(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 103975318 B
(45)授权公告日 2017.06.09

(21)申请号 201280059770.2
(22)申请日 2012.12.05
(65)同一申请的已公布的文献号
申请公布号 CN 103975318 A
(43)申请公布日 2014.08.06
(30)优先权数据
61/567,282 2011.12.06 US
13/677,922 2012.11.15 US
(85)PCT国际申请进入国家阶段日
2014.06.05
(86)PCT国际申请的申请数据
PCT/US2012/067923 2012.12.05
(87)PCT国际申请的公布数据
WO2013/085981 EN 2013.06.13
(73)专利权人 博科通迅系统有限公司
地址 美国加利福尼亚
(72)发明人 R·赫格兰 A·萨巴
S·阿迪拉朱
(74)专利代理机构 中国国际贸易促进委员会专利商标事务所 11038
代理人 陈新
(51)Int.Cl.
G06F 15/16(2006.01)
(56)对比文件
US 2006218210 A1,2006.09.28,
CN 1577314 A,2005.02.09,
CN 1906593 A,2007.01.31,
US 7197661 B1,2007.03.27,
CN 1812342 A,2006.08.02,
US 6687222 B1,2004.02.03,
审查员 黄文琪

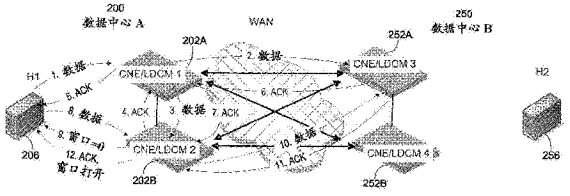
权利要求书3页 说明书12页 附图18页

(54)发明名称

用于镜像设备的无损连接故障切换

(57)摘要

WAN优化设备延迟ACK，直到在打开TCP窗口的同时从目标实际接收到ACK。当ACK被接收并转发时，TCP窗口的大小被减小。如果存在镜像的WAN优化设备，则起始WAN优化设备发送数据报通过WAN，并同时发送数据报到镜像WAN优化设备。当镜像WAN优化设备对镜像的数据报ACK时，起始WAN优化设备对主机ACK。当通过WAN的ACK被接收时，镜像WAN优化设备获得转发的ACK并删除镜像的数据报。在失去设备时，TCP连接转换到镜像WAN优化设备，镜像WAN优化设备关闭LAN TCP窗口并发送所有未经ACK的数据。接着，一经发送成功，镜像WAN优化设备重新打开LAN TCP窗口。



CN 103975318 B

1.一种广域网(WAN)通信方法,包括:

在主控广域网(WAN)设备处从源接收第一TCP数据;

由主控WAN设备转发所述第一TCP数据到WAN;

由所述主控WAN设备转发所述第一TCP数据到镜像WAN设备,所述镜像WAN设备与所述主控WAN设备在一位置处并行;

由所述镜像WAN设备存储所述第一TCP数据;

由所述镜像WAN设备提供对所述第一TCP数据的第一ACK;

在所述主控WAN设备处从所述镜像WAN设备接收对所述第一TCP数据的所述第一ACK;

在从所述镜像WAN设备接收到所述第一ACK之后,从所述主控WAN设备提供第二ACK给所述源;

在所述主控WAN设备处从所述WAN接收对所述第一TCP数据的第三ACK;

由所述主控WAN设备将从所述WAN接收到的所述第三ACK转发到所述镜像WAN设备;和

在接收到转发的、从所述WAN接收到的第三ACK之后,由所述镜像WAN设备删除所述第一TCP数据。

2.根据权利要求1所述的方法,进一步包括:

在来自所述主控WAN设备的TCP连接的控制变化之后,在所述镜像WAN设备处从所述源接收第二TCP数据;

由所述镜像WAN设备关闭针对所述源的TCP窗口;

由所述镜像WAN设备将从所述主控WAN设备接收到的TCP数据转储到所述WAN;

在所述镜像WAN设备处接收用于从所述镜像WAN设备转储的所述TCP数据的第四ACK;和

在接收到用于所转储的TCP数据的所述第四ACK之后,由所述镜像WAN设备打开针对所述源的TCP窗口。

3.一种广域网(WAN)通信方法,包括:

在用作主控的第一广域网(WAN)设备处从源接收第一TCP数据;

由第一WAN设备转发所述第一TCP数据到WAN;

由所述第一WAN设备转发所述第一TCP数据到用作镜像的第二WAN设备,所述第二WAN设备与所述第一WAN设备在一位置处并行;

在所述第一WAN设备处从所述第二WAN设备接收对所述第一TCP数据的第一ACK;

在从所述第二WAN设备接收到所述第一ACK之后,从所述第一WAN设备提供第二ACK到所述源;

在所述第一WAN设备处从所述WAN接收对所述第一TCP数据的第三ACK;和

由所述第一WAN设备将从所述WAN接收到的所述第三ACK转发到所述第二WAN设备。

4.根据权利要求3所述的方法,进一步包括:

由用作镜像的所述第一WAN设备存储从用作主控的第三WAN设备转发的第二TCP数据,所述第三WAN设备与所述第一WAN设备在一位置处并行;

由所述第一WAN设备提供对所述第二TCP数据的第四ACK;

接收从所述WAN接收到的并由所述第三WAN设备转发的针对所述第二TCP数据的第五ACK;和

响应于所转发的第五ACK,由所述第一WAN设备删除所述第二TCP数据。

5.根据权利要求4所述的方法，进一步包括：

在来自用作主控的所述第三WAN设备的TCP连接的控制变化之后，在用作镜像的所述第一WAN设备处从所述源接收第三TCP数据；

由所述第一WAN设备关闭针对所述源的TCP窗口；

由所述第一WAN设备将从所述第三WAN设备接收到的TCP数据转储到所述WAN；

在所述第一WAN设备处接收用于从所述第一WAN设备转储的所述TCP数据的第六ACK；和

在接收到用于所转储的TCP数据的所述第六ACK之后，由所述第一WAN设备打开针对所述源的TCP窗口。

6.根据权利要求5所述的方法，其中，所述第二WAN设备和所述第三WAN设备是同一设备。

7.根据权利要求4所述的方法，其中，所述第二WAN设备和所述第三WAN设备是同一设备。

8.一种广域网WAN设备，其是在第一位置处的多个并行的WAN设备中的第一WAN设备，所述WAN设备包括：

用于耦合到在所述第一位置处的所述多个并行的WAN设备中的其它WAN设备、耦合到WAN以及耦合到TCP数据的源的多个网络端口；

耦合到所述多个网络端口的处理器；和

耦合到所述处理器并存储软件的存储器，所述软件使得所述处理器执行以下操作：

在用作主控时，响应于从源接收到第一TCP数据，转发所述第一TCP数据到所述WAN；

转发所述第一TCP数据到所述多个并行的WAN设备中的用作镜像的第二WAN设备；

响应于从所述第二WAN设备接收到对所述第一TCP数据的第一ACK，提供第二ACK到所述源；和

响应于从所述WAN接收到对所述第一TCP数据的第三ACK，转发从所述WAN接收到的所述第三ACK到所述第二WAN设备。

9.根据权利要求8所述的WAN设备，其中，所述软件进一步使得所述处理器执行以下操作：

在用作镜像时，存储从所述多个并行的WAN设备中的用作主控的第三WAN设备转发的第二TCP数据；

提供对所述第二TCP数据的第四ACK；和

响应于接收到从所述WAN接收到的并由所述第三WAN设备转发的用于所述第二TCP数据的第五ACK，删除所述第二TCP数据。

10.根据权利要求9所述的WAN设备，其中，所述软件进一步使得所述处理器执行以下操作：

在用作镜像时，响应于在来自作为主控的所述第三WAN设备的TCP连接的控制变化之后从所述源接收到第三TCP数据，关闭针对所述源的TCP窗口；

将从所述第三WAN设备接收到的TCP数据转储到所述WAN；和

响应于接收到用于所转储的TCP数据的第六ACK，打开针对所述源的TCP窗口。

11.根据权利要求10所述的WAN设备，其中，所述第二WAN设备和所述第三WAN设备是同一设备。

12.根据权利要求9所述的WAN设备，其中，所述第二WAN设备和所述第三WAN设备是同一设备。

13.一种网络，包括：

在第一位置处的多个并行的广域网(WAN)设备，所述WAN设备中的第一WAN设备和第二WAN设备耦合在一起，并耦合到WAN；和

连接到所述第一WAN设备和所述第二WAN设备的TCP数据的源，

其中，所述第一WAN设备用作主控并接收来自所述源的第一TCP数据，

其中，主控WAN设备转发所述第一TCP数据到所述WAN，

其中，所述主控WAN设备转发所述第一TCP数据到用作镜像的第二WAN设备，

其中，所述镜像WAN设备存储所述第一TCP数据，

其中，所述镜像WAN设备提供对所述第一TCP数据的第一ACK，

其中，所述主控WAN设备从所述镜像WAN设备接收对所述第一TCP数据的所述第一ACK；

其中，在从所述镜像WAN设备接收到所述第一ACK之后，所述主控WAN设备提供第二ACK到所述源，

其中，所述主控WAN设备从所述WAN接收对所述第一TCP数据的第三ACK，

其中，所述主控WAN设备将从所述WAN接收到的所述第三ACK转发到所述镜像WAN设备，以及

其中，在接收到转发的、从所述WAN接收到的第三ACK之后，所述镜像WAN设备删除所述第一TCP数据。

14.根据权利要求13所述的网络，其中，在来自所述主控WAN设备的TCP连接的控制变化之后，所述镜像WAN设备从所述源接收第二TCP数据，

其中，所述镜像WAN设备关闭针对所述源的TCP窗口，

其中，所述镜像WAN设备将从所述主控WAN设备接收到的TCP数据转储到所述WAN，

其中，所述镜像WAN设备接收用于所转储的TCP数据的第四ACK，以及

其中，在接收到用于所转储的TCP数据的所述第四ACK之后，所述镜像WAN设备打开针对所述源的TCP窗口。

用于镜像设备的无损连接故障切换

[0001] 相关申请交叉引用

[0002] 本申请根据美国法典第35章第119条e款，要求2011年12月6日提交的题为“Lossless Connection Failover for Single and Mirrored Devices”的美国临时专利申请序列号61/567,282的权利，其通过引用并入于此。

[0003] 此申请还涉及题为“Lossless Connection Failover for Single Devices”、代理人案卷号为112-0690US的美国专利申请序列号13/677,929，题为“TCP Connection Relocation”、代理人案卷号为112-0691US的美国专利申请序列号13/677,909以及题为“Flow-Based TCP”、代理人案卷号为112-0692US的美国专利申请序列号13/678,032，三者均与此同时提交，其通过引用并入于此。

技术领域

[0004] 本发明涉及网络设备，更特别地涉及网络设备的故障。

背景技术

[0005] WAN优化设备可从最接近客户端的本地WAN优化设备进行本地TCP确认。本地WAN优化设备缓冲数据并将数据传递到远程WAN优化设备，远程WAN优化设备又将数据发送到服务器。如果WAN优化设备使得已使用本地TCP确认被确认的数据失效，那么问题就出现了。该数据将丢失并且不能恢复。

发明内容

[0006] 在根据本发明的某些实施例中，WAN优化设备不进行可能导致数据丢失的本地或早的ACK。取而代之，WAN优化设备延迟ACK，直到从目标实际接收到ACK，但打开TCP窗口以允许来自客户端的改善的流，而没有节流(throttling)问题。作为例子，如果连接以64K的窗口开始并且接收到32K的数据，那么TCP窗口通常将是32K。WAN优化设备将延迟对数据确认，并且将再次打开TCP窗口为64K。当ACK最终被接收并转发时，TCP窗口的大小减小到适当的大小，在这个例子中将是64K。

[0007] 如果有镜像的CNE/LDCM设备(WAN优化设备的一种形式)，则可使用根据本发明的第二实施例。在本实施例中，主机CNE/LDCM设备发送数据报通过WAN，并同时发送数据报到镜像CNE/LDCM设备。当镜像CNE/LDCM设备对镜像的数据报进行ACK时，起始的CNE/LDCM设备对主机进行ACK，这相对于WAN延时是短的时间。当ACK通过WAN被接收到时，镜像CNE/LDCM设备得到从主机CNE/LDCM设备转发的ACK并删除镜像的数据报。在失去设备时，TCP连接转换到镜像CNE/LDCM设备，该镜像CNE/LDCM设备辨识到这一点，关闭LAN TCP窗口，并发送所有未经ACK的数据。接着，一经发送成功，镜像CNE/LDCM设备则重新打开LAN TCP窗口，允许数据继续。

附图说明

[0008] 并入本说明书并构成本说明书一部分的附图示出了与本发明一致的装置和方法的实现，附图和详细说明一起用于解释与本发明一致的优点和原理。

[0009] 图1是根据本发明的两个连接的数据中心的实施例的框图。

[0010] 图2示出了根据本发明的一个实施例的包括用于便利跨数据中心的通信的CNE设备的示例性网络体系结构。

[0011] 图3示出了根据本发明的一个实施例的CNE使能(CNE-enabled)的VCS的示例性实现。

[0012] 图4A呈现了根据本发明的一个实施例的示出CNE设备如何处理跨数据中心的广播、未知单播和多播(BUM)业务的示意图。

[0013] 图4B呈现了根据本发明的一个实施例的示出CNE设备如何处理跨数据中心的单播业务的示意图。

[0014] 图5示出了根据本发明的实施例的其中两个CNE设备被用于构造vLAG的例子。

[0015] 图6是根据本发明的LDCM装置的实施例的框图。

[0016] 图7是根据本发明的各方面的被修改以进行操作的图1的数据中心的框图。

[0017] 图8A和图8B是图6的LDCM装置的功能框的框图。

[0018] 图9是根据本发明的Hyper-TCP会话创建和关闭处理的阶梯图。

[0019] 图10是根据本发明的Hyper-TCP数据传送操作的阶梯图。

[0020] 图11是示出根据本发明的Hyper-TCP的操作的框图。

[0021] 图12是根据本发明第一实施例的WAN连接的框图。

[0022] 图13是示出根据本发明的第一实施例的WAN连接阶段的阶梯图。

[0023] 图14是示出根据本发明的第一实施例的WAN数据阶段的阶梯图。

[0024] 图15是根据本发明的第二实施例的HA配置的框图。

[0025] 图16是示出根据本发明的第二实施例的故障切换的框图。

具体实施方式

[0026] 参照图1，示出了描述根据本发明的部分的网络。第一数据中心700被示出为具有三个单独的内部网络：TRILL网络702，正常的以太网生成树协议(STP)网络704和存储区域网络(SAN)706。应用服务器708连接到TRILL网络702，而应用服务器710连接到STP网络704和SAN706。存储设备712被示出为连接到SAN706。每个网络702、704和706都连接有融合网络扩展(CNE)设备714、716和718。CNE设备714、716和718连接到路由器720，路由器720又连接到WAN722。第二数据中心750是相似的，具有VCS以太网架构网络752和SAN754。应用服务器756连接到每个网络752和754，存储设备连接到SAN754。CNE设备760、762连接到每个网络752和754并连接到路由器764，路由器764也连接到WAN722，以允许数据中心700和750进行通信。CNE设备714-718和760-762的操作产生有效的CNE重叠(overlay)网络766以及从每个CNE设备到CNE重叠网络766的虚拟链路。

[0027] 本发明实施例的一个目标是跨数据中心延伸VCS和TRILL网络并满足部署所需要的可扩展性要求。CNE设备可按两个盒子的解决方案来实现，其中一个盒子能够实现L2/L3/FCoE交换，并且是VCS的一部分，另一个盒子便利WAN隧道以在WAN上传输以太网和/或FC业务。CNE设备也可按一个盒子的解决方案来实现，其中单件网络设备组合L2/L3/FCoE交换和

WAN隧道的功能。

[0028] VCS作为二层交换机，使用TRILL作为其交换机之间的连接性，并提供单个逻辑二层交换机的概念。该单个逻辑二层交换机提供透明的LAN服务。VCS的所有边缘端口支持标准协议和特征，如链路聚合控制协议(LACP)、链路层发现协议(LLDP)、VLAN、MAC学习等。VCS使用以太网名称服务(eNS)实现分布式MAC地址数据库，并尝试尽可能避免泛洪。VCS还提供各种智能服务，诸如虚拟链路聚合组(vLAG)、预先端口配置管理(APPM)、端到端FCoE、和边缘环路检测等。有关VCS的更多细节可在以下文件中获得：2011年4月29日提交的、题为“Converged Network Extension”的美国专利申请序列号13/098,360；2010年3月16日提交的、题为“Redundant Host Connection in a Routed Network”的美国专利申请序列号12/725,249；2011年4月14日提交的、题为“Virtual Cluster Switching”的美国专利申请序列号13/087,239；2011年4月22日提交的、题为“Fabric Formation for Virtual Cluster Switching”的美国专利申请序列号13/092,724；2011年4月22日提交的、题为“Distributed Configuration Management for Virtual Cluster Switching”的美国专利申请序列号13/092,580；2011年3月7日提交的、题为“Port Profile Management for Virtual Cluster Switching”的美国专利申请序列号13/042,259；2011年4月22日提交的、题为“Advanced Link Tracking for Virtual Cluster Switching”的美国专利申请序列号13/092,460；2011年4月22日提交的、题为“Virtual Port Grouping for Virtual Cluster Switching”的美国专利申请序列号13/092,701；2011年4月22日提交的、题为“Name Services for Virtual Cluster Switching”的美国专利申请序列号13/092,752；2011年4月22日提交的、题为“Traffic Management for Virtual Cluster Switching”的美国专利申请序列号13/092,877；2011年4月22日提交的、题为“Method and System for Link Aggregation Across Multiple Switches”的美国专利申请序列号13/092,864，以上全部通过引用并入于此。

[0029] 在本发明的实施例中，出于跨数据中心通信的目的，每个数据中心被表示为单个逻辑RBridge。该逻辑RBridge可被分配虚拟RBridge ID或使用执行WAN隧道的CNE设备的RBridge ID。

[0030] 图2示出了根据本发明的一个实施例的包括用于便利跨数据中心通信的CNE设备的示例性网络体系结构。在这个例子中，两个数据中心844和846分别经由网关路由器824和828耦合到WAN826。数据中心844包括VCS816，VCS816经由其成员交换机(诸如交换机810)耦合到若干主机(诸如主机801)。主机801包括两个虚拟机(VM)802和804，它们以双归属配置耦合到虚拟交换机806和808。在一个实施例中，虚拟交换机806和808存在于主机801上的两个网络接口卡上。虚拟交换机806和808耦合到VCS成员交换机810。VCS816中还包括CNE设备818。CNE设备818被配置为从成员交换机810既经由以太网(或TRILL)链路812接收以太网(或TRILL)业务，又经由FC链路814接收FC业务。也耦合到CNE设备818的是目标存储设备820和克隆的目标存储设备822(由虚线表示)。CNE设备818维持经由网关路由器824和828通过WAN826的、到数据中心846的FCIP隧道。

[0031] 类似地，数据中心846包括VCS842，VCS842又包括成员交换机832。成员交换机832耦合到主机841，主机841包括VM834和836，它们二者都耦合到虚拟交换机838和840。VCS842中还包括CNE设备830。CNE设备经由以太网(TRILL)链路和FC链路耦合到成员交换机832。

CNE设备830还耦合到目标存储设备822和目标存储设备的克隆820。

[0032] 在操作期间，假设VM802需要从主机801移动到主机841。要注意，该移动之前是不可能的，因为虚拟机只在同一二层网络域内可见。一旦二层网络域由三层设备(诸如网关路由器824)终结，特定的虚拟机的所有标识信息(其在二层报头中携带)都被丢失。然而，在本发明的实施例中，因为CNE设备将二层域从VCS816延伸到VCS842，所以VM802从数据中心844到数据中心846的移动现在是可能的，原因在于上述基本要求得到满足。

[0033] 当将TRILL帧从数据中心844转发到数据中心846时，CNE设备818修改出口TRILL帧的报头，以使得目的地RBridge标识符是分配给数据中心846的RBridge标识符。CNE设备818接着使用FCIP隧道来将这些TRILL帧传递到CNE设备830，CNE设备830又将这些TRILL帧转发到它们各自的二层目的地。

[0034] VCS使用FC控制平面以自动形成架构，并将RBridge标识符分配给每个成员交换机。在一个实施例中，CNE体系结构在数据中心之间将TRILL和SAN架构保持为分开。从TRILL的角度，每个VCS(其对应于相应的数据中心)表示为单个虚拟RBridge。另外，CNE设备可用TRILL链路和FC链路两者耦合到VCS成员交换机。CNE设备可经由TRILL链路加入VCS。然而，由于CNE设备将TRILLVCS架构和SAN(FC)架构保持为分开，所以CNE设备与成员交换机之间的FC链路被配置用于FC多架构(multi-fabric)。

[0035] 如图3中所示，数据中心908经由网关路由器910耦合到WAN，数据中心920经由网关路由器912耦合到WAN。数据中心908包括VCS906，VCS906包括成员交换机904。数据中心908中还包括CNE设备902。CNE设备902经由TRILL链路和FC链路耦合到VCS成员交换机904。CNE设备902可经由该TRILL链路加入VCS。然而，该FC链路允许CNE设备902维持和VCS成员交换机904的分开的FC架构以携带FC业务。在一个实施例中，CNE设备902上的FC端口是FC EX_port。成员交换机904上对应的端口是FC E_Port。CNE设备902上在WAN侧的端口(耦合到网关路由器910)是FCIP VE_Port。数据中心920具有类似的配置。

[0036] 在一个实施例中，每个数据中心的VCS包括指定为根RBridge的节点，用于多播目的。在初始设置期间，VCS中的CNE设备交换每个VCS的根RBridge标识符。另外，CNE设备还交换每个数据中心的RBridge标识符。要注意，该RBridge标识符代表整个数据中心。关于数据中心RBridge标识符的信息被分布作为到本地VCS中所有节点的静态路由。

[0037] 图4A呈现了根据本发明的一个实施例的示出CNE设备如何处理跨数据中心的广播、未知单播和多播(BUM)业务的示意图。在这个例子中，两个数据中心DC-1和DC-2经由核心IP路由器耦合到IP WAN。DC-1中的CNE设备具有RBridge标识符RB4，DC-2中的CNE设备具有RBridge标识符RB6。而且，在DC-1中的VCS中，成员交换机RB1耦合到主机A。在DC-2中的VCS中，成员交换机RB5耦合到主机Z。

[0038] 假设主机A需要发送多播业务到主机Z，并且主机A已经具有主机Z的MAC地址的知识。在操作期间，主机A组装以太网帧1002，以太网帧1002具有主机Z的MAC地址(表示为MAC-Z)作为其目的地地址(DA)，和主机A的MAC地址(表示为MAC-A)作为其源地址(SA)。基于帧1002，成员交换机RB1组装TRILL帧1003，其TRILL报头1006包括为数据中心DC-1的根RBridge(表示为“DC1-ROOT”)的RBridge标识符作为目的地RBridge，和RB1作为源RBridge。(也就是说，在DC-1内，多播业务在本地多播树上分布。)帧1003的外以太网报头1004具有CNE设备RB4的MAC地址(表示为MAC-RB4)作为DA，以及成员交换机RB1的MAC地址(表示为

MAC-RB1)作为SA。

[0039] 当帧1003到达CNE设备RB4时,CNE设备RB4进一步修改该帧的TRILL报头以产生帧1005。CNE设备RB4用数据中心DC-2的根RBridge标识符DC2-ROOT替换TRILL报头1010中的目的地RBridge标识符。源RBridge标识符改变为数据中心DC-1的虚拟RBridge标识符DC1-RB(这允许数据中心DC-2学习数据中心DC-1的RBridge标识符)。外以太网报头1008具有核心路由器的MAC地址(MAC-RTR)作为其DA,以及CNE设备RB4的MAC地址(MAC-DC-1)作为其SA。

[0040] 帧1005随后在FCIP隧道中传输通过IP WAN并到达CNE设备RB6。对应地,CNE设备RB6更新报头以产生帧1007。帧1007的TRILL报头1014与帧1005保持相同。外以太网报头1012现在具有成员交换机RB5的MAC地址——MAC-RB5——作为其DA,以及CNE设备RB6的MAC地址——MAC-RB6——作为其SA。一旦帧1007到达成员交换机RB5,TRILL报头就被移除,并且内以太网帧被传递到主机Z。

[0041] 在各种实施例中,CNE设备可被配置为允许或不允许未知单播、广播(例如ARP)或多播(例如IGMP探听的)业务跨过数据中心的边界。通过具有这些选项,可限制跨数据中心的BUM业务的量。要注意,数据中心之间的所有TRILL封装BUM业务可带有远程数据中心的根RBridge标识符而被发送。这种转译在FCIP隧道的终结点处完成。

[0042] 可实现附加的机制以使跨数据中心的BUM业务最少。例如,CNE设备和任何VCS成员交换机之间的TRILL端口可被配置为不参与任何的VLAN MGID。另外,两个VCS上的eNS可被配置为同步它们所学习的MAC地址数据库以使具有未知MAC DA的业务最少。(要注意,在一个实施例中,在所学习的MAC地址数据库在不同的VCS中同步之前,具有未知MAC DA的帧仅在本地数据中心内泛洪。)

[0043] 为了进一步使BUM业务最少,可通过探听ARP响应以在VCS成员交换机上构建ARP数据库来减少诸如ARP业务的广播业务。所学习的ARP数据库接着使用eNS跨不同的数据中心交换和同步。基于代理的ARP用于对VCS中的所有已知ARP请求进行响应。而且,可通过跨数据中心分布多播组成员关系(通过经由eNS共享IGMP探听信息)来减少跨数据中心的多播业务。

[0044] 在数据中心之间转发单播业务的处理描述如下。在FCIP隧道的形成期间,表示数据中心的逻辑RBridge标识符被交换。当TRILL帧到达FCIP隧道的进入节点时,其中,TRILL目的地RBridge被设定为远程数据中心的RBridge标识符,TRILL报头中的源RBridge被转译为分配给本地数据中心的逻辑RBridge标识符。当帧离开FCIP隧道时,TRILL报头中的目的地RBridge字段被设定为本地(即目的地)数据中心的虚拟RBridge标识符。内以太网报头中的MAC DA和VLAN ID接着用于查找对应的目的地RBridge(即目的地主机所附接到的成员交换机的RBridge标识符),并且TRILL报头中的目的地RBridge字段被相应地更新。

[0045] 在目的地数据中心中,基于入口帧,所有VCS成员交换机学习MAC SA(在帧的内以太网报头中)和TRILL源RBridge(其为分配给源数据中心的虚拟RBridge标识符)之间的映射。这允许去往该MAC地址的将来的出口帧将被发送到正确的远程数据中心。要注意,因为分配给给定的数据中心的RBridge标识符并不对应于物理RBridge,在一个实施例中,使用静态路由以将远程数据中心RBridge 标识符映射到本地CNE设备。

[0046] 图4B呈现了根据本发明的一个实施例的示出CNE设备如何处理跨数据中心的单播业务的示意图。假设主机A需要发送单播业务到主机Z,而且主机A已经具有主机Z的MAC地址

的知识。在操作期间，主机A组装以太网帧1002，以太网帧1002具有主机Z的MAC地址(MAC-Z)作为其DA，以及主机A的MAC地址(MAC-A)作为其SA。基于帧1002，成员交换机RB1组装TRILL帧1003，其TRILL报头1009包括为数据中心DC-2的虚拟Rbridge(表示为“DC2-RB”)的RBridge标识符作为目的地RBridge，以及RB1作为源RBridge。帧1003的外以太网报头1004具有CNE设备RB4的MAC地址(MAC-RB4)作为DA，以及成员交换机RB1的MAC地址(MAC-RB1)作为SA。

[0047] 当帧1003到达CNE设备RB4时，CNE设备RB4进一步修改该帧的TRILL报头以产生帧1005。CNE设备RB4用数据中心DC-1的虚拟RBridge标识符DC1-RB替换TRILL报头1011中的源RBridge标识符(这允许数据中心DC-2学习数据中心DC-1的RBridge标识符)。外以太网报头1008具有核心路由器的MAC地址(MAC-RTR)作为其DA，以及CNE设备RB4的MAC地址(MAC-DC-1)作为其SA。

[0048] 帧1005随后在FCIP隧道中传输通过IP WAN并到达CNE设备RB6。对应地，CNE设备RB6更新报头以产生帧1007。帧1007的TRILL报头1015具有更新的目的地RBridge标识符，其为RB5，DC-2中耦合到主机Z的VCS成员交换机。外以太网报头1012现在具有成员交换机RB5的MAC地址——MAC-RB5——作为其DA，以及CNE设备RB6的MAC地址——MAC-RB6——作为其SA。一旦帧1007到达成员交换机RB5，TRILL报头就被移除，并且内以太网帧被传递到主机Z。

[0049] 具有未知MAC DA的帧的跨数据中心的泛洪是数据中心学习另一数据中心中的MAC地址的一种方式。所有未知的SA在RBridge后面作为MAC被学习，这对CNE设备也不例外。在一个实施例中，eNS可用于分布所学习的MAC地址数据库，这降低跨数据中心的泛洪的量。

[0050] 为了优化转储(flush)，即使在RBridge后面学习了MAC地址，与MAC地址相关联的实际VCS边缘端口也存在于eNS MAC更新中。然而，边缘端口ID可能不再是跨数据中心唯一。要解决此问题，跨数据中心的所有eNS更新将用数据中心的RBridge标识符限制MAC条目。该配置允许端口转储的跨数据中心传播。

[0051] 在本文所描述的体系结构中，不同数据中心中的VCS不加入彼此；因此，分布式配置被保持分开。然而，为了允许虚拟机跨数据中心移动，将会存在需要跨数据中心被同步的一些配置数据。在一个实施例中，出于CNE的目的，创建特殊的模块(软件或硬件的)。该模块被配置为对便利虚拟机跨数据中心的移动所需的配置信息进行检索，并且其在两个或更多个VCS之间同步。

[0052] 在一个实施例中，所学习的MAC地址数据库跨数据中心分布。同样，边缘端口状态变更通知(SCN)也跨数据中心分布。当物理RBridge正停止工作时，SCN被转换为数据中心之间的FCIP链路上的多个端口SCN。

[0053] 为了保护数据中心之间的连接性，VCS可在两个或更多个CNE设备之间形成vLAG。在该模型中，vLAG RBridge标识符被用作数据中心RBridge标识符。FCIP控制平面被配置为知晓这种布置，并在这样的情况下交换vLAG RBridge标识符。

[0054] 图5示出了根据本发明的实施例的其中两个CNE设备被用于构造vLAG的例子。在该例子中，VCS1100包括两个CNE设备1106和1108。CNE设备1106和1108二者形成vLAG1110，vLAG1110耦合到核心IP路由器。vLAG1110被分配虚拟RBridge标识符，其也被用作用于VCS1100的数据中心RBridge标识符。而且，vLAG1110可针对VCS1100内的任何成员交换机便利入口和出口二者的负载均衡(例如基于等价多路径(ECMP))。

[0055] 图6示出了CNE/LDCM设备1200，LDCM特征优选被添加到CNE设备以创建单个设备。具有多个CPU1204的片上系统(SOC)1202提供主处理能力。若干以太网连接1206优选包括在SOC1202上以用作WAN链路，虽然如果期望的话，可使用单独的以太网设备。FC交换芯片1208连接到SOC1202以提供到FC SAN的连接。CEE交换芯片1210连接到SOC1202以允许到VCS或到以太网LAN的附接。压缩引擎1212与SOC1202一起被提供来提供压缩和重复数据删除(dedup)能力，以减少WAN链路上的业务。出于安全目的，加密引擎1214被提供，因为为了安全性，FCIP隧道优选被加密。

[0056] 各种软件模块1216存在于CNE/LDCM设备1200中的存储器中。这些包括底层操作系统1218、管理与VCS的交互的控制平面模块1220、控制平面之上的用于TRILL功能的TRILL管理模块1222、管理WAN上的FCIP隧道的FCIP管理模块1224、与FC SAN交互的FC管理模块1226、以及地址管理模块1228。另外的模块是高可用性(HA)模块1230，其又包括连接故障切换子模块1232。连接故障切换子模块1232中的软件在CPU1204中执行，以执行下面关于图12-16所描述的连接故障切换操作。

[0057] 图7示出了添加有CNE/LDCM设备1302和1352的数据中心。示出了两个数据中心100和150。每个都具有一系列执行实际应用的应用服务器集群102和152(诸如以SaaS(软件即服务)体系结构)。数据存储在存储架构104和154中。到应用服务器集群102和152的访问被示出为通过web服务器集群106和156，虽然在LAN层的更直接的访问是常见的。站点负载均衡器108和158跨web服务器集群106和156中的web服务器分布传入的请求。全局负载均衡器110连接到因特网112以在数据中心100和150之间均衡负载。CNE/LDCM设备1302和1352在它们自身之间创建云虚拟互连(CVI)1304，其实际上是通过WAN1306的FCIP隧道。CVI1304用于数据中心100和150之间的VM移动性、应用负载均衡和存储复制。

[0058] 云虚拟互连1304优选包括以下组件。如在2010年9月13日提交的、题为“FCIP Communications with Load Sharing and Failover”的美国专利申请序列号12/880,495(其通过引用并入于此)中更充分地描述的，FCIP汇聚聚合多个TCP连接以支持从100Mbps上至20Gbps的很宽的WAN带宽范围。其还支持多归属，并使得能够在冗余网络路径之间透明地故障切换。

[0059] 自适应速率限制(ARL)在TCP连接上执行以改变数据通过TCP连接发送的速率。ARL使用来自TCP连接的信息来动态地确定和调整对TCP连接的速率限制。这将允许TCP连接利用最大可用带宽。其还提供了灵活数目的用于限定策略的优先级，并且用户可以限定所需要的优先级。

[0060] 高带宽TCP(HBTCP)被设计用于长肥网络上的高吞吐量应用，诸如虚拟机和存储迁移。其克服了WAN中传统TCP/IP的负面影响的挑战。为了优化性能，已经做出了以下更改。

[0061] 1)扩展的窗口：在HBTCP中，扩展的窗口用于支持高达350ms或更多的WAN延时。最大可消耗存储器将按每会话分配，以维持线路速率。

[0062] 2)优化的重排序耐受性：HBTCP对重复确认具有更大的耐受性，并需要更多的重复ACK以触发快速重传。

[0063] 3)优化的快速恢复：在HBTCP中，作为将拥塞窗口(cwnd)减少一半的替代，拥塞窗口所减少的大大小于50％，来为其中做出大量的网络重排序的情况作准备。

[0064] 4)快启动：慢启动阶段被修改为快启动，在快启动中，初始吞吐量被设定为很大的

值，相比拥塞事件之前的吞吐量，吞吐量仅最低限度地减少。

[0065] 5)拥塞避免：通过精心地将所发送的数据量匹配到网络速度，避免了拥塞，而不是注入更多的业务并造成拥塞事件，使得拥塞避免可能被禁用。

[0066] 6)优化的慢恢复：HBTCP中的重传计时器(150ms)到期相比传统TCP中的快得多，并在快速重传不能提供恢复时使用。当拥塞事件发生时，这更早地触发慢启动阶段。

[0067] 7)丢失分组连续重试：作为等待对SACK重传的分组的ACK的替代，连续地重传该分组以改善慢恢复，如在2010年12月20日提交的、题为“Repeated Lost Packet Retransmission in a TCP/IP Network”的美国专利申请序列号12/972,713中所更详细地描述的，其通过引用并入于此。

[0068] 在用于VMware系统的VM移动性中使用的vMotion迁移数据通过CEE交换芯片1210的LAN以太网链路进入CNE/LDCM设备1302，经压缩、加密的数据使用WAN上行链路(其使用SOC1202的以太网端口1206)在WAN基础结构上被发送。存储迁移也类似，来自FC交换芯片1208所提供的SAN FC链路的数据使用WAN上行链路迁移到迁移存储。控制平面模块1220负责建立、维护和终结与应用服务器和目的地LDCM服务器的TCP会话。

[0069] 图8A和8B示出了CNE/LDCM设备的功能框和模块。LAN终结1402和SAN终结1404通过应用模块1408、数据精简引擎1410和高可靠性传递应用(HRDA)层1412互连到CVI1406。

[0070] LAN终结1402具有连接到LAN端口的二层(以太网或CEE)模块1420。IP虚拟边缘路由模块1422将二层模块1420连接到Hyper-TCP模块1424。Hyper-TCP模块1424操作在下面更详细地描述，其包括连接到虚拟边缘路由模块1422的TCP分类器1426。TCP分类器1426连接到数据处理模块1428和会话管理器1430。事件管理器1432连接到数据处理模块1428和会话管理器1430。事件管理器1432、数据处理模块1428和会话管理器1430都连接到套接字层1434，套接字层1434用作Hyper-TCP模块1424和LAN终结1402到应用模块1408的接口。

[0071] SAN终结1404具有连接到SAN端口的FC二层模块1436。批处理/解批处理模块1438将FC二层模块1436连接到路由模块1440。为FICON业务1442、FCP业务1444和F_Class业务1446提供分开的模块，每个模块连接到路由模块1440，并用作SAN终结1404和应用模块1408之间的接口。

[0072] 应用模块1408具有三个主应用：管理程序1448、web/安全1452和存储1454。管理程序应用1448与各种管理程序移动功能(诸如vMotion、XenMotion和MS实时迁移)协作。高速缓存子系统1450与管理程序应用1448一起被设置用于在移动操作期间高速缓存数据。web/安全应用1452与VPN、防火墙和入侵系统协作。存储应用1454处理iSCSI、NAS和SAN业务，并具有随附的高速缓存1456。

[0073] 数据精简引擎1410使用压缩引擎1212来处理压缩/解压缩和重复数据删除操作，以允许WAN链路的改进的效率。

[0074] HRDA层1412的主要功能是确保网络层面以及传输层面上的通信可靠性。如所示出的，数据中心通过在IP上延伸L2TRILL网络通过WAN基础结构而得以巩固。提供冗余链路用作备份路径。万一主路径失效，HRDA层1412就执行到备份路径的无缝切换。通过在备份路径上重传任何未确认的分段，防止在主路径上运行的HBTCP会话经历任何拥塞事件。对未确认分段的确认和未确认的分段本身被假定为丢失了。HRDA层1412也确保单个路径内的TCP会话的可靠性。万一HBTCP会话失效，使用HBTCP会话的任何迁移应用也将失效。为了防止应用

失效，HRDA层1412透明地切换到备份HBTCP会话。

[0075] CVI1406包括连接到WAN链路的IP模块1466。提供IPSEC模块1464用于链路安全性。提供HBTCP模块1462以允许上述HBTCP操作。QoS/ARL模块1460处理上述QoS和ARL功能。汇聚模块1458处理上述汇聚操作。

[0076] Hyper-TCP是对实时进行的(live)服务和应用在长距离网络上的迁移进行加速的组件。简单地说，应用客户端和服务器之间的TCP会话被本地终结，并且通过利用数据中心之间的高带宽传输技术，应用迁移被加速。

[0077] Hyper-TCP主要支持两种操作模式：

[0078] 1)数据终结模式(DTM)：在数据终结模式中，终端设备TCP会话不改变，但数据被本地确认，并且数据序列的完整性得以维持。

[0079] 2)完全终结模式(CTM)：在完全终结模式中，终端设备TCP会话完全由LDCM终结。数据序列未在终端设备之间维持，但数据完整性得以保障。

[0080] Hyper-TCP中主要有三个阶段。它们是会话建立、数据传送和会话终结。下面说明这三个阶段。

[0081] 1)会话建立：在该阶段期间，连接建立分组被探听，如连接端节点、窗口大小、MTU和序列号的TCP会话数据被高速缓存。如MAC地址的二层信息也被高速缓存。Hyper-TCP服务器上的TCP会话状态与应用服务器的相同，Hyper-TCP客户端的TCP会话状态与应用客户端相同。使用高速缓存的TCP状态信息，Hyper-TCP设备可本地终结应用客户端和服务器之间的TCP连接，并可本地确认数据分组的接收。因此，由应用计算的RTT(往返时延)将被掩盖而不包括WAN延时，这产生更好的性能。

[0082] 图9中示出了会话创建处理。应用客户端发送SYN，SYN由Hyper-TCP服务器探听。Hyper-TCP服务器将SYN转发到Hyper-TCP客户端，可能在TCP报头选项字段中带有种子值。种子值可指示这是否是Hyper-TCP会话、终结模式和Hyper-TCP版本等。种子值由各种模块(诸如数据精简引擎1410和CVI1406)使用以确定会话加速的需要和水平。Hyper-TCP客户端探听SYN并将其转发到应用服务器。应用服务器用SYN+ACK进行响应，Hyper-TCP客户端探听SYN+ACK并将其转发到Hyper-TCP服务器。Hyper-TCP服务器探听SYN+ACK并将其转发到应用客户端。应用客户端用ACK进行响应，Hyper-TCP服务器将ACK转发到Hyper-TCP客户端，Hyper-TCP客户端又将ACK提供给应用服务器。这产生已创建的TCP会话。

[0083] 2)数据传送处理：一旦会话已经建立，数据传送始终在Hyper-TCP设备和终端设备之间本地处理。用作应用客户端的代理目的地服务器的Hyper-TCP服务器本地确认数据分组，并且TCP会话状态被更新。数据被移交给Hyper-TCP客户端和服务器之间的HBTCP会话。HBTCP会话压缩数据并将其转发到Hyper-TCP客户端。这减小了应用客户端和源所“看到”的RTT，因为其掩盖了在网络上发生的延时。在Hyper-TCP客户端处所接收的数据如同该数据是已由该Hyper-TCP客户端生成的而被对待，并且该数据被交给在Hyper-TCP客户端和应用服务器之间运行的Hyper-TCP处理。在网络中有拥塞时，从Hyper-TCP套接字(socket)取得的数据量被控制。

[0084] 图10示出了该处理。数据从应用客户端提供到Hyper-TCP服务器，Hyper-TCP服务器按照期望地对数据ACK，从而在Hyper-TCP服务器处本地终结连接。LDCM设备聚合并精简所接收的数据以减少WAN业务，并将数据发送到另一LDCM设备中的Hyper-TCP客户端。接收

LDCM设备解精简以及解聚合数据，并将其提供给Hyper-TCP客户端，Hyper-TCP客户端又将数据提供给应用服务器，应用服务器周期性地对数据ACK。如果应用服务器需要将数据发送到应用客户端，则该处理基本是逆转的。通过使Hyper-TCP服务器和客户端本地对所接收的数据进行响应从而本地终结连接，应用服务器和客户端不知晓由Hyper-TCP服务器和客户端之间的WAN链路产生的延迟。

[0085] 3)会话终结：所接收的FIN/RST像会话建立分组一样被透明地发送通过。完成这一点以确保两个终端设备之间的数据完整性和一致性。只有当在接收到FIN之前所接收的所有分组都已经被本地确认并发送到Hyper-TCP客户端时，在Hyper-TCP服务器处所接收的FIN/RST才会被透明地发送通过。如果已经在Hyper-TCP客户端上接收到FIN/RST分组，那么在所有的排队数据已经被发送并由应用服务器确认之后，该分组将被透明地转发。在任一方向上，一旦FIN已经被接收并转发，进一步的分组传送透明地完成，并且不被本地终结。

[0086] 图9中更详细地示出了这一点。应用客户端提供FIN到Hyper-TCP服务器。如果Hyper-TCP服务器没有接收到任何数据，Hyper-TCP服务器则将从应用客户端恢复数据并将其提供给Hyper-TCP客户端。Hyper-TCP服务器接着将FIN转发到Hyper-TCP客户端，Hyper-TCP客户端转储Hyper-TCP客户端中任何剩余的数据，并接着将FIN转发到应用服务器。应用服务器以对转储数据的ACK答复，并接着以FIN答复。Hyper-TCP客户端接着从应用服务器接收任何未完成的数据，并对应用服务器恢复数据。ACK和数据被转发到Hyper-TCP服务器。在数据被传送之后，Hyper-TCP客户端将FIN转发到Hyper-TCP服务器。当接收到ACK时，Hyper-TCP服务器转发ACK并转储任何剩余的数据到应用客户端。在完成这些之后，Hyper-TCP服务器转发FIN，并且会话结束。

[0087] 图11示出了Hyper-TCP服务器和客户端在CVI1712上的有效操作。一系列的应用1702-1至1702-n分别与应用1704-1至1704-n通信。Hyper-TCP服务器代理1706与应用1702协作，而Hyper-TCP代理1708与应用1704协作。在图示中，示出了四个不同的Hyper-TCP会话H1、H2、H3和Hn(1710-1至1710-n)，这些会话使用CVI1712穿过WAN。

[0088] 图12示出了用于CNE/LDCM设备的基本高可用性(HA)配置。每个数据中心200(250)包括两个并行的CNE/LDCM设备202A和202B(252A和252B)，以及连接数据中心200和250的负载均衡器/路由器204(254)。

[0089] WAN优化中的主要问题之一是提供无损的故障切换。可建立多个TCP连接以提供用于客户端-服务器对的WAN优化。如图12所示，建立三个TCP连接TCP1300、TCP2302和TCP3304以提供用于客户端306和服务器308之间的TCP连接的WAN业务优化。

[0090] WAN优化设备(诸如CNE/LDCM202)通常从最接近客户端306的WAN优化设备202进行本地TCP确认。该WAN优化设备202缓冲数据并将其传递到远程WAN优化设备252，远程WAN优化设备252将数据发送到服务器308。如果WAN优化设备252使已经由WAN优化设备202使用TCP确认而确认的数据失效，那么问题就产生了。数据将丢失并且不能恢复。

[0091] 该问题可分为两个阶段：连接建立阶段和数据阶段。在LAN站点200上有客户端306和CNE/LDCM设备202之间的LAN连接，在另一LAN站点250中有CNE/LDCM设备252和服务器308之间的连接。在连接建立阶段，如图13所示的以下序列保障在连接建立阶段CNE/LDCM设备202和252之间的同步。

[0092] 1.客户端306发送去往远程服务器308的SYN请求。

[0093] 2.本地侧的CNE/LDCM设备202截获SYN请求并隧道传输该请求到远程CNE/LDCM设备252。

[0094] 3.远程CNE/LDCM设备252使用和其从客户端306所接收的相同的源和目的地IP地址、TCP端口和序列号发送SYN请求到远程服务器308。

[0095] 4.服务器308发送SYN/ACK，SYN/ACK将由远程CNE/LDCM设备252截获。

[0096] 5.远程CNE/LDCM设备252隧道传输SYN/ACK到本地CNE/LDCM设备202。

[0097] 6.本地CNE/LDCM设备202用相同的源/目的地IP地址、TCP端口和序列号发送SYN/ACK到客户端306。

[0098] 7.客户端306发送ACK，ACK由本地站点200中的CNE/LDCM设备202截获。

[0099] 8.本地CNE/LDCM设备202使用已经创建的隧道将ACK隧道传输到远程CNE/LDCM设备252。

[0100] 9.远程CNE/LDCM设备252使用和其所接收的相同的源/目的地IP地址、TCP端口和序列号转发ACK到服务器308。

[0101] 在一些变型中，可在步骤4之后进行步骤9。在该情况下，更接近服务器308的CNE/LDCM设备252可能需要缓冲在发送ACK之后所接收的任何数据分组，直到其接收到来自远程CNE/LDCM设备202的ACK。服务器侧的CNE/LDCM设备252在其接收到一定量的分组之后可关闭它的TCP窗口。这可限制该CNE/LDCM设备在接收到来自CNE/LDCM设备202的ACK前需要缓冲的分组的数量。

[0102] 为了能够提供无损的故障切换，来自更接近客户端306的WAN优化设备202的确认应当与来自服务器308的确认同步。也就是说，CNE/LDCM设备202不提供ACK到客户端306，直到其接收到来自服务器308的转发的ACK。由于WAN延时，该同步可能降低WAN310的吞吐量。WAN延时可导致TCP窗口被关闭，直到从远程侧CNE/LDCM设备252接收到确认。根据本发明，为了改善LAN延时，在本地侧CNE/LDCM202正确地接收到数据之后，本地侧CNE/LDCM202中的TCP窗口被打开，而不发送对该数据的确认。通过这样做，由于客户端306和CNE/LDCM设备202之间的TCP1300连接的TCP窗口于在WAN优化设备202处接收到该数据之后打开，所以在WAN310上有LAN一样的性能。同时，WAN优化设备202从不对未被传递到服务器308或其它端节点的数据确认，直到它实际接收到来自服务器308的转发的ACK。图14中针对两个数据传送示出了这一点。

[0103] 在根据本发明的一个变型中，用于TCP2302连接的WAN TCP窗口可被设定为大的大小，以从一开始就适应WAN延时，而不是每次顺序数据被接收时都打开TCP窗口。

[0104] 同步确认和打开TCP窗口的组合允许无损的操作以及对WAN连接的更好使用。

[0105] 这可以改述为：不进行早的ACK，其可导致数据丢失。相反，延迟ACK直到实际从目标接收到ACK，但打开TCP窗口以允许没有节流问题的改善的流。作为例子，如果连接以64K的窗口开始并接收到32K的数据，则TCP窗口将是32K。CNE/LDCM设备202将立即转发数据并将再次打开TCP窗口为64K。当从服务器308接收到ACK并将其转发到客户端306时，TCP窗口大小将通过指定64K而不是96K被减小，如果TCP窗口大小未被改变，其会是96K。

[0106] 如果有镜像的CNE/LDCM设备，则可使用第二实施例，如图15所示。在该实施例中，主控CNE/LDCM设备202A发送数据报通过WAN310，并同时发送该数据报到镜像CNE/LDCM设备202B。当镜像CNE/LDCM设备202B对镜像的数据报ACK时，主控CNE/LDCM设备202A对主机206

进行ACK，这相比WAN延时是短的时间。当ACK通过WAN被接收时，镜像CNE/LDCM设备202B获得从主控CNE/LDCNM设备202A转发的ACK并删除镜像的数据报。在失去主控CNE/LDCM设备202A时，TCP连接转换到镜像CNE/LDCM设备202B。镜像CNE/LDCM设备202B辨识到该转变，关闭LAN TCP窗口并发送所有未经ACK的数据。接着，一经发送成功，镜像CNE/LDCM设备202B重新打开LAN TCP窗口，允许数据继续。如果两个CNE/LDCM设备202同时失效，则该实施例仍具有潜在的数据丢失，但它确实提供较早的ACK到主机。图16中示出了该序列。

[0107] 因此，通过转向具有镜像的WAN优化设备的冗余环境，可完成到本地客户端的快速确认和无损的操作两者。

[0108] 系统设计者在设计网络时可基于冗余以及故障预期和风险而在替换方案之间选择。另外，当失去镜像时，第二实施例可退回到第一实施例。

[0109] 上面的描述旨在是说明性的而非限制性的。例如，上述实施例可彼此组合使用。对于本领域技术人员来说，在查看上面的描述后，许多其它的实施例将是显而易见的。因此，本发明的范围应当参照权利要求以及这样的权利要求所享有的等同物的全部范围来确定。在权利要求中，术语“包括(including)”和“其中(in which)”用作相应的术语“包括(comprising)”和“其中(wherein)”的通俗英语等同物。

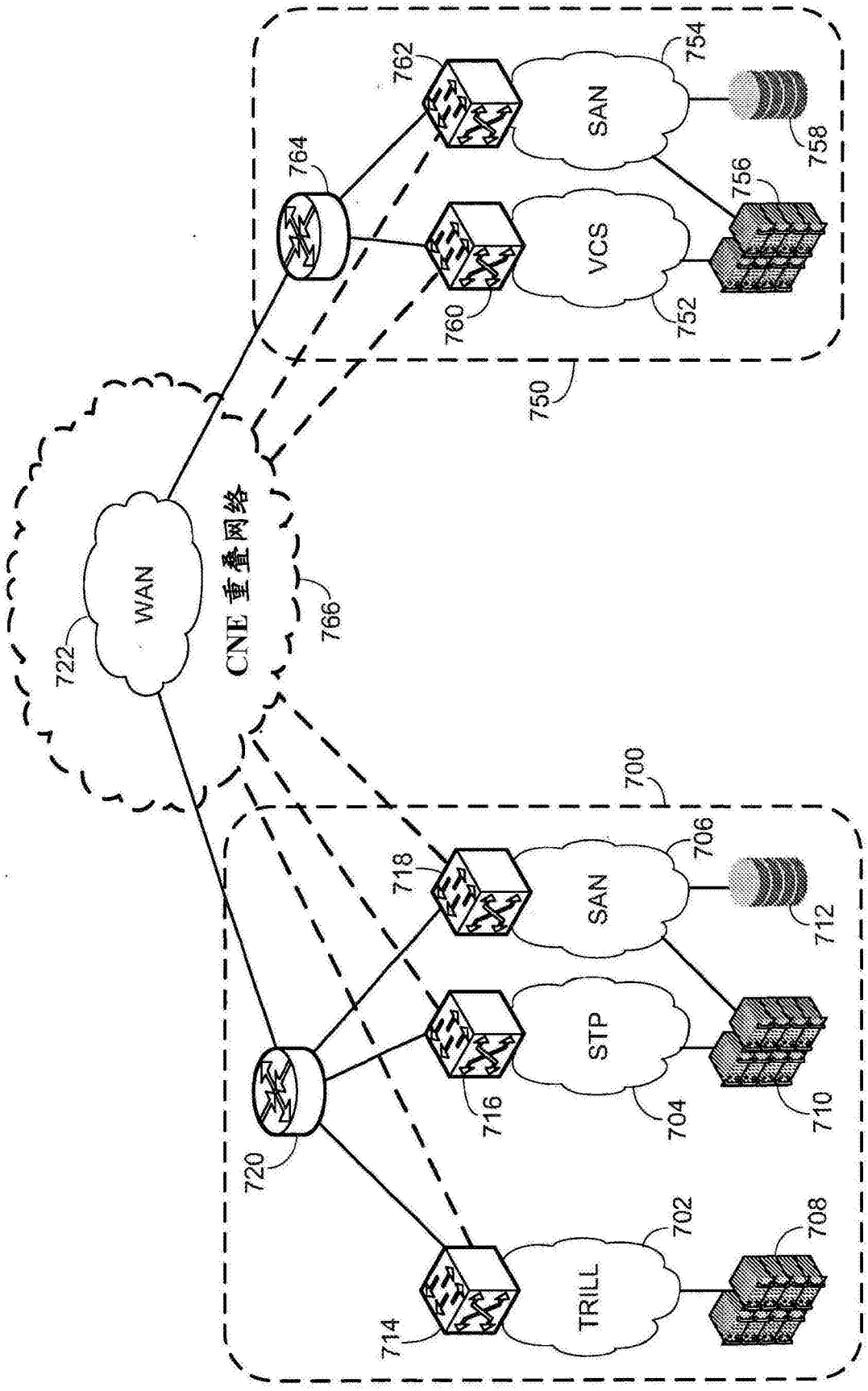

762
754
SAN
758
764
756
VCS
760
752
750
WAN
722
CNE 重叠网络
766
700
718
706
SAN
712
STP
716
704
710
720
TRILL
702
708
714

图1

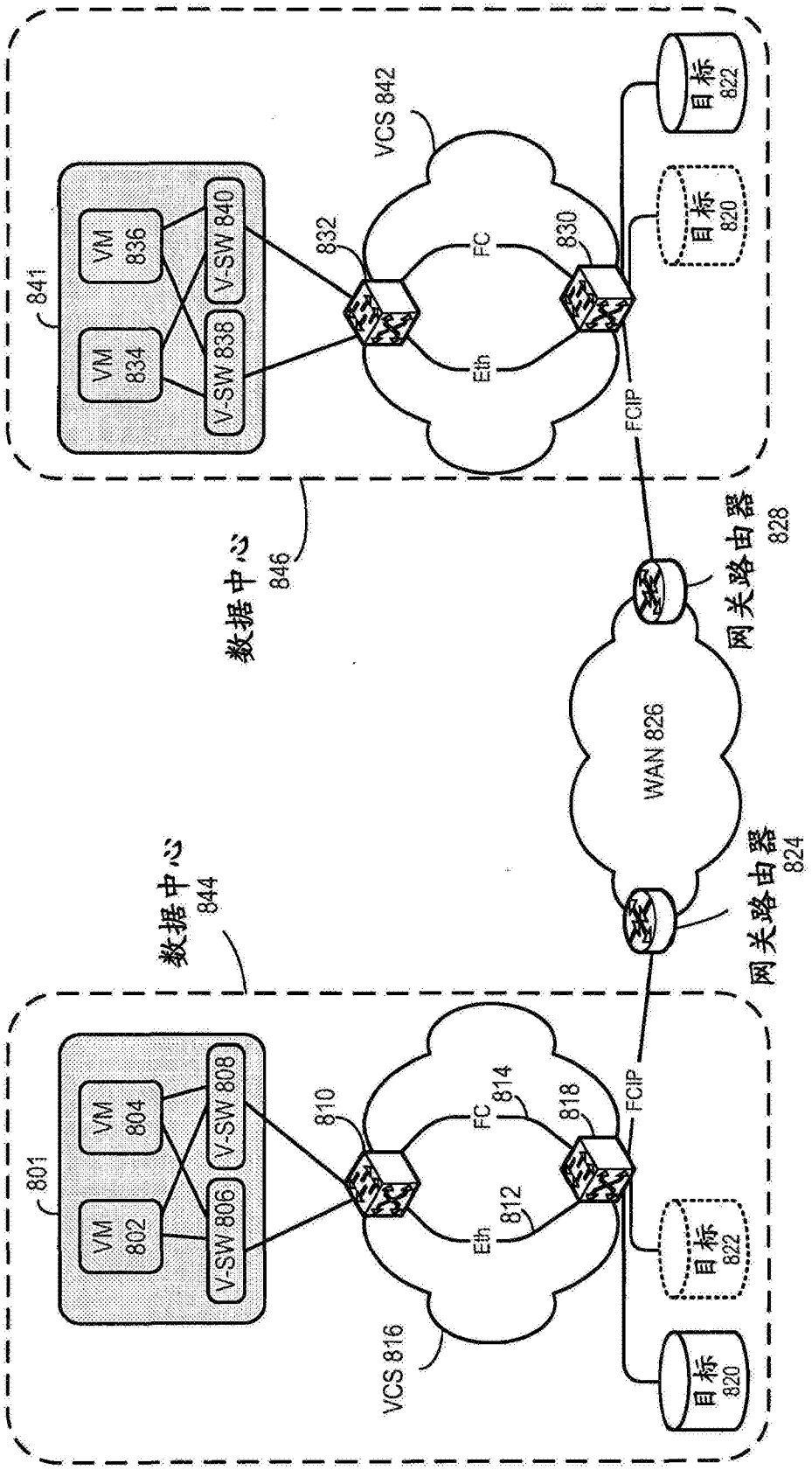

数据中心
846
841
VM
836
VM
834
V-SW 840
V-SW 838
832
VCS 842
FC
Eth
830
目标
822
目标
820
FCIP
网关路由器
828
WAN 826
网关路由器
824
数据中心
844
801
VM
804
VM
802
V-SW 808
V-SW 806
810
FC
814
812
Eth
818
FCIP
目标
822
目标
820
VCS 816

图2

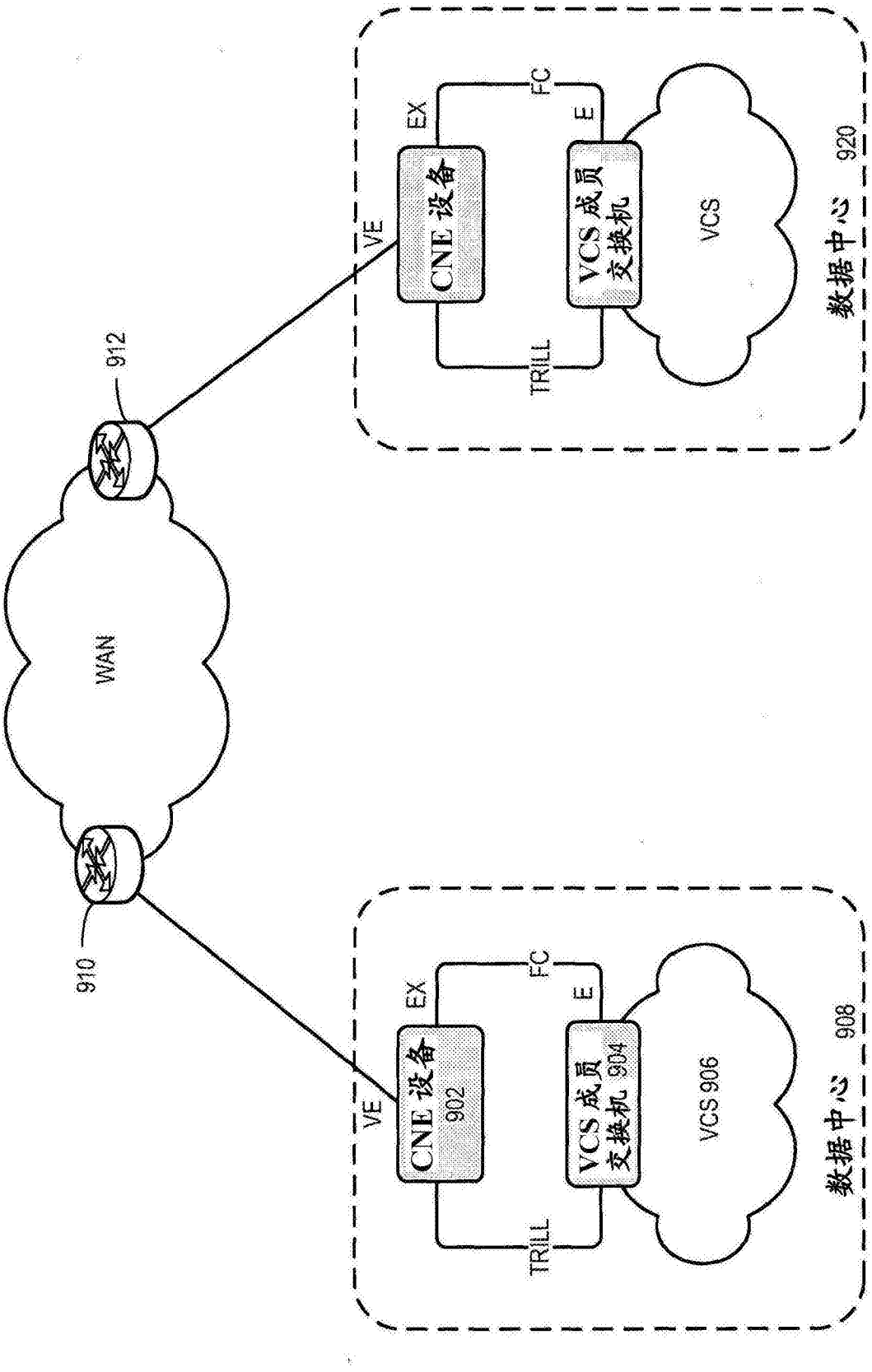

912
910
WAN
VE
EX
FC
E
TRILL
CNE 设备
VCS 成员
交换机
VCS
数据中心 920
VE
EX
FC
E
TRILL
CNE 设备
902
VCS 成员
交换机 904
VCS 906
数据中心 908

图3

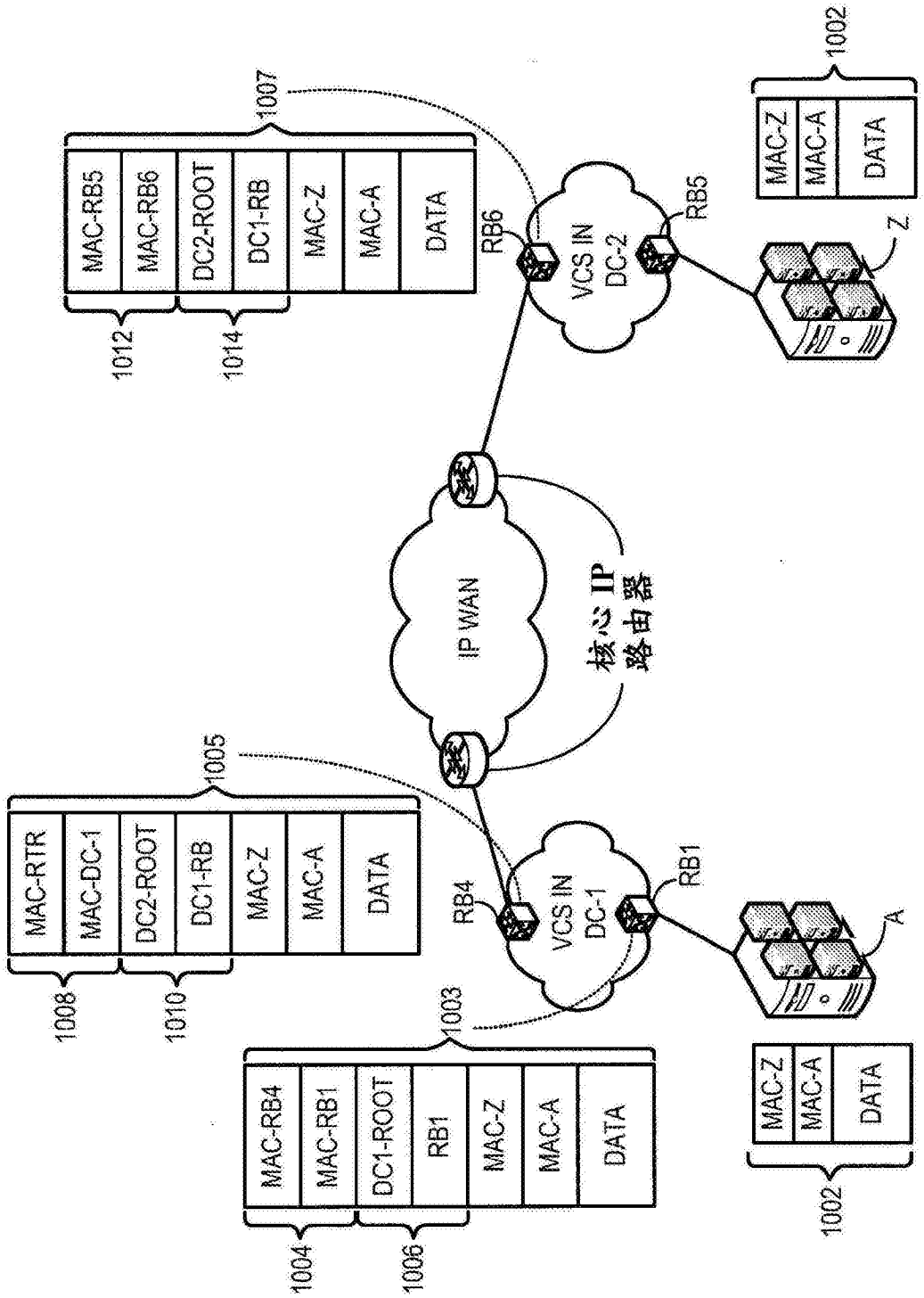

1007
MAC-RB5
MAC-RB6
DC2-ROOT
DC1-RB
MAC-Z
MAC-A
DATA
1012
1014
RB6
VCS IN
DC-2
RB5
1002
MAC-Z
MAC-A
DATA
Z
IP WAN
核心 IP
路由器
1005
MAC-RTR
MAC-DC-1
DC2-ROOT
DC1-RB
MAC-Z
MAC-A
DATA
1008
1010
RB4
VCS IN
DC-1
RB1
A
1003
MAC-RB4
MAC-RB1
DC1-ROOT
RB1
MAC-Z
MAC-A
DATA
1004
1006
1002
MAC-Z
MAC-A
DATA

图4A

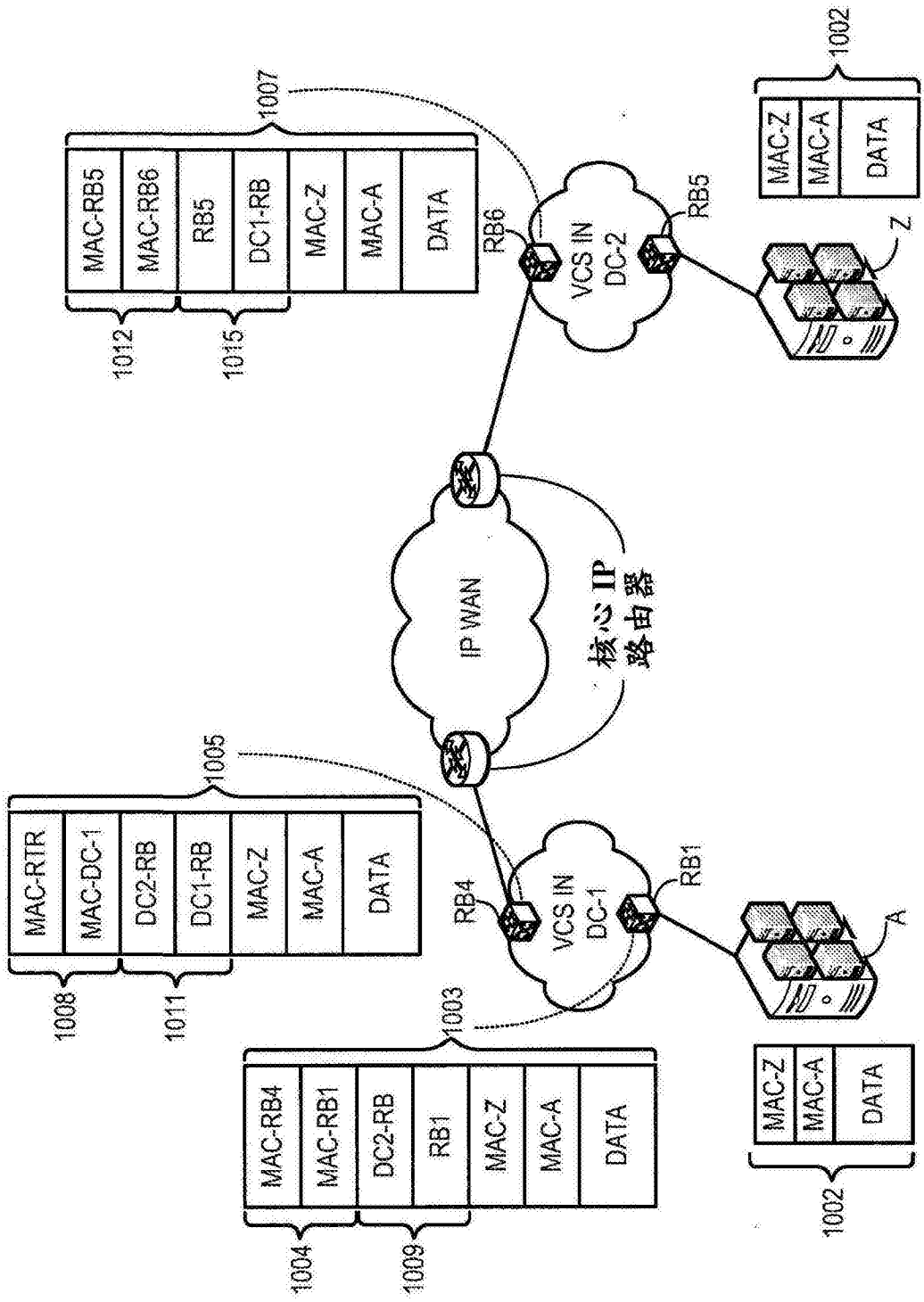

1007
MAC-RB5
MAC-RB6
RB5
DC1-RB
MAC-Z
MAC-A
DATA
1012
1015
RB6
VCS IN
DC-2
RB5
1002
MAC-Z
MAC-A
DATA
Z
IP WAN
核心IP
路由器
1005
MAC-RTR
MAC-DC-1
DC2-RB
DC1-RB
MAC-Z
MAC-A
DATA
1008
1011
RB4
VCS IN
DC-1
RB1
A
1003
MAC-RB4
MAC-RB1
DC2-RB
RB1
MAC-Z
MAC-A
DATA
1004
1009
1002
MAC-Z
MAC-A
DATA

图4B

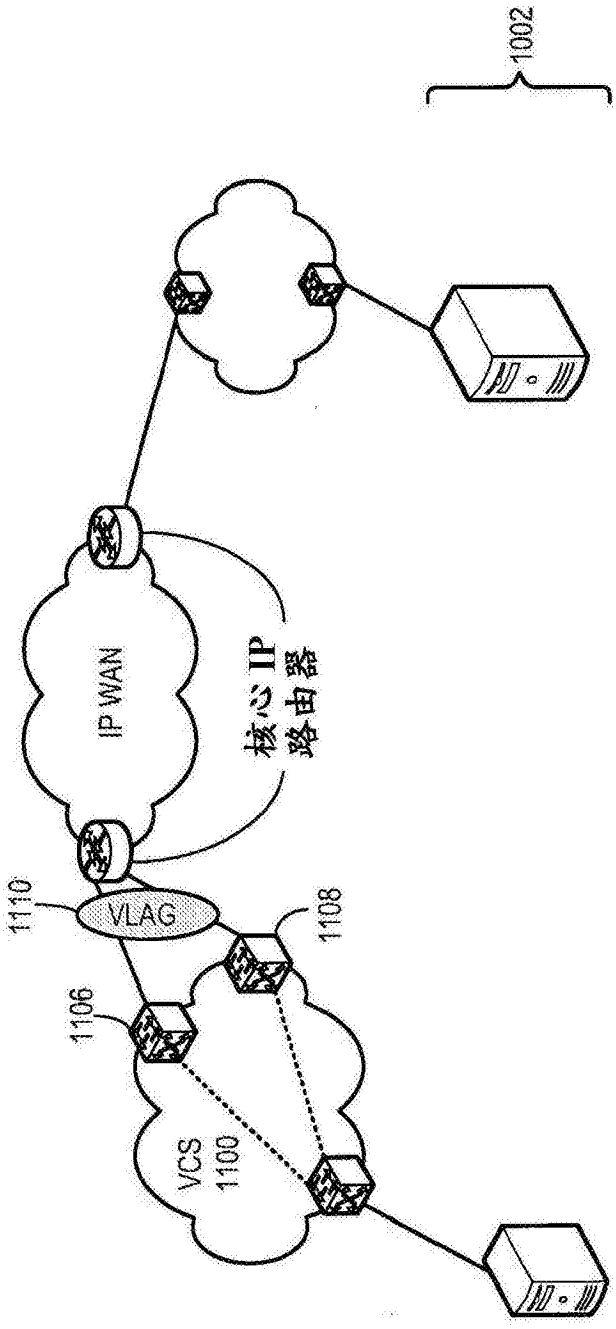


图5

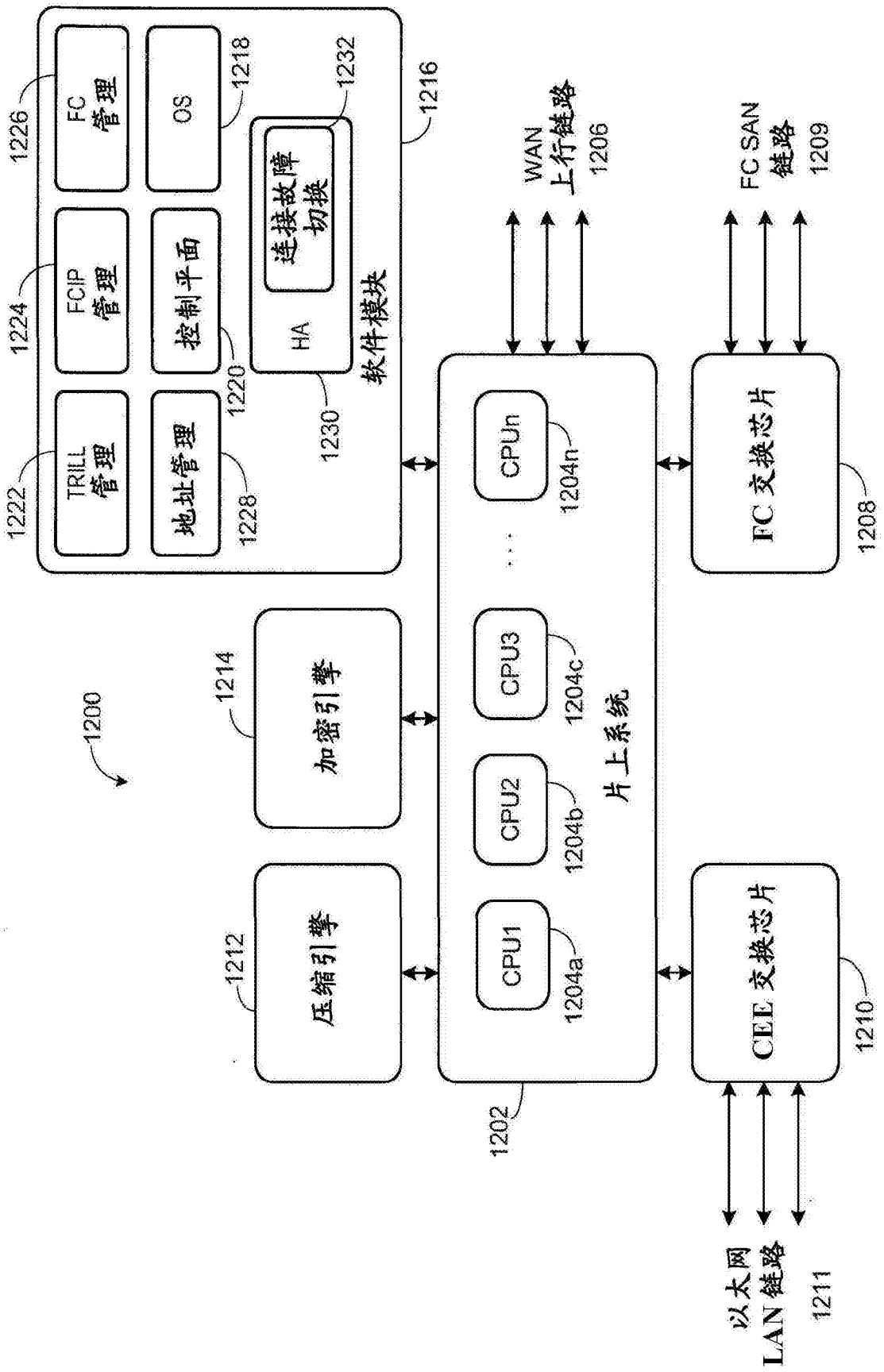

1200
1202
片上系统
CPU1
1204a
CPU2
1204b
CPU3
1204c
CPUn
1204n
1212
压缩引擎
1214
加密引擎
1216
软件模块
1218
OS
1220
控制平面
1222
TRILL 管理
1224
FCIP 管理
1226
FC 管理
1228
地址管理
1230
HA
1232
连接故障切换
1206
WAN 上行链路
1208
FC 交换芯片
1209
FC SAN 链路
1210
CEE 交换芯片
1211
以太网 LAN 链路

图6

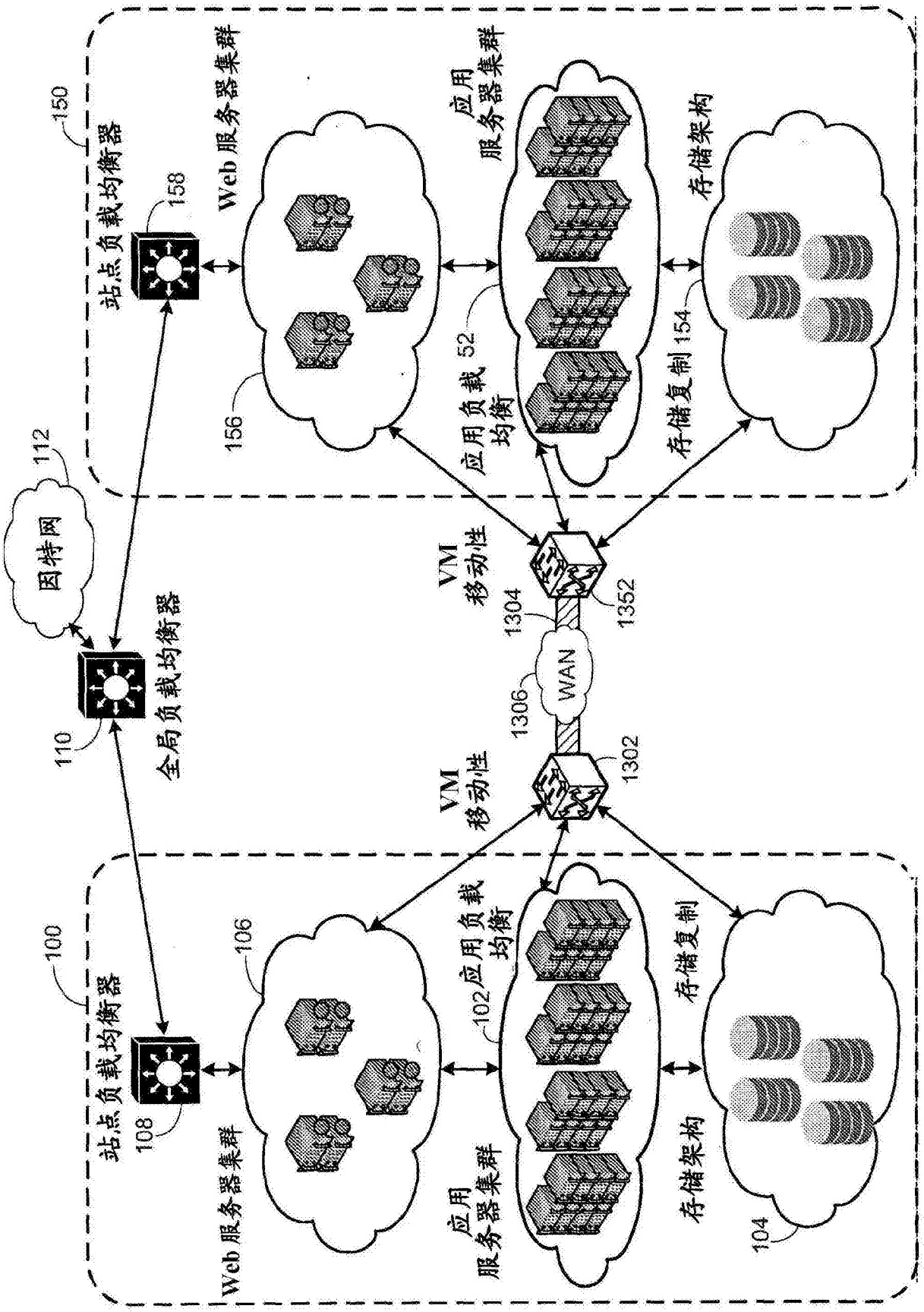

150
站点负载均衡器
158
Web 服务器集群
156
应用
服务器集群
52
应用负载
均衡
存储架构
存储复制 154
112
因特网
110
全局负载均衡器
VM
移动性
1304
1352
WAN
1306
1302
VM
移动性
100
站点负载均衡器
108
Web 服务器集群
106
102
应用负载
均衡
应用
服务器集群
存储复制
存储架构
104

图7

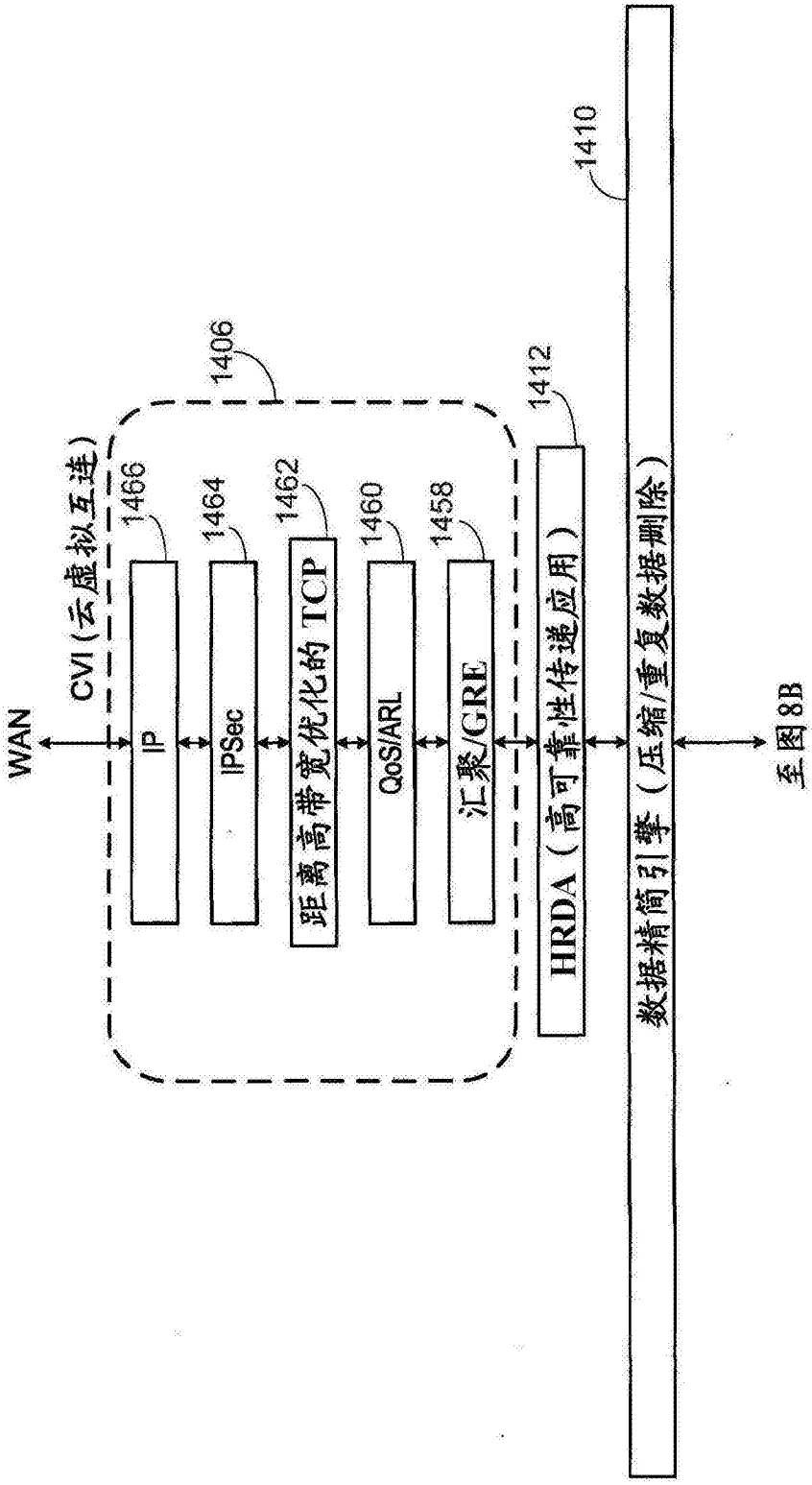

WAN
CVI(云虚拟互连)
IP
IPSec
距离高带宽优化的 TCP
QoS/ARL
汇聚/GRE
HRDA(高可靠性传递应用)
数据精简引擎(压缩/重复数据删除)
至图8B
1406
1466
1464
1462
1460
1458
1412
1410

图8A

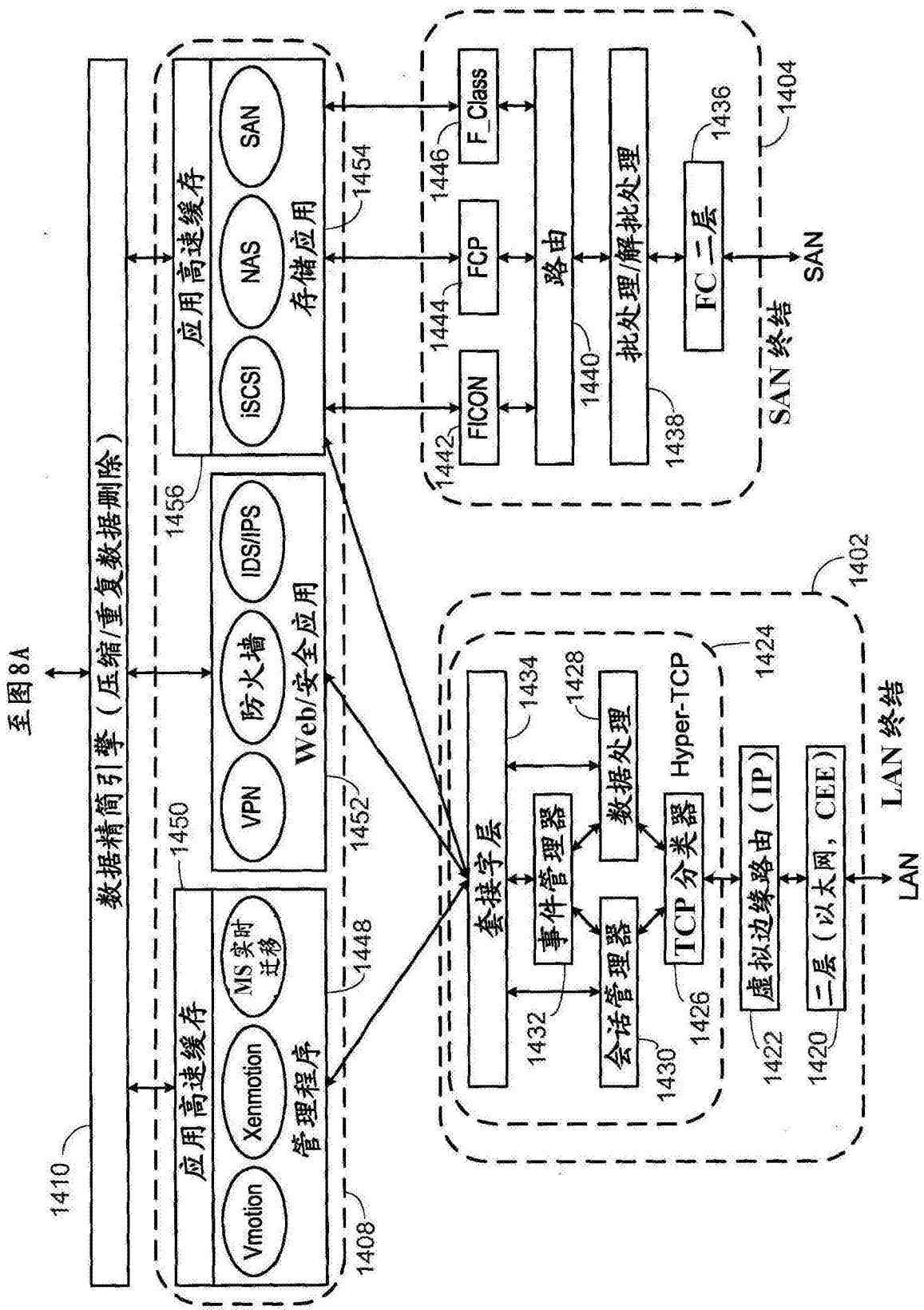

至图8A
数据精简引擎（压缩/重复数据删除）
1410
应用高速缓存
Vmotion
Xenmotion
MS实时迁移
管理程序
1408
1448
1450
应用高速缓存
VPN
防火墙
IDS/IPS
Web/安全应用
1452
1456
应用高速缓存
iSCSI
NAS
SAN
存储应用
1454
FICON
FCP
F_Class
1442
1444
1446
路由
1440
批处理/解批处理
1438
FC二层
1436
SAN终结
1404
SAN
1402
套接字层
1434
事件管理器
1428
数据处理
1432
会话管理器
1430
TCP分类器
Hyper-TCP
1426
1424
虚拟边缘路由（IP）
1422
二层（以太网，CEE）
1420
LAN终结
LAN

图8B

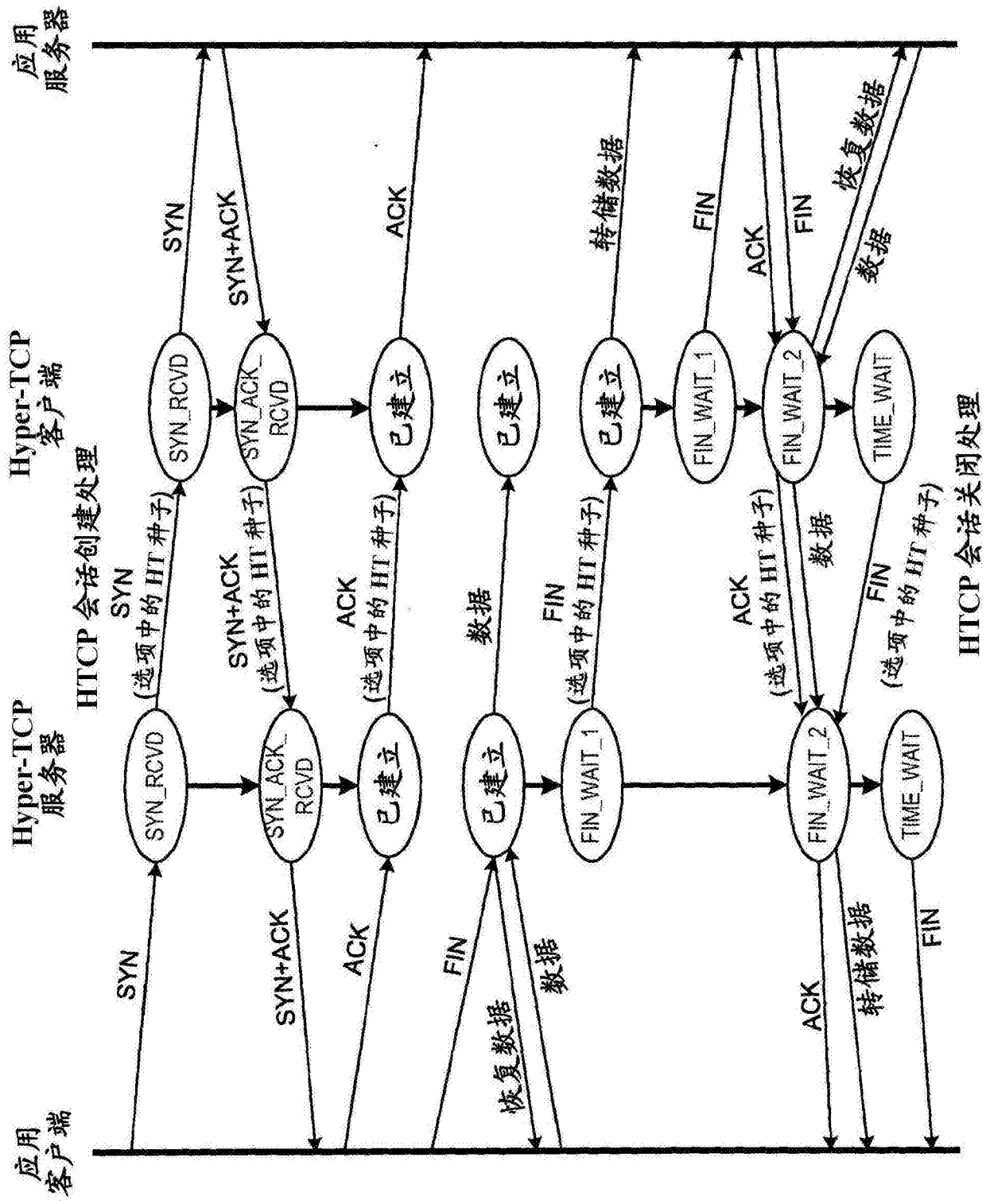

应用
客户端
Hyper-TCP
服务器
HTCP 会话创建处理
SYN
SYN_RCVD
SYN+ACK
SYN_ACK_
RCVD
ACK
已建立
数据
FIN
FIN_WAIT_1
FIN_WAIT_2
TIME_WAIT
恢复数据
转储数据
(选项中的 HTT 种子)
HTCP 会话关闭处理
Hyper-TCP
客户端
应用
服务器

图9

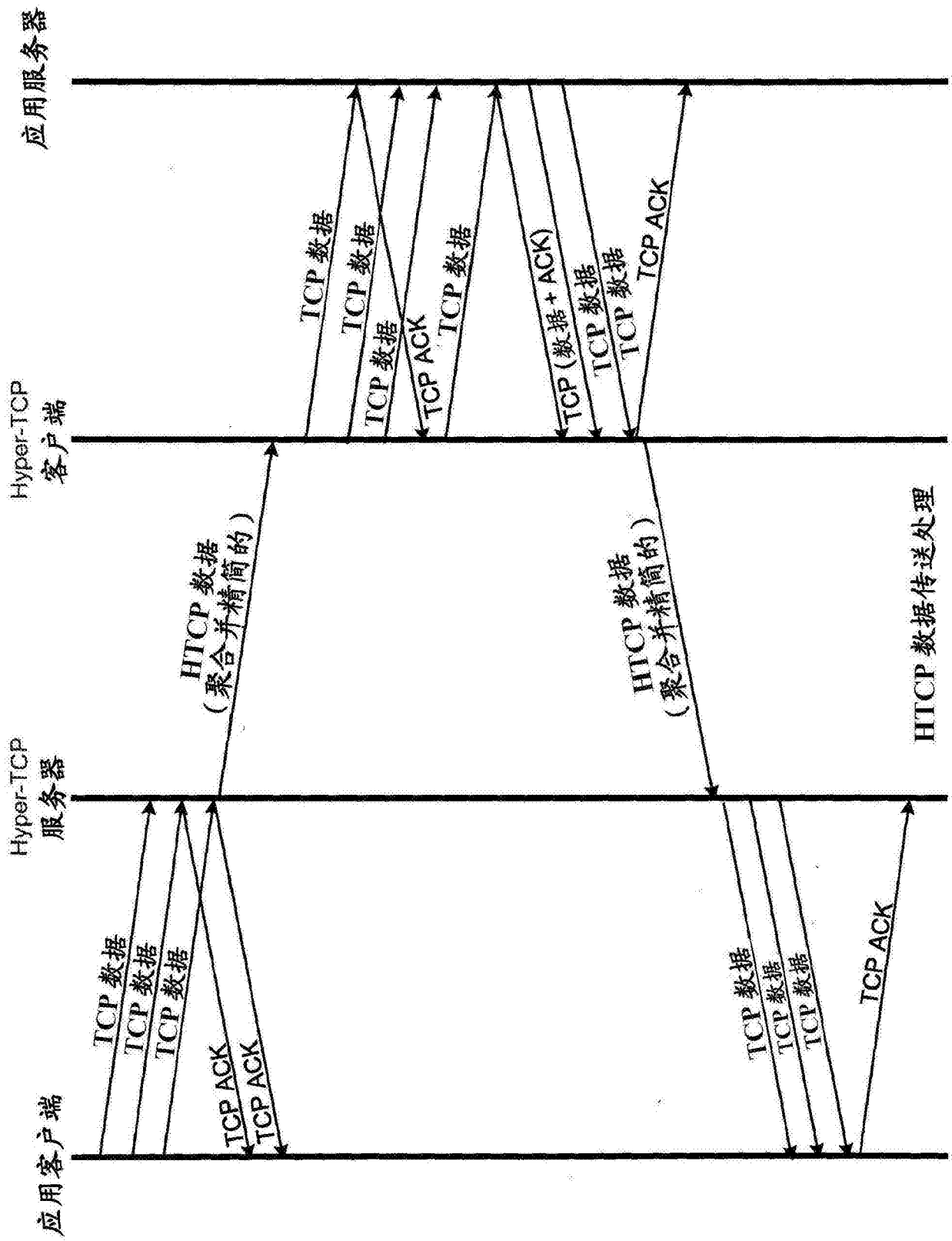

应用服务器
Hyper-TCP
客户端
Hyper-TCP
服务器
应用客户端
TCP 数据
TCP ACK
TCP (数据 + ACK)
HTCP 数据
(聚合并精简的)
HTCP 数据传送处理

图10

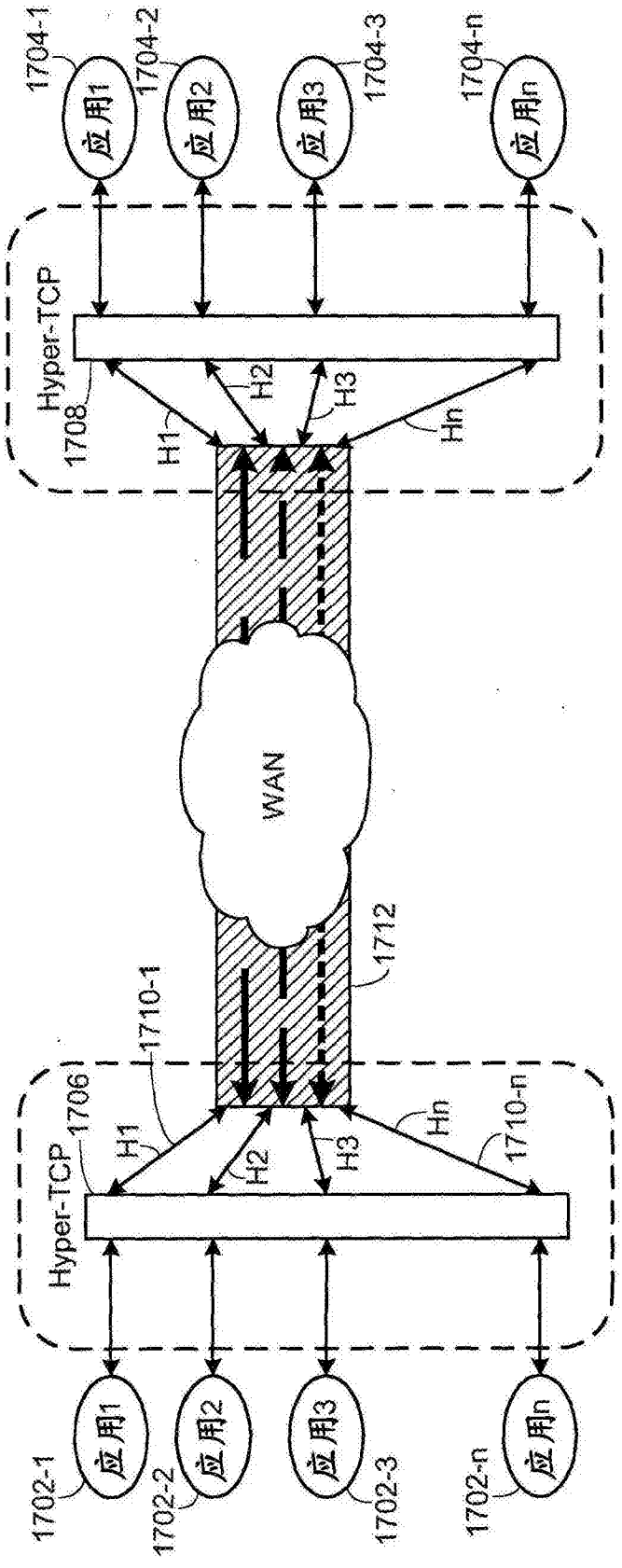

1704-1
1704-2
1704-3
1704-n
应用1
应用2
应用3
应用n
Hyper-TCP
1708
H1
H2
H3
Hn
WAN
1712
1710-1
1706
1710-n
1702-1
1702-2
1702-3
1702-n

图11

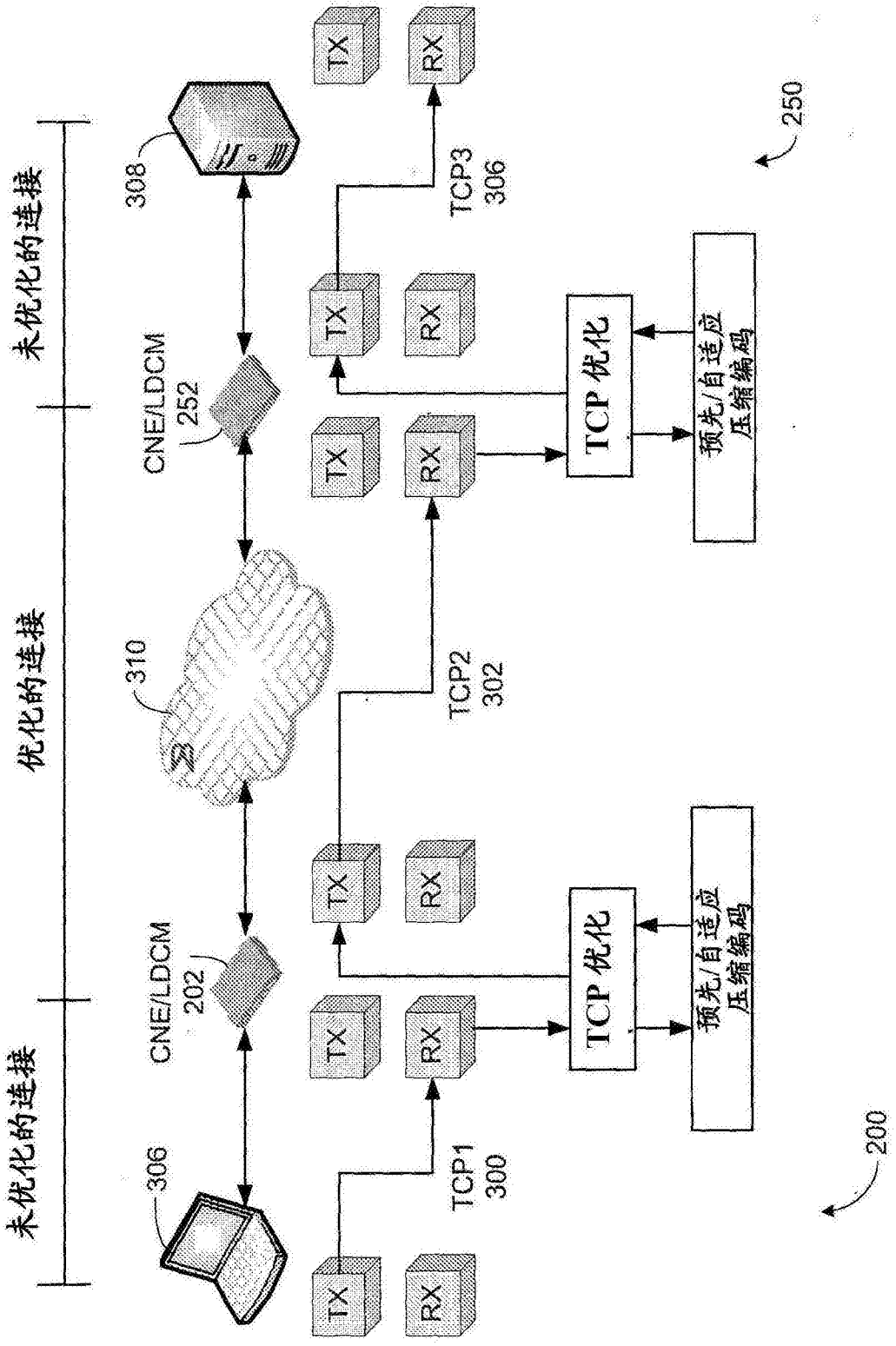

未优化的连接
优化的连接
未优化的连接
306
CNE/LDCM
202
310
CNE/LDCM
252
308
TX
RX
TCP1
300
TCP2
302
TCP3
306
TCP 优化
预先/自适应
压缩编码
250
200

图12

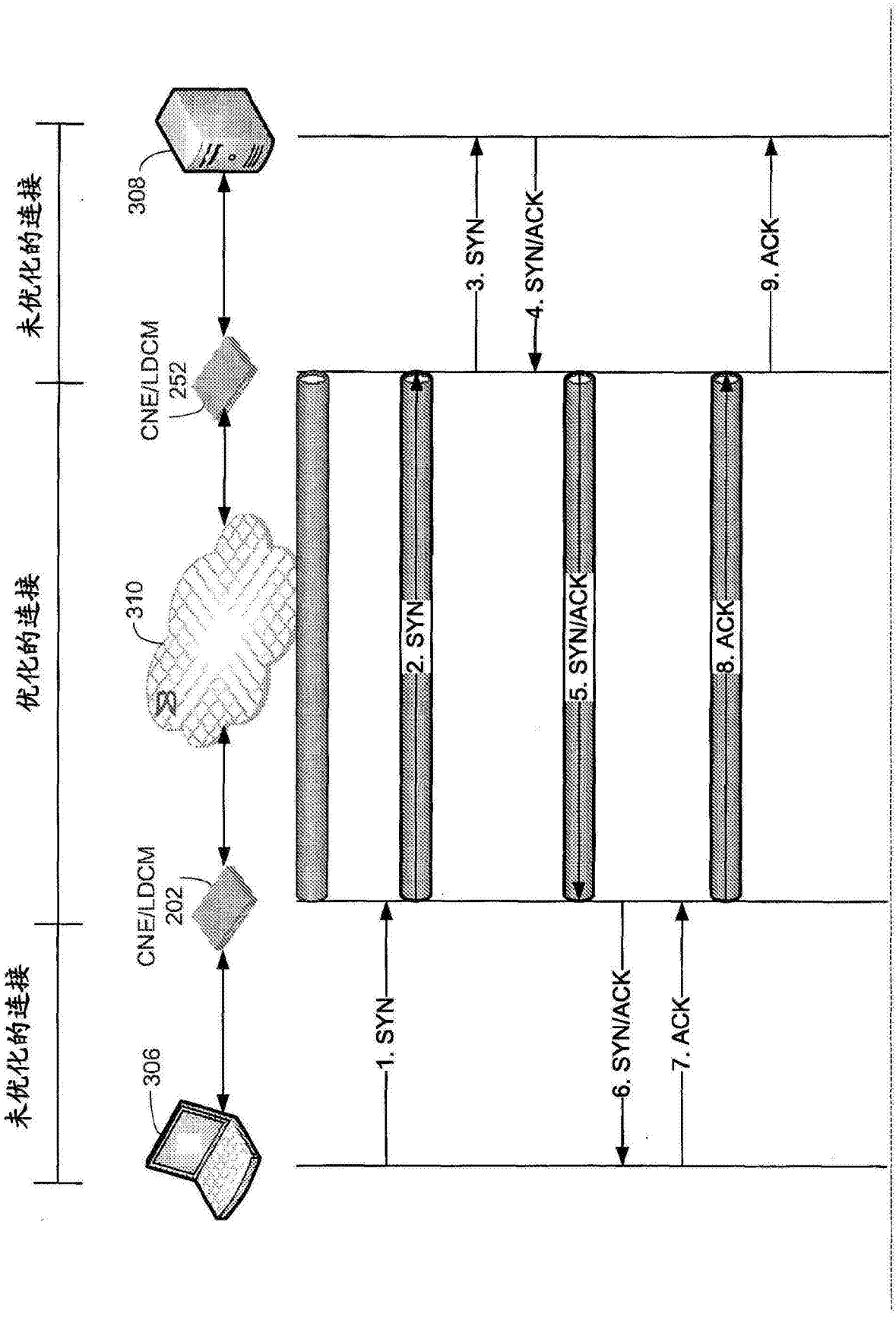

未优化的连接
优化的连接
未优化的连接
308
CNE/LDCM
252
310
CNE/LDCM
202
306
1. SYN
2. SYN
3. SYN
4. SYN/ACK
5. SYN/ACK
6. SYN/ACK
7. ACK
8. ACK
9. ACK

图13

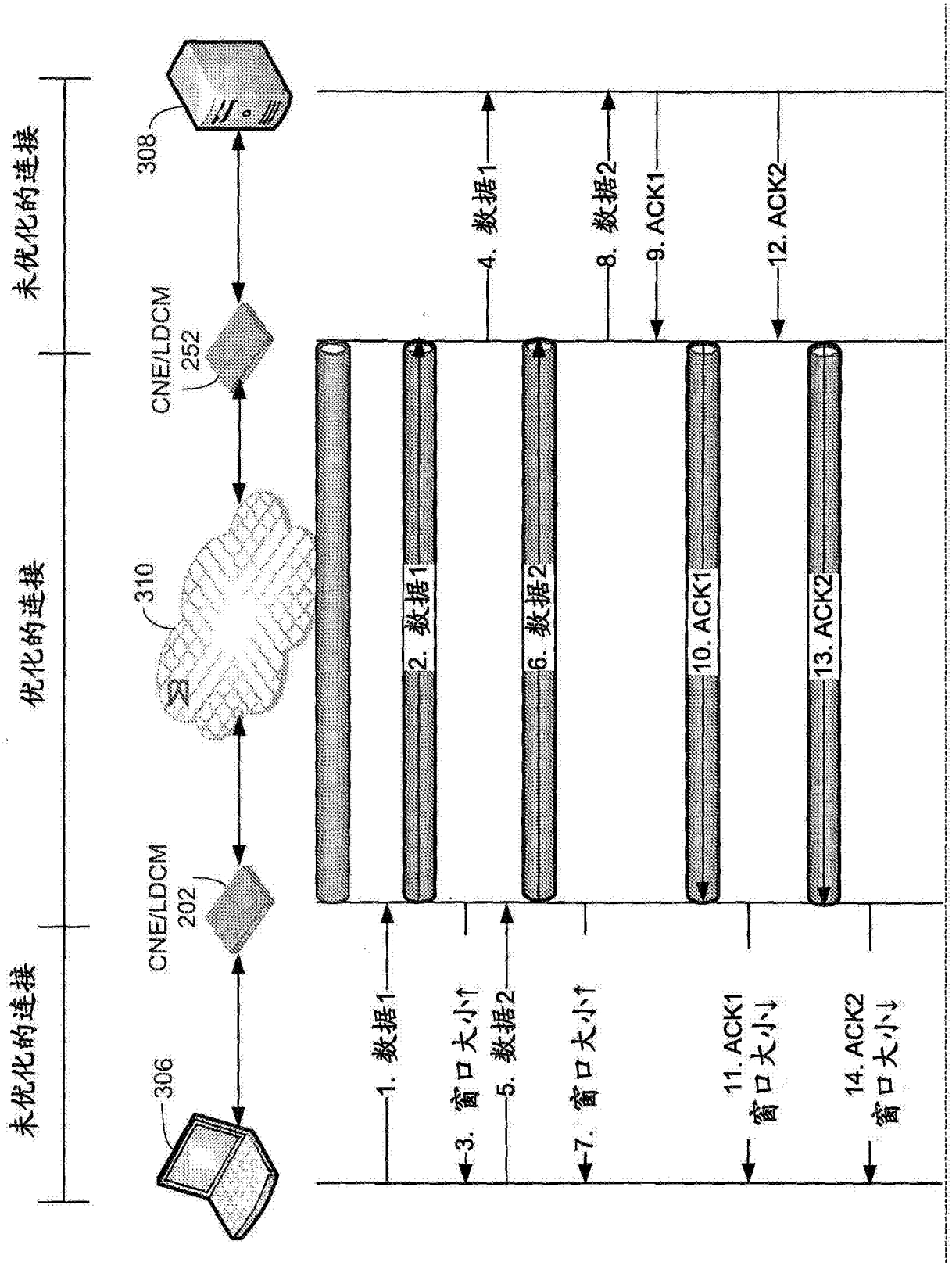

未优化的连接
优化的连接
未优化的连接
308
CNE/LDCM
252
310
CNE/LDCM
202
306
1. 数据1
2. 数据1
3. 窗口大小↑
4. 数据1
5. 数据2
6. 数据2
7. 窗口大小↑
8. 数据2
9. ACK1
10. ACK1
11. ACK1
窗口大小↓
12. ACK2
13. ACK2
14. ACK2
窗口大小↓

图14

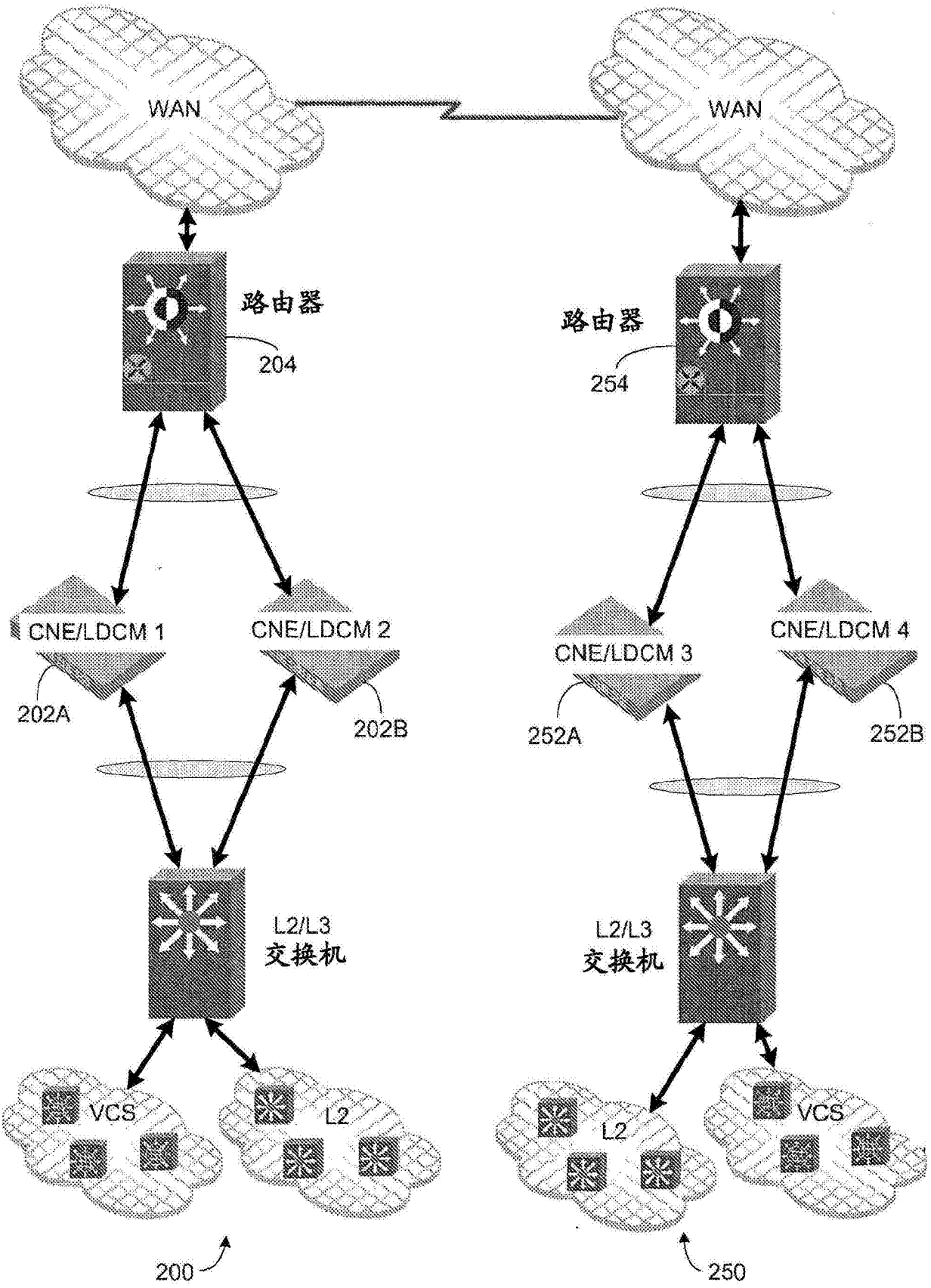

WAN
WAN
路由器
204
路由器
254
CNE/LDCM 1
CNE/LDCM 2
CNE/LDCM 3
CNE/LDCM 4
202A
202B
252A
252B
L2/L3
交换机
L2/L3
交换机
VCS
L2
L2
VCS
200
250

图15

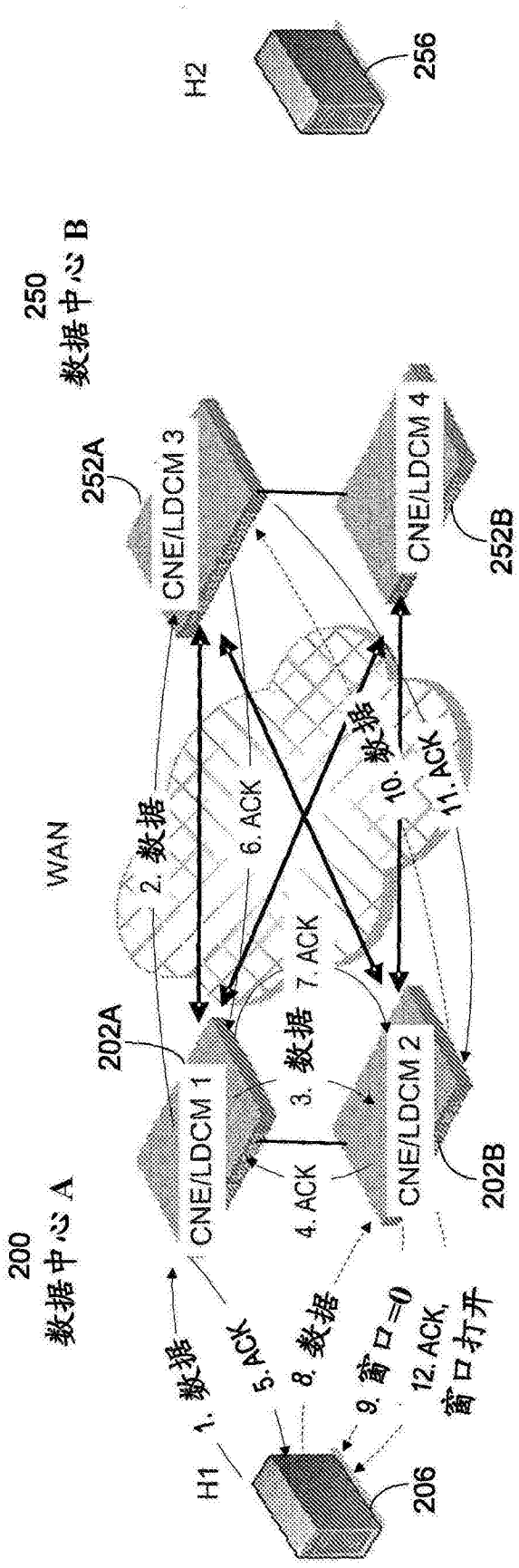

H2
256
250
数据中心 B
252A
CNE/LDCM 3
CNE/LDCM 4
252B
WAN
2. 数据
6. ACK
10. 数据
11. ACK
202A
CNE/LDCM 1
3. 数据
7. ACK
CNE/LDCM 2
202B
200
数据中心 A
4. ACK
1. 数据
5. ACK
8. 数据
9. 窗口=0
12. ACK,
窗口打开
H1
206

图16