(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 108141415 B
(45)授权公告日 2021.01.08

(21)申请号 201780003595.8
(22)申请日 2017.01.26
(65)同一申请的已公布的文献号
申请公布号 CN 108141415 A
(43)申请公布日 2018.06.08
(30)优先权数据
62/303,646 2016.03.04 US
15/413,143 2017.01.23 US
(85)PCT国际申请进入国家阶段日
2018.04.17
(86)PCT国际申请的申请数据
PCT/US2017/015167 2017.01.26
(87)PCT国际申请的公布数据
WO2017/151249 EN 2017.09.08
(73)专利权人 甲骨文国际公司
地址 美国加利福尼亚
(72)发明人 B·D·约翰森 D·G·莫克斯纳斯
B·博格丹斯基 P·文卡泰什
L·霍雷恩
(74)专利代理机构 中国贸促会专利商标事务所有限公司 11038
代理人 边海梅
(51)Int.Cl.
H04L 12/931(2006.01)
H04L 12/713(2006.01)
(56)对比文件
CN 103597795 A,2014.02.19
US 2005117598 A1,2005.06.02
WO 2013074697 A1,2013.05.23
CN 104079491 A,2014.10.01
审查员 张宇

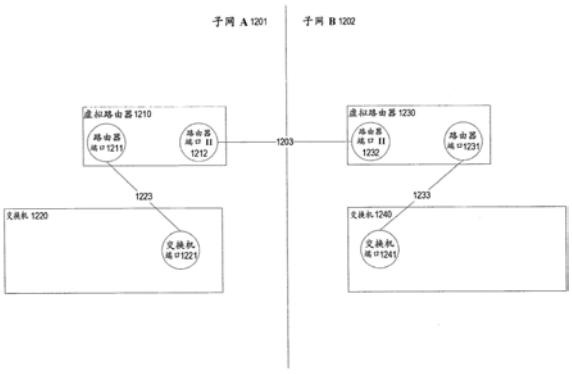
权利要求书3页 说明书16页 附图13页

(54)发明名称

用于支持高性能计算环境中的双端口虚拟路由器的系统和方法

(57)摘要

用于支持高性能计算环境中的双端口虚拟路由器的系统和方法。根据实施例，双端口路由器抽象可以提供简单的方式用于使得能够基于交换机硬件实现来定义子网到子网的路由器功能。虚拟双端口路由器可以逻辑上连接到对应交换机端口的外部。这个虚拟双端口路由器可以向标准管理实体(诸如子网管理器)提供符合InfiniBand规范的视图。根据实施例，双端口路由器模型意味着可以以每个子网完全控制分组的转发以及到子网的入口路径中的地址映射的方式来连接不同的子网。



CN 108141415 B

1.一种用于支持高性能计算环境中的双端口虚拟路由器的系统,包括:

一个或多个微处理器;

第一子网,第一子网包括:

多个交换机,所述多个交换机至少包括叶交换机,其中所述多个交换机中的每个交换机包括多个交换机端口,

多个主机通道适配器,每个主机通道适配器包括至少一个主机通道适配器端口,

多个端节点,其中端节点中的每个端节点与所述多个主机通道适配器中的至少一个主机通道适配器相关联,以及

子网管理器,所述子网管理器在所述多个交换机和所述多个主机通道适配器中的一个上运行;以及

第二子网,第二子网包括:

第二子网的多个交换机,所述第二子网的多个交换机至少包括第二子网的叶交换机,其中所述第二子网的多个交换机中的每个交换机包括第二子网的多个交换机端口,

第二子网的多个主机通道适配器,第二子网的每个主机通道适配器包括至少一个主机通道适配器端口,

第二子网的多个端节点,其中第二子网的端节点中的每个端节点与多个主机通道适配器中的至少一个主机通道适配器相关联,以及

第二子网的子网管理器,所述第二子网的子网管理器在所述第二子网的多个交换机和所述第二子网的多个主机通道适配器中的一个上运行;

其中所述多个交换机中的交换机上的所述多个交换机端口中的交换机端口被配置为路由器端口;

其中被配置为所述路由器端口的所述交换机端口逻辑上连接到虚拟路由器的第一虚拟路由器端口,所述虚拟路由器包括至少两个虚拟路由器端口;

其中,所述第二子网的多个交换机中的交换机上的所述第二子网的多个交换机端口中的第二子网的交换机端口被配置为第二子网的路由器端口;

其中,被配置为所述第二子网的路由器端口的所述第二子网的交换机端口逻辑上连接到第二子网的虚拟路由器的第一虚拟路由器端口,所述第二子网的虚拟路由器包括至少两个虚拟路由器端口;并且

其中,第一子网经由物理链路与第二子网互连,其中所述物理链路的第一端附连到所述第一子网的虚拟路由器的至少第二虚拟路由器端口,并且其中所述物理链路的第二端附连到所述第二子网的虚拟路由器的至少第二虚拟路由器端口。

2.如权利要求1所述的系统,其中所述子网管理器检测所述至少两个虚拟路由器端口中的第一虚拟路由器端口作为所述子网的端点。

3.如权利要求1或2所述的系统,其中所述第二子网的子网管理器检测所述第二子网的第一虚拟路由器端口作为第二子网的端点。

4.一种用于支持高性能计算环境中的双端口虚拟路由器的方法,包括:

在包括一个或多个微处理器的一个或多个计算机处提供:

第一子网,第一子网包括:

多个交换机,所述多个交换机至少包括叶交换机,其中所述多个交换机中的每个交换

机包括多个交换机端口，

多个主机通道适配器，每个主机通道适配器包括至少一个主机通道适配器端口，

多个端节点，其中端节点中的每个端节点与所述多个主机通道适配器中的至少一个主机通道适配器相关联，以及

子网管理器，所述子网管理器在所述多个交换机和所述多个主机通道适配器中的一个上运行；以及

第二子网，第二子网包括：

第二子网的多个交换机，所述第二子网的多个交换机至少包括第二子网的叶交换机，其中所述第二子网的多个交换机中的每个交换机包括第二子网的多个交换机端口，

第二子网的多个主机通道适配器，第二子网的每个主机通道适配器包括至少一个主机通道适配器端口，

第二子网的多个端节点，其中第二子网的端节点中的每个端节点与多个主机通道适配器中的至少一个主机通道适配器相关联，以及

第二子网的子网管理器，所述第二子网的子网管理器在所述第二子网的多个交换机和所述第二子网的多个主机通道适配器中的一个上运行；

将所述多个交换机中的交换机上的所述多个交换机端口中的交换机端口配置为路由器端口；

将被配置为所述路由器端口的所述交换机端口逻辑上连接到虚拟路由器的第一虚拟路由器端口，所述虚拟路由器包括至少两个虚拟路由器端口；

将所述第二子网的多个交换机中的交换机上的所述第二子网的多个交换机端口中的第二子网的交换机端口被配置为第二子网的路由器端口；以及

将被配置为所述第二子网的路由器端口的所述第二子网的交换机端口逻辑上连接到第二子网的虚拟路由器的第一虚拟路由器端口，所述第二子网的虚拟路由器包括至少两个虚拟路由器端口；

其中，第一子网经由物理链路与第二子网互连，其中所述物理链路的第一端附连到所述第一子网的虚拟路由器的至少第二虚拟路由器端口，并且其中所述物理链路的第二端附连到所述第二子网的虚拟路由器的至少第二虚拟路由器端口。

5.如权利要求4所述的方法，还包括：

由所述子网管理器检测所述至少两个虚拟路由器端口中的第一虚拟路由器端口作为所述子网的端点。

6.如权利要求4或5所述的方法，还包括：

由所述第二子网的子网管理器检测所述第二子网的第一虚拟路由器端口作为第二子网的端点。

7.一种非瞬态计算机可读存储介质，包括存储在其上用于支持高性能计算环境中的双端口虚拟路由器的指令，所述指令在由一个或多个计算机读取和执行时，使得所述一个或多个计算机执行包括以下的步骤：

在包括一个或多个微处理器的一个或多个计算机处提供：

第一子网，第一子网包括：

多个交换机，所述多个交换机至少包括叶交换机，其中所述多个交换机中的每个交换

机包括多个交换机端口，

多个主机通道适配器，每个主机通道适配器包括至少一个主机通道适配器端口，

多个端节点，其中端节点中的每个端节点与所述多个主机通道适配器中的至少一个主机通道适配器相关联，以及

子网管理器，所述子网管理器在所述多个交换机和所述多个主机通道适配器中的一个上运行；以及

第二子网，第二子网包括：

第二子网的多个交换机，所述第二子网的多个交换机至少包括第二子网的叶交换机，其中所述第二子网的多个交换机中的每个交换机包括第二子网的多个交换机端口，

第二子网的多个主机通道适配器，第二子网的每个主机通道适配器包括至少一个主机通道适配器端口，

第二子网的多个端节点，其中第二子网的端节点中的每个端节点与多个主机通道适配器中的至少一个主机通道适配器相关联，以及

第二子网的子网管理器，所述第二子网的子网管理器在所述第二子网的多个交换机和所述第二子网的多个主机通道适配器中的一个上运行；

将所述多个交换机中的交换机上的所述多个交换机端口中的交换机端口配置为路由器端口；

将被配置为所述路由器端口的所述交换机端口逻辑上连接到虚拟路由器的第一虚拟路由器端口，所述虚拟路由器包括至少两个虚拟路由器端口；

将所述第二子网的多个交换机中的交换机上的所述第二子网的多个交换机端口中的第二子网的交换机端口被配置为第二子网的路由器端口；以及

将被配置为所述第二子网的路由器端口的所述第二子网的交换机端口逻辑上连接到第二子网的虚拟路由器的第一虚拟路由器端口，所述第二子网的虚拟路由器包括至少两个虚拟路由器端口；

其中，第一子网经由物理链路与第二子网互连，其中所述物理链路的第一端附连到所述第一子网的虚拟路由器的至少第二虚拟路由器端口，并且其中所述物理链路的第二端附连到所述第二子网的虚拟路由器的至少第二虚拟路由器端口。

8.如权利要求7所述的非瞬态计算机可读存储介质，所述步骤还包括：

由所述子网管理器检测所述至少两个虚拟路由器端口中的第一虚拟路由器端口作为所述子网的端点；以及

由所述第二子网的子网管理器检测所述第二子网的第一虚拟路由器端口作为第二子网的端点。

9.一种用于支持高性能计算环境中的双端口虚拟路由器的装置，包括用于执行如权利要求4至6中任一项所述的方法的部件。

用于支持高性能计算环境中的双端口虚拟路由器的系统和方法

[0001] 版权声明

[0002] 本专利文献的公开内容的一部分包含受版权保护的材料。版权所有者不反对任何人对专利文献或专利公开内容如同其在专利和商标局专利文件或记录中所呈现的那样进行传真复制,但是以其他方式保留所有版权权利。

技术领域

[0003] 本发明一般而言涉及计算机系统,并且特别地涉及支持高性能计算环境中的双端口虚拟路由器。

背景技术

[0004] 随着更大的云计算体系架构的引入,与传统网络和存储装置相关联的性能和管理瓶颈已成为重要的问题。人们对使用诸如InfiniBand(IB)技术的高性能无损互连作为云计算架构的基础越来越感兴趣。这是本发明的实施例旨在解决的一般领域。

发明内容

[0005] 本文描述的是用于支持高性能计算环境中的双端口虚拟路由器的系统和方法。示例性方法可以在包括一个或多个微处理器的一个或多个计算机处提供第一子网,第一子网包括:多个交换机,该多个交换机至少包括叶交换机,其中该多个交换机中的每个交换机包括多个交换机端口;多个主机通道适配器,每个主机通道适配器包括至少一个主机通道适配器端口;多个端节点,其中端节点中的每个端节点与多个主机通道适配器中的至少一个主机通道适配器相关联;以及子网管理器,该子网管理器在多个交换机和多个主机通道适配器中的一个上运行。该方法可以将多个交换机中的交换机上的多个交换机端口中的交换机端口配置为路由器端口。被配置为路由器端口的交换机端口可以逻辑上连接到虚拟路由器,该虚拟路由器包括至少两个虚拟路由器端口。

[0006] 根据实施例,被配置为路由器端口的交换机端口可以逻辑上连接到至少两个虚拟路由器端口中的第一虚拟路由器端口。

[0007] 根据实施例,示例性方法可以提供第二子网。第二子网可以包括:第二子网的多个交换机,第二子网的多个交换机至少包括第二子网的叶交换机,其中第二子网的多个交换机中的每个交换机包括第二子网的多个交换机端口;第二子网的多个主机通道适配器,第二子网的每个主机通道适配器包括第二子网的至少一个主机通道适配器端口;以及第二子网的子网管理器,第二子网的子网管理器在第二子网的多个交换机和第二子网的多个主机通道适配器中的一个上运行。该方法可以将第二子网的多个交换机中的交换机上的第二子网的多个交换机端口中的第二子网的交换机端口配置为第二子网的路由器端口。被配置为第二子网的路由器端口的第二子网的交换机端口可以逻辑上连接到第二子网的虚拟路由器,第二子网的虚拟路由器包括第二子网的至少两个虚拟路由器端口。最后,第一子网可以

经由物理链路与第二子网互连。

[0008] 根据实施例,(第一子网或第二子网中的任一个的)多个主机通道适配器中的一个或多个主机通道适配器可以包括至少一个虚拟功能、至少一个虚拟交换机和至少一个物理功能。(第一子网或第二子网的)多个端节点可以包括物理主机、虚拟机或者物理主机与虚拟机的组合,其中虚拟机与至少一个虚拟功能相关联。

附图说明

[0009] 图1示出了根据实施例的InfiniBand环境的图示。

[0010] 图2示出了根据实施例的分区集群环境的图示。

[0011] 图3示出了根据实施例的网络环境中的树形拓扑的图示。

[0012] 图4示出了根据实施例的示例性共享端口体系架构。

[0013] 图5示出了根据实施例的示例性vSwitch体系架构。

[0014] 图6示出了根据实施例的示例性vPort体系架构。

[0015] 图7示出了根据实施例的具有预填充的LID的示例性vSwitch体系架构。

[0016] 图8示出了根据实施例的具有动态LID分配的示例性vSwitch体系架构。

[0017] 图9示出了根据实施例的具有动态LID分配和预填充的LID的vSwitch的示例性vSwitch体系架构。

[0018] 图10示出了根据实施例的示例性多子网InfiniBand架构。

[0019] 图11示出了根据实施例的在高性能计算环境中的两个子网之间的互连。

[0020] 图12示出了根据实施例的在高性能计算环境中经由双端口虚拟路由器配置的两个子网之间的互连。

[0021] 图13示出了根据实施例的用于支持高性能计算环境中的双端口虚拟路由器的方法的流程图。

具体实施方式

[0022] 在附图的图中通过示例的方式而非限制的方式图示了本发明,附图中相同的标号指示类似的元件。应当注意的是,在本公开中对实施例或“一个”实施例或“一些”实施例的引用不一定是相同的实施例,并且这种引用意味着至少一个。虽然讨论了特定的实现,但是应当理解的是,仅为了说明性目的而提供特定的实现。相关领域的技术人员将认识到,在不脱离本发明的范围和精神的情况下,可以使用其它组件和配置。

[0023] 贯穿附图和具体实施方式可以使用共同的附图标记来指示相同的元件;因此,如果在其它地方描述了元件,则在图中使用的附图标记可以或可以不在特定于该图的具体实施方式中被引用。

[0024] 本文描述的是支持高性能计算环境中的双端口虚拟路由器的系统和方法。

[0025] 本发明的以下描述使用InfiniBand™(IB)网络作为高性能网络的示例。贯穿以下描述,可以对InfiniBand™规范(也被不同地称为InfiniBand规范、IB规范或传统IB规范)进行参考。这样的参考被理解为是指可在http://www.inifinibandta.org获得的于2015年3月发布的**InfiniBand®**贸易协会体系架构规范(**InfiniBand®** Trade Association Architecture Specification),卷1,版本1.3,该规范的全文通过引用并入本文。对于本领

域技术人员来说显而易见的是,可以在没有限制的情况下使用其它类型的高性能网络。以下的描述还使用胖树(fat-tree)拓扑作为架构拓扑的示例。对于本领域技术人员来说显而易见的是,可以在没有限制的情况下使用其它类型的架构拓扑。

[0026] 为了满足当前时代(例如,Exascale时代)的云的需求,期望虚拟机能够利用诸如远程直接存储器访问(RDMA)的低开销网络通信范例。RDMA绕过OS堆栈并且直接与硬件通信,因此可以使用像单根I/O虚拟化(SR-IOV)网络适配器这样的透传(pass-through)技术。根据实施例,对于高性能无损互连网络中的适用性可以提供虚拟交换机(vSwitch)SR-IOV体系架构。由于网络重新配置时间对于使实时迁移成为实用选项是至关重要的,因此除了网络体系架构之外,还可以提供可伸缩的和拓扑无关的动态重新配置机制。

[0027] 根据实施例,并且此外,可以提供使用vSwitch的虚拟化环境的路由策略,并且可以提供用于网络拓扑(例如,胖树拓扑)的高效路由算法。动态重新配置机制可以被进一步调整以使胖树中的施加开销最小化。

[0028] 根据本发明的实施例,虚拟化可以有益于云计算中的高效资源利用和弹性资源分配。实时迁移使得有可能通过以应用透明的方式在物理服务器之间移动虚拟机(VM)来优化资源使用。因此,虚拟化可以通过实时迁移实现整合、资源的按需供给以及弹性。

[0029] InfiniBand™

[0030] InfiniBand™(IB)是由InfiniBand™贸易协会开发的开放标准无损网络技术。该技术基于供应高吞吐量和低延时通信的串行点对点全双工互连,特别地面向高性能计算(HPC)应用和数据中心。

[0031] InfiniBand™体系架构(IBA)支持双层拓扑划分。在下层,IB网络被称为子网,其中子网可以包括使用交换机和点对点链路进行互连的一组主机。在上层,IB架构构成可以使用路由器进行互连的一个或多个子网。

[0032] 在子网内,主机可以使用交换机和点对点链路来连接。另外,可以存在主管理实体:子网管理器(SM),该子网管理器驻留在子网中的指定设备上。子网管理器负责配置、激活和维护IB子网。另外,子网管理器(SM)可以负责在IB架构中执行路由表计算。这里,例如,IB网络的路由旨在在本地子网中的所有源和目的地对之间进行适当的负载平衡。

[0033] 通过子网管理接口,子网管理器与子网管理代理(SMA)交换被称为子网管理分组(SMP)的控制分组。子网管理代理驻留在每个IB子网设备上。通过使用SMP,子网管理器能够发现架构、对端节点和交换机进行配置,并从SMA接收通知。

[0034] 根据实施例,IB网络中的子网内路由可以基于存储在交换机中的线性转发表(LFT)。LFT由SM根据所使用的路由机制来计算。在子网中,交换机和端节点上的主机通道适配器(HCA)端口使用本地标识符(LID)进行寻址。线性转发表(LFT)中的每个条目包括目的地LID(DLID)和输出端口。仅支持表中每LID一个条目。当分组到达交换机时,通过在交换机的转发表中查找DLID来确定其输出端口。路由是确定性的,因为分组在网络中在给定的源-目的地对(LID对)之间采用相同的路径。

[0035] 一般而言,除了主子网管理器之外的所有其它子网管理器在备用模式中活动用于容错。但是,在主子网管理器发生故障的情况下,由备用子网管理器协商新的主子网管理器。主子网管理器还执行子网的周期性扫描,以检测任何拓扑变化并相应地对网络进行重新配置。

[0036] 此外,可以使用本地标识符(LID)来对子网内的主机和交换机进行寻址,并且可以将单个子网限制为49151个单播LID。除了作为在子网内有效的本地地址的LID之外,每个IB设备还可以具有64位全局唯一标识符(GUID)。GUID可以用于形成作为IB层三(L3)地址的全局标识符(GID)。

[0037] 在网络初始化时,SM可以计算路由表(即,子网内每对节点之间的连接/路由)。此外,每当拓扑改变时,都可以更新路由表以便确保连接性和最佳性能。在正常操作期间,SM可以执行网络的周期性轻扫描(light sweep)以检查拓扑变化。如果在轻扫描期间发现变化,或者如果SM接收到发信号通知网络变化的信息(陷阱),则SM可以根据所发现的变化来对网络进行重新配置。

[0038] 例如,当网络拓扑改变时(诸如当链路断开时、当设备被添加时或者当链路被移除时),SM可以对网络进行重新配置。重新配置步骤可以包括在网络初始化期间执行的步骤。此外,重新配置可以具有限于其中发生网络改变的子网的本地范围。此外,用路由器对大型架构进行分段可以限制重新配置的范围。

[0039] 图1中示出了示例InfiniBand架构,图1示出了根据实施例的InfiniBand环境100的图示。在图1中示出的示例中,节点A-E 101-105使用InfiniBand架构120经由相应的主机通道适配器111-115进行通信。根据实施例,各种节点(例如,节点A-E 101-105)可以由各种物理设备来表示。根据实施例,各种节点(例如,节点A-E 101-105)可以由诸如虚拟机的各种虚拟设备来表示。

[0040] 在InfiniBand中分区

[0041] 根据实施例,IB网络可以支持分区作为安全机制,以提供对共享网络架构的系统的逻辑组的隔离。架构中的节点上的每个HCA端口可以是一个或多个分区的成员。分区成员资格由集中式分区管理器管理,集中式分区管理器可以是SM的一部分。SM可以将每个端口的分区成员资格信息配置为16位分区密钥(P_Key)的表。SM还可以用包含与通过交换机和路由器端口发送或接收数据业务的端节点相关联的P_Key信息的分区实施表来对这些端口进行配置。此外,在一般情况下,交换机端口的分区成员资格可以表示与在出口(朝链路)方向经由端口进行路由的LID间接相关联的所有成员资格的联合。

[0042] 根据实施例,分区是端口的逻辑组,使得组的成员只能与相同逻辑组的其它成员进行通信。在主机通道适配器(HCA)和交换机处,可以使用分区成员资格信息对分组进行过滤以实施隔离。一旦分组到达传入端口,就可以丢弃具有无效分区信息的分组。在分区的IB系统中,分区可以用于创建租户集群。在分区实施就位的情况下,节点不能与属于不同租户集群的其它节点进行通信。以这种方式,即使存在受损或恶意的租户节点,系统的安全性也能够得到保证。

[0043] 根据实施例,对于节点之间的通信,除管理队列对(QP0和QP1)以外,队列对(QP)和端到端上下文(EEC)可以被分配给特定的分区。然后,可以将P_Key信息添加到所发送的每个IB传输分组。当分组到达HCA端口或交换机时,可以针对由SM配置的表验证该分组的P_Key值。如果找到无效的P_Key值,则立即丢弃该分组。以这种方式,仅在共享分区的端口之间才允许通信。

[0044] 图2中示出了IB分区的示例,图2示出了根据实施例的分区的集群环境的图示。在图2中示出的示例中,节点A-E 101-105使用InfiniBand架构120经由相应的主机通道适配

器111-115进行通信。节点A-E被布置到分区中，即分区1 130、分区2 140和分区3 150。分区1包括节点A 101和节点D 104。分区2包括节点A 101、节点B 102和节点C 103。分区3包括节点C 103和节点E 105。由于分区的布置，节点D 104和节点E 105不被允许通信，因为这些节点不共享分区。同时，例如，节点A 101和节点C 103被允许通信，因为这些节点都是分区2 140的成员。

[0045] InfiniBand中的虚拟机

[0046] 在过去的十年中，虚拟化高性能计算(HPC)环境的前景已经有相当大的提高，因为通过硬件虚拟化支持已经实际上去除了CPU开销；通过对存储器管理单元进行虚拟化已经显著降低了存储器开销；通过使用快速SAN存储装置或分布式网络文件系统已经减少了存储开销；并且通过使用像单根输入/输出虚拟化(SR-IOV)的设备透传技术已经减少了网络I/O开销。现在，云有可能使用高性能互连解决方案来容纳虚拟HPC(vHPC)集群，并且递送必要的性能。

[0047] 但是，当与诸如InfiniBand(IB)的无损网络耦合时，由于在这些解决方案中使用的复杂寻址方案和路由方案，某些云功能(诸如虚拟机(VM)的实时迁移)仍然是个问题。IB是供应高带宽和低延时的互连网络技术，因此非常适合于HPC和其它通信密集型工作负载。

[0048] 用于将IB设备连接到VM的传统方法是通过利用具有直接分配的SR-IOV。但是，使用SR-IOV实现分配有IB主机通道适配器(HCA)的VM的实时迁移已经被证明是有挑战性的。每个IB连接的节点具有三个不同的地址：LID、GUID和GID。当发生实时迁移时，这些地址中的一个或多个改变。与迁移中的VM(VM-in-migration)进行通信的其它节点会丢失连接性。当发生这种情况时，通过向IB子网管理器(SM)发送子网管理(SA)路径记录查询来定位要重新连接到的虚拟机的新地址，可以尝试更新丢失的连接。

[0049] IB使用三种不同类型的地址。第一种类型的地址是16位本地标识符(LID)。SM向每个HCA端口和每个交换机分配至少一个唯一的LID。LID用于在子网内对业务进行路由。由于LID为16位长，因此可以做出65536个唯一地址组合，这些组合中只有49151个(0x0001-0xBFFF)可以用作单播地址。因此，可用单播地址的数量限定了IB子网的最大尺寸。第二种类型的地址是由制造商分配给每个设备(例如，HCA和交换机)和每个HCA端口的64位全局唯一标识符(GUID)。SM可以向HCA端口分配附加的子网唯一GUID，这在使用SR-IOV时是有用的。第三种类型的地址是128位全局标识符(GID)。GID是有效的IPv6单播地址，并且向每个HCA端口分配至少一个GID。GID通过组合由架构管理员分配的全局唯一64位前缀和每个HCA端口的GUID地址来形成。

[0050] 胖树(FTree)拓扑和路由

[0051] 根据实施例，基于IB的HPC系统中的一些采用胖树拓扑来利用胖树供应的有用属性。由于每个源目的地对之间有多条路径可用，因此这些属性包括完全的二分带宽和固有的容错。胖树背后的最初想法是，当树朝着拓扑的根移动时，在节点之间采用具有更多可用带宽的较胖链路。较胖链路可以帮助避免上层交换机中的拥塞并且维持二分带宽。

[0052] 图3示出了根据实施例的网络环境中的树形拓扑结构的图示。如图3所示，可以在网络架构200中连接一个或多个端节点201-204。网络架构200可以基于包括多个叶交换机211-214和多个主干交换机或根交换机231-234的胖树拓扑。此外，网络架构200可以包括一个或多个中间交换机，诸如交换机221-224。

[0053] 还如图3所示，端节点201-204中的每一个端节点可以是多宿主节点，即，通过多个端口连接到网络架构200的两个或更多个部分的单个节点。例如，节点201可以包括端口H1和H2，节点202可以包括端口H3和H4，节点203可以包括端口H5和H6，并且节点204可以包括端口H7和H8。

[0054] 此外，每个交换机可以具有多个交换机端口。例如，根交换机231可以具有交换机端口1-2，根交换机232可以具有交换机端口3-4，根交换机233可以具有交换机端口5-6，并且根交换机234可以具有交换机端口7-8。

[0055] 根据实施例，胖树路由机制是用于基于IB的胖树拓扑的最流行的路由算法中的一个路由算法。胖树路由机制也在OFED(开放架构企业分发-用于构建和部署基于IB的应用的标准软件堆栈)子网管理器OpenSM中实现。

[0056] 胖树路由机制旨在生成在网络架构中跨链路均匀传播(spread)最短路径路由的LFT。该机制按索引次序遍历架构并将端节点的目标LID(以及因此对应的路由)分配给每个交换机端口。对于连接到相同的叶交换机的端节点，索引次序可以取决于端节点连接到的交换机端口(即端口编号序列)。对于每个端口，该机制可以维护端口使用计数器，并且可以在每次添加新路由时使用这个端口使用计数器来选择最少使用的端口。

[0057] 根据实施例，在分区的子网中，不允许不作为公共分区的成员的节点进行通信。在实践中，这意味着由胖树路由算法分配的一些路由不用于用户业务。当胖树路由机制以与它针对其它功能路径所做的相同的方式为那些路由生成LFT时，会出现该问题。由于节点按索引的次序进行路由，因此这种行为可能导致链路上的恶化的平衡。由于路由可以在与分区无关的情况下执行，因此，一般而言，胖树路由的子网提供分区之间较差的隔离。

[0058] 根据实施例，胖树是可以利用可用网络资源进行伸缩的分层网络拓扑。而且，使用放置在不同级别层次上的商用交换机容易构建胖树。通常可以获得胖树的不同变体，包括k-ary-n-trees、扩展广义胖树(XGFT)、并行端口广义胖树(PGFT)和现实生活胖树(RLFT)。

[0059] k-ary-n-tree是具有k^n个端节点和$n \cdot k^{n-1}$个交换机的n级胖树，每个交换机具有2k个端口。每个交换机在树中具有相同数量的上连接和下连接。XGFT胖树通过允许交换机不同数量的上连接和下连接以及在树中每个级别的不同数量的连接来扩展k-ary-n-tree。PGFT定义进一步拓宽了XGFT拓扑，并且允许交换机之间的多个连接。可以使用XGFT和PGFT来定义各种各样的拓扑。但是，为了实践的目的，引入了作为PGFT受限版本的RLFT来定义当今HPC集群中常见的胖树。RLFT在胖树的所有级别使用相同端口计数的交换机。

[0060] 输入/输出(I/O)虚拟化

[0061] 根据实施例，I/O虚拟化(IOV)可以通过允许虚拟机(VM)访问底层物理资源来提供I/O的可用性。存储业务和服务器间通信的组合施加了可能淹没(overwhelm)单个服务器的I/O资源的增加的负载，从而导致积压(backlog)以及由于处理器在等待数据而导致空闲处理器。随着I/O请求数量的增加，IOV可以提供可用性；并且可以提高(虚拟化)I/O资源的性能、可伸缩性和灵活性以现代CPU虚拟化中所见到的性能水平相匹配。

[0062] 根据实施例，IOV是所期望的，因为它可以允许共享I/O资源并且提供从VM对资源的受保护的访问。IOV将暴露于VM的逻辑设备与IOV的物理实现解除耦合。当前，可以存在不同类型的IOV技术，诸如仿真、半虚拟化、直接分配(DA)和单根I/O虚拟化(SR-IOV)。

[0063] 根据实施例，一种类型的IOV技术是软件仿真。软件仿真可以允许解除耦合的前

端/后端软件体系架构。前端可以是置于VM中的设备驱动程序，从而前端与由管理程序实现的后端进行通信以提供I/O访问。物理设备共享比率高，并且VM的实时迁移可能只需几毫秒的网络停机时间。但是，软件仿真引入了附加的非期望的计算开销。

[0064] 根据实施例，另一种类型的IOV技术是直接设备分配。直接设备分配涉及将I/O设备耦合到VM，其中在VM之间没有设备共享。直接分配或设备透传以最小的开销提供接近本地性能。物理设备绕过管理程序并且直接附连到VM。但是，这种直接设备分配的缺点是有限的可伸缩性，因为在虚拟机之间不存在共享-一个物理网卡与一个VM耦合。

[0065] 根据实施例，单根IOV(SR-IOV)可以允许物理设备通过硬件虚拟化表现为相同设备的多个独立的轻量级实例。这些实例可以被分配给VM作为透传设备，并作为虚拟功能(VF)被访问。管理程序通过唯一的(每设备)、全特征物理功能(PF)访问设备。SR-IOV使纯直接分配的可伸缩性问题变得容易。但是，SR-IOV呈现出的问题是SR-IOV可能影响VM迁移。在这些IOV技术中，SR-IOV可通过允许从多个VM直接访问单个物理设备同时维持接近本地性能来扩展PCI快速(PCI Express，PCIe)规范。因此，SR-IOV可以提供良好的性能和可伸缩性。

[0066] SR-IOV允许PCIe设备通过向每个访客(guest)分配一个虚拟设备来暴露可以在多个访客之间被共享的多个虚拟设备。每个SR-IOV设备具有至少一个物理功能(PF)以及一个或多个相关联的虚拟功能(VF)。PF是由虚拟机监视器(VMM)或管理程序控制的正常PCIe功能，而VF是轻量级的PCIe功能。每个VF都具有其自己的基地址(BAR)，并被分配有唯一的请求者ID，该请求者ID使得I/O存储器管理单元(IOMMU)能够区分来自/去往不同VF的业务流。IOMMU还在PF和VF之间应用存储器和中断转换。

[0067] 但是，令人遗憾的是，直接设备分配技术在数据中心优化期望虚拟机的透明实时迁移的情况下对云提供商施加障碍。实时迁移的本质是将VM的存储器内容复制到远程管理程序。然后VM在源管理程序处被暂停，并且VM的操作在目的地处恢复。当使用软件仿真方法时，网络接口是虚拟的，因此网络接口的内部状态被存储到存储器中并且也被复制。因此，可以使停机时间下降到几毫秒。

[0068] 但是，当使用诸如SR-IOV的直接设备分配技术时，迁移变得更加困难。在这种情况下，不能复制网络接口的完整内部状态，因为网络接口与硬件绑定。被分配给VM的SR-IOV VF反而被分离，实时迁移将运行，并且新的VF将在目的地处被附连。在InfiniBand和SR-IOV的情况下，该处理可能引入秒的数量级的停机时间。而且，在SR-IOV共享端口模型中，在迁移之后VM的地址将改变，从而导致SM中的附加开销并对底层网络架构的性能造成负面影响。

[0069] InfiniBand SR-IOV体系架构-共享端口

[0070] 可以存在不同类型的SR-IOV模型，例如共享端口模型、虚拟交换机模型和虚拟端口模型。

[0071] 图4示出了根据实施例的示例性共享端口体系架构。如图所绘出的，主机300(例如，主机通道适配器)可以与管理程序310进行交互，管理程序310可以将各种虚拟功能330、340、350分配给多个虚拟机。而且，可以由管理程序310处理物理功能。

[0072] 根据实施例，当使用诸如图4所绘出的共享端口体系架构时，主机(例如，HCA)在网络中出现为具有在物理功能320和虚拟功能330、350、350之间的共享队列对(QP)空间和单

个共享LID的单个端口。但是,每个功能(即,物理功能和虚拟功能)可以具有其自身的GID。

[0073] 如图4所示,根据实施例,可以将不同的GID分配给虚拟功能和物理功能,并且由物理功能拥有特殊队列对QP0和QP1(即,用于InfiniBand管理分组的专用队列对)。这些QP也被暴露给VF,但是不允许VF使用QP0(从VF到QP0的所有SMP被丢弃),并且QP1可以充当由PF所拥有的实际QP1的代理。

[0074] 根据实施例,共享端口体系架构可以允许不受(通过被分配给虚拟功能而附连到网络的)VM的数量限制的高度可伸缩的数据中心,因为LID空间仅被网络中的物理机器和交换机消耗。

[0075] 但是,共享端口体系架构的缺点是无法提供透明的实时迁移,从而阻碍了灵活VM放置的潜力。由于每个LID与特定管理程序相关联,并且在驻留在管理程序上的所有VM之间被共享,因此正迁移的VM(即,正迁移到目的地管理程序的虚拟机)必须将该VM的LID改变为目的地管理程序的LID。此外,由于受限的QP0访问,子网管理器不能在VM内部运行。

[0076] InfiniBand SR-IOV体系架构模型-虚拟交换机(vSwitch)

[0077] 图5示出了根据实施例的示例性vSwitch体系架构。如图所示,主机400(例如,主机通道适配器)可以与管理程序410进行交互,管理程序410可以将各种虚拟功能430、440、450分配给多个虚拟机。而且,可以由管理程序410处理物理功能。也可以由管理程序401处理虚拟交换机415。

[0078] 根据实施例,在vSwitch体系架构中,每个虚拟功能430、440、450是完整的虚拟主机通道适配器(vHCA),意味着分配给VF的VM被分配了完整的一组IB地址(例如,GID、GUID、LID)和硬件中的专用QP空间。对于网络的其余部分和SM,HCA 400经由虚拟交换机415看起来像具有连接的附加节点的交换机。管理程序410可以使用PF 420,并且(附连到虚拟功能的)VM使用VF。

[0079] 根据实施例,vSwitch体系架构提供透明的虚拟化。但是,由于每个虚拟功都被分配唯一的LID,因此可用LID的数量被迅速消耗。而且,在使用许多LID地址(即,每个物理功能和每个虚拟功能各有一个LID地址)的情况下,需要由SM计算更多的通信路径,并且需要将更多的子网管理分组(SMP)发送到交换机,以便更新交换机的LFT。例如,通信路径的计算在大型网络中可能花费几分钟。因为LID空间限于49151个单播LID,并且由于(经由VF的)每个VM、物理节点和交换机每个占用一个LID,因此网络中的物理节点和交换机的数量限制了活动VM的数量,反之亦然。

[0080] InfiniBand SR-IOV体系架构模型-虚拟端口(vPort)

[0081] 图6示出了根据实施例的示例性vPort概念。如图所绘出的,主机300(例如,主机通道适配器)可以与管理程序410进行交互,管理程序410可以将各种虚拟功能330、340、350分配给多个虚拟机。而且,可以由管理程序310处理物理功能。

[0082] 根据实施例,vPort概念被松散地定义以便赋予供应商实现的自由(例如,定义不规定实现必须是特定于SRIOV的),并且vPort的目标是使VM在子网中的处理方式标准化。利用vPort概念,可以定义可在空间域和性能域都更可伸缩的类似SR-IOV共享端口和类似vSwitch体系架构或两者的组合。vPort支持可选的LID,并且与共享端口不同,即使vPort不使用专用LID,SM也知道子网中可用的所有vPort。

[0083] InfiniBand SR-IOV体系架构模型-具有预填充LID的vSwitch

[0084] 根据实施例，本公开提供了用于提供具有预填充的LID的vSwitch体系架构的系统和方法。

[0085] 图7示出了根据实施例的具有预填充的LID的示例性vSwitch体系架构。如图所绘出的，多个交换机501-504可以在网络交换环境600(例如，IB子网)内提供在架构(诸如InfiniBand架构)的成员之间的通信。该架构可以包括多个硬件设备，诸如主机通道适配器510、520、530。主机通道适配器510、520、530中的每个适配器进而可以分别与管理程序511、521和531进行交互。每个管理程序进而可以结合与该管理程序进行交互的主机通道适配器建立多个虚拟功能514、515、516、524、525、526、534、535、536并将这些虚拟功能分配给多个虚拟机。例如，虚拟机1 550可以由管理程序511分配给虚拟功能1 514。管理程序511可以附加地将虚拟机2 551分配给虚拟功能2 515，并且将虚拟机3 552分配给虚拟功能3 516。管理程序531进而可以将虚拟机4 553分配给虚拟功能1 534。管理程序可以通过主机通道适配器中的每个主机通道适配器上的全特征物理功能513、523、533来访问主机通道适配器。

[0086] 根据实施例，交换机501-504中的每个交换机可以包括多个端口(未示出)，这些端口用于设置线性转发表以便在网络交换环境600内引导业务。

[0087] 根据实施例，虚拟交换机512、522和532可以由它们相应的管理程序511、521、531处理。在这样的vSwitch体系架构中，每个虚拟功能是完整的虚拟主机通道适配器(vHCA)，意味着被分配给VF的VM被分配了完整的一组IB地址(例如，GID、GUID、LID)和硬件中的专用QP空间。对于网络的其余部分和SM(未示出)，HCA 510、520和530经由虚拟交换机看起来像具有连接的附加节点的交换机。

[0088] 根据实施例，本公开提供了用于提供具有预填充的LID的vSwitch体系架构的系统和方法。参考图7，LID被预填充到各种物理功能513、523、533以及虚拟功能514-516、524-526、534-536(甚至当前未与活动虚拟机相关联的那些虚拟功能)。例如，用LID1预填充物理功能513，而用LID 10预填充虚拟功能1 534。当启动网络时，在启用SR-IOV vSwitch的子网中预填充LID。即使当并非所有的VF都在网络中被VM占用时，填充后的VF也被分配有LID，如图7所示。

[0089] 根据实施例，非常类似于物理主机通道适配器可以具有多于一个的端口(两个端口用于冗余是常见的)，虚拟HCA也可以用两个端口表示，并且经由一个、两个或更多个虚拟交换机连接到外部IB子网。

[0090] 根据实施例，在具有预填充LID的vSwitch体系架构中，每个管理程序可以通过PF为自己消耗一个LID，并为每个附加的VF消耗更多一个LID。在IB子网中的所有管理程序中可用的所有VF的总和给出了允许在子网中运行的VM的最大数量。例如，在子网中的每管理程序具有16个虚拟功能的IB子网中，则每个管理程序在子网中消耗17个LID(16个虚拟功能中的每个虚拟功能消耗一个LID加上用于物理功能的一个LID)。在这种IB子网中，由可用单播LID的数量规定针对单个子网的理论管理程序极限，并且该理论管理程序极限是：2891(49151个可用LID除以每管理程序的17个LID)，并且VM的总数(即，极限)是46256(2891个管理程序乘以每管理程序的16个VF)。(实际上，这些数字实际上较小，因为IB子网中的每个交换机、路由器或专用SM节点也消耗LID)。要注意的是，vSwitch不需要占用附加的LID，因为它可以与PF共享LID。

[0091] 根据实施例，在具有预填充LID的vSwitch体系架构中，第一次启动网络时为所有

LID计算通信路径。当需要启动新的VM时，系统不需要在子网中添加新的LID，否则该动作将造成网络的完全重新配置，包括作为最耗时部分的路径重新计算。相反，VM的可用端口(即，可用的虚拟功能)定位于管理程序中的一个管理程序中，并且虚拟机被附连到该可用的虚拟功能。

[0092] 根据实施例，具有预填充LID的vSwitch体系架构还允许计算和使用不同路径以到达由相同管理程序托管的不同VM的能力。本质上，这允许这样的子网和网络使用类似LID掩码控制(LMC)的特征来向一个物理机器提供替代路径，而不受要求LID必须是连续的LMC的限制约束。当需要迁移VM并将该VM相关联的LID携带到目的地时，自由使用非连续LID尤其有用。

[0093] 根据实施例以及以上示出的具有预填充LID的vSwitch体系架构的益处，可以进行某些考虑。例如，由于当启动网络时LID在启用SR-IOV vSwitch的子网中被预填充，因此(例如，启动时的)初始路径计算可能比如果LID没有被预填充的情况花费更长的时间。

[0094] InfiniBand SR-IOV体系架构模型-具有动态LID分配的vSwitch

[0095] 根据实施例，本公开提供了用于提供具有动态LID分配的vSwitch体系架构的系统和方法。

[0096] 图8示出了根据实施例的具有动态LID分配的示例性vSwitch体系架构。如图所绘出的，多个交换机501-504可以在网络交换环境700(例如，IB子网)内提供架构(诸如InfiniBand架构)的成员之间的通信。该架构可以包括多个硬件设备，诸如主机通道适配器510、520、530。主机通道适配器510、520、530中的每个主机通道适配器进而可以分别与管理程序511、521、531进行交互。每个管理程序进而可以结合与该管理程序进行交互的主机通道适配器建立多个虚拟功能514、515、516、524、525、526、534、535、536并将这些虚拟功能分配给多个虚拟机。例如，虚拟机1 550可以由管理程序511分配给虚拟功能1 514。管理程序511可以附加地将虚拟机2 551分配给虚拟功能2 515，并且将虚拟机3 552分配给虚拟功能3 516。管理程序531进而可以将虚拟机4 553分配给虚拟功能1 534。管理程序可以通过主机通道适配器中的每个主机通道适配器上的全特征物理功能513、523、533来访问主机通道适配器。

[0097] 根据实施例，交换机501-504中的每个交换机可以包括多个端口(未示出)，这些端口用于设置线性转发表以便在网络交换环境700内引导业务。

[0098] 根据实施例，虚拟交换机512、522和532可以由它们相应的管理程序511、521、531处理。在这样的vSwitch体系架构中，每个虚拟功能是完整的虚拟主机通道适配器(vHCA)，意味着被分配给VF的VM被分配了完整的一组IB地址(例如，GID、GUID、LID)和硬件中的专用QP空间。对于网络的其余部分和SM(未示出)，HCA 510、520和530经由虚拟交换机看起来像具有连接的附加节点的交换机。

[0099] 根据实施例，本公开提供了用于提供具有动态LID分配的vSwitch体系架构的系统和方法。参考图8，LID被动态分配给各种物理功能513、523、533，其中物理功能513接收LID 1、物理功能523接收LID 2并且物理功能533接收LID 3。与活动虚拟机相关联的那些虚拟功能也可以接收动态分配的LID。例如，由于虚拟机1550是活动的并且与虚拟功能1 514相关联，因此虚拟功能514可以被分配LID 5。同样，虚拟功能2 515、虚拟功能3 516和虚拟功能1534每个与活动虚拟功能相关联。由此，这些虚拟功能按以下方式被分配LID，其中LID 7被

分配给虚拟功能2 515、LID 11被分配给虚拟功能3 516、并且LID 9被分配给虚拟功能1 534。与具有预填充LID的vSwitch不同，当前未与活动虚拟机相关联的那些虚拟功能不接收LID分配。

[0100] 根据实施例，利用动态LID分配，可以显著减少初始路径计算。当第一次启动网络并且不存在VM时，则可以使用相对较少数量的LID用于初始路径计算和LFT分发。

[0101] 根据实施例，非常类似于物理主机通道适配器可以具有多于一个的端口(两个端口用于冗余是常见的)，虚拟HCA也可以用两个端口表示，并且经由一个、两个或更多个虚拟交换机连接到外部IB子网。

[0102] 根据实施例，当在利用具有动态LID分配的vSwitch的系统中创建新的VM时，找到空闲的VM时隙以便决定在哪个管理程序上启动新添加的VM，并且也找到唯一未使用的单播LID。但是，不存在网络中的已知路径和交换机的LFT用于处理新添加的LID。在每分钟可以启动几个VM的动态环境中，为了处理新添加的VM而计算新的一组路径是不期望的。在大型IB子网中，计算新的一组路由可以花费几分钟，并且这个过程将需要在每次启动新的VM时重复。

[0103] 有利地，根据实施例，由于管理程序中的所有VF与PF共享相同的上行链路，因此不需要计算新的一组路由。只需要遍历网络中所有物理交换机的LFT、将转发端口从属于(创建VM的)管理程序的PF的LID条目复制到新添加的LID、并且发送单个SMP以更新特定交换机的对应LFT块。因此，该系统和方法避免了需要计算新的一组路由。

[0104] 根据实施例，在具有动态LID分配的vSwitch体系架构中分配的LID不一定是连续的。当将在具有预填充LID的vSwitch中与在具有动态LID分配的vSwitch中的每个管理程序上的VM上分配的LID进行比较时，应当注意的是，在动态LID分配体系架构中分配的LID是非连续的，而被预填充的那些LID本质上是连续的。在vSwitch动态LID分配体系架构中，当创建新的VM时，在VM的整个生命期中使用下一个可用的LID。相反，在具有预填充LID的vSwitch中，每个VM继承已经被分配给对应VF的LID，并且在没有实时迁移的网络中，连续附连到给定VF的VM获得相同的LID。

[0105] 根据实施例，以一些附加的网络和运行时SM开销为代价，具有动态LID分配的vSwitch体系架构可以解决具有预填充LID的vSwitch体系架构模型的缺点。每次创建VM时，用与所创建的VM相关联的新添加的LID来更新子网中的物理交换机的LFT。对于这个操作，需要发送每交换机的一个子网管理分组(SMP)。因为每个VM正在使用与该VM的主机管理程序相同的路径，因此类似LMC的功能也是不可用的。但是，对所有管理程序中存在的VF的总量没有限制，并且VF的数量可以超过单播LID极限的数量。当然，如果是这种情况，则并非所有VF都被允许同时附连到活动VM上，但是，当接近单播LID极限进行操作时，具有更多的空闲管理程序和VF增加了灾难恢复和分段网络优化的灵活性。

[0106] InfiniBand SR-IOV体系架构模型-具有动态LID分配和预填充LID的vSwitch

[0107] 图9示出了根据实施例的具有动态LID分配和预填充LID的vSwitch的示例性vSwitch体系架构。如图所绘出的，多个交换机501-504可以在网络交换环境800(例如，IB子网)内提供架构(诸如InfiniBand架构)的成员之间的通信。该架构可以包括多个硬件设备，诸如主机通道适配器510、520、530。主机通道适配器510、520、530中的每个主机通道适配器进而可以分别与管理程序511、521和531进行交互。每个管理程序进而可以结合与该管理程

序进行交互的主机通道适配器建立多个虚拟功能514、515、516、524、525、526、534、535、536并将这些虚拟功能分配给多个虚拟机。例如，虚拟机1 550可以由管理程序511分配给虚拟功能1 514。管理程序511可以附加地将虚拟机2 551分配给虚拟功能2 515。管理程序521可以将虚拟机3 552分配给虚拟功能3 526。管理程序531进而可以将虚拟机4 553分配给虚拟功能2 535。管理程序可以通过主机通道适配器的每个主机通道适配器上的全特征物理功能513、523、533来访问主机通道适配器。

[0108] 根据实施例，交换机501-504中的每个交换机可以包括多个端口(未示出)，这些端口用于设置线性转发表以便在网络交换环境800内引导业务。

[0109] 根据实施例，虚拟交换机512、522和532可以由它们相应的管理程序511、521、531处理。在这样的vSwitch体系架构中，每个虚拟功能是完整的虚拟主机通道适配器(vHCA)，意味着被分配给VF的VM被分配了完整的一组IB地址(例如，GID、GUID、LID)和硬件中的专用QP空间。对于网络的其余部分和SM(未示出)，HCA 510、520和530经由虚拟交换机看起来像具有连接的附加节点的交换机。

[0110] 根据实施例，本公开提供了用于提供具有动态LID分配和预填充LID的混合vSwitch体系架构的系统和方法。参考图9，管理程序511可以用具有预填充LID体系架构的vSwitch进行布置，而管理程序521可以用具有预填充LID和动态LID分配的vSwitch进行布置。管理程序531可以用具有动态LID分配的vSwitch进行布置。因此，物理功能513和虚拟功能514-516使它们的LID被预填充(即，甚至未附连到活动虚拟机的那些虚拟功能被分配了LID)。物理功能523和虚拟功能1 524可以使它们的LID被预填充，而虚拟功能2525和虚拟功能3 526使它们的LID被动态分配(即，虚拟功能2525可用于动态LID分配，并且虚拟功能3 526由于附连虚拟机3552而具有动态分配的LID 11)。最后，与管理程序3 531相关联的功能(物理功能和虚拟功能)可以使它们的LID被动态分配。这使得虚拟功能1 534和虚拟功能3 536可用于动态LID分配，而虚拟功能2 535由于虚拟机4 553被附连到虚拟功能2而具有动态分配的LID 9。

[0111] 根据诸如图9所绘出的实施例，其中(在任何给定管理程序内独立地或组合地)利用了具有预填充LID的vSwitch和具有动态LID分配的vSwitch两者，每主机通道适配器的预填充LID的数量可以由架构管理员定义，并且可以在(每主机通道适配器)0<＝预填充VF<＝总VF的范围内，并且可以通过从(每主机通道适配器的)VF的总数减去预填充VF的数量来找到可用于动态LID分配的VF。

[0112] 根据实施例，非常类似于物理主机通道适配器可以具有多于一个的端口(两个端口用于冗余是常见的)，虚拟HCA也可以用两个端口表示，并且经由一个、两个或更多个虚拟交换机连接到外部IB子网。

[0113] InfiniBand-子网间通信(架构管理器)

[0114] 根据实施例，除了在单个子网内提供InfiniBand架构之外，本公开的实施例还可以提供跨越两个或更多个子网的InfiniBand架构。

[0115] 图10示出了根据实施例的示例性多子网InfiniBand架构。如图所绘出的，在子网A1000内，多个交换机1001-1004可以在子网A1000(例如，IB子网)内提供架构(诸如InfiniBand架构)的成员之间的通信。该架构可以包括多个硬件设备，诸如，例如通道适配器1010。主机通道适配器1010进而可以与管理程序1011进行交互。管理程序进而可以结合

与该管理程序进行交互的主机通道适配器建立多个虚拟功能1014。管理程序可以附加地将虚拟机分配给每个虚拟功能，诸如虚拟机1 1015被分配给虚拟功能1 1014。管理程序可以通过主机通道适配器中的每个主机通道适配器上的全特征物理功能(诸如物理功能1013)来访问该管理程序相关联的主机通道适配器。在子网B 1040内，多个交换机1021-1024可以在子网B 1040(例如，IB子网)内提供架构(诸如InfiniBand架构)的成员之间的通信。该架构可以包括多个硬件设备，诸如，例如通道适配器1030。主机通道适配器1030进而可以与管理程序1031进行交互。管理程序进而可以结合与该管理程序进行交互的主机通道适配器建立多个虚拟功能1034。管理程序可以附加地将虚拟机分配给虚拟功能中的每个虚拟功能，诸如虚拟机2 1035被分配给虚拟功能2 1034。管理程序可以通过主机通道适配器中的每个主机通道适配器上的全特征物理功能(诸如物理功能1033)来访问该管理程序相关联的主机通道适配器。要注意的是，虽然在每个子网(即，子网A和子网B)内仅示出了一个主机通道适配器，但是应该理解的是，每个子网内可以包括多个主机通道适配器及这些主机通道适配器的对应组件。

[0116] 根据实施例，主机通道适配器中的每个主机通道适配器可以附加地与虚拟交换机(诸如虚拟交换机1012和虚拟交换机1032)相关联，并且可以用不同的体系架构模型来建立每个HCA，如上所述。虽然图10内的两个子网都被显示为使用具有预填充LID的vSwitch体系架构模型，但这并不意味着暗示所有这种子网配置必须遵循相似的体系架构模型。

[0117] 根据实施例，每个子网内的至少一个交换机可以与路由器相关联，诸如子网A 1000内的交换机1002与路由器1005相关联，并且子网B 1040内的交换机1021与路由器1006相关联。

[0118] 根据实施例，至少一个设备(例如，交换机、节点等)可以与架构管理器(未示出)相关联。架构管理器可以用于例如发现子网间架构拓扑、创建架构简档(例如，虚拟机架构简档)、构建虚拟机相关的数据库对象，该数据库对象形成用于构建虚拟机架构简档的基础。此外，架构管理器可以根据哪些子网被允许经由哪些路由器端口使用哪些分区号进行通信来限定合法的子网间连接性。

[0119] 根据实施例，当始发源(诸如子网A内的虚拟机1)处的业务被寻址到不同子网中的目的地(诸如子网B内的虚拟机2)时，该业务可以被寻址到子网A内的路由器，即路由器1005，路由器1005然后可以经由它与路由器1006的链路将业务传递到子网B。

[0120] 虚拟双端口路由器

[0121] 根据实施例，双端口路由器抽象可以提供简单的方式用于使得能够基于如下交换机硬件实现来定义子网到子网的路由器功能，该交换机硬件实现除了执行正常的基于LRH(本地路由报头)的交换之外，还具有进行GRH(全局路由报头)到LRH转换的能力。

[0122] 根据实施例，虚拟双端口路由器可以在对应交换机端口的外部被逻辑上连接。这种虚拟双端口路由器可以向标准管理实体(诸如子网管理器)提供符合InfiniBand规范的视图。

[0123] 根据实施例，双端口路由器模型意味着可以以在不影响任一个不正确连接的子网内的路由和逻辑连接性的情况下每个子网完全控制分组的转发以及到子网的入口路径中的地址映射的方式来连接不同的子网。

[0124] 根据实施例，在涉及不正确连接的架构的情况下，使用虚拟双端口路由器抽象还

可以允许诸如子网管理器和IB诊断软件的管理实体在存在与远程子网的非预期物理连接性时正确地表现。

[0125] 图11示出了根据实施例的在高性能计算环境中的两个子网之间的互连。在用虚拟双端口路由器进行配置之前，子网A 1101中的交换机1120可以通过交换机1120的交换机端口1121经由物理连接1110经由子网B 1102中的交换机1130的交换机端口1131连接到交换机1130。在这样的实施例中，每个交换机端口(1121和1131)既可以充当交换机端口又可以充当路由器端口。

[0126] 根据实施例，这种配置的问题在于诸如InfiniBand子网中的子网管理器的管理实体不能区分既是交换机端口又是路由器端口的物理端口。在这种情况下，SM可以将交换机端口视为具有连接到该交换机端口的路由器端口。但是，如果交换机端口经由例如物理链路连接到具有另一个子网管理器的另一个子网，则子网管理器可以能够在物理链路上发送发现消息。但是，这样的发现消息在另一个子网处可能不被允许。

[0127] 图12示出了根据实施例的在高性能计算环境中经由双端口虚拟路由器配置的两个子网之间的互连。

[0128] 根据实施例，在配置之后，可以提供双端口虚拟路由器配置使得子网管理器看到正确的端节点，从而表明子网管理器负责的子网的端部。

[0129] 根据实施例，在子网A 1201中的交换机1220处，交换机端口可以经由虚拟链路1223连接(即，逻辑上连接)到虚拟路由器1210中的路由器端口1211。虚拟路由器1210(例如，双端口虚拟路由器)(虽然该虚拟路由器1210被显示为在交换机1220的外部，但在实施例中该虚拟路由器1210可以逻辑上被包含在交换机1220内)还可以包括第二路由器端口：路由器端口II 1212。根据实施例，可以具有两个端部的物理链路1203可以经由路由器端口II 1212和被包含在子网B 1202中的虚拟路由器1230中的路由器端口II 1232，经由物理链路的第一端将子网A 1201经由物理链路的第二端与子网B 1202连接。虚拟路由器1230可以附加地包括路由器端口1231，路由器端口1231可以经由虚拟链路1233连接(即，逻辑上连接)到交换机1240上的交换机端口1241。

[0130] 根据实施例，子网A上的子网管理器(未示出)可以检测虚拟路由器1210上的路由器端口1211作为子网管理器控制的子网的端点。双端口虚拟路由器抽象可以允许子网A上的子网管理器以通常的方式(例如，如按照InfiniBand规范所限定的方式)来处理子网A。在子网管理代理级别处，可以提供双端口虚拟路由器抽象使得SM看到正常的交换机端口，然后在SMA级别处，可以提供连接到交换机端口的另一个端口的抽象，并且这个端口是双端口虚拟路由器上的路由器端口。在本地SM中，可以继续使用常规的架构拓扑(在拓扑中SM将端口视为标准交换机端口)，并且因此SM将路由器端口视为端部端口。物理连接可以在也被配置为两个不同子网中的路由器端口的两个交换机端口之间进行。

[0131] 根据实施例，双端口虚拟路由器还可以解决物理链路可能被错误地连接到相同子网中的某个其它交换机端口或者被错误地连接到不旨在提供到另一个子网的连接的交换机端口的问题。因此，本文所描述的方法和系统还提供了关于在子网的外部上是什么的表示。

[0132] 根据实施例，在子网(诸如子网A)内，本地SM确定交换机端口，然后确定连接到该交换机端口的路由器端口(例如，经由虚拟链路1223连接到交换机端口1221的路由器端口

1211)。由于SM将路由器端口1211视为SM管理的子网的端部,因此SM不能超过这个点(例如,向路由器端口II 1212)发送发现消息和/或管理消息。

[0133] 根据实施例,上述双端口虚拟路由器提供的益处是双端口虚拟路由器抽象完全由双端口虚拟路由器所属的子网内的管理实体(例如,SM或SMA)来管理。通过允许仅在本地侧进行管理,系统不需要提供外部的独立的管理实体。即,子网到子网连接的每一侧都可以负责对该侧自己的双端口虚拟路由器进行配置。

[0134] 根据实施例,在被寻址到远程目的地(即,在本地子网外部)的诸如SMP的分组到达未经由上述双端口虚拟路由器进行配置的本地目标端口的情况下,则本地端口可以返回指定该本地端口不是路由器端口的消息。

[0135] 本发明的许多特征可以在硬件、软件、固件或它们的组合中、利用硬件、软件、固件或它们的组合、或者在硬件、软件、固件或它们的组合的辅助下执行。因此,可以使用(例如包括一个或多个处理器的)处理系统来实现本发明的特征。

[0136] 图13示出了根据实施例的用于支持高性能计算环境中的双端口虚拟路由器的方法。在步骤1310处,该方法可以提供第一子网,第一子网包括:多个交换机,该多个交换机至少包括叶交换机,其中该多个交换机中的每个交换机包括多个交换机端口;多个主机通道适配器,每个主机通道适配器包括至少一个主机通道适配器端口;多个端节点,其中端节点中的每个端节点与多个主机通道适配器中的至少一个主机通道适配器相关联;以及子网管理器,该子网管理器在多个交换机中的交换机和多个主机通道适配器中的一个上运行。

[0137] 在步骤1320处,该方法可以将多个交换机中的交换机上的多个交换机端口中的交换机端口配置为路由器端口。

[0138] 在步骤1330处,该方法可以将被配置为路由器端口的交换机端口逻辑上连接到虚拟路由器,该虚拟路由器包括至少两个虚拟路由器端口。

[0139] 本发明的特征可以在计算机程序产品中、利用计算机程序产品、或者在计算机程序产品的辅助下实现,其中计算机程序产品是具有其上/其中存储有可用于对处理系统编程以执行本文所呈现的任何特征的指令的存储介质或计算机可读介质。存储介质可以包括但不限于,任何类型的盘(包括软盘、光盘、DVD、CD-ROM、微驱动器、以及磁光盘)、ROM、RAM、EPROM、EEPROM、DRAM、VRAM、闪存存储器设备、磁卡或光卡、纳米系统(包括分子存储器IC)、或适于存储指令和/或数据的任何类型的介质或设备。

[0140] 存储在任何一种机器可读介质上的本发明的特征可以被并入到软件和/或固件中,用于控制处理系统的硬件,并且用于使处理系统能够利用本发明的结果与其它机制进行交互。这种软件或固件可以包括但不限于,应用代码、设备驱动程序、操作系统和执行环境/容器。

[0141] 本发明的特征还可以使用例如硬件组件(诸如专用集成电路(ASIC))在硬件中实现。硬件状态机的实现以使执行本文所描述的功能对相关领域的技术人员来说将是显而易见的。

[0142] 此外,本发明可以方便地使用一个或多个常规的通用数字计算机或专用数字计算机、计算设备、机器或微处理器来实现,其中包括一个或多个处理器、存储器和/或根据本公开内容的教导进行编程的计算机可读存储介质。适当的软件编码可以容易地由熟练的程序员基于本公开的教导来准备,如对软件领域的技术人员将显而易见的。

[0143] 虽然以上已经描述了本发明的各种实施例，但是应该理解的是，这些实施例以示例的方式而非限制的方式被呈现。对相关领域的技术人员来说将显而易见的是，在不背离本发明的精神和范围的情况下，可以在本发明中做出各种形式和细节上的改变。

[0144] 已经借助说明具体功能及具体功能的关系的执行的功能构建块描述了本发明。这些功能构建块的边界在本文中通常是为了方便描述而任意定义的。可以定义替代边界，只要适当地执行具体功能及具体功能的关系即可。任何这种替代边界因此在本发明的范围和精神之内。

[0145] 已经为了说明和描述的目的提供了本发明的前面描述。本发明的前面描述不旨在是穷尽的或者将本发明限制到所公开的精确形式。本发明的广度和范围不应该由任何上述示例性实施例来限制。许多修改和变化对本领域技术人员来说将显而易见。这些修改和变化包括所公开特征的任何相关组合。实施例的选择与描述是为了最好地解释本发明的原理及其实践应用，从而使本领域其它技术人员能够理解本发明的各种实施例以及适于所设想的特定用途的各种修改。旨在由以下权利要求及这些权利要求的等价物来限定本发明的范围。

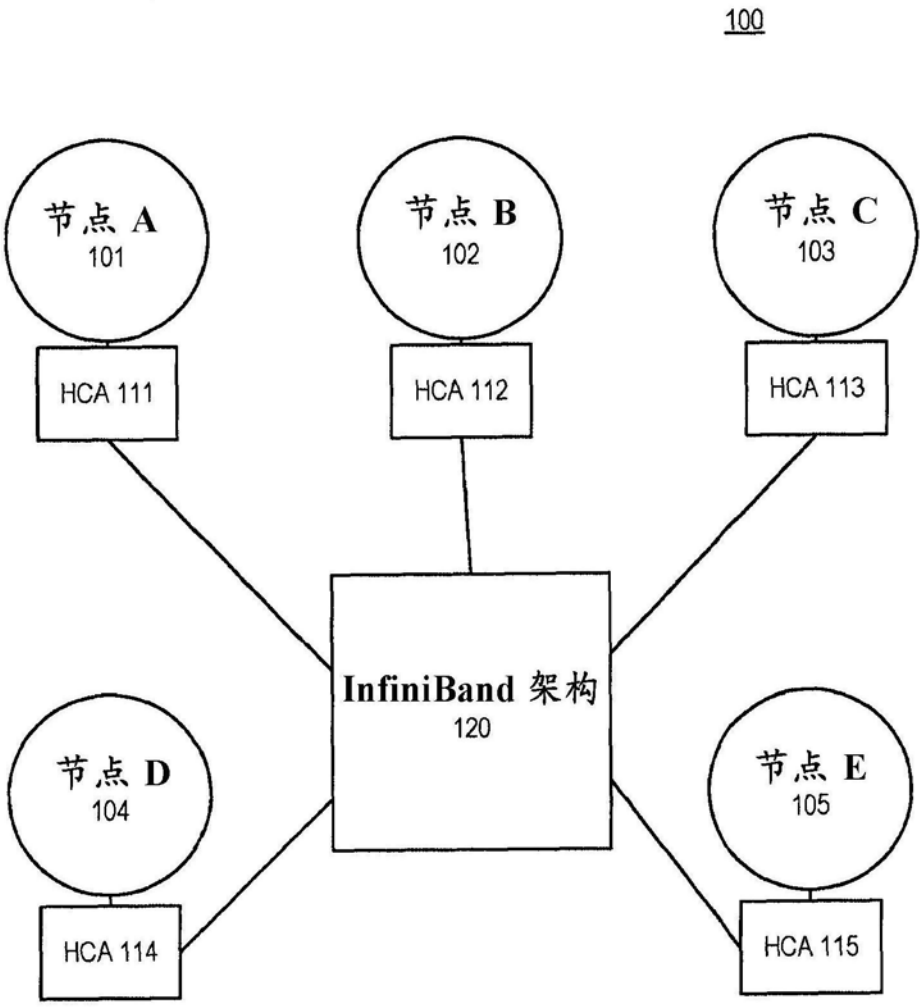

100
节点 A
101
节点 B
102
节点 C
103
HCA 111
HCA 112
HCA 113
InfiniBand 架构
120
节点 D
104
节点 E
105
HCA 114
HCA 115

图1

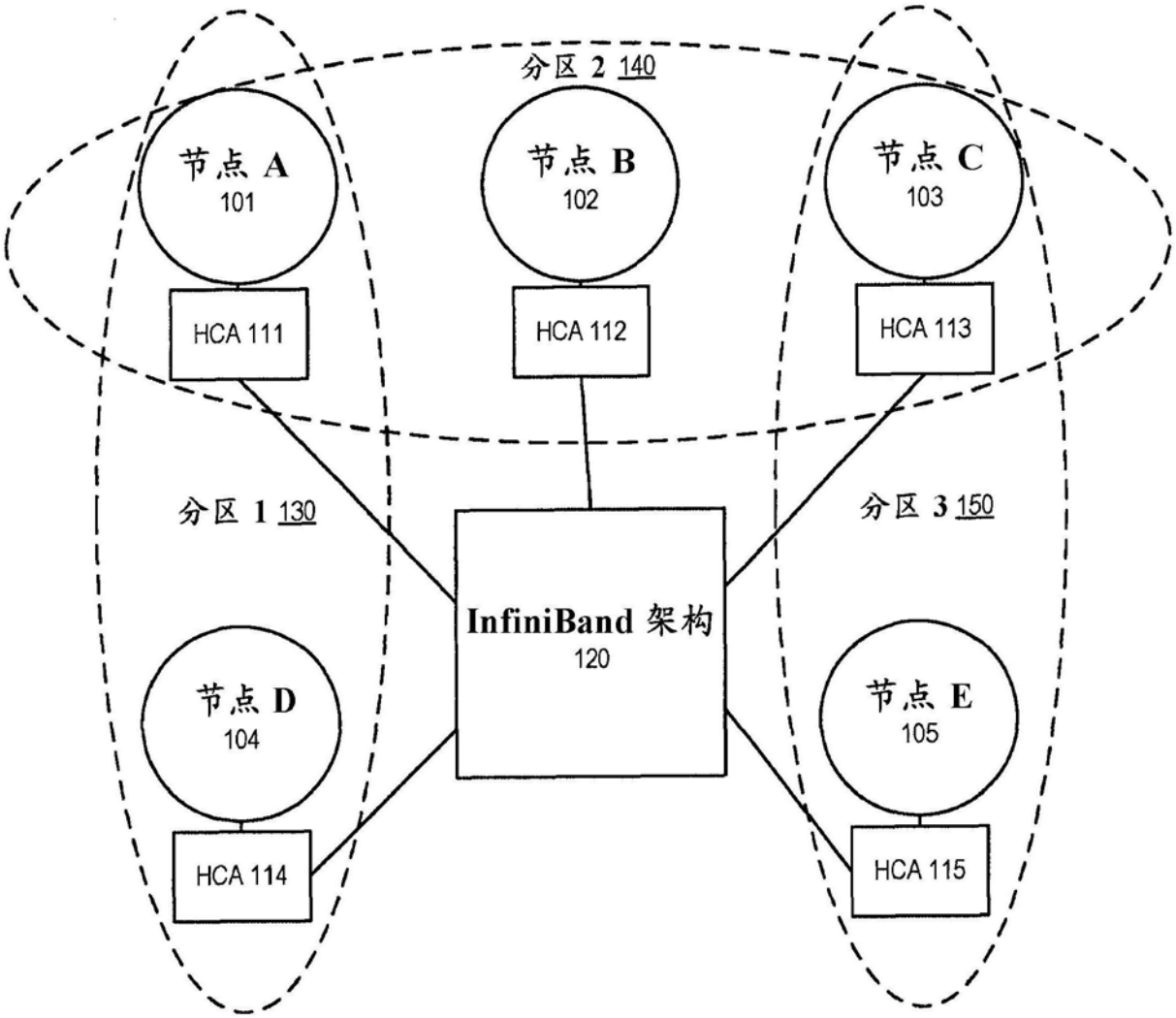

分区 2 140
节点 A
101
节点 B
102
节点 C
103
HCA 111
HCA 112
HCA 113
分区 1 130
分区 3 150
InfiniBand 架构
120
节点 D
104
节点 E
105
HCA 114
HCA 115

图2

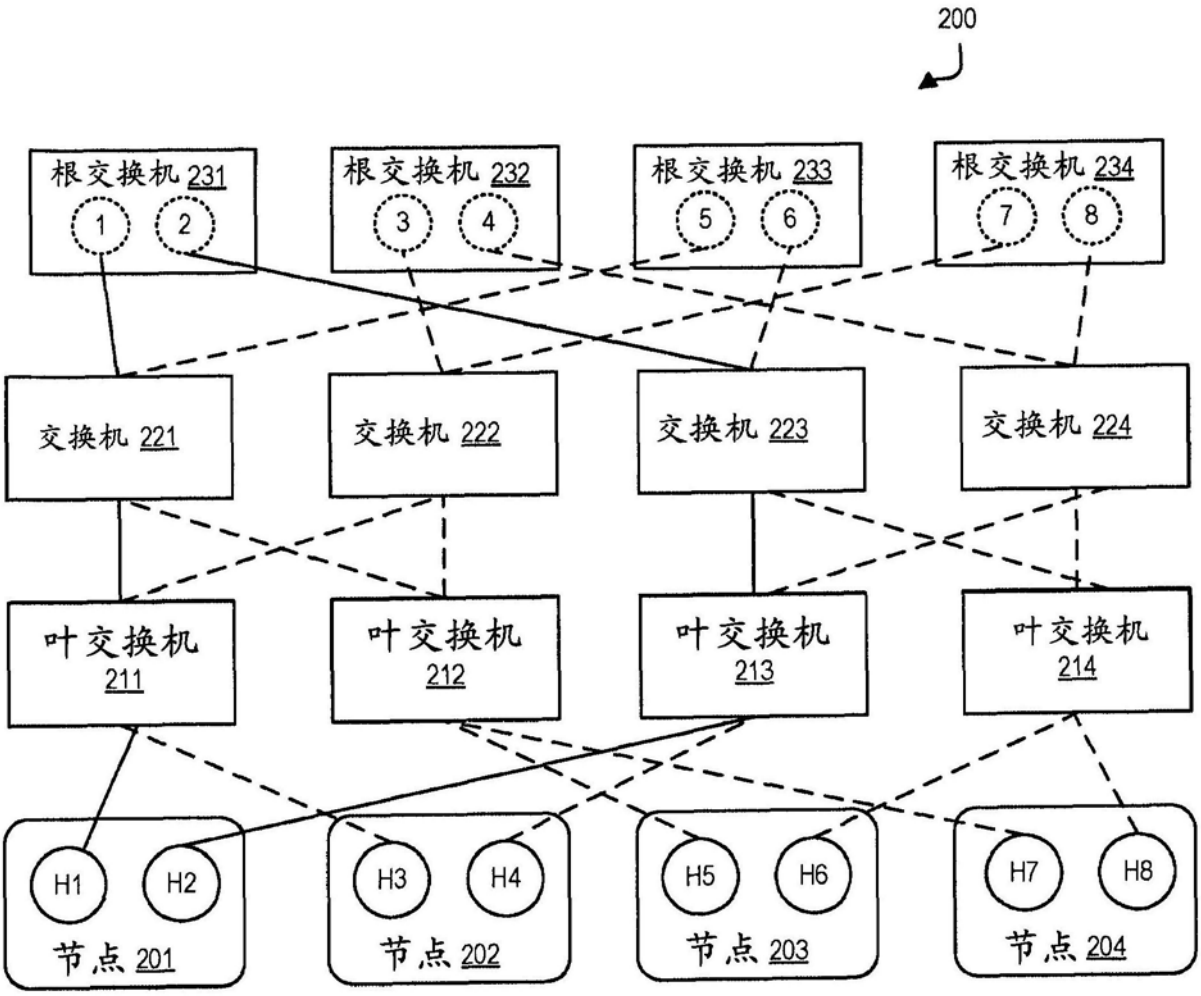

200
根交换机 231
1
2
根交换机 232
3
4
根交换机 233
5
6
根交换机 234
7
8
交换机 221
交换机 222
交换机 223
交换机 224
叶交换机 211
叶交换机 212
叶交换机 213
叶交换机 214
H1
H2
节点 201
H3
H4
节点 202
H5
H6
节点 203
H7
H8
节点 204

图3

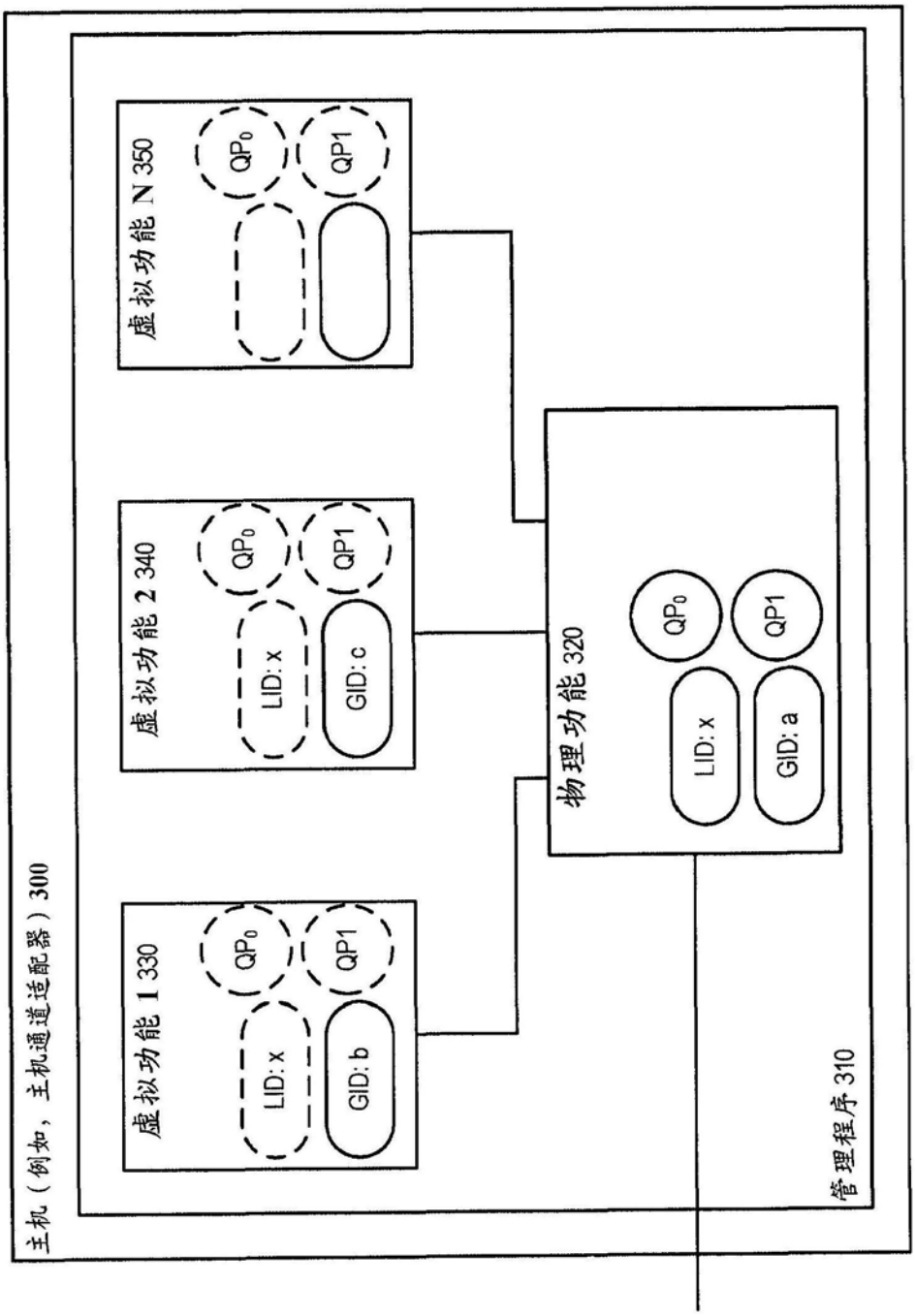

主机（例如，主机通道适配器）300
虚拟功能 1 330
QP0
QP1
LID: x
GID: b
虚拟功能 2 340
QP0
QP1
LID: x
GID: c
虚拟功能 N 350
QP0
QP1
物理功能 320
QP0
QP1
LID: x
GID: a
管理程序 310

图4

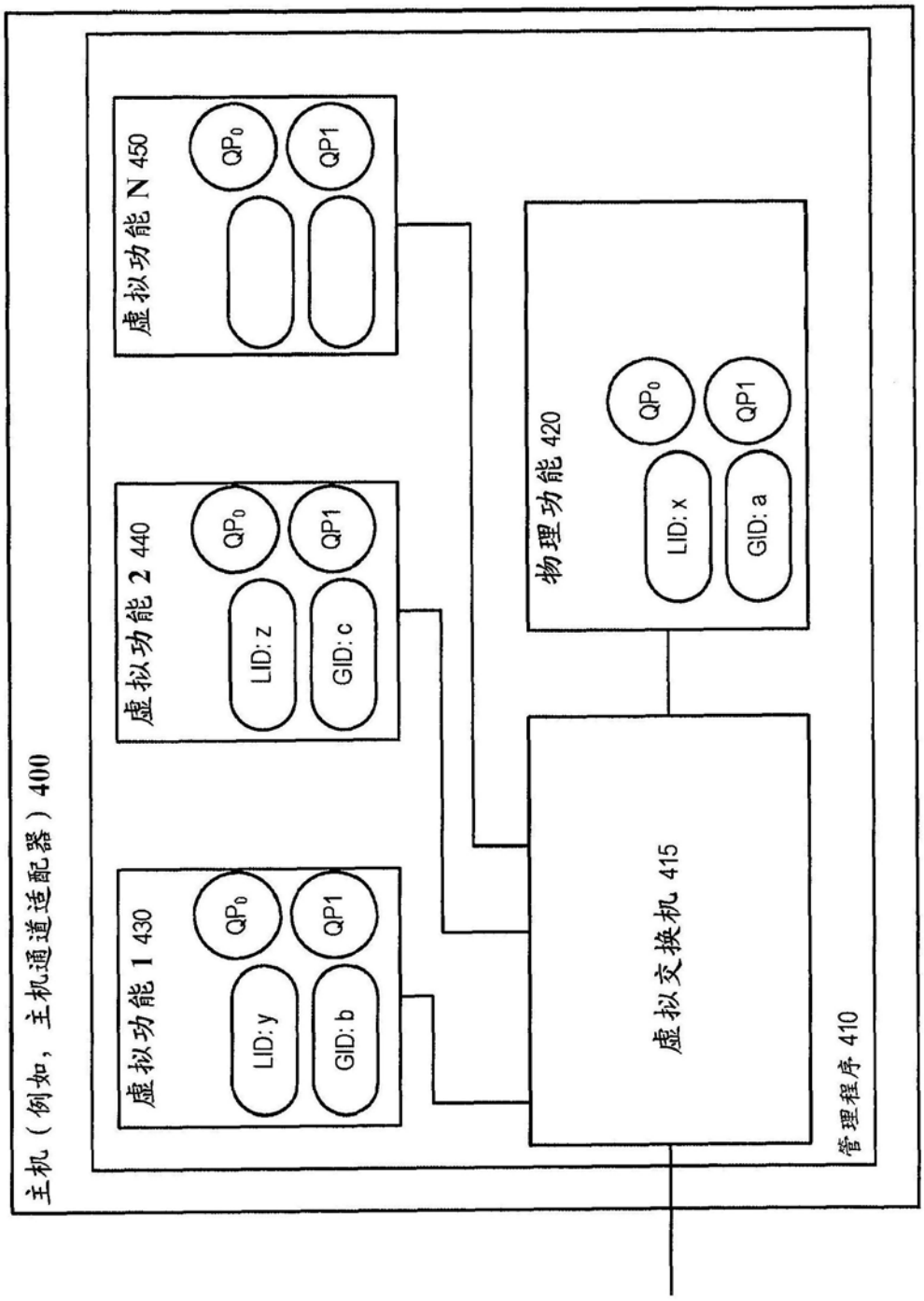

主机（例如，主机通道适配器）400
虚拟功能 1 430
QP0
LID: y
QP1
GID: b
虚拟功能 2 440
QP0
LID: z
QP1
GID: c
虚拟功能 N 450
QP0
QP1
虚拟交换机 415
物理功能 420
QP0
LID: x
QP1
GID: a
管理程序 410

图5

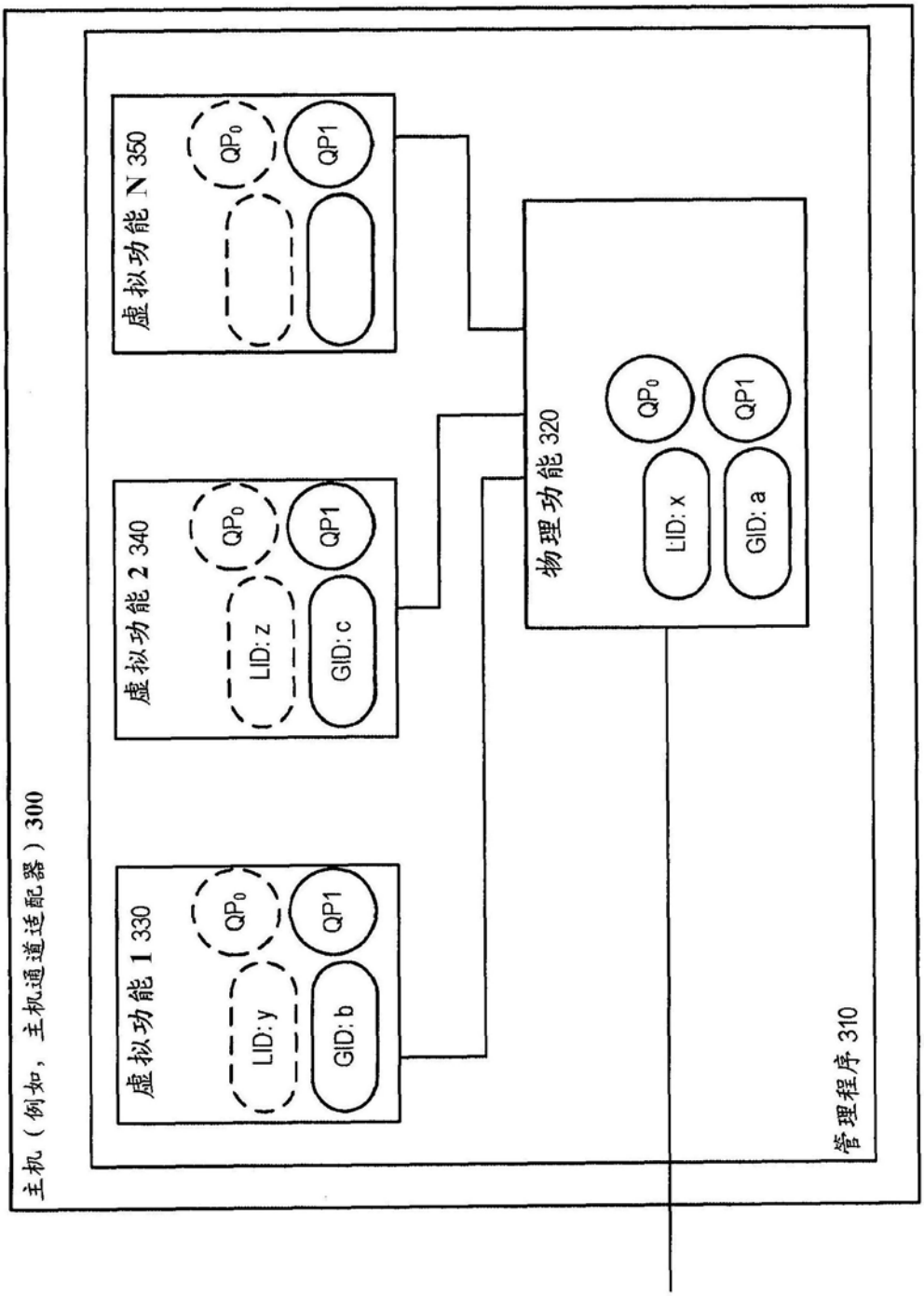

主机（例如，主机通道适配器）300
虚拟功能 1 330
LID: y
GID: b
QP0
QP1
虚拟功能 2 340
LID: z
GID: c
QP0
QP1
虚拟功能 N 350
QP0
QP1
物理功能 320
LID: x
GID: a
QP0
QP1
管理程序 310

图6

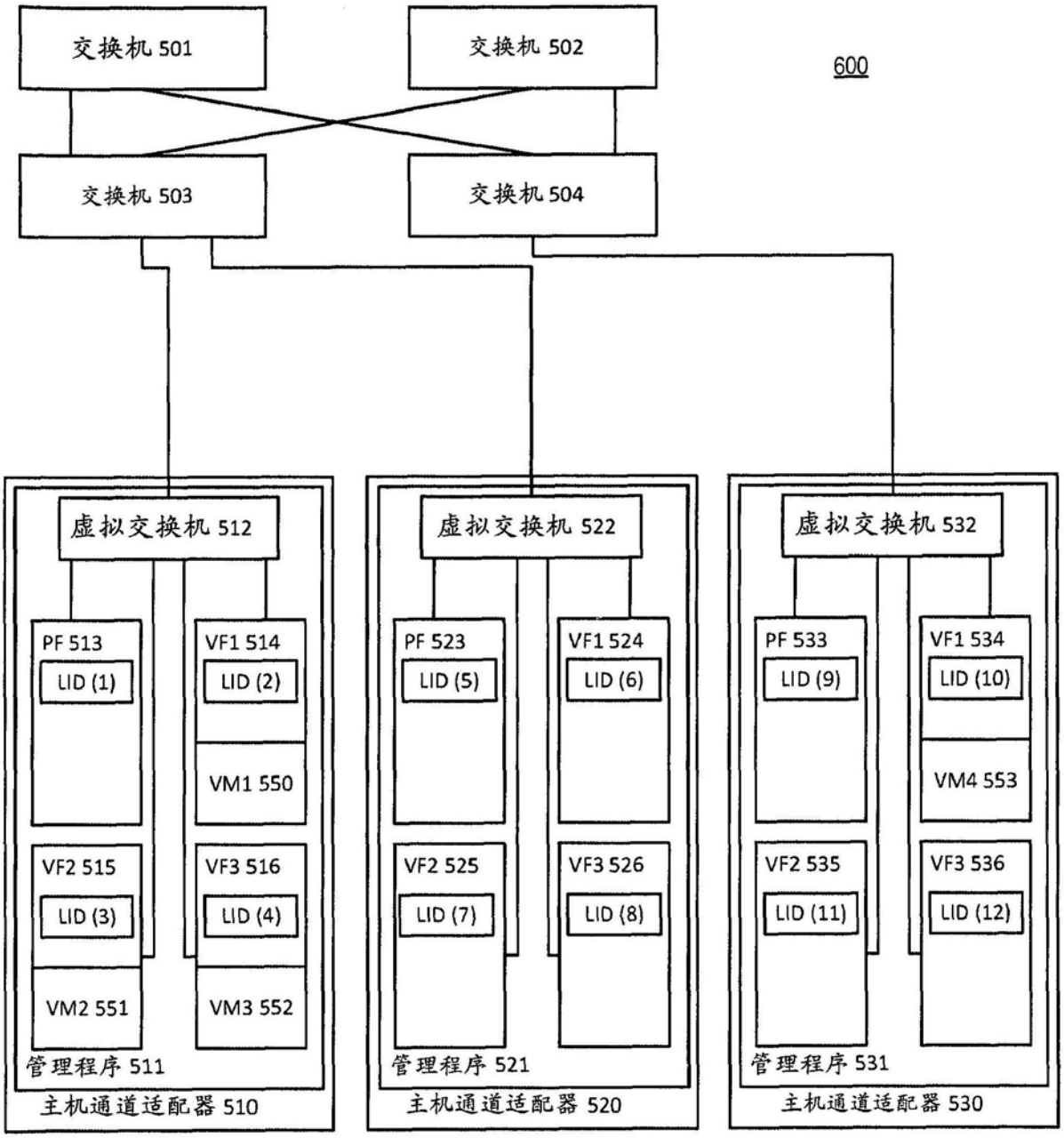

交换机 501
交换机 502
600
交换机 503
交换机 504
虚拟交换机 512
PF 513
LID (1)
VF1 514
LID (2)
VM1 550
VF2 515
LID (3)
VF3 516
LID (4)
VM2 551
VM3 552
管理程序 511
主机通道适配器 510
虚拟交换机 522
PF 523
LID (5)
VF1 524
LID (6)
VF2 525
LID (7)
VF3 526
LID (8)
管理程序 521
主机通道适配器 520
虚拟交换机 532
PF 533
LID (9)
VF1 534
LID (10)
VM4 553
VF2 535
LID (11)
VF3 536
LID (12)
管理程序 531
主机通道适配器 530

图7

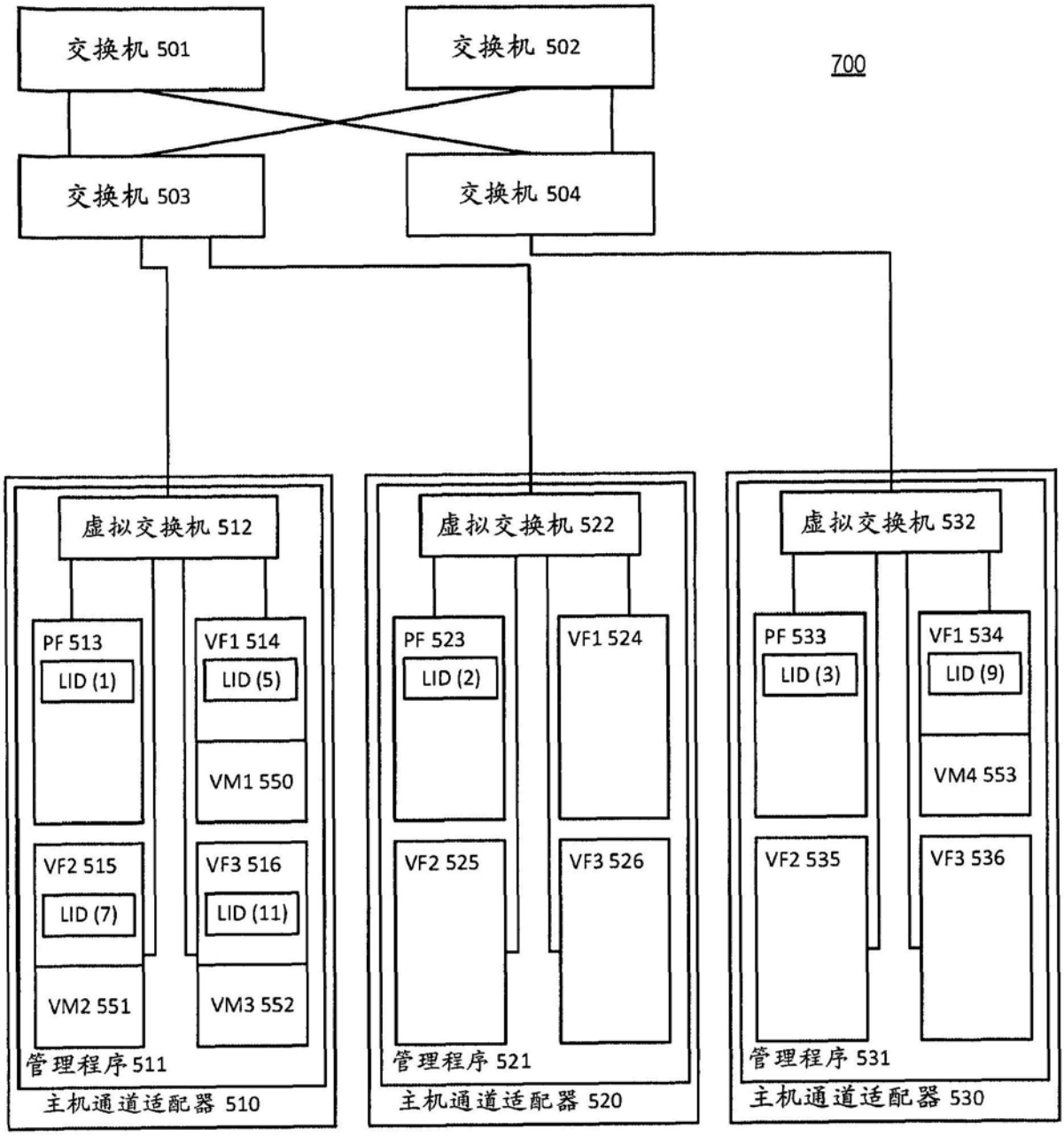

交换机 501
交换机 502
700
交换机 503
交换机 504
虚拟交换机 512
PF 513
LID (1)
VF1 514
LID (5)
VM1 550
VF2 515
LID (7)
VF3 516
LID (11)
VM2 551
VM3 552
管理程序 511
主机通道适配器 510
虚拟交换机 522
PF 523
LID (2)
VF1 524
VF2 525
VF3 526
管理程序 521
主机通道适配器 520
虚拟交换机 532
PF 533
LID (3)
VF1 534
LID (9)
VM4 553
VF2 535
VF3 536
管理程序 531
主机通道适配器 530

图8

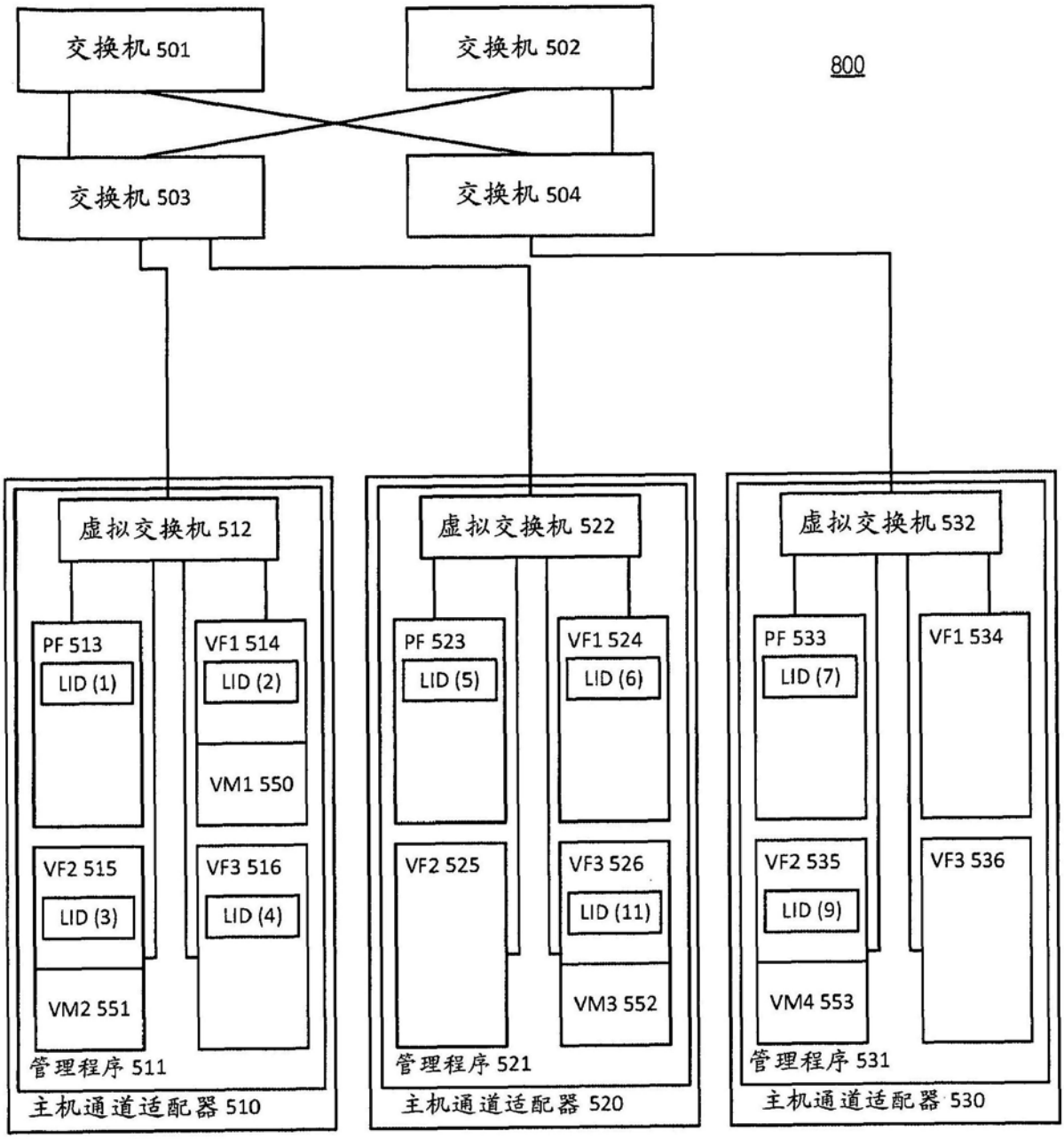

交换机 501
交换机 502
800
交换机 503
交换机 504
虚拟交换机 512
PF 513
LID (1)
VF1 514
LID (2)
VM1 550
VF2 515
LID (3)
VF3 516
LID (4)
VM2 551
管理程序 511
主机通道适配器 510
虚拟交换机 522
PF 523
LID (5)
VF1 524
LID (6)
VF2 525
VF3 526
LID (11)
VM3 552
管理程序 521
主机通道适配器 520
虚拟交换机 532
PF 533
LID (7)
VF1 534
VF2 535
LID (9)
VF3 536
VM4 553
管理程序 531
主机通道适配器 530

图9

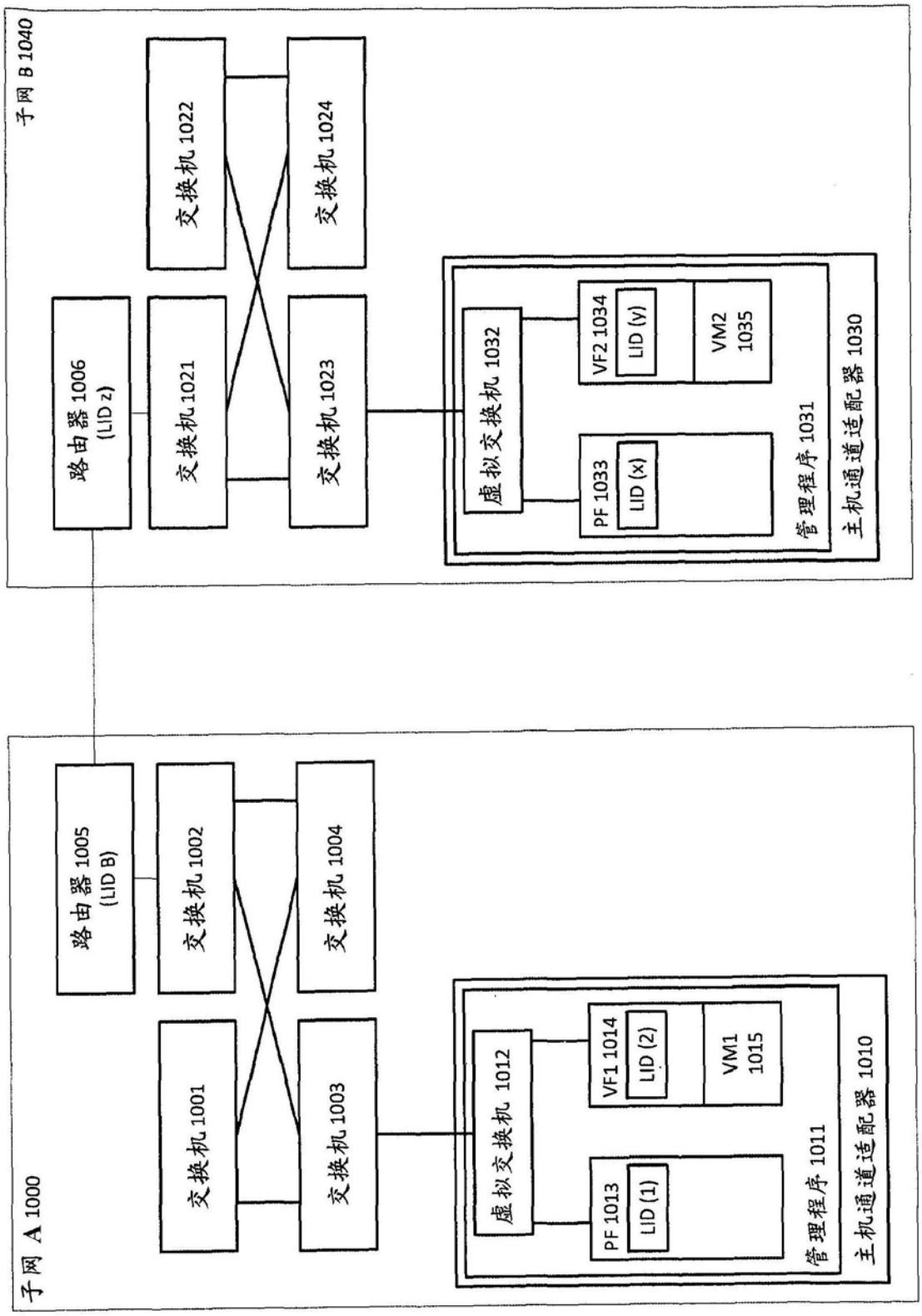

子网 A 1000
路由器 1005
(LID B)
交换机 1001
交换机 1002
交换机 1003
交换机 1004
虚拟交换机 1012
PF 1013
LID (1)
VF1 1014
LID (2)
VM1
1015
管理程序 1011
主机通道适配器 1010
子网 B 1040
路由器 1006
(LID z)
交换机 1021
交换机 1022
交换机 1023
交换机 1024
虚拟交换机 1032
PF 1033
LID (x)
VF2 1034
LID (y)
VM2
1035
管理程序 1031
主机通道适配器 1030

图10

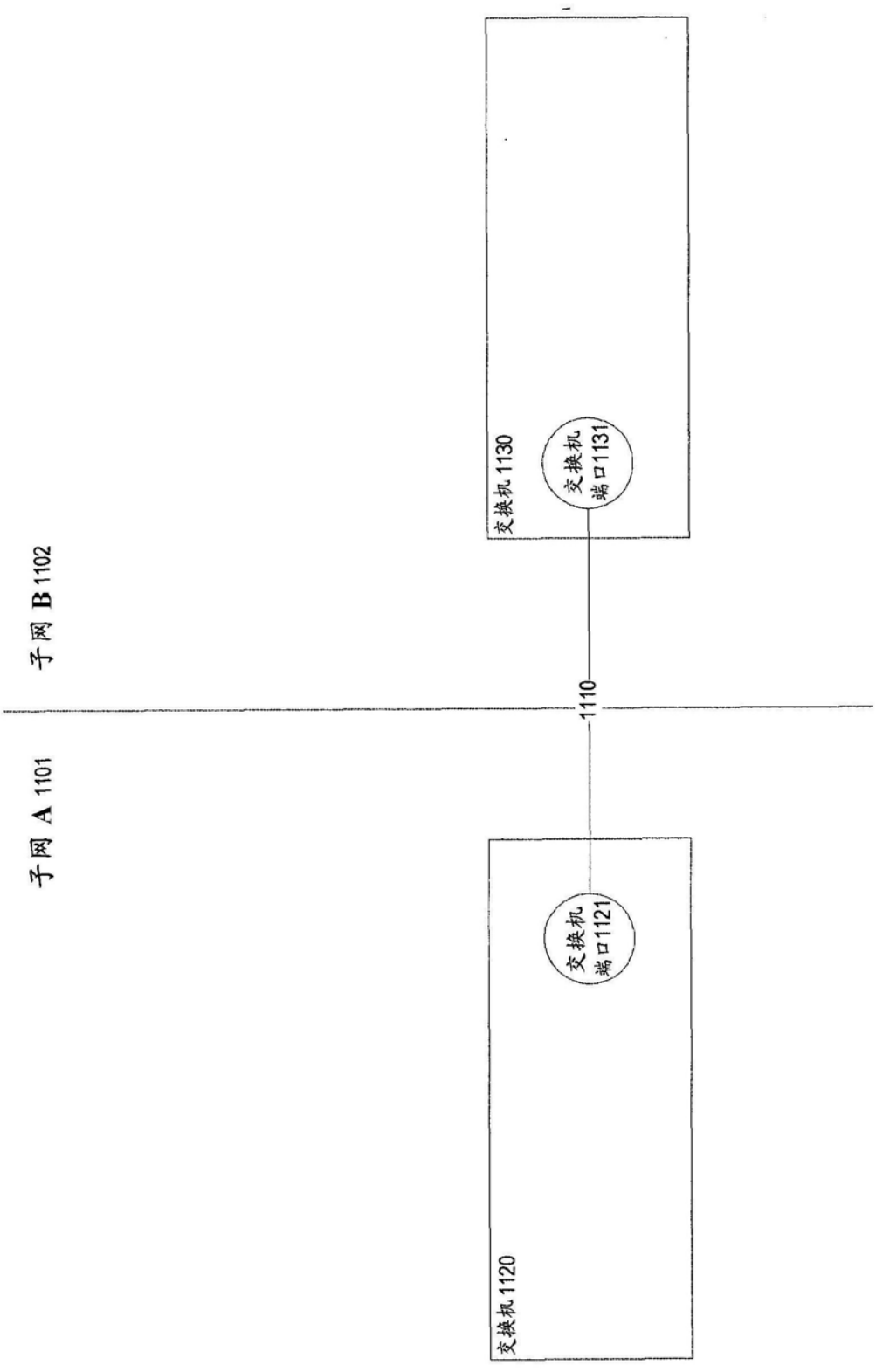

子网 B 1102
子网 A 1101
1110
交换机 1130
交换机端口1131
交换机端口1121
交换机 1120

图11

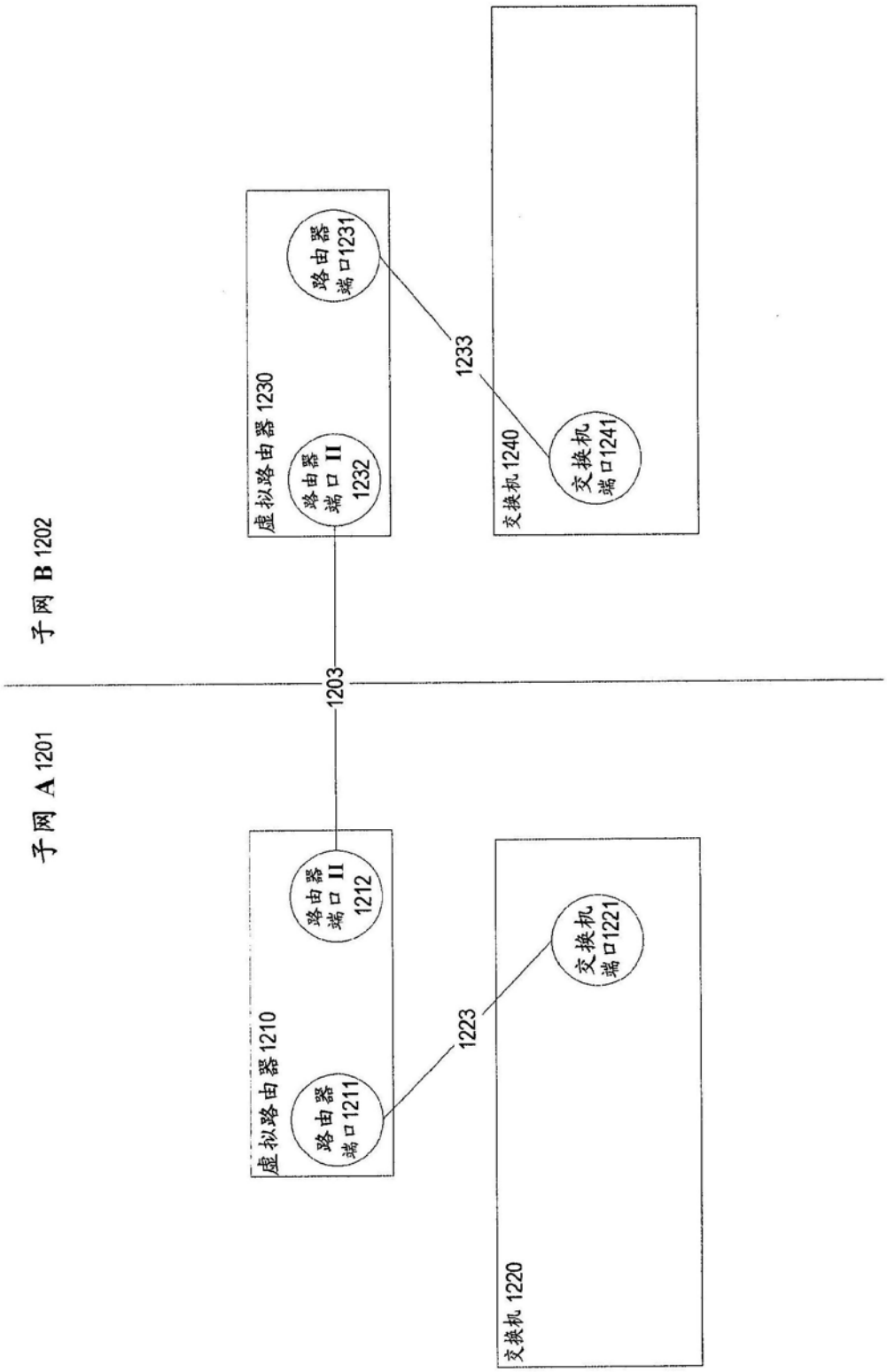

子网 B 1202
子网 A 1201
1203
虚拟路由器 1230
路由器端口 II 1232
路由器端口1231
1233
交换机 1240
交换机端口1241
虚拟路由器1210
路由器端口 II 1212
路由器端口1211
1223
交换机端口1221
交换机 1220

图12

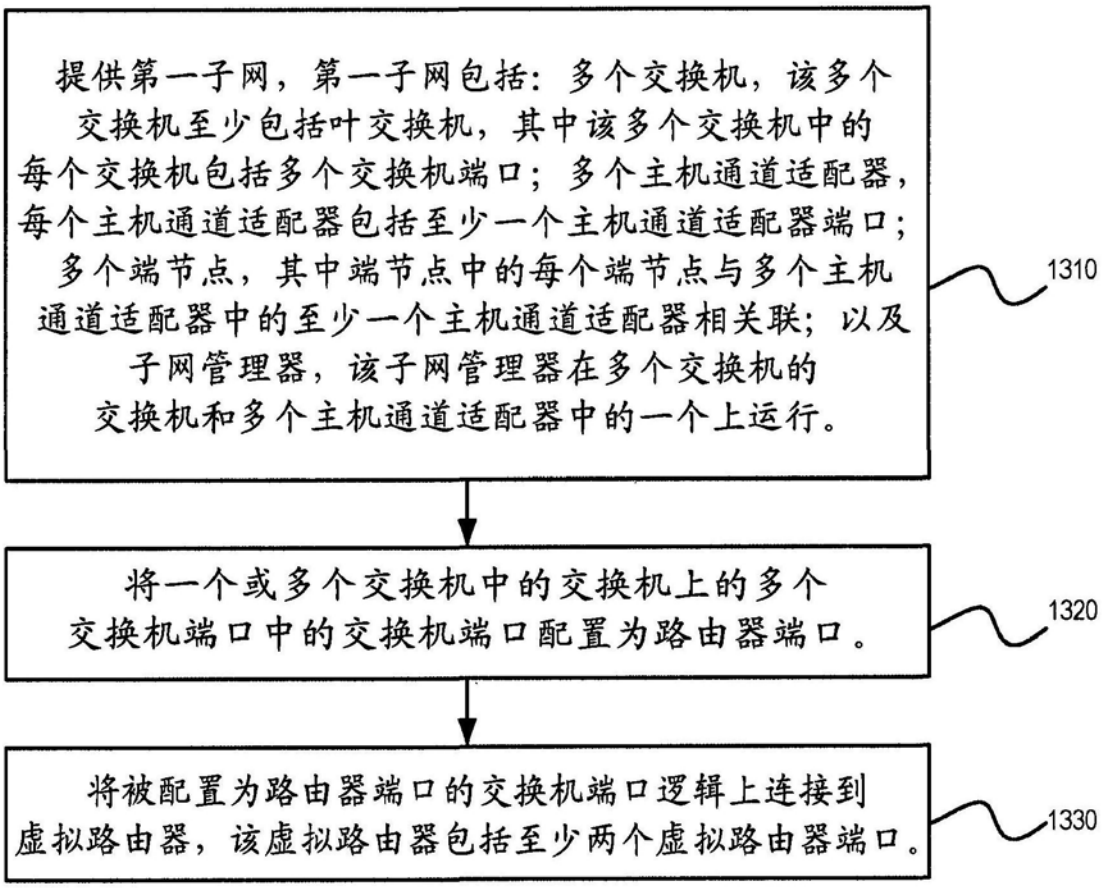

提供第一子网，第一子网包括：多个交换机，该多个交换机至少包括叶交换机，其中该多个交换机中的每个交换机包括多个交换机端口；多个主机通道适配器，每个主机通道适配器包括至少一个主机通道适配器端口；多个端节点，其中端节点中的每个端节点与多个主机通道适配器中的至少一个主机通道适配器相关联；以及子网管理器，该子网管理器在多个交换机的交换机和多个主机通道适配器中的一个上运行。
1310
将一个或多个交换机中的交换机上的多个交换机端口中的交换机端口配置为路由器端口。
1320
将被配置为路由器端口的交换机端口逻辑上连接到虚拟路由器，该虚拟路由器包括至少两个虚拟路由器端口。
1330

图13