订阅,开启缓存(1009)并随后存储订阅详情(1011)。如果未订阅,存储订阅详情(1011)(而不开启缓存)。

[0113] 图11示出了REDO缓存优化的流程图。

[0114] 此处,一旦MDER(主数据库节点边缘路由器)确定包为REDO包,MDER尝试对该包进行最优缓存。如果该MDER-SDER路径上存在多个订阅者节点,那么该包必定重复。MDER检查缓存是否在该MDER-SDER路径上开启。如果开启,那么MDER继续如以上流程图所述处理包。要注意的另一关键点在于给缓存的包贴上标签。该标签可以与SDER预协商或由SDER/MDER静态地配置。一旦SDER接收具有该标签的包,SDER将指纹转换为实际包并将实际的包转发到订户。需要周期性地调整主-S1和主-S2之间的TCP流。如果不调整,那么由于网络流可能具有偏置,并且缓存失败。

[0115] 参照图11,MDER(主DB节点边缘路由器)确定从主数据库接收(1102)的数据包是否为REDO数据(1103)。如果接收的数据为REDO数据,从数据包中提取目的地址并将目的地址与查找边缘路由器信息(1105)进行匹配。如果确定还没有为边缘路由器缓存数据包,转发数据包(1108)。

[0116] 否则,如果已经为边缘路由器缓存数据包,为整个包生成指纹(1107)。如果指纹已经存在(1109),存储指纹和数据包(1110)并转发这两者(1111)。

[0117] 如果指纹已不存在,给数据包贴上标签并仅转发指纹。

[0118] 在本发明的范围内对数据包进行指纹识别是一种使用包的特点创建唯一密钥的方法,通过该方法可以在缓存中唯一确定包。

[0119] 因此,对IP包/数据包进行优化缓存基于开启缓存。对物理复制日志(DB中的REDO)以及逻辑复制日志(DB中的SQL)进行缓存。

[0120] 图12示出了参照图11如上所述的用于优化的REDO缓存的相应序列图。

[0121] 参照图12的序列图,主数据库(DB)向主DB边缘路由器发送第一REDO-1(1200)。主DB边缘路由器随后缓存所述接收到的第一REDO-1(1202)。之后,主DB边缘路由器将第一REDO-1以及指纹转发到订户数据库(DB)路由器(1203)。订户数据库路由器将REDO-1包转发到订户数据库(DB)(1)。同时,(由于不知道底层拓扑的主数据库存在两个订户数据库,)主数据库(DB)将第二REDO-1发送到主DB边缘路由器(1201)。主DB边缘路由器随后提取指纹(1204)并将指纹(1205)转发到订户数据库(DB)路由器(1205)。之后,订户数据库路由器将REDO-1转发到订户数据库2(1207)。

[0122] 图13示出了基于缓存的最优路由的示例性网络图。

[0123] 在本发明的典型实施例中,MDER在确定DB(数据库)集群的拓扑之后尝试最优地使用缓存。如果新SDER订阅到MDER并且SDER仅拥有一名订户,MDER尝试找出是否存在任意已订阅的邻近SDER。如果存在,MDER尝试创建“交换路径”用于优化MDER。

[0124] 在图13中,发往S1(订户1)的第一REDO通过RE发送(路由器2)。标记该包进行缓存。当相同的REDO被发送到S2(订户2)时,指纹切换到R2(路由器2)。随后,R2(路由器2)检索实际的数据包并将该数据包转发到R3(路由器3),R3随后将该数据包转发到S2(订户2)。

[0125] 确切地说,参照图13,本发明的分布式网络的示例性实施例包括主数据库(M)、两名订户(S1、S2)、一个边缘路由器(R1)以及两个订户路由器(R2、R3)。

[0126] 主数据库(M)为其两个订户(S1、S2)分别发送一个REDO(1301)。边缘路由器(R1)接收这两个REDO,但是仅为订户2(S2)将指纹发送给路由器2(R2)。

[0127] 路由器R2随后为S1(订户1)发送REDO并且还将一个单独的REDO转发给路由器3(R3)。实际上,路由器3(R3)将REDO转发给订户2(S2)。

[0128] 图14示出了基于缓存的最优路由的示例性路径创建。

[0129] 在该实施例中,使路由器知道多种类型的DB复制流量,流量在DB(数据库)节点之间流动。每种类型的流量示为一个信道。数据复制流量可以为常规REDO(同步)流量或初始(或完整)镜像或优先级包(特定类型的DDL)。对于每种类型的流量,在路由器中配置多种策略。本发明的路由器在路由器中创建规则,例如开启压缩的完全冗余消除。MDER还创建特殊路径或标签以适配这些信道。对于初始(或完整)镜像,路由器创建具有最大MTU(最大传输单元)的路径。类似地,MDER为优先信道创建到目的地的另一标记的路径。MDER从来不为优先级信道开启压缩。

[0130] MDER(主数据库边缘路由器)检测信道的机制包括但不限于IP+Port配置或使用DPI。

[0131] 另外,DB(数据库)节点主动与路由器进行交互以提供多种输入,这样路由器可以最优操作以建立拓扑。

[0132] 通过实施配置或SNMP或Netconf等管理协议可配置路由器管理流。

[0133] 因此,本质上数据库节点与边缘路由器交互以帮助边缘路由器确定各种包的性质,路由器可以很有效地优化这些包。

[0134] 另外,数据库节点可提供复制拓扑洞察,例如当数据库节点确定对等节点服务中断时,随后数据库节点向路由器指示该信息。这使得数据库对拓扑的实际情况作出更快地响应。

[0135] 本发明的技术方案有利地消除了参照本发明的图4所述的现有系统/方法的这种缺点以及更多缺点。图4本质上示出了本发明的示例性系统架构。

[0136] 因此,本发明配置了数据库节点和存储器节点与边缘路由器交互的能力以提供正在复制的数据的“性质”的清晰视界,这样路由器可以更好地优化数据。

[0137] 同样,数据库节点和存储器节点还可用于与边缘路由器交互以输入集群拓扑中的变化,这样路由器可以进行适当的调整。

[0138] 包括但不限于本发明的路由器的示例性节点的典型架构参照图15进行解释。

[0139] CPU总线(1402)本质上为所有子系统连接的互联线。一般而言,每次仅一对设备可相互对话,所以必须协调总线的通信以防止消息碰撞。这种协调通常由CPU(1401)处理。

[0140] 中央处理器(CPU)(1401)执行存储器(1403)中包含的指令。这些指令以计算机时钟(1404)指定的速率执行。

[0141] CPU(1401)需要访问两种不同类型的存储器(1403)以执行程序。微控制器中使用了两种类型的存储器,即只读存储器(ROM)(1405)和随机存取存储器(RAM)(1406)。

[0142] 在微控制器中,只读存储器(ROM)(1406)用于存储永久程序、运行的驱动程序以及数据。许多微控制器使用可擦除可编程只读存储器(EPROM)或电可擦除可编程只读存储器(EEPROM)来存储程序、运行的驱动程序以及数据。EPROM和EEPROM为非易失性存储器。

[0143] 随机存取存储器或RAM(1406)用于临时存储数据和指令。

[0144] 本发明的路由器和数据库节点(404)的相关部件选择性地包括:

[0145] 信号控制单元(未示出):

[0146] 设备:主要包括CPU+存储器中的软件+rf部分,用于控制设备中的带宽使用。

[0147] 业务提供商网络:主要包括服务器+存储器中的软件+rf部分,用于控制网络中的带宽使用。

[0148] 存储器单元:

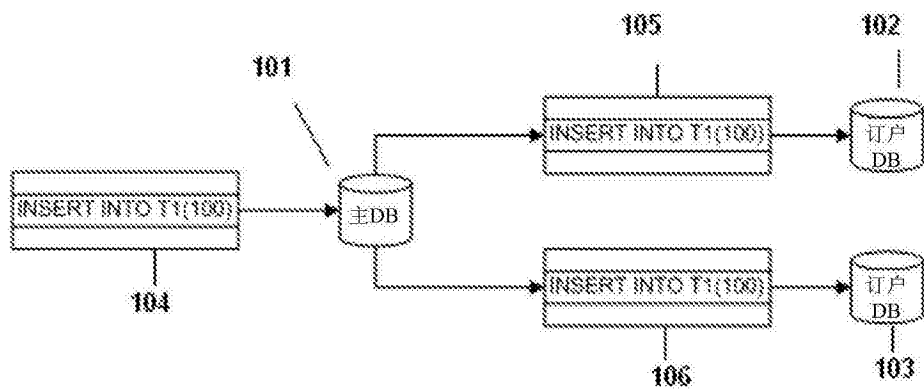
[0149] 设备:主要包括存储器,用于存储软件和数据,这些数据与一个或多个服务/任务/操作相关联并且由所述信号控制单元收发。

[0150] 业务提供商网络:主要包括存储器,用于存储软件和数据,这些数据与一个或多个服务/任务/操作相关联并且由所述信号控制单元收发。

[0151] 信号处理单元:

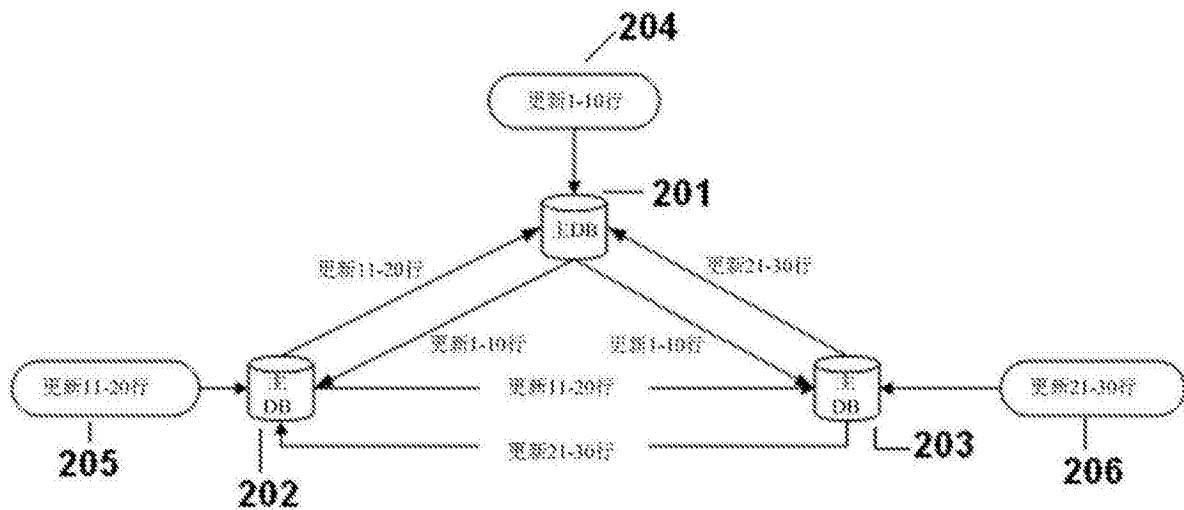
[0152] 设备:主要包括CPU+存储器中的软件+扬声器,用于处理短切换触发数据脉冲信号以通过重新调用来自设备的相应服务内存之后向扬声器进行输出来完成操作并向网络供应商/操作员进行确认。

[0153] 业务提供商网络:主要包括服务器和存储器中的软件,用于处理短切换触发数据脉冲信号以通过向设备发送信号来完成操作并完成来自设备的确认。



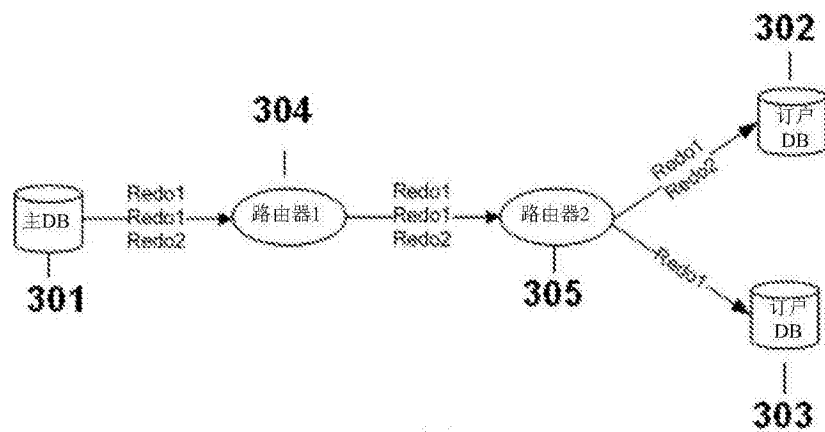
当前单主复制

图1



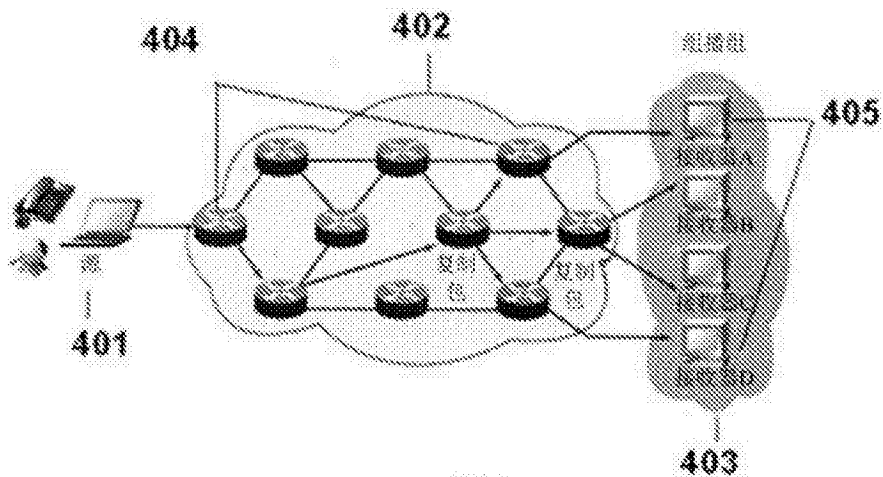
当前多主复制

图2



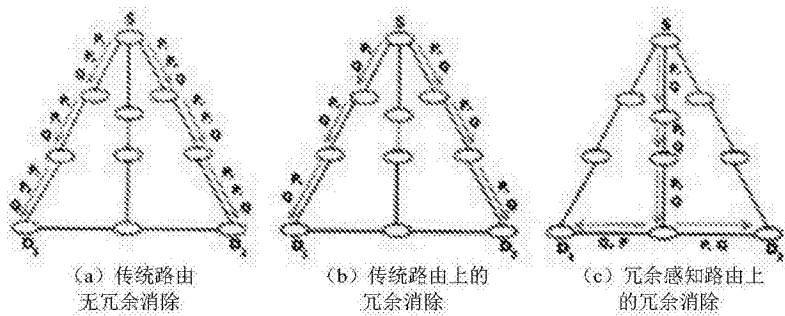
传统复制中的冗余

图3



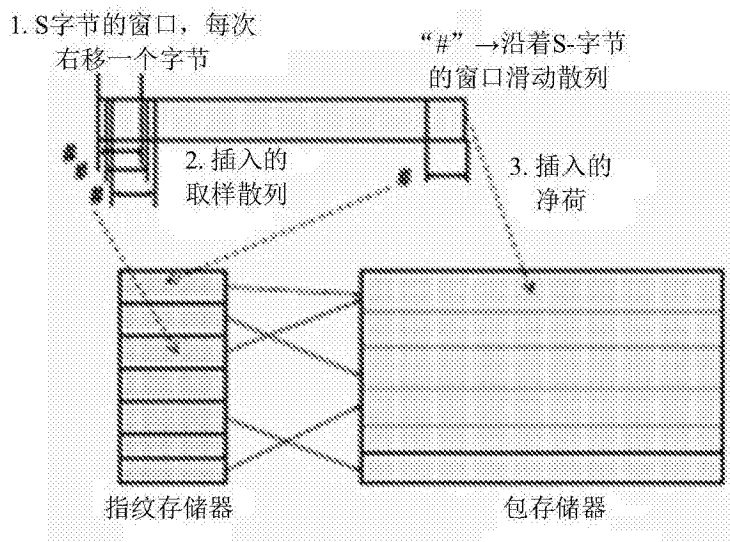
用于解决过度订阅的组播

图4



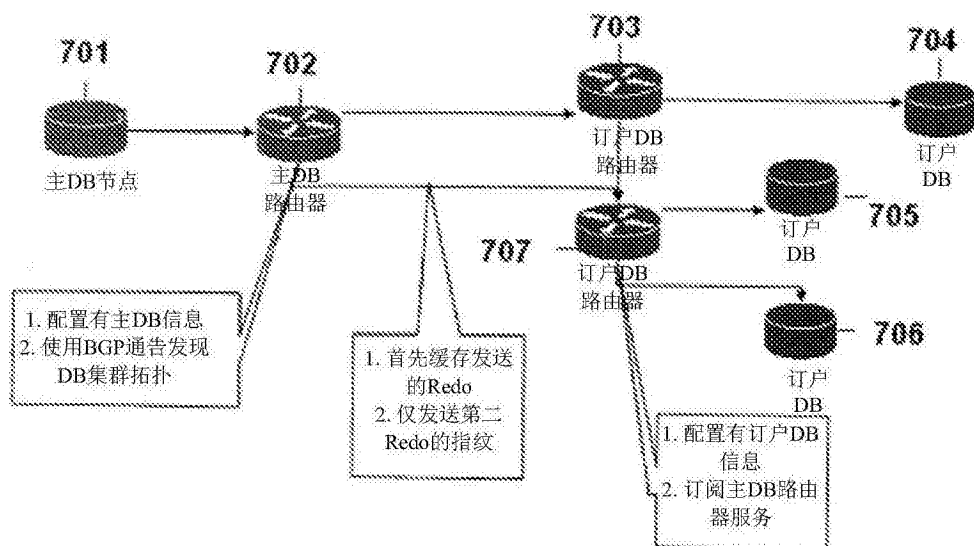
冗余消除

图5



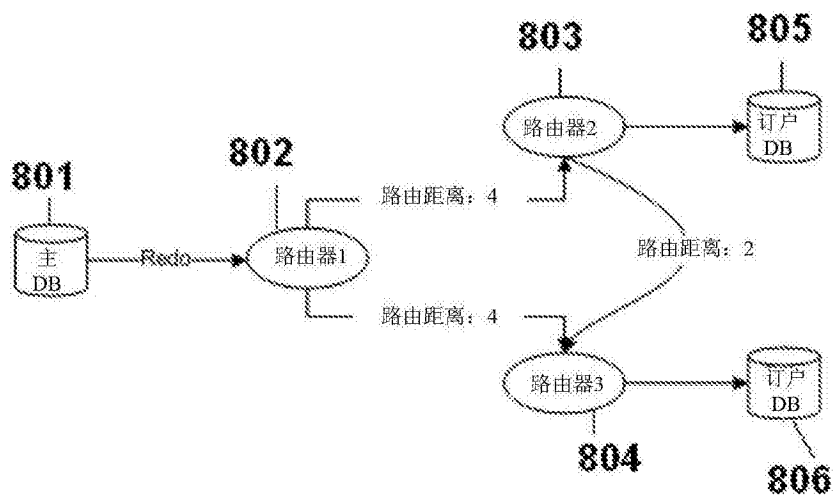
包级冗余检测

图6



示例性复制数据库

图7



基于缓存和拓扑感知的最优路由方法

图8



示例性MDER（主数据库边缘路由器）中配置的数据

图9

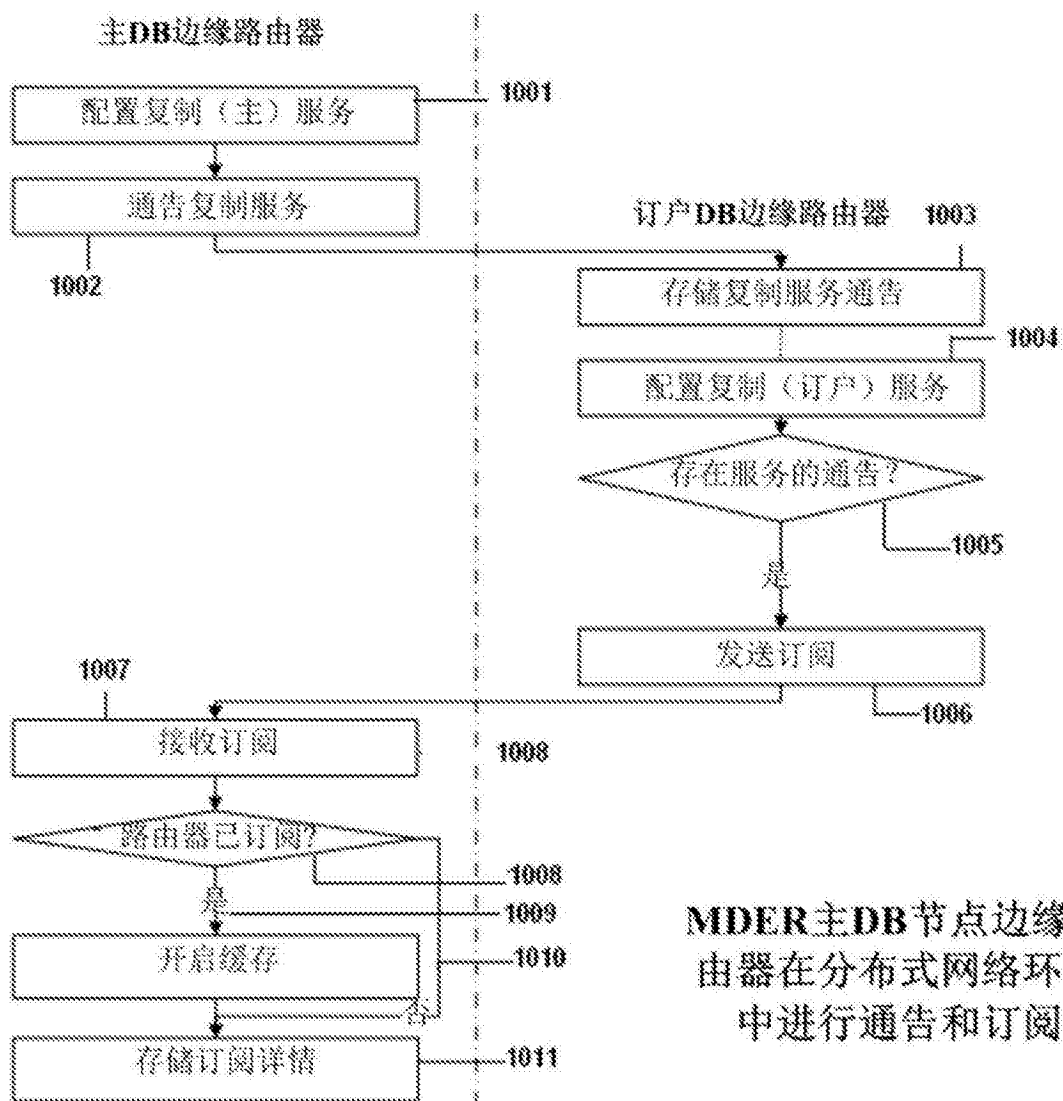
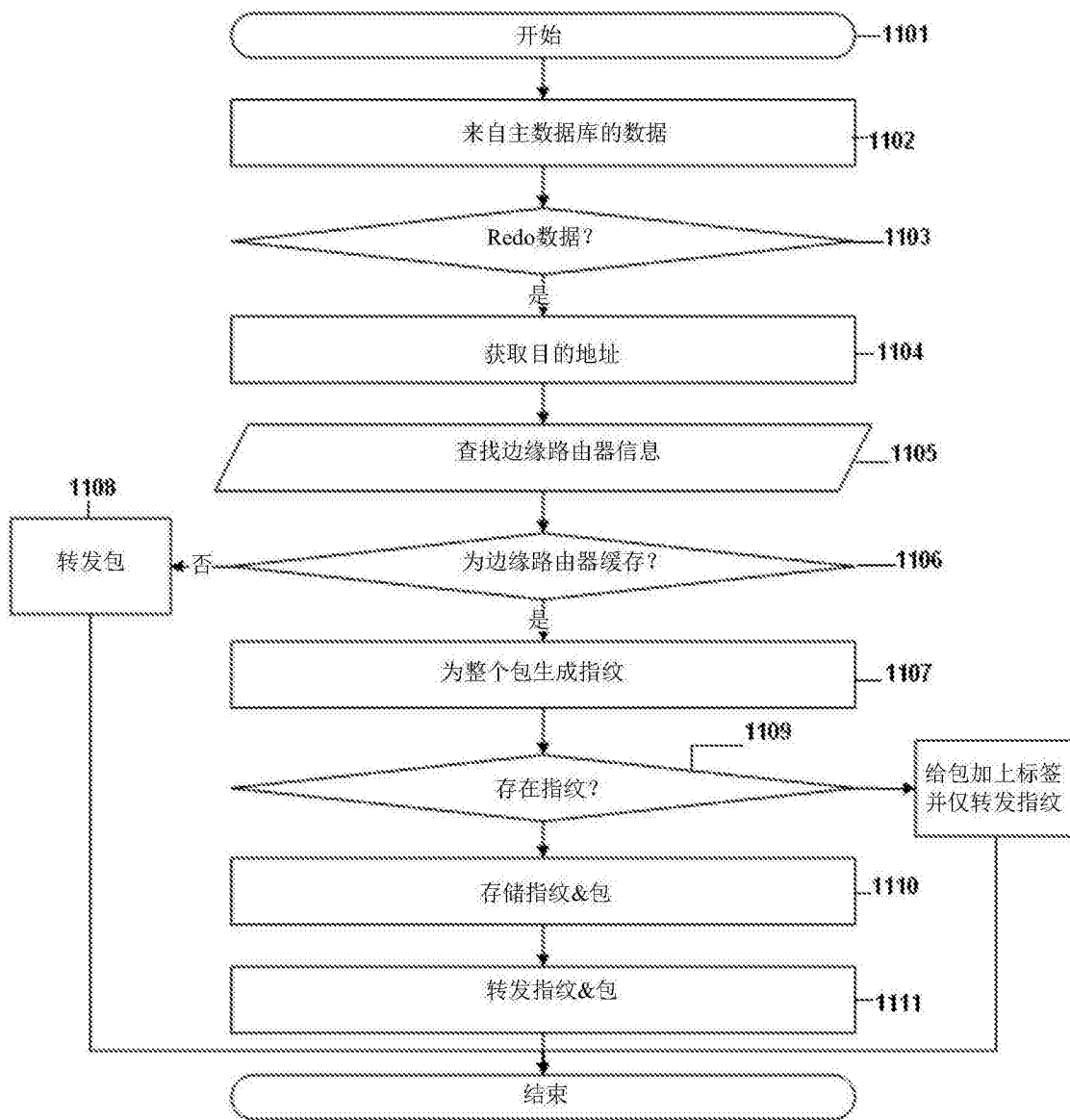
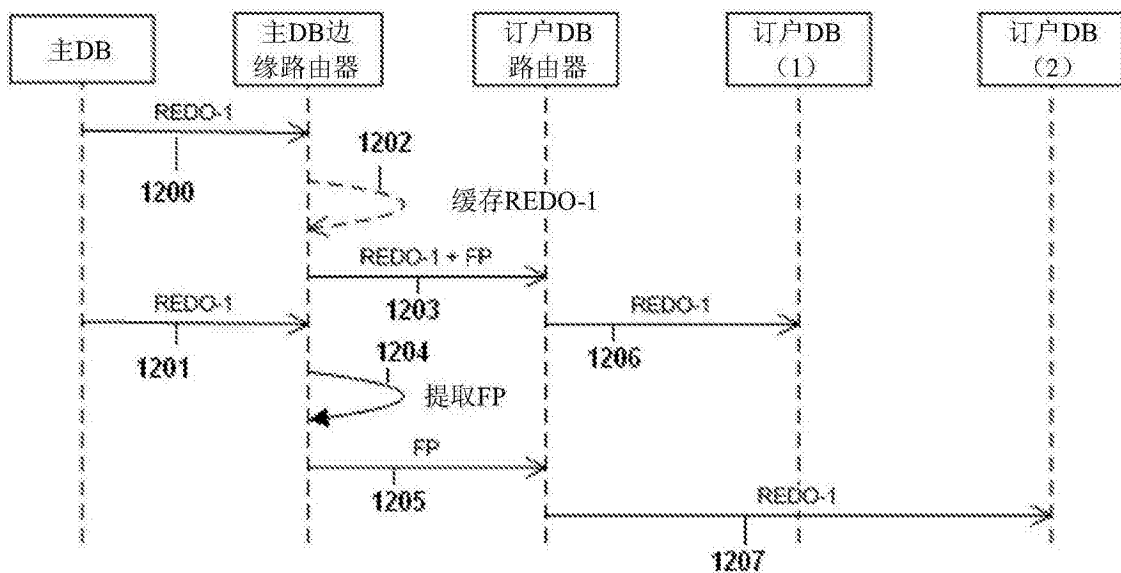


图10



REDO缓存优化

图11



REDO缓存的序列图

图12

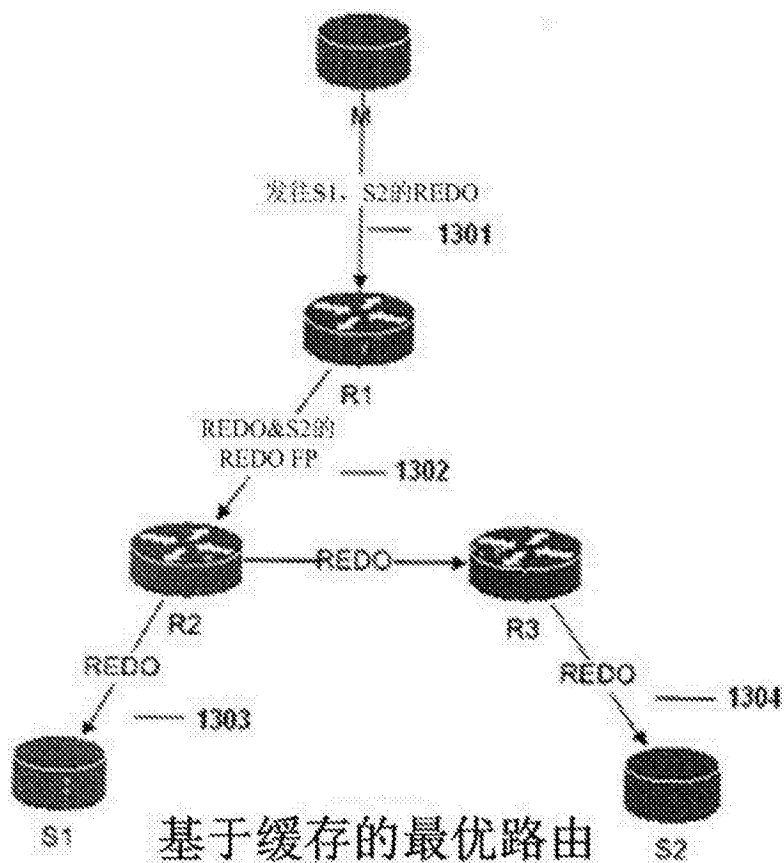
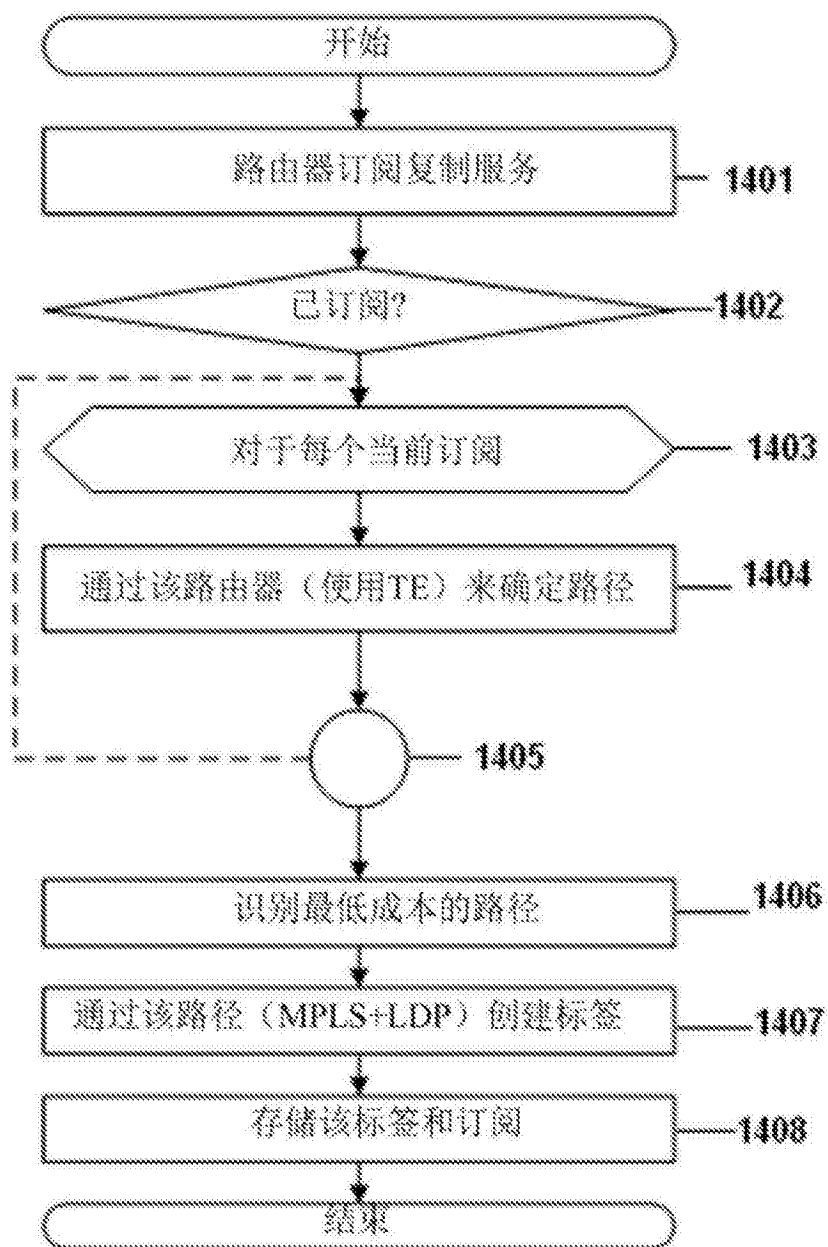
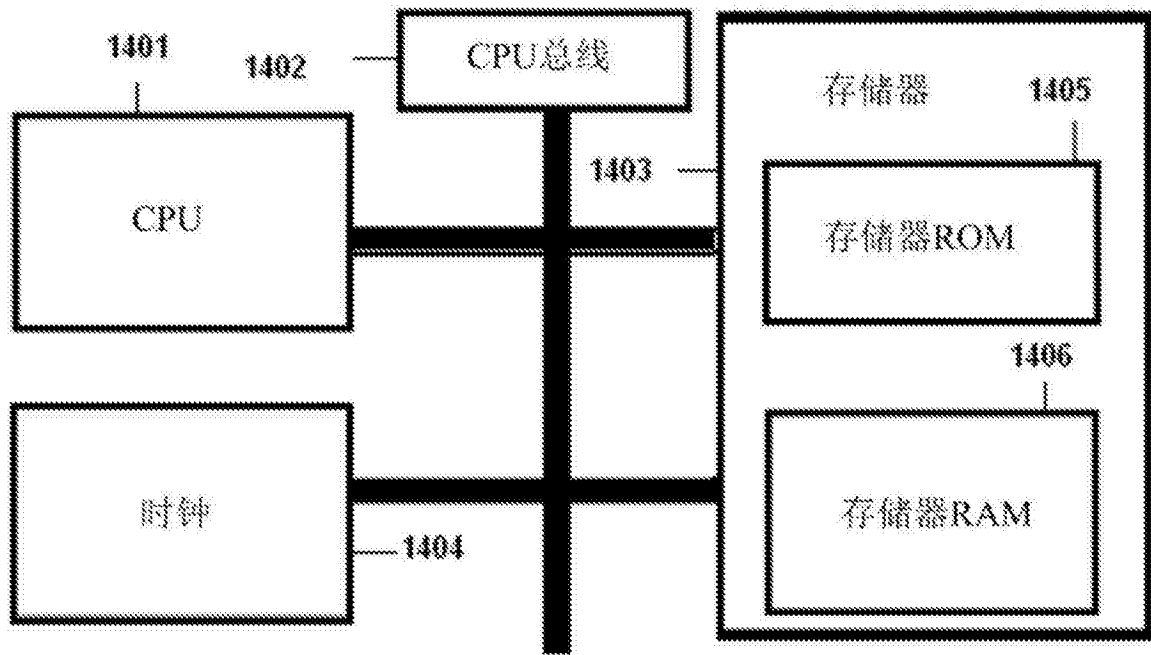


图13



用于基于缓存的最优 路由的路径创建

图14



路由器设备和数据库节点
的示例性硬件架构

图15