(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 107852377 B
(45)授权公告日 2021.06.25

(21)申请号 201780002356.0
(22)申请日 2017.01.25
(65)同一申请的已公布的文献号
申请公布号 CN 107852377 A
(43)申请公布日 2018.03.27
(30)优先权数据
62/287,704 2016.01.27 US
15/412,995 2017.01.23 US
15/413,075 2017.01.23 US
(85)PCT国际申请进入国家阶段日
2018.01.09
(86)PCT国际申请的申请数据
PCT/US2017/014963 2017.01.25
(87)PCT国际申请的公布数据
WO2017/132271 EN 2017.08.03
(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 D·G·莫克斯纳斯 L·霍雷恩 B·D·约翰森
(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(2013.01)
G06F 9/455(2006.01)
(56)对比文件
US 2004030763 A1,2004.02.12
JP H1056464 A,1998.02.24
US 7200704 B2,2007.04.03
NAKAMURA Minoru.InfiniBand程序所需的基本概念.《www.nminoru.jp》.2014,全文.
审查员 王勇

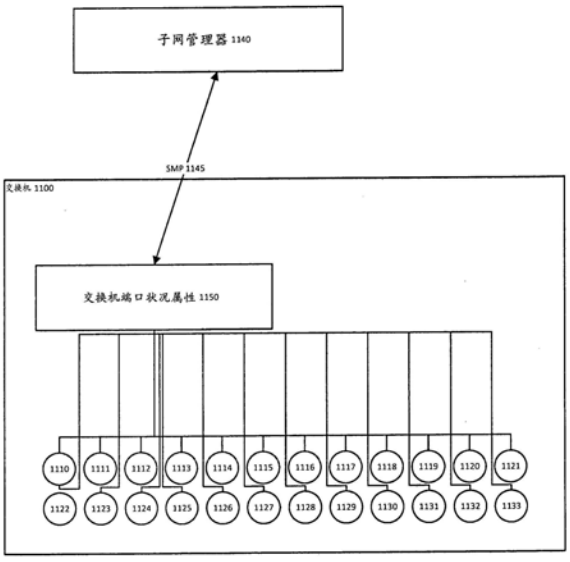
权利要求书3页 说明书20页 附图17页

(54)发明名称

用于在高性能计算环境中支持交换机端口状况的可伸缩表示的系统和方法

(57)摘要

用于在高性能计算环境中支持链路稳定性和可用性的可伸缩表示的系统和方法。该方法可以在子网中的每个节点处提供属性,其中该属性在每个节点处提供单个位置以供子网管理器查询连接到被查询节点的每个链路的稳定性和可用性。该属性可以由驻留在节点处的子网管理代理填充和维护。



CN 107852377 B

1.一种用于在高性能计算环境中支持交换机端口状况的可伸缩表示的系统,包括:

一个或多个微处理器;

至少一个子网,所述至少一个子网包括:

一个或多个交换机,所述一个或多个交换机至少包括叶子交换机,其中所述一个或多个交换机中的每个交换机包括多个端口,多个主机通道适配器,其中所述多个主机通道适配器经由所述一个或多个交换机互连,

多个端节点,所述多个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,所述子网管理器运行在所述一个或多个交换机中的一个交换机上或所述多个主机通道适配器中的一个主机通道适配器上;

其中所述一个或多个交换机中的每个交换机包括固定大小的消息;

其中所述一个或多个交换机上的所述多个端口中的每个端口与交换机端口状况相关联;

其中与每个交换机上的所述多个端口中的每个端口相关联的每个交换机端口状况被表示在相关联的交换机处的所述固定大小的消息中;

并且,其中所述子网管理器使用一个操作来确定所述一个或多个交换机中的一个交换机上的所有端口的交换机端口状况。

2.如权利要求1所述的系统,还包括:

其中所述一个或多个交换机上的所述多个交换机端口中的每个交换机端口与扩展链路状况相关联;

并且,其中与每个交换机上的所述多个交换机端口中的每个交换机端口相关联的每个扩展链路状况被表示在相关联的交换机处的所述固定大小的消息中。

3.如权利要求1所述的系统,其中所述一个操作是子网管理分组。

4.如权利要求2所述的系统,其中所述子网管理器使用一个操作来确定所述一个或多个交换机中的一个交换机上的每个端口的扩展链路状况。

5.如权利要求4所述的系统,其中所述一个操作是子网管理分组。

6.如权利要求1-5中任一项所述的系统,其中所述多个主机通道适配器中的主机通道适配器包括虚拟交换机,所述虚拟交换机包括多个虚拟交换机端口;

其中每个虚拟交换机端口与虚拟交换机端口状况相关联;以及

其中与每个虚拟交换机端口相关联的每个虚拟交换机端口状况被表示在所述虚拟交换机处的固定大小的消息中。

7.一种用于在高性能计算环境中支持交换机端口状况的可伸缩表示的方法,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:

至少一个子网,所述至少一个子网包括:

一个或多个交换机,所述一个或多个交换机至少包括叶子交换机,其中所述一个或多个交换机中的每个交换机包括多个端口,并且其中所述一个或多个交换机中的每个交换机包括固定大小的消息,

多个主机通道适配器,其中所述多个主机通道适配器经由所述一个或多个交换机互连,

多个端节点,所述多个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,所述子网管理器运行在所述一个或多个交换机中的一个交换机上或所述多个主机通道适配器中的一个主机通道适配器上;

将所述一个或多个交换机上的所述多个端口中的每个端口与交换机端口状况相关联;

将与每个交换机上的所述多个端口中的每个端口相关联的每个交换机端口状况表示在相关联的交换机处的所述固定大小的消息中;以及

使用一个操作来确定所述一个或多个交换机中的一个交换机上的所有端口的交换机端口状况。

8.如权利要求7所述的方法,还包括:

将所述一个或多个交换机上的所述多个端口中的每个端口与扩展链路状况相关联;

将与每个交换机上的所述多个端口中的每个端口相关联的每个扩展链路状况表示在所述相关联的交换机处的所述固定大小的消息中。

9.如权利要求7所述的方法,其中所述一个操作是子网管理分组。

10.如权利要求8所述的方法,还包括:

由所述子网管理器使用一个操作来确定所述一个或多个交换机中的一个交换机上的每个端口的扩展链路状况。

11.如权利要求10所述的方法,其中所述一个操作是子网管理分组。

12.如权利要求7-11中任一项所述的方法,其中所述多个主机通道适配器中的主机通道适配器包括虚拟交换机,所述虚拟交换机包括多个虚拟交换机端口;

其中每个虚拟交换机端口与虚拟交换机端口状况相关联;并且

其中与每个虚拟交换机端口相关联的每个虚拟交换机端口状况被表示在所述虚拟交换机处的固定大小的消息中。

13.一种非暂态计算机可读存储介质,包括存储在其上的指令,所述指令用于在高性能计算环境中支持交换机端口状况的可伸缩表示,所述指令当被一个或多个计算机读取并执行时,使所述一个或多个计算机执行包括以下各项的步骤:

在包括一个或多个微处理器的一个或多个计算机处提供:

至少一个子网,所述至少一个子网包括:

一个或多个交换机,所述一个或多个交换机至少包括叶子交换机,其中所述一个或多个交换机中的每个交换机包括多个端口,并且其中所述一个或多个交换机中的每个交换机包括固定大小的消息,

多个主机通道适配器,其中所述多个主机通道适配器经由所述一个或多个交换机互连,

多个端节点,所述多个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,所述子网管理器运行在所述一个或多个交换机中的一个交换机上或所述多个主机通道适配器中的一个主机通道适配器上;

将所述一个或多个交换机上的所述多个端口中的每个端口与交换机端口状况相关联;

将与每个交换机上的所述多个端口中的每个端口相关联的每个交换机端口状况表示

在相关联的交换机处的所述固定大小的消息中;以及

使用一个操作来确定所述一个或多个交换机中的一个交换机上的所有端口的交换机端口状况。

14.如权利要求13所述的非暂态计算机可读存储介质,所述步骤还包括:

将所述一个或多个交换机上的所述多个端口中的每个端口与扩展链路状况相关联;

将与每个交换机上的所述多个端口中的每个端口相关联的每个扩展链路状况表示在所述相关联的交换机处的所述固定大小的消息中。

15.如权利要求13所述的非暂态计算机可读存储介质,其中所述一个操作是子网管理分组。

16.如权利要求14所述的非暂态计算机可读存储介质,所述步骤还包括:

由所述子网管理器使用一个操作来确定所述一个或多个交换机中的一个交换机上的每个端口的扩展链路状况。

17.如权利要求16所述的非暂态计算机可读存储介质,其中所述一个操作是子网管理分组。

18.一种包括用于执行如权利要求7-12中任一项所述的方法的部件的装置。

用于在高性能计算环境中支持交换机端口状况的可伸缩表示的系统和方法

[0001] 版权声明

[0002] 本专利文档公开的一部分包含受版权保护的素材。版权拥有者不反对任何人对专利文档或专利公开按照在专利商标局的专利文件或记录中出现得那样进行的传真复制，但是除此之外在任何情况下都保留所有版权。

技术领域

[0003] 本发明一般而言涉及计算机系统，并且具体而言涉及在高性能计算环境中支持交换机端口状况的可伸缩表示(scalable representation)。

背景技术

[0004] 随着更大的云计算体系架构被引入，与传统网络和存储相关联的性能和管理瓶颈成为重要的问题。人们越来越感兴趣使用诸如InfiniBand(IB)(无限带宽)技术的高性能无损互连作为用于云计算架构的基础。这是本发明的实施例旨在解决的一般领域。

发明内容

[0005] 本文描述了用于在高性能计算环境中支持交换机端口状况的可伸缩表示的系统和方法。一种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至少一个子网，该至少一个子网包括：一个或多个交换机、多个主机通道适配器、多个端节点以及子网管理器。该一个或多个交换机至少包括叶子交换机，其中该一个或多个交换机中的每个交换机包括多个端口，并且其中该一个或多个交换机中的每个交换机包括至少一个属性；其中该多个主机通道适配器经由该一个或多个交换机互连；该多个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；该子网管理器运行在一个或多个交换机中的一个交换机或多个主机通道适配器中的一个主机通道适配器上。该方法可以将一个或多个交换机上的多个端口中的每个端口与交换机端口状况相关联。该方法可以将与每个交换机上的多个端口中的每个端口相关联的每个交换机端口状况表示在相关联的交换机处的至少一个属性中。

[0006] 本文描述了用于在高性能计算环境中支持链路稳定性和可用性属性的系统和方法。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至少一个子网，至少一个或多个交换机，该一个或多个交换机至少包括叶子交换机，其中该一个或多个交换机中的每个交换机包括多个端口，并且其中该一个或多个交换机中的每个交换机包括至少一个属性；多个主机通道适配器，其中该多个主机通道适配器经由一个或多个交换机互连；多个端节点，该多个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，该子网管理器在一个或多个交换机中的一个交换机或多个主机通道适配器中的一个主机通道适配器上运行。该方法可以在一个或多个交换机中的每个交换机处提供至少一个属性。该方法可以在一个或多个交换机中的交换机处提

供多个子网管理代理(SMA)中的子网管理代理。该方法可以通过一个或多个交换机中的交换机的SMA来监视该交换机的多个端口中的每个端口处的链路稳定性和该交换机处的多个端口中的每个端口处的链路可用性中的至少一个。

[0007] 根据实施例,多个主机通道适配器中的一个或多个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。多个端节点可以包括物理主机、虚拟机、或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

[0008] 图1示出了根据实施例的InfiniBand环境的图示。

[0009] 图2示出了根据实施例的分区集群环境的图示。

[0010] 图3示出了根据实施例的网络环境中的树形拓扑的图示。

[0011] 图4示出了根据实施例的示例性共享端口体系架构。

[0012] 图5示出了根据实施例的示例性vSwitch体系架构。

[0013] 图6示出了根据实施例的示例性vPort体系架构。

[0014] 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。

[0015] 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。

[0016] 图9示出了根据实施例的具有vSwitch的示例性vSwitch体系架构,其中vSwitch具有动态LID分配和预填充的LID。

[0017] 图10示出了根据实施例的示例性多子网InfiniBand架构。

[0018] 图11示出了根据实施例的交换机端口状况的可伸缩表示。

[0019] 图12示出了根据实施例的链路状况的可伸缩表示。

[0020] 图13是根据实施例的用于在高性能计算环境中支持交换机端口状况的可伸缩表示的方法的流程图。

[0021] 图14是根据实施例的用于在高性能计算环境中支持交换机端口状况的可伸缩表示的方法的流程图。

[0022] 图15示出了根据实施例的链路稳定性的可伸缩表示。

[0023] 图16示出了根据实施例的链路可用性的可伸缩表示。

[0024] 图17是根据实施例的用于在高性能计算环境中支持链路稳定性和可用性的可伸缩表示的示例性方法的流程图。

具体实施方式

[0025] 通过示例而非限制的方式在附图的各图中图示了本发明,附图中相似的标号指示类似的元件。应当注意的是,在本公开中对“一个”或“一些”实施例的引用不一定是对相同的实施例,并且这种引用意味着至少一个实施例。虽然讨论了特定的实现方案,但是应当理解的是,这些特定的实现方案仅仅是为了说明性目的而提供。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,可以使用其它部件和配置。

[0026] 在整个附图和具体实施方式中,可以使用共同的标号来指示相同的元素;因此,如果在其它地方描述了元素,那么在图中使用的标号可以或可以不在特定于该图的具体描述中被引用。

[0027] 本文描述的是用于在高性能计算环境中支持交换机端口状况的可伸缩表示的系统和方法。

[0028] 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,可以参考InfiniBand™规范(也被不同地称为InfiniBand规范、IB规范或传统IB规范)。这样的参考被理解为是指可在http://www.inifinibandta.org处获得的于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,其全部内容通过引用被结合于此。对于本领域技术人员来说将清楚的是,可以使用其它类型的高性能网络而没有限制。以下描述还使用胖树(fat-tree)拓扑作为架构拓扑的示例。对于本领域技术人员来说将清楚的是,可以使用其它类型的架构拓扑而没有限制。

[0029] 为了满足当前时代(例如,Exascale时代)对云的需求,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式(paradigm)。RDMA绕过OS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)网络适配器这样的直通技术。根据实施例,虚拟交换机(vSwitch)SR-IOV体系架构可以被提供以用于高性能无损互连网络中的适用性。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并且拓扑无关(topology-agnostic)的动态重新配置机制。

[0030] 并且除此之外,根据实施例,可以提供针对使用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微调以使胖树中施加的开销最小化。

[0031] 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来实现整合、资源的按需供给以及弹性。

[0032] InfiniBand™

[0033] InfiniBand™(IB)是由InfiniBand™贸易协会开发的开放标准无损网络技术。该技术基于提供高吞吐量和低延迟通信的串行点对点全双工互连,特别针对高性能计算(HPC)应用和数据中心。

[0034] 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互连的主机集合。在上层,IB架构构成可以使用路由器互连的一个或多个子网。

[0035] 在子网内,可以使用交换机和点对点链路来连接主机。此外,可以存在主管理实体,即子网管理器(SM),其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。此外,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网络的路由的目的在于本地子网中所有源和目的地对之间的恰当的负载平衡。

[0036] 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、配置端节点和交换机,并从SMA接收通知。

[0037] 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LFT。LFT由SM根据使用中的路由机制来计算。在子网中,使用本地标识符(LID)对端节点和交换机上的主机

通道适配器(HCA)端口进行寻址。LFT中的每个条目包括目的地LID(DLID)和输出端口。表中只支持每LID一个条目。当分组到达交换机时,通过在该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,因为分组在给定的源-目的地对(LID对)之间采取网络中的相同路径。

[0038] 一般而言,除了主子网管理器之外的所有其它子网管理器都在备用模式下起作用以用于容错。然而,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对子网的周期性扫描,以检测任何拓扑改变并相应地重新配置网络。

[0039] 此外,可以使用本地标识符(LID)来寻址子网内的主机和交换机,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3(L3)地址的全局标识符(GID)。

[0040] SM可以在网络初始化时计算路由表(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可以执行对网络的周期性轻扫描以检查拓扑改变。如果在轻扫描期间发现改变,或者如果SM接收到发信号通知网络改变的信息(俘获(trap)),那么SM可以根据所发现的改变来重新配置网络。

[0041] 例如,当网络拓扑改变时,诸如当链路断开时、当添加设备时或者当链路被移除时,SM可以重新配置网络。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局部范围。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。

[0042] 图1中示出了示例InfiniBand架构,其示出了根据实施例的InfiniBand环境100的图示。在图1所示的示例中,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,各个节点(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。

[0043] 在InfiniBand中分区

[0044] 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可以将每个端口上的分区成员资格信息配置为16位分区键(P_Key)的表。SM还可以用分区实施表来配置交换机和路由器端口,其中分区实施表包含与通过这些端口发送或接收数据流量的端节点相关联的P_Key信息。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)方向上经由该端口路由的LID间接相关联的所有成员资格的联合。

[0045] 根据实施例,分区是端口的逻辑组,使得组的成员只能与同一逻辑组的其它成员通信。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,具有无效分区信息的分组就可以被丢弃。在分区的IB系统中,可以使用分区来创建租户集群。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。

[0046] 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配置的表来验证该分组的P_Key值。如果找到无效的P_Key值,那么立即丢弃该分组。以这种方式,只有在共享分区的端口之间才允许通信。

[0047] 图2中示出了IB分区的示例,其示出了根据实施例的分区的集群环境的图示。在图2所示的示例中,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。

[0048] InfiniBand中的虚拟机

[0049] 在过去的十年中,虚拟化高性能计算(HPC)环境的前景已得到相当大的提高,因为已通过硬件虚拟化支持实际上移除了CPU开销;存储器开销已通过虚拟化存储器管理单元被显著降低;已通过使用快速SAN存储装置或分布式联网文件系统减少了存储开销;并且已通过使用像单根输入/输出虚拟化(SR-IOV)这样的设备直通技术减少了网络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)集群并提供必要的性能。

[0050] 然而,当与诸如InfiniBand(IB)的无损网络耦合时,某些云功能(诸如虚拟机(VM)的实时迁移)仍然是个问题,这是由于在这些解决方案中使用的复杂的寻址和路由方案。IB是提供高带宽和低延迟的互连网络技术,因此非常适合HPC和其它通信密集型工作负载。

[0051] 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。然而,使用SR-IOV来实现被分配有IB主机通道适配器(HCA)的VM的实时迁移已被证明是具有挑战性的。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)通信的其它节点会丢失连接性。当发生这种情况时,可以通过向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,来尝试更新丢失的连接。

[0052] 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用于在子网内路由流量。由于LID为16位长,因此可以做出65536个唯一的地址组合,其中只有49151个(0x0001-0xBFFF)可以用作单播地址。因此,可用的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,其在使用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是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地址而形成的。

[0053] 胖树(FTree)拓扑和路由

[0054] 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提供的有用属

性。这些属性包括完全的二分带宽和固有的容错性，这是由于每个源目的地对之间有多个路径可用。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。

[0055] 图3示出了根据实施例的网络环境中的树形拓扑的图示。如图3所示，一个或多个端节点201-204可以在网络架构200中被连接。网络架构200可以基于包括多个叶子交换机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。

[0056] 同样如图3所示，端节点201-204中的每一个可以是多宿主节点，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。

[0057] 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。

[0058] 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之一。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的应用的标准软件栈)子网管理器OpenSM中实现。

[0059] 胖树路由机制的目的在于生成在网络架构中跨链路均匀散布最短路径路由的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。对于连接到相同叶子交换机的端节点，索引次序可以取决于该端节点连接到的交换机端口(即端口编号顺序)。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。

[0060] 根据实施例，在分区的子网中，不允许不是共同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没有被用于用户流量。当胖树路由机制以与它为其它功能路径所做的那样相同的方式为这些路由生成LFT时，会出现问题。由于节点是按索引的次序进行路由的，因此这种行为会导致链路上的平衡恶化。由于路由可以在分区不知情(oblivious)的情况下执行，因此，一般而言，胖树路由的子网提供分区间不好的隔离。

[0061] 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。胖树的不同变体通常是可用的，包括k元n级树(k-ary-n-tree)、扩展的广义胖树(XGFT)、平行端口广义胖树(PGFT)和现实生活胖树(RLFT)。

[0062] k元n级树是具有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的不同数量的上连接和下连接以及在树中每个级别处的不同数量的连接来扩展k元n级胖树。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来定义各种拓扑。然而，为了实用目的，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在胖树中的所有级别处使用相同端口计数的交换机。

[0063] 输入/输出(I/O)虚拟化

[0064] 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。存储流量和服务器间通信的组合施加可能压倒单个服务器的I/O资源的增加的负载,从而导致处理器在等待数据时的积压和空闲处理器。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(虚拟化的)I/O资源的性能、可伸缩性和灵活性,以匹配现代CPU虚拟化中所看到的性能水平。

[0065] 根据实施例,IOV是期望的,因为它可以允许共享I/O资源并且提供对来自VM的资源的受保护的访问。IOV将暴露于VM的逻辑设备与其物理实现方案解耦。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)。

[0066] 根据实施例,一种类型的IOV技术是软件仿真。软件仿真可以允许经解耦的前端/后端软件体系架构。前端可以是放置在VM中的、与由管理程序实现的后端进行通信以提供I/O访问的设备驱动器。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。然而,软件仿真引入了附加的、不期望的计算开销。

[0067] 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。直接分配或设备直通以最小的开销提供接近本地(near to native)的性能。物理设备绕过管理程序并且直接附连到VM。然而,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。

[0068] 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同一设备的多个独立的轻量级实例。这些实例可以被分配给VM作为直通设备,并作为虚拟功能(VF)被访问。管理程序通过唯一的(每设备的)、全特征(fully featured)物理功能(PF)来访问设备。SR-IOV使纯直接分配的可伸缩性问题变得容易。然而,SR-IOV呈现的问题是它可能会影响VM迁移。在这些IOV技术中,SR-IOV可以扩展PCI Express(PCIe)规范,这意味着允许从多个VM直接访问单个物理设备同时维持接近本地的性能。因此,SR-IOV可以提供良好的性能和可伸缩性。

[0069] SR-IOV允许PCIe设备通过向每个顾客分配一个虚拟设备来暴露可以在多个顾客之间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)和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,该唯一的请求者ID使得I/O存储器管理单元(IOMMU)能够区分来自/去往不同VF的流量流。IOMMU还在PF和VF之间应用存储器和中断转换。

[0070] 然而,令人遗憾的是,对于在其中数据中心优化期望虚拟机的透明实时迁移的情况,直接设备分配技术对云提供商造成障碍。实时迁移的实质是将VM的存储器内容复制到远程管理程序。然后在源管理程序处暂停VM,并且在目的地处恢复VM的操作。当使用软件仿真方法时,网络接口是虚拟的,因此其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。

[0071] 然而,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,网络接口的完整内部状态不能被复制,因为它与硬件绑定。作为替代,分配给VM的SR-IOV VF被分离(detach),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,该过程会引入在秒的数量级上的停机时间。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中的附加开销以及对底层网络架构的性能的负面影响。

[0072] InfiniBand SR-IOV体系架构——共享端口

[0073] 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。

[0074] 图4示出了根据实施例的示例性共享端口体系架构。如图所示,主机300(例如,主机通道适配器)可以与管理程序310交互,管理程序310可以将各个虚拟功能330、340、350分配给多个虚拟机。同样,物理功能可以由管理程序310处置。

[0075] 根据实施例,当使用诸如图4所描绘的共享端口体系架构时,主机(例如,HCA)在网络中表现为具有单个共享LID和在物理功能320与虚拟功能330、350、350之间的共享队列对(QP)空间的单个端口。然而,每个功能(即,物理功能和虚拟功能)可以具有其自己的GID。

[0076] 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,并且特殊队列对QP0和QP1(即,用于InfiniBand管理分组的专用队列对)由物理功能拥有。这些QP也被暴露给VF,但是VF不被允许使用QP0(从VF到QP0的所有SMP都被丢弃),并且QP1可以充当由PF拥有的实际QP1的代理。

[0077] 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。

[0078] 然而,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放置的可能性。由于每个LID与具体管理程序相关联,并且在驻留于该管理程序上的所有VM之间共享,因此迁移的VM(即,迁移到目的地管理程序的虚拟机)必须将其LID改变为目的地管理程序的LID。此外,作为受限的QP0访问的结果,子网管理器不能在VM内部运行。

[0079] InfiniBand SR-IOV体系架构模型——虚拟交换机(vSwitch)

[0080] 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交互,管理程序410可以将各个虚拟功能430、440、450分配给多个虚拟机。同样,物理功能可以由管理程序410处置。虚拟交换机415也可以由管理程序401处置。

[0081] 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),这意味着分配给VF的VM被分配完整的IB地址集合(例如,GID、GUID、LID)以及硬件中的专用QP空间。对于网络的其余部分和SM,HCA 400看起来像经由虚拟交换机415、具有连接到它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。

[0082] 根据实施例,vSwitch体系架构提供透明的虚拟化。然而,由于每个虚拟功能都被分配唯一的LID,因此可用LID的数量被迅速消耗。同样,在使用许多LID地址(即,每个物理功能和每个虚拟功能都有一个LID地址)的情况下,必须由SM计算更多的通信路径,并且必须将更多的子网管理分组(SMP)发送到交换机以便更新其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,并且由于每个VM(经由VF)、物理节点和交换机各自占用一个LID,所以网络中的物理节点和交换机的数量限制了活动VM的数量,并且反之亦然。

[0083] InfiniBand SR-IOV体系架构模型——虚拟端口(vPort)

[0084] 图6示出了根据实施例的示例性vPort概念。如图所示，主机300(例如，主机通道适配器)可以与管理程序410交互，管理程序410可以将各个虚拟功能330、340、350分配给多个虚拟机。同样，物理功能可以由管理程序310来处置。

[0085] 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，定义不规定实现方案必须是特定于SRIOV的)，并且vPort的目标是使在子网中处置VM的方式标准化。利用vPort概念，可以定义可在空间和性能域二者中都更可伸缩的、类似SR-IOV共享端口的体系架构和类似vSwitch的体系架构二者或这二者的组合。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。

[0086] InfiniBand SR-IOV体系架构模型——具有预填充LID的vSwitch

[0087] 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。

[0088] 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。如图所示，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中的每个又可以分别与管理程序511、521和531交互。每个管理程序又可以与和其交互的主机通道适配器结合设立多个虚拟功能514、515、516、524、525、526、534、535、536并将这多个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又可以将虚拟机4 553分配给虚拟功能1 534。在主机通道适配器中的每一个上，管理程序可以通过全特征物理功能513、523、533来访问主机通道适配器。

[0089] 根据实施例，交换机501-504中的每一个可以包括多个端口(未示出)，这些端口在设置线性转发表中被使用以便导引网络交换环境600内的流量。

[0090] 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处置。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)，这意味着分配给VF的VM被分配完整的IB地址(例如，GID、GUID、LID)集合以及硬件中的专用QP空间。对于网络的其余部分和SM(未示出)，HCA 510、520和530看起来像经由虚拟交换机的、具有连接到它们的附加节点的交换机。

[0091] 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被预填充到各个物理功能513、523、533以及虚拟功能514-516、524-526、534-536(甚至那些当前未与活动虚拟机相关联的虚拟功能)。例如，物理功能513被预填充有LID 1，而虚拟功能1 534被预填充有LID 10。当网络被引导(root)时，LID被预填充在启用SR-IOV vSwitch的子网中。即使并非所有的VF都被网络中的VM占用时，所填充的VF也被分配有LID，如图7所示。

[0092] 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(为了冗余，两个端口是常见的)，虚拟HCA也可以用两个端口来表示，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。

[0093] 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消耗再多一个LID。在IB子网中的所有管理程序中可用的所有VF的总和给出了被允许在该子网中运行的VM的最大量。例如,在子网中每管理程序具有16个虚拟功能的IB子网中,那么每个管理程序在该子网中消耗17个LID(16个虚拟功能中的每个虚拟功能一个LID加上用于物理功能的一个LID)。在这种IB子网中,单个子网的理论管理程序极限取决于可用单播LID的数量并且是:2891个(49151个可用LID除以每管理程序17个LID),并且VM的总数(即,极限)是46256个(2891个管理程序乘以每管理程序16个VF)。(实际上,这些数字在实际中更小,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)。注意的是,vSwitch不需要占用附加的LID,因为它可以与PF共享LID。

[0094] 根据实施例,在具有预填充LID的vSwitch体系架构中,当网络第一次被引导时,为所有LID计算通信路径。当需要启动新的VM时,系统不必在子网中添加新的LID,否则该动作将导致网络完整的重新配置,包括路径重新计算,这是最耗时的部分。作为替代,在管理程序之一中定位用于VM的可用端口(即,可用的虚拟功能),并且将该虚拟机附连到该可用的虚拟功能。

[0095] 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同一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)的特征提供朝向一个物理机器的替代路径,而不受LMC的限制(其要求LID必须是顺序的)约束。当需要迁移VM并将其相关联的LID携带到目的地时,自由使用非顺序的LID是尤其有用的。

[0096] 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,可以考虑某些注意事项。例如,由于LID在网络被引导时被预填充在启用SR-IOV vSwitch的子网中,所以(例如在启动时的)初始路径计算会比不预填充LID花费更长的时间。

[0097] InfiniBand SR-IOV体系架构模型——具有动态LID分配的vSwitch

[0098] 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。

[0099] 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。如图所示,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中的每一个又可以分别与管理程序511、521、531交互。每个管理程序又可以与和其交互的主机通道适配器结合来设立多个虚拟功能514、515、516、524、525、526、534、535、536并将这多个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又可以将虚拟机4 553分配给虚拟功能1 534。在主机通道适配器中的每一个上,管理程序可以通过全特征物理功能513、523、533来访问主机通道适配器。

[0100] 根据实施例,交换机501-504中的每一个可以包括多个端口(未示出),这些端口在设置线性转发表中被使用以便导引网络交换环境700内的流量。

[0101] 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处置。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),这意味着分配给VF的VM被分配完整的IB地址(例如,GID、GUID、LID)集合以及硬件中的专用

QP空间。对于网络的其余部分和SM(未示出),HCA 510、520和530看起来像经由虚拟交换机的、具有连接到它们的附加节点的交换机。

[0102] 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被动态分配给各个物理功能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这些虚拟功能也可以接收动态分配的LID。例如,由于虚拟机1 550是活动的并且与虚拟功能1 514相关联,所以虚拟功能514可以被分配LID 5。同样地,虚拟功能2 515、虚拟功能3 516和虚拟功能1 534各自与活动的虚拟功能相关联。由此,这些虚拟功能被分配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,那些当前未与活动虚拟机相关联的虚拟功能不接收LID分配。

[0103] 根据实施例,利用动态LID分配,可以实质上减少初始路径计算。当网络第一次被引导并且不存在VM时,那么可以使用相对较少数量的LID来用于初始路径计算和LFT分发。

[0104] 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(为了冗余两个端口是常见的),虚拟HCA也可以用两个端口来表示,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。

[0105] 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槽以便决定在哪个管理程序上引导新添加的VM,并且也找到未使用的唯一的单播LID。然而,在网络和交换机的LFT中没有用于处置新添加的LID的已知路径。在其中每分钟可以引导几个VM的动态环境中,为了处置新添加的VM而计算新的一组路径是非期望的。在大型IB子网中,计算新的一组路由会花费几分钟,并且每次引导新的VM时这个过程都将不得不重复。

[0106] 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,所以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、将转发端口从属于管理程序(VM被创建在该管理程序处)的PF的LID条目复制到新添加的LID、以及发送单个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计算新的一组路由的需要。

[0107] 根据实施例,在具有动态LID分配体系架构的vSwitch中分配的LID不必是顺序的。当将具有预填充LID的vSwitch中的每个管理程序上的VM上所分配的LID与具有动态LID分配的vSwitch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是非顺序的,而被预填充的那些LID本质上是顺序的。在vSwitch动态LID分配体系架构中,当创建新的VM时,在该VM的整个寿命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继承已经分配给对应VF的LID,并且在没有实时迁移的网络中,连续地附连到给定VF的VM获得相同的LID。

[0108] 根据实施例,具有动态LID分配体系架构的vSwitch可以以一些附加的网络和运行时SM开销为代价,解决具有预填充LID体系架构模型的vSwitch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,需要发送每交换机一个子网管理分组(SMP)。因为每个VM正在使用与其主机管理程序相同的路径,所以类似LMC的功能也不可用。然而,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,那么并不是所有VF都被允

许同时附连到活动VM上，但是，当操作接近单播LID极限时，具有更多的备用管理程序和VF增加了分段网络的优化和灾难恢复的灵活性。

[0109] InfiniBand SR-IOV体系架构模型——具有动态LID分配和预填充LID的vSwitch

[0110] 图9示出了根据实施例的具有vSwitch的示例性vSwitch体系架构，其中vSwitch具有动态LID分配和预填充LID。如图所示，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。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中的每一个又可以分别与管理程序511、521和531交互。每个管理程序又可以与和其交互的主机通道适配器结合来设立多个虚拟功能514、515、516、524、525、526、534、535、536并将这多个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又可以将虚拟机4 553分配给虚拟功能2 535。在主机通道适配器中的每一个上，管理程序可以通过全特征物理功能513、523、533来访问主机通道适配器。

[0111] 根据实施例，交换机501-504中的每一个可以包括多个端口(未示出)，这些端口在设置线性转发表中被使用以便导引网络交换环境800内的流量。

[0112] 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处置。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)，这意味着分配给VF的VM被分配完整的IB地址(例如，GID、GUID、LID)集合以及硬件中的专用QP空间。对于网络的其余部分和SM(未示出)，HCA 510、520和530看起来像经由虚拟交换机的、具有连接到它们的附加节点的交换机。

[0113] 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可以被布置为带有具有预填充LID体系架构的vSwitch，而管理程序521可以被布置为带有具有预填充LID和动态LID分配的vSwitch。管理程序531可以被布置为带有具有动态LID分配的vSwitch。因此，物理功能513和虚拟功能514-516使其LID被预填充(即，即使那些未附连到活动虚拟机的虚拟功能也被分配LID)。物理功能523和虚拟功能1 524可以使其LID被预填充，而虚拟功能2 525和虚拟功能3 526使其LID被动态分配(即，虚拟功能2 525可用于动态LID分配，并且虚拟功能3 526由于虚拟机3 552被附连而具有动态分配的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)可以使其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被附连到那里所以具有动态分配的LID 9。

[0114] 根据诸如图9所示的在其中(在任何给定管理程序内独立地或组合地)利用具有预填充LID的vSwitch和具有动态LID分配的vSwitch两者的实施例，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0<＝预填充的VF<＝(每主机通道适配器的)总VF的范围内，并且，可用于动态LID分配的VF可以通过从(每主机通道适配器)VF的总数量减去预填充VF的数量而得出。

[0115] 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(为了冗余两个端口是常见的)，虚拟HCA也可以用两个端口来表示，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。

[0116] InfiniBand——子网间通信(架构管理器)

[0117] 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两个或更多个子网的InfiniBand架构。

[0118] 图10示出了根据实施例的示例性多子网InfiniBand架构。如图所示,在子网A 1000内,多个交换机1001-1004可以在子网A 1000(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又可以与管理程序1011交互。该管理程序又可以与和其交互的主机通道适配器结合来设立多个虚拟功能1014。该管理程序可以附加地将虚拟机分配给虚拟功能中的每一个,诸如虚拟机1 10105被分配给虚拟功能1 1014。在主机通道适配器中的每一个上,管理程序可以通过全特征物理功能(诸如物理功能1013)来访问其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又可以与管理程序1031交互。该管理程序又可以与和其交互的主机通道适配器结合来设立多个虚拟功能1034。管理程序可以附加地将虚拟机分配给虚拟功能中的每一个,诸如虚拟机2 1035被分配给虚拟功能2 1034。在主机通道适配器中的每一个上,管理程序可以通过全特征物理功能(诸如物理功能1033)访问其相关联的主机通道适配器。应当注意的是,虽然在每个子网(即,子网A和子网B)内仅示出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其对应的部件。

[0119] 根据实施例,主机通道适配器中的每一个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,并且每个HCA可以被设立为具有如上所述的不同的体系架构模型。虽然图10内的两个子网都被显示为使用具有预填充LID体系架构模型的vSwitch,但这并不意味着暗示所有此类子网配置都必须遵循相似的体系架构模型。

[0120] 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。

[0121] 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)、构建与虚拟机相关的数据库对象,这些数据库对象形成用于构建虚拟机架构简档的基础。此外,架构管理器可以针对哪些子网被允许经由哪些路由器端口使用哪些分区编号进行通信来定义合法的子网间连接性。

[0122] 根据实施例,当始发源(诸如子网A内的虚拟机1)处的流量被寻址到不同子网处的目的地(诸如子网B内的虚拟机2)时,该流量可以被寻址到子网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的链路将该流量传递到子网B。

[0123] 交换机端口状况的可伸缩表示

[0124] 根据实施例,在IB规范下,为了观察链路状况改变,IB规范定义每个端口(例如,在任何给定交换机或虚拟交换机处的端口中的每个端口)处的、可以指示何时任何端口状态改变的属性。为了让SM确定架构内的任何端口处的状况是否已经改变状态,SM必须为每个

端口发送子网管理分组,以便确定端口状况是否已经改变。

[0125] 根据实施例,虽然上面定义的用于确定架构内的端口状况的方法对于大多数是静态的架构(例如,由其中端口状况改变不经常发生的物理端节点构成的那些架构)工作良好,但是对于那些已经被虚拟化的架构(例如,在引入由动态创建的虚拟机使用的虚拟HCA以及其中vSwitch体系架构被用于互连虚拟HCA端口的情况下),以及对于非常大的物理架构配置,该方法不能很好地伸缩。

[0126] 根据实施例,可以提供交换机端口状况的可伸缩表示。通过在每个交换机(物理交换机和虚拟交换机二者)处添加交换机端口状况的可伸缩表示——而不是单独地获得所有交换机端口状况改变,交换机端口状况的可伸缩表示可以通过仅使用用于每个端口状况的几个位的信息来组合可以进行伸缩的多个端口。

[0127] 根据实施例,交换机端口状况的可伸缩表示可以是在每个交换机处的固定大小的消息,该固定大小的消息可以表示在与该固定大小的消息相关联的交换机中的端口的全部或子集中的所有端口状况信息。这在使用虚拟化的架构中尤其重要,因为可伸缩表示可以动态地表示虚拟交换机及其相关联的端口。如上面所讨论的,虚拟机(即,虚拟端节点)能够并且被设计为进行迁移(例如,为了性能益处),这可能意味着频繁地改变架构内的物理交换机和虚拟交换机上的端口状况。传统规范依赖于不频繁的改变——即,当SM检查自从上次检查以来是否发生了任何状态改变时,不太可能发生任何改变(默认没有改变),因此SM可以在单次操作中接收任何端口是否存在任何改变的指示,并且如果没有改变,则可以继续到下一个交换机。然而,每当任何端口发生任何改变时,那么所有端口都必须被SM单独核查。然而,在虚拟机的情况下,SM可以预期检测到架构中的端口处的更频繁的状态改变。

[0128] 根据实施例,交换机端口状况的可伸缩表示可以是固定大小的消息(例如,交换机属性),该消息可以向SM提供在一个操作(即,一个SMP)中观察交换机处的所有端口的所有状态改变的手段。这减少了SM以其它方式将会遇到的开销,并且优化了查询每个交换机以确定哪些端口需要进一步处置的SM机制。

[0129] 根据实施例,通常,观察/检查各个交换机端口的链路状况需要多个SMP操作,因为必须为交换机处的每个端口发送一个SMP。然而,通过将每个链路状况组合成一个属性,SM可以向每个交换机发送较少的SMP来发现在每个端口处的链路状况,从而减少拓扑发现所需的开销。同样,也同样地减少了SM确定它是否需要对端口执行用于检索更多信息或设立新的配置参数的附加操作的开销。

[0130] 根据实施例,交换机端口状况的可伸缩表示可以是在其中端口/链路状况被表示为标量对象(一位值或多位值)的属性。其中包含标量对象的该属性可以提供获取(虚拟)交换机的逻辑链路状态的压缩方式。这种属性可以附加地被路由算法使用,以便在平衡通过架构的各个路由时忽略虚拟链路。

[0131] 根据实施例,64位掩码值足以覆盖所有交换机(现有交换机通常具有少于64个端口)。然而,如果使用更大的交换机,那么可以通过使用索引到更高端口范围中的属性修饰符来扩展位掩码值上的这个上限(cap)。

[0132] 根据实施例,除了减少SM确定架构中的端口/链路处是否已经发生任何状态改变所需的SMP的数量之外,交换机端口状况的可伸缩表示还可以优化对架构的拓扑的SM发现,因为每个端口的链路状况可以被表示为属性内的标量对象。

[0133] 图11示出了根据实施例的交换机端口状况的可伸缩表示。更具体而言，图11图示了具有表示交换机端口状况的可伸缩表示的属性的交换机。

[0134] 根据实施例，并且如图11所示，交换机1100可以包括多个端口，诸如端口1110-1133(注意的是，图11中所示出的端口的数量既不是说明也不是指示架构(诸如InfiniBand架构)内给定交换机处的端口的通常数量)。交换机1100还包括交换机端口状况属性1150，交换机端口状况属性1150可以是表示交换机1100中的交换机端口1110-1133的交换机端口状况信息的固定大小的消息。

[0135] 根据实施例，诸如子网管理器1140的管理模块可以发送一个SMP 1145来查询交换机端口状况属性1150，而不是针对交换机1100内的每个端口发送一个SMP来确定每个端口的状况。SMP可以在检查时中继(relay)每个端口1110-1133的状况。

[0136] 图12示出了根据实施例的扩展链路状况的可伸缩表示。更具体而言，图12图示了具有表示扩展链路状况的可伸缩表示的属性的交换机。

[0137] 根据实施例，并且如图12所示，交换机1100可以包括多个端口，诸如端口1110-1133(注意的是，图12中所示出的端口的数量既不是说明也不是指示架构(诸如InfiniBand架构)内给定交换机处的端口的通常数量)。交换机1100还包括扩展链路状况属性1250，该扩展链路状况属性1250可以是表示连接到交换机1100中的交换机端口1110-1133的任何链路的状况的固定大小的消息。

[0138] 根据实施例，诸如子网管理器1140的管理模块可以发送一个SMP 1245来查询扩展链路状况属性1250，而不是针对交换机1100内的每个端口发送一个SMP来确定每个端口处的扩展链路状况。SMP可以在检查时中继每个端口1110-1133状况的链路状况。

[0139] 图13是根据实施例的用于在高性能计算环境中支持交换机端口状况的可伸缩表示的方法的流程图。

[0140] 在步骤1310处，诸如InfiniBand子网管理器的管理实体可以向交换机发送管理分组，该管理分组请求该管理分组被发送到的交换机处的每个端口的交换机端口状况。

[0141] 在步骤1320处，管理分组被发送到的交换机可以接收该管理分组。

[0142] 在步骤1330处，交换机可以经由包含该交换机处的每个端口的交换机端口状况的属性来提供该交换机的交换机端口中的每一个的状况。

[0143] 在步骤1340处，所请求的每个交换机端口的状况可以经由管理分组被中继到诸如该InfiniBand子网管理器的管理实体。

[0144] 图14是根据实施例的用于在高性能计算环境中支持交换机端口状况的可伸缩表示的方法的流程图。在步骤1410处，该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至少一个子网，该至少一个子网包括：一个或多个交换机、多个主机通道适配器、多个端节点，以及子网管理器。该一个或多个交换机至少包括叶子交换机，其中该一个或多个交换机中的每个交换机包括多个端口，并且其中该一个或多个交换机中的每个交换机包括至少一个属性；其中该多个主机通道适配器经由一个或多个交换机互连；该多个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，该子网管理器运行在一个或多个交换机中的一个交换机上或多个主机通道适配器中的一个主机通道适配器上。

[0145] 在步骤1420处，该方法可以将一个或多个交换机上的多个端口中的每个端口与交

换机端口状况相关联。

[0146] 在步骤1430处，该方法可以将与每个交换机上的多个端口中的每个端口相关联的每个交换机端口状况表示在相关联的交换机处的至少一个属性中。

[0147] 可伸缩链路稳定性属性

[0148] 图15示出了根据实施例的链路稳定性的可伸缩表示。更具体而言，图15图示了具有表示链路稳定性的可伸缩表示的属性的交换机。

[0149] 根据实施例，并且如图15所示，交换机1100可以包括多个端口，诸如端口1110-1133(注意的是，图15中所示出的端口的数量既不是说明也不是指示架构(诸如InfiniBand架构)内给定交换机处的端口的通常数量)。交换机1100还包括链路稳定性属性1550，该链路稳定性属性1550可以是表示连接到交换机1100中的交换机端口1110-1133的任何链路的稳定性的固定大小的消息。

[0150] 根据实施例，诸如子网管理器1140的管理模块可以发送一个SMP 1545来查询链路稳定性属性1550，而不是针对交换机1100内的每个端口发送一个SMP来确定每个端口处的链路稳定性。SMP可以在检查时中继每个端口1110-1133的链路稳定性。

[0151] 根据实施例，子网管理代理(SMA)1555可以在(例如，可变的或固定的)时间段内监视连接到交换机端口1110-1133的链路的稳定性。这种监视可以包括例如连接到交换机处的每个端口的每个链路在所设置的时间段期间遇到的错误的数量。

[0152] 根据实施例，由SMA 1555在交换机内的任何给定端口处发现的链路错误的数量可以被用于连续地更新链路稳定性属性1550，链路稳定性属性1550可以被来自子网管理器的单个SMP查询。有利地，通过具有这种可伸缩的链路稳定性属性，SM可以经由一个SMP从子网中的任何给定交换机收集链路稳定性信息，而不是发送多个SMP并且每个SMP用于检查节点处的每个链路。

[0153] 根据实施例，所公开的实施例附加地允许连续地监视和更新系统中的任何给定节点处的链路稳定性属性(即，经由每个节点的SMA)，使得SM可以(经由例如Get()操作)收集连接到SM所管理的子网中的节点的每个链路的链路稳定性信息。

[0154] 可伸缩链路可用性属性

[0155] 图16示出了根据实施例的链路可用性的可伸缩表示。更具体而言，图16图示了具有表示链路可用性的可伸缩表示的属性的交换机。

[0156] 根据实施例，并且如图16所示，交换机1100可以包括多个端口，诸如端口1110-1133(注意的是，图16中所示出的端口的数量既不是说明也不是指示架构(诸如InfiniBand架构)内给定交换机处的端口的通常数量)。交换机1100还包括链路可用性属性1650，链路可用性属性1650可以是表示连接到交换机1100中的交换机端口1110-1133的任何链路的可用性的固定大小的消息。

[0157] 根据实施例，诸如子网管理器1140的管理模块可以发送一个SMP 1645来查询链路可用性属性1650，而不是针对交换机1100内的每个端口发送一个SMP来确定每个端口处的链路可用性。SMP可以在检查时中继每个端口1110-1133的链路可用性。

[0158] 根据实施例，子网管理代理(SMA)1655可以在(例如，可变的或固定的)时间段内监视连接到交换机端口1110-1133的链路的可用性。这种监视可以包括例如连接到交换机的每个端口的每个链路上的拥塞水平。

[0159] 根据实施例，如由SMA 1655确定的每个链路上的拥塞水平可以被用于连续地更新链路可用性属性1650，链路可用性属性1650可以被来自子网管理器的单个SMP查询。有利地，通过具有这种可伸缩的链路可用性属性，SM可以经由一个SMP从子网中的任何给定交换机收集链路可用性信息，而不是发送多个SMP并且每一个SMP用于检查交换机/节点处的每个链路。

[0160] 根据实施例，所公开的实施例附加地允许连续地监视和更新系统中的任何给定节点处的链路可用性(即，经由每个节点处的SMA)，使得SM可以(经由例如Get()操作)收集连接到SM所管理的子网中的节点的每个链路的链路可用性信息。

[0161] 图17是根据实施例的用于在高性能计算环境中支持链路稳定性和可用性的可伸缩表示的示例性方法的流程图。

[0162] 在步骤1710处，该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至少一个子网，该至少一个子网包括：一个或多个交换机、多个主机通道适配器、多个端节点、以及子网管理器。该一个或多个交换机至少包括叶子交换机，其中该一个或多个交换机中的每个交换机包括多个端口，并且其中该一个或多个交换机中的每个交换机包括至少一个属性；其中该多个主机通道适配器经由一个或多个交换机互连；该多个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；该子网管理器运行在一个或多个交换机中的一个交换机上或多个主机通道适配器中的一个主机通道适配器上。

[0163] 在步骤1720处，该方法可以在一个或多个交换机中的每个交换机处提供至少一个属性。

[0164] 在步骤1730处，该方法可以在一个或多个交换机中的交换机处提供多个子网管理代理(SMA)中的子网管理代理。

[0165] 在步骤1740处，该方法可以由一个或多个交换机中的交换机的交换机的SMA来监视交换机处的多个端口中的每个端口处的链路稳定性和链路可用性中的至少一个。

[0166] 根据实施例，一种用于在高性能计算环境中支持链路稳定性和可用性的可伸缩表示的系统包括：一个或多个微处理器；至少一个子网，该至少一个子网包括：一个或多个交换机、多个主机通道适配器、多个端节点、以及子网管理器。该一个或多个交换机至少包括叶子交换机，其中该一个或多个交换机中的每个交换机包括多个端口，并且其中该一个或多个交换机中的每个交换机包括至少一个属性；其中该多个主机通道适配器经由该一个或多个交换机互连；该多个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；该子网管理器运行在一个或多个交换机中的一个交换机上或多个主机通道适配器中的一个主机通道适配器上；其中一个或多个交换机中的每个交换机包括至少一个属性；其中在一个或多个交换机中的交换机处提供多个子网管理代理(SMA)中的子网管理代理；其中一个或多个交换机中的交换机的子网管理代理监视该交换机的多个端口中的每个端口处的链路稳定性和该交换机处的多个端口中的每个端口处的链路可用性中的至少一个。

[0167] 根据实施例，上述系统还包括，由一个或多个交换机中的交换机的SMA监视该交换机的多个端口中的每个端口处的链路稳定性包括：在监视时间段内，对附连到交换机的多个端口中的每个端口的每个链路处的错误的数量进行计数；以及，其中在由SMA监视交换机

的多个端口中的每个端口处的链路稳定性之后，SMA将所计数的每个链路的错误的表示填充在该至少一个属性中。

[0168] 根据实施例，在上述系统中，子网管理器使用一个操作来确定一个或多个交换机中的交换机上的每个端口的链路稳定性。

[0169] 根据实施例，在上述系统中，该一个操作包括子网管理分组。

[0170] 根据实施例，上述系统还包括，由一个或多个交换机中的交换机的SMA监视交换机的多个端口中的每个端口处的链路可用性包括：在监视时间段内，观察附连到交换机的多个端口中的每个端口的每个链路处的流量负载；并且在由SMA监视交换机的多个端口中的每个端口处的链路可用性之后，SMA将所观察到的每个链路的流量负载的表示填充在该至少一个属性中。

[0171] 根据实施例，在上述系统中，子网管理器使用一个操作来确定一个或多个交换机中的交换机上的每个端口的链路可用性。

[0172] 根据实施例，在上述系统中，该一个操作包括子网管理分组。

[0173] 根据实施例，一种用于在高性能计算环境中支持链路稳定性和可用性的可伸缩表示的方法包括：在包括一个或多个微处理器的一个或多个计算机处提供至少一个子网，该至少一个子网包括：一个或多个交换机、多个主机通道适配器、多个端节点以及子网管理器，该一个或多个交换机至少包括叶子交换机，其中该一个或多个交换机中的每个交换机包括多个端口，其中该多个主机通道适配器经由该一个或多个交换机互连，该多个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，该子网管理器运行在一个或多个交换机中的一个交换机上或多个主机通道适配器中的一个主机通道适配器上；在一个或多个交换机中的每个交换机处提供至少一个属性；在一个或多个交换机中的交换机处提供多个子网管理代理(SMA)中的子网管理代理；由一个或多个交换机中的交换机的SMA监视交换机的多个端口中的每个端口处的链路稳定性和交换机处的多个端口中的每个端口处的链路可用性中的至少一个。

[0174] 根据实施例，上述方法还包括，由一个或多个交换机中的交换机的SMA监视交换机的多个端口中的每个端口处的链路稳定性包括：在监视时间段内，对附连到交换机的多个端口中的每个端口的每个链路处的错误的数量进行计数；以及在完成由一个或多个交换机中的交换机的SMA对多个端口中的每个端口处的链路稳定性的监视之后，SMA将所计数的每个链路的错误的表示填充在该至少一个属性中。

[0175] 根据实施例，上述方法包括由子网管理器使用一个操作来确定一个或多个交换机中的交换机上的每个端口的链路稳定性。

[0176] 根据实施例，在上述方法中，该一个操作包括子网管理分组。

[0177] 根据实施例，上述方法还包括，由一个或多个交换机中的交换机的SMA监视交换机的多个端口中的每个端口处的链路可用性包括：在监视时间段内，观察附连到交换机的多个端口中的每个端口的每个链路处的流量负载；以及在监视时间段内完成监视交换机的多个端口中的每个端口处的链路可用性之后，由SMA将所观察到的每个链路的流量负载的表示填充在该至少一个属性中。

[0178] 根据实施例，在上述方法中，子网管理器使用一个操作来确定一个或多个交换机中的一个交换机上的每个端口的链路可用性状况。

[0179] 根据实施例，在上述方法中，该一个操作包括子网管理分组。

[0180] 根据实施例，一种非暂态计算机可读存储介质，包括存储在其上指令，该指令用于在高性能计算环境中支持链路稳定性和可用性的可伸缩表示，该指令当由一个或多个计算机读取并执行时，使一个或多个计算机执行包括以下各项的步骤：在包括一个或多个微处理器的一个或多个计算机处提供至少一个子网，该至少一个子网包括：一个或多个交换机、多个主机通道适配器、多个端节点以及子网管理器，该一个或多个交换机至少包括叶子交换机，其中该一个或多个交换机中的每个交换机包括多个端口，其中该多个主机通道适配器经由一个或多个交换机互连，该多个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，该子网管理器运行在一个或多个交换机中的一个交换机上或多个主机通道适配器中的一个主机通道适配器上；在一个或多个交换机中的每个交换机处提供至少一个属性；在一个或多个交换机中的交换机处提供多个子网管理代理(SMA)中的子网管理代理；由一个或多个交换机中的交换机的SMA监视交换机的多个端口中的每个端口处的链路稳定性和交换机处的多个端口中的每个端口处的链路可用性中的至少一个。

[0181] 根据实施例，上述非暂态计算机可读存储介质还包括，由一个或多个交换机中的交换机的SMA监视交换机的多个端口中的每个端口处的链路稳定性包括：在监视时间段内，对附连到交换机的多个端口中的每个端口的每个链路处的错误的数量进行计数；以及在完成由一个或多个交换机中的交换机的SMA对多个端口中的每个端口处的链路稳定性的监视之后，SMA将所计数的每个链路的错误的表示填充在该至少一个属性中。

[0182] 根据实施例，上述非暂态计算机可读存储介质还包括：由子网管理器使用一个操作来确定一个或多个交换机中的交换机上的每个端口的链路稳定性。

[0183] 根据实施例，在上述非暂态计算机可读存储介质中，该一个操作包括子网管理分组。

[0184] 根据实施例，上述非暂态计算机可读存储介质还包括，由一个或多个交换机中的交换机的SMA监视交换机的多个端口中的每个端口处的链路可用性包括：在监视时间段内，观察附连到交换机的多个端口中的每个端口的每个链路处的流量负载；以及在监视时间段内完成监视交换机的多个端口中的每个端口处的链路可用性之后，由SMA将所观察到的每个链路的流量负载填充在至少一个属性中。

[0185] 根据实施例，在上述非暂态计算机可读存储介质中，子网管理器使用一个操作来确定一个或多个交换机中的一个交换机上的每个端口的链路可用性状况，该一个操作包括子网管理分组。

[0186] 根据实施例，一种计算机程序包括以机器可读格式的程序指令，该程序指令当由计算机系统执行时使计算机系统执行上述方法。

[0187] 根据实施例，一种计算机程序包括存储在非暂态机器可读数据存储介质中的上述计算机程序。

[0188] 本发明的许多特征可以在硬件、软件、固件或其组合中执行，利用硬件、软件、固件或其组合来执行，或者在硬件、软件、固件或其组合的帮助下执行。因此，本发明的特征可以利用(例如，包括一个或多个处理器的)处理系统来实现。

[0189] 本发明的特征可以在计算机程序产品中实现、利用计算机程序产品实现、或者在

计算机程序产品的帮助下实现，其中计算机程序产品是具有存储于其上/其中的指令的(一个或多个)存储介质或(一个或多个)计算机可读介质，该指令可用来编程处理系统以执行本文所呈现的任何特征。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媒体或设备。

[0190] 在被存储在(一个或多个)机器可读介质中的任何一个的情况下，本发明的特征可以被结合到软件和/或固件中，以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、设备驱动器、操作系统和执行环境/容器。

[0191] 本发明的特征也可以利用例如诸如专用集成电路(ASIC)的硬件部件在硬件中实现。实现硬件状态机以执行本文所描述的功能对相关领域的技术人员而言将是清楚的。

[0192] 此外，本发明可以方便地利用一个或多个常规的通用或专用数字计算机、计算设备、机器或微处理器来实现，其包括一个或多个处理器、存储器和/或根据本公开的教导编程的计算机可读存储介质。如对软件领域的技术人员而言将清楚的，熟练的程序员基于本公开的教导可以容易地准备适当的软件编码。

[0193] 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，它们是作为示例而不是限制给出的。对相关领域的技术人员将清楚的是，在不背离本发明的精神和范围的情况下，其中可以做出各种形式和细节上的改变。

[0194] 本发明已经借助说明指定功能的执行及其关系的功能构建块进行了描述。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。可以定义替代的边界，只要指定的功能及其关系被适当地执行。任何这种可替代的边界因此在本发明的范围和精神之内。

[0195] 本发明的以上描述是为了说明和描述的目的而提供的。它不旨在是穷尽的或者要把本发明限定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。许多修改和变形对本领域技术人员来说将是清楚的。这些修改和变形包括所公开特征的任何相关组合。实施例被选择和描述以便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用于各种实施例并且可以进行适于预期特定用途的各种修改。本发明的范围旨在由以下权利要求及其等同物来定义。

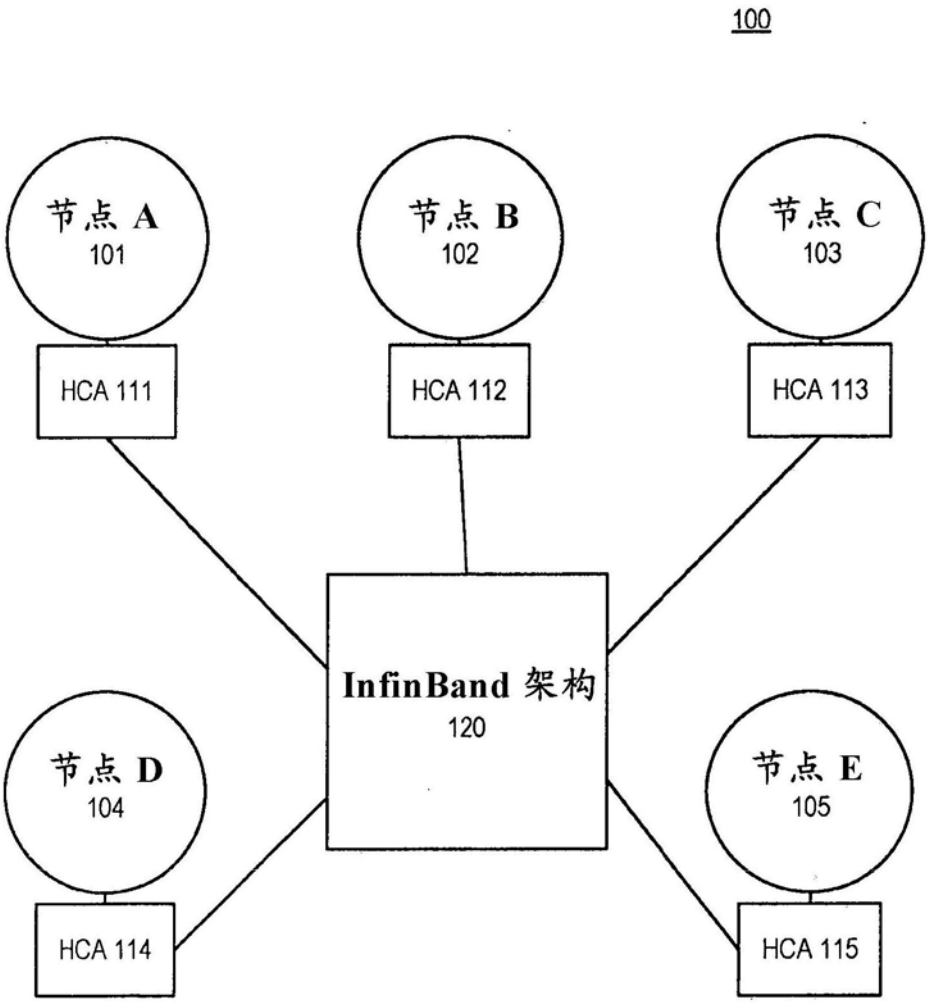

100
节点 A
101
节点 B
102
节点 C
103
HCA 111
HCA 112
HCA 113
InfinBand 架构
120
节点 D
104
节点 E
105
HCA 114
HCA 115

图1

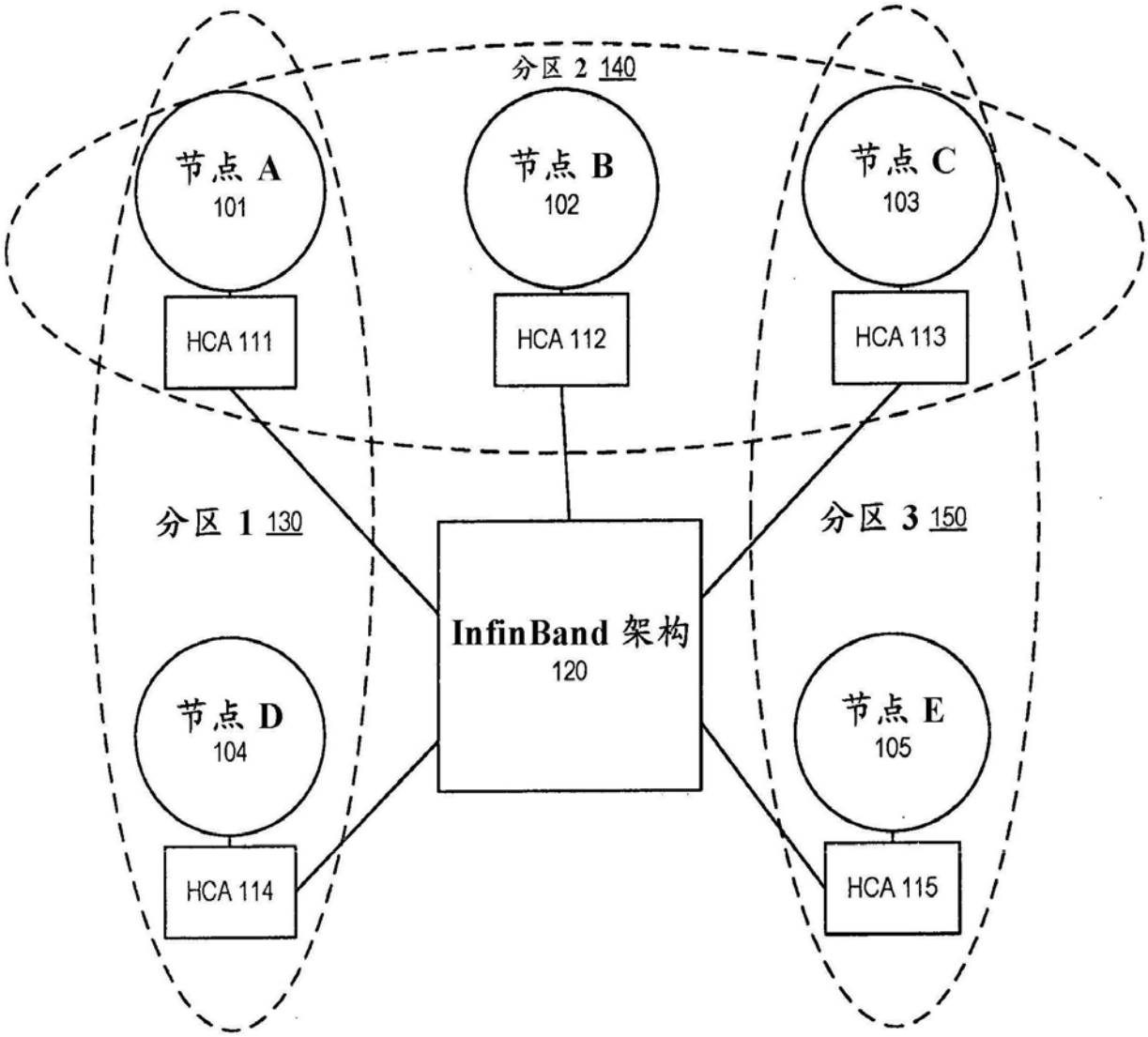

分区 2 140
节点 A
101
节点 B
102
节点 C
103
HCA 111
HCA 112
HCA 113
分区 1 130
分区 3 150
InfinBand 架构
120
节点 D
104
节点 E
105
HCA 114
HCA 115

图2

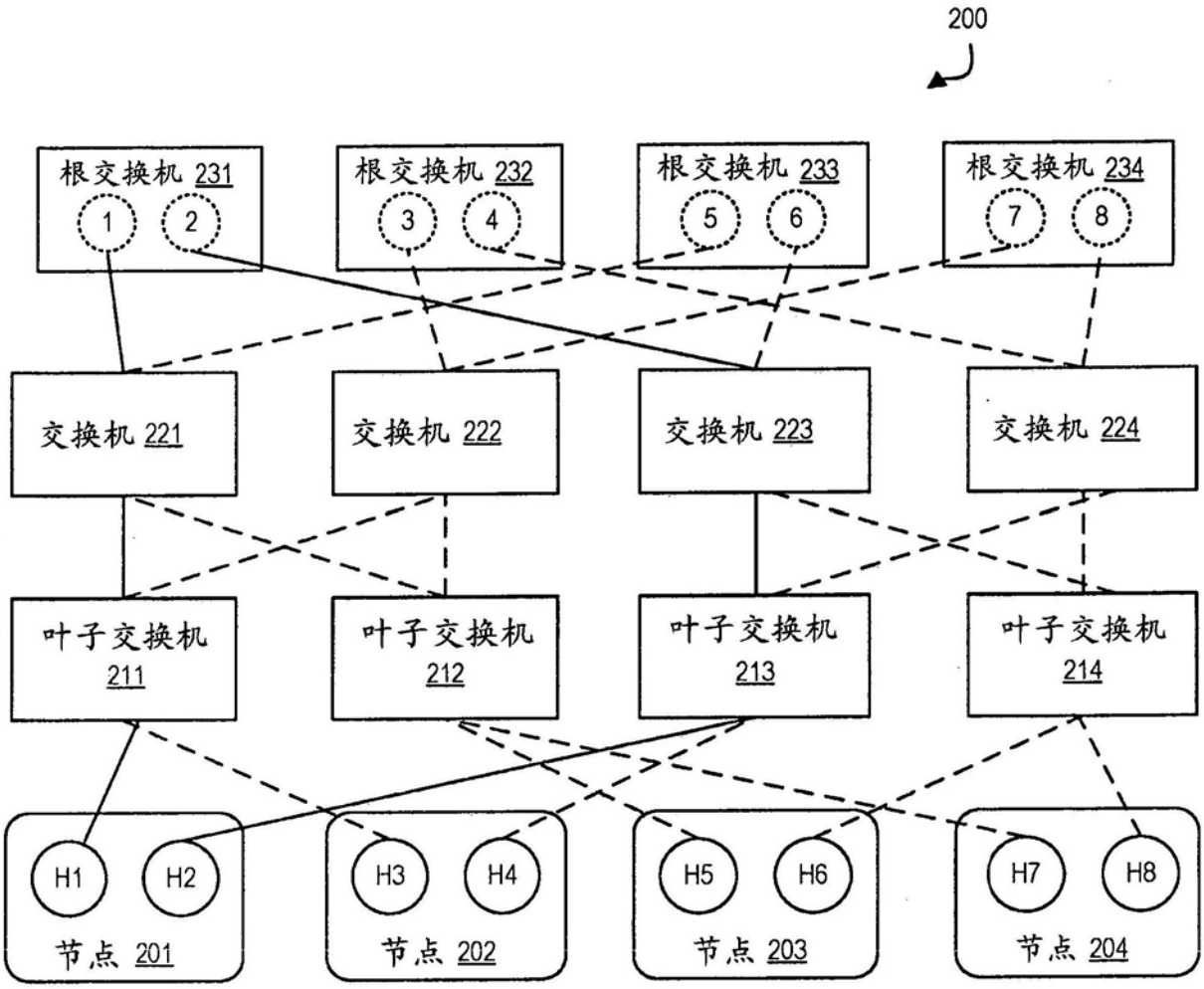

200
根交换机 231
1
2
根交换机 232
3
4
根交换机 233
5
6
根交换机 234
7
8
交换机 221
交换机 222
交换机 223
交换机 224
叶子交换机 211
叶子交换机 212
叶子交换机 213
叶子交换机 214
H1
H2
节点 201
H3
H4
节点 202
H5
H6
节点 203
H7
H8
节点 204

图3

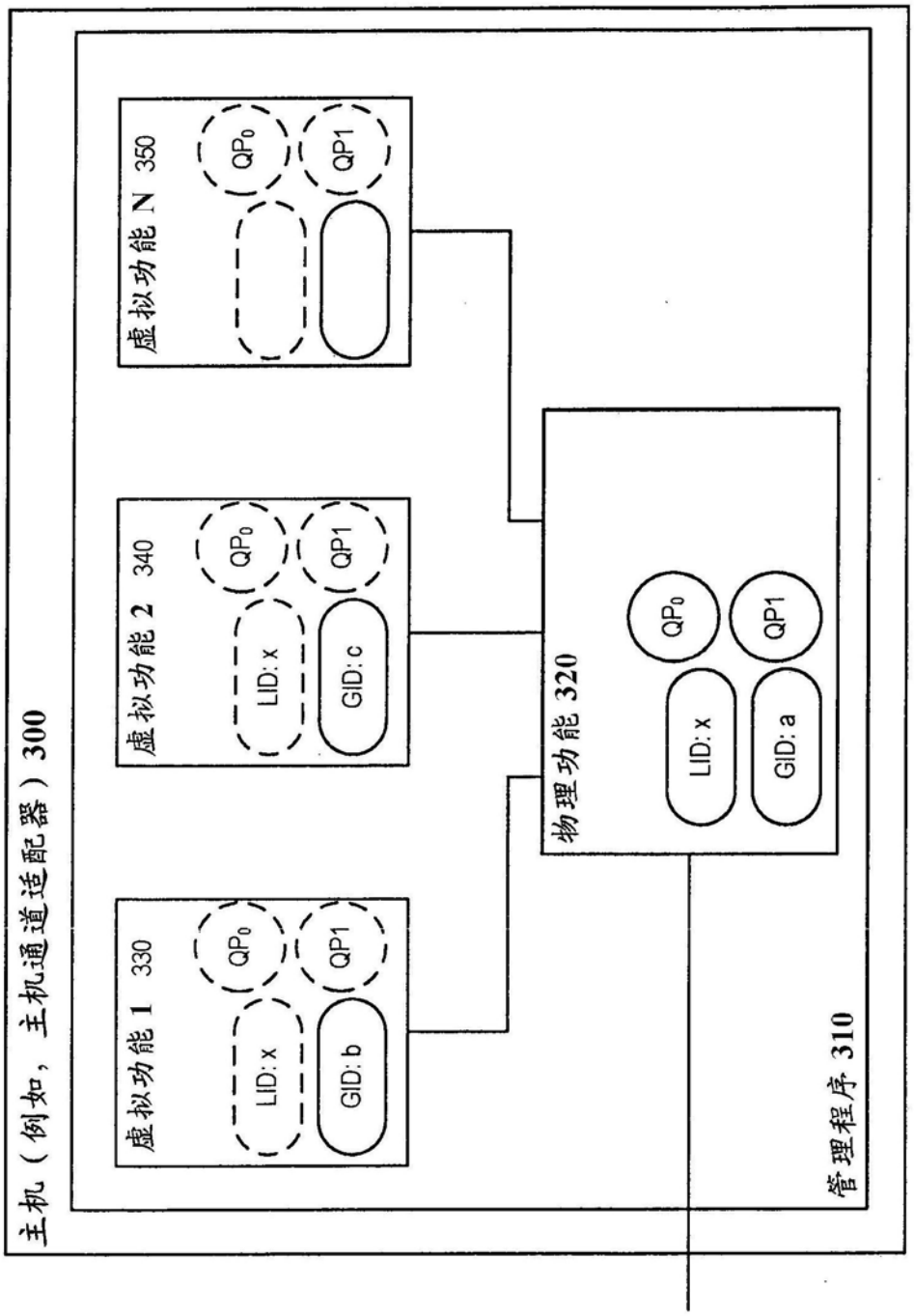

主机（例如，主机通道适配器）300
虚拟功能 1 330
QP0
LID: x
QP1
GID: b
虚拟功能 2 340
QP0
LID: x
QP1
GID: c
虚拟功能 N 350
QP0
QP1
物理功能 320
QP0
LID: x
QP1
GID: a
管理程序 310

图4

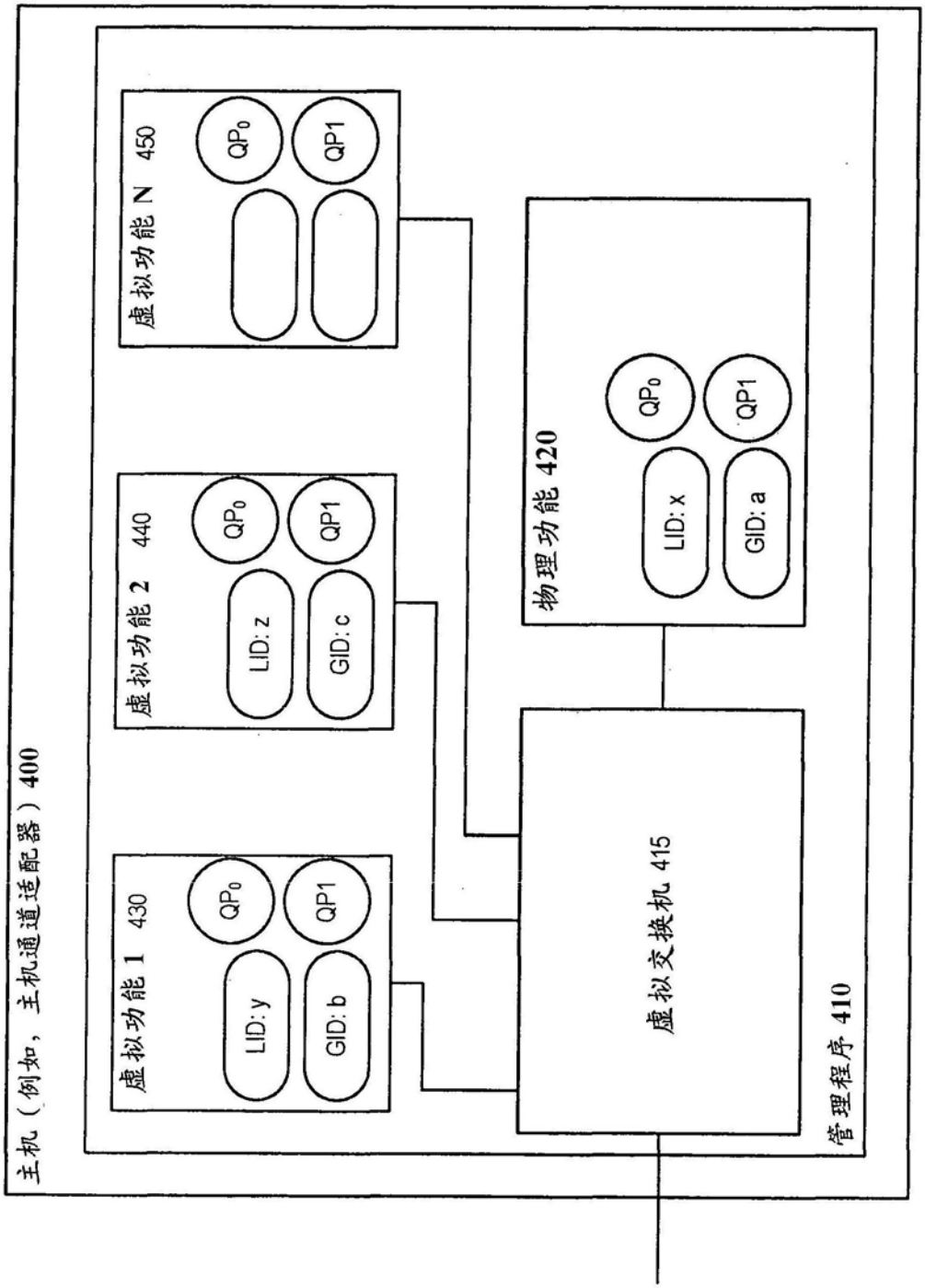

主机（例如，主机通道适配器）400
虚拟功能 1 430
QP0
QP1
LID: y
GID: b
虚拟功能 2 440
QP0
QP1
LID: z
GID: c
虚拟功能 N 450
QP0
QP1
物理功能 420
QP0
QP1
LID: x
GID: a
虚拟交换机 415
管理程序 410

图5

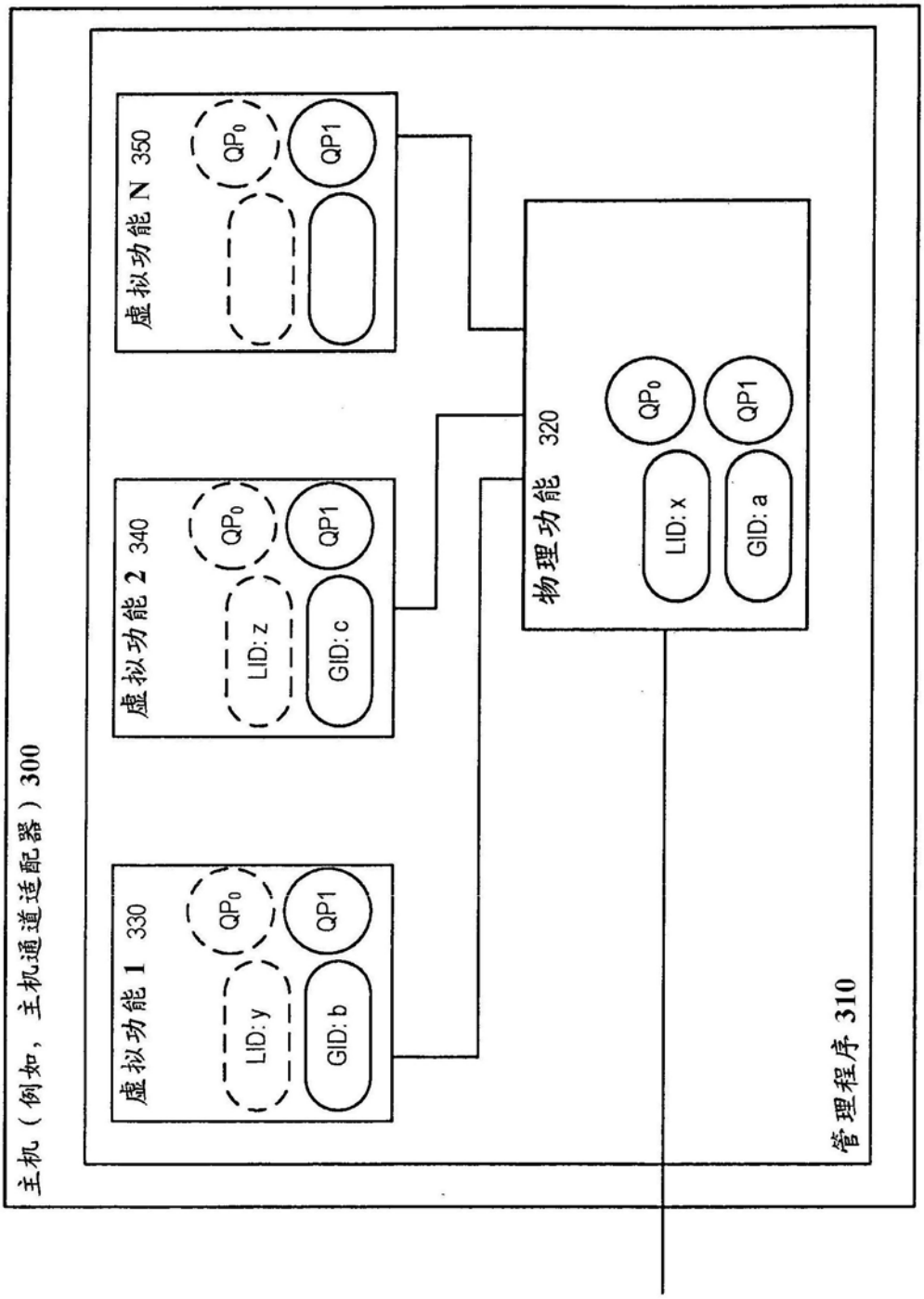

主机（例如，主机通道适配器）300
虚拟功能 1 330
QP0
LID: y
GID: b
QP1
虚拟功能 2 340
QP0
LID: z
GID: c
QP1
虚拟功能 N 350
QP0
QP1
物理功能 320
QP0
LID: x
GID: a
QP1
管理程序 310

图6

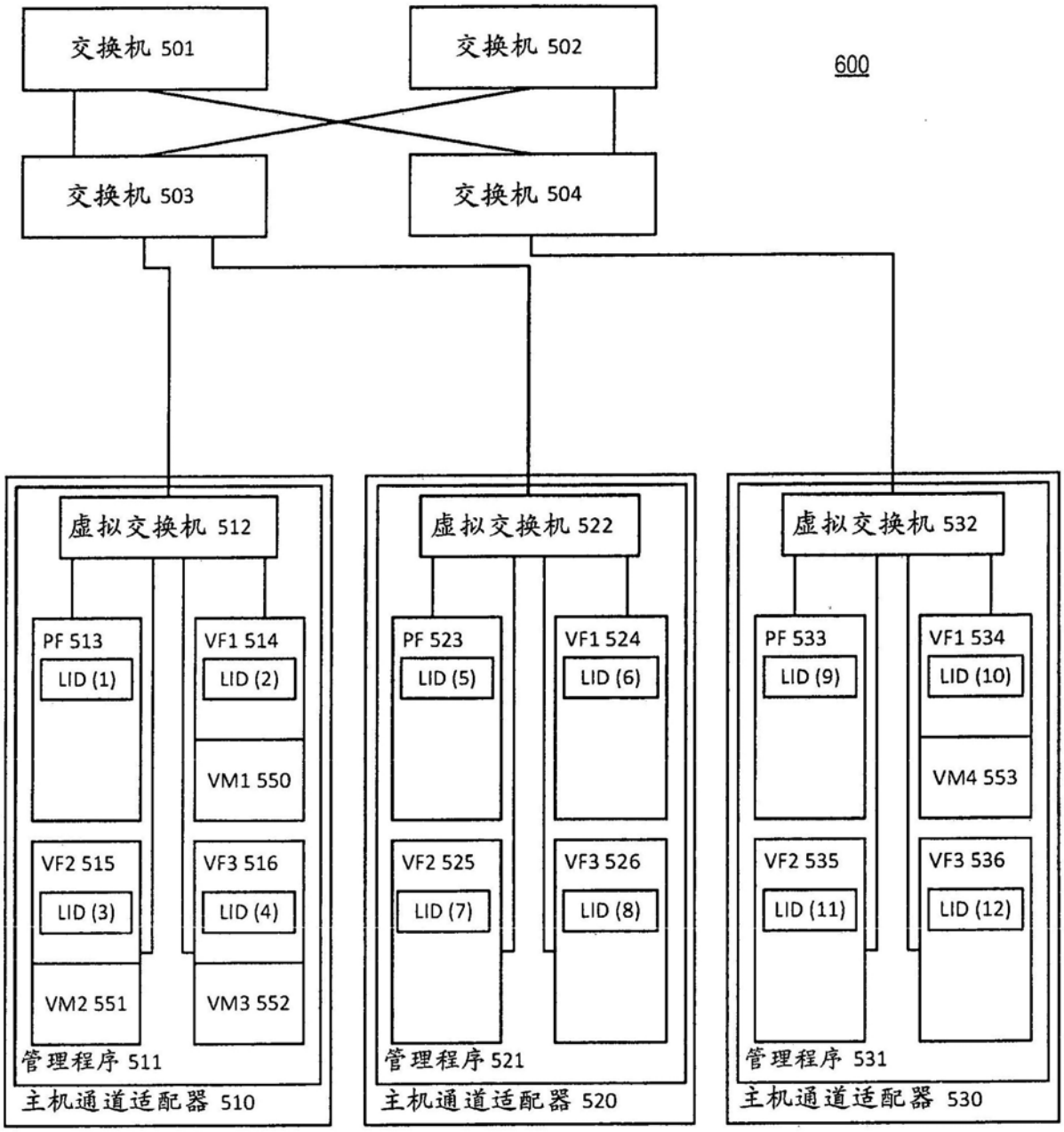

交换机 501
交换机 502
600
交换机 503
交换机 504
虚拟交换机 512
PF 513
LID (1)
VF1 514
LID (2)
VM1 550
VF2 515
LID (3)
VF3 516
LID (4)
VM2 551
VM3 552
管理程序 511
主机通道适配器 510
虚拟交换机 522
PF 523
LID (5)
VF1 524
LID (6)
VF2 525
LID (7)
VF3 526
LID (8)
管理程序 521
主机通道适配器 520
虚拟交换机 532
PF 533
LID (9)
VF1 534
LID (10)
VM4 553
VF2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

图7

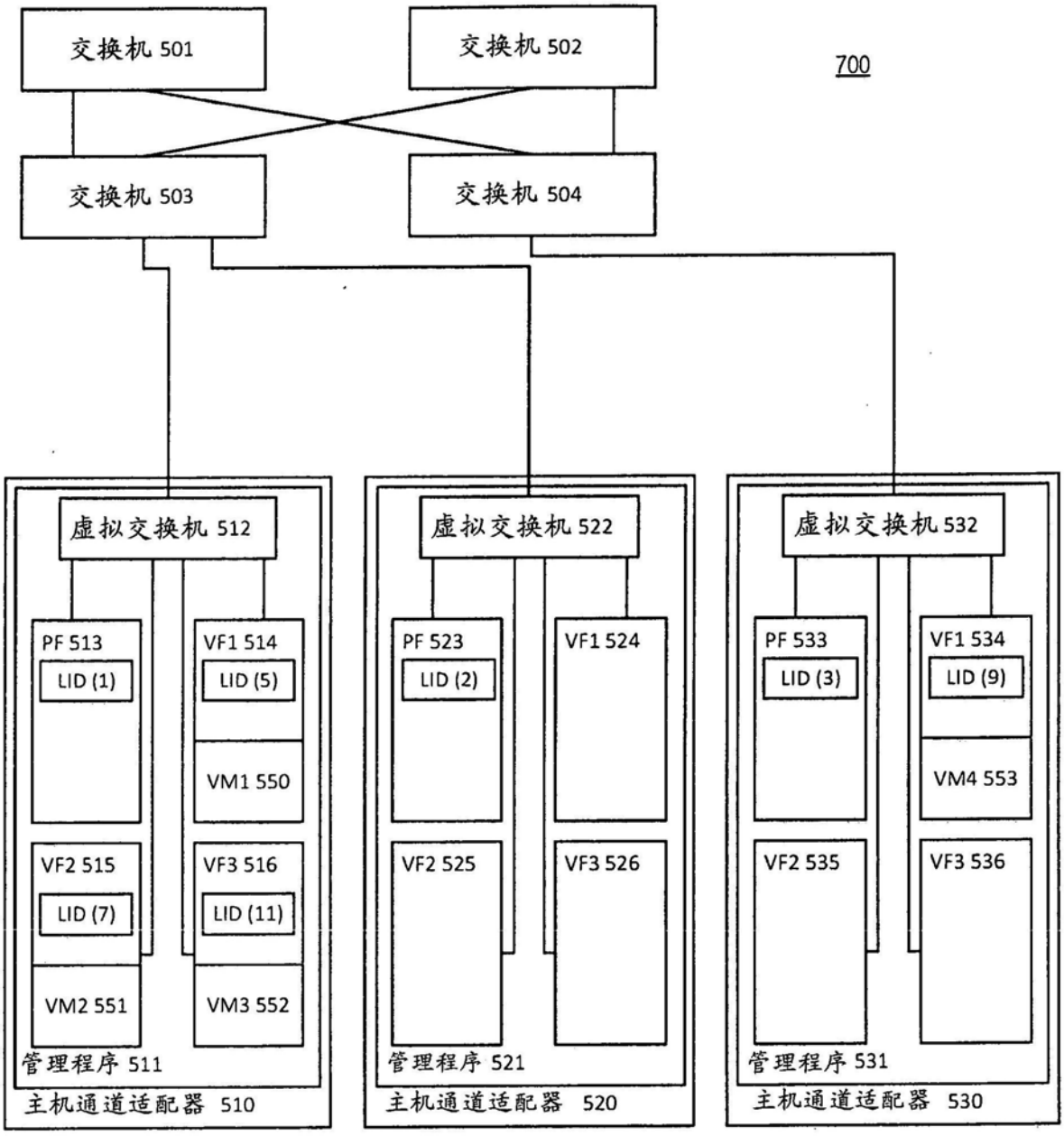

交换机 501
交换机 502
700
交换机 503
交换机 504
虚拟交换机 512
PF 513
LID (1)
VF1 514
LID (5)
VM1 550
VF2 515
LID (7)
VF3 516
LID (11)
VM2 551
VM3 552
管理程序 511
主机通道适配器 510
虚拟交换机 522
PF 523
LID (2)
VF1 524
VF2 525
VF3 526
管理程序 521
主机通道适配器 520
虚拟交换机 532
PF 533
LID (3)
VF1 534
LID (9)
VM4 553
VF2 535
VF3 536
管理程序 531
主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

图8

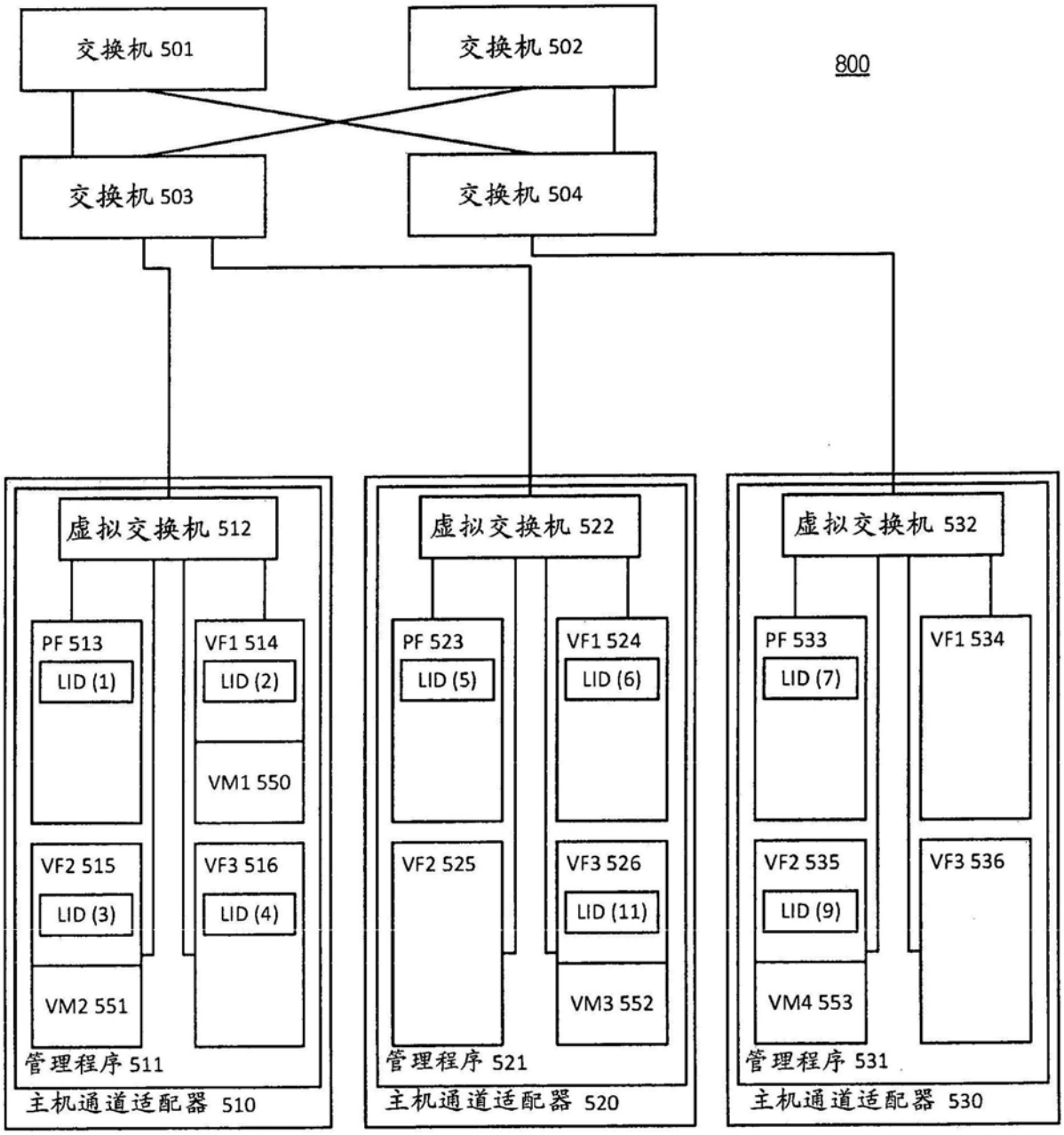

交换机 501
交换机 502
800
交换机 503
交换机 504
虚拟交换机 512
PF 513
LID (1)
VF1 514
LID (2)
VM1 550
VF2 515
LID (3)
VF3 516
LID (4)
VM2 551
管理程序 511
主机通道适配器 510
虚拟交换机 522
PF 523
LID (5)
VF1 524
LID (6)
VF2 525
VF3 526
LID (11)
VM3 552
管理程序 521
主机通道适配器 520
虚拟交换机 532
PF 533
LID (7)
VF1 534
VF2 535
LID (9)
VF3 536
VM4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

图9

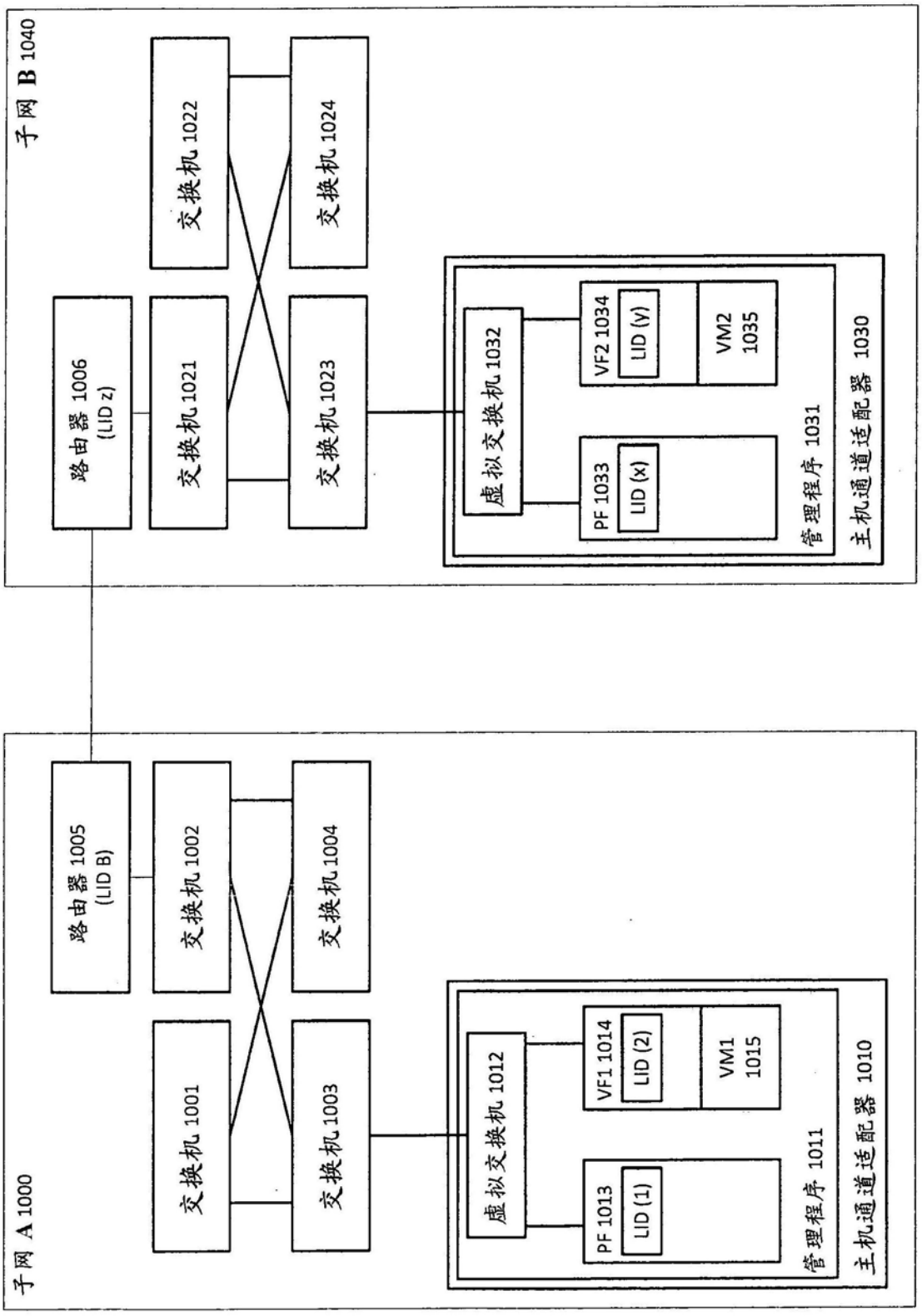

子网 A 1000
路由器 1005
(LID B)
交换机 1001
交换机 1002
交换机 1003
交换机 1004
主机通道适配器 1010
虚拟交换机 1012
PF 1013
LID (1)
VF1 1014
LID (2)
VM1
1015
管理程序 1011
子网 B 1040
路由器 1006
(LID z)
交换机 1021
交换机 1022
交换机 1023
交换机 1024
主机通道适配器 1030
虚拟交换机 1032
PF 1033
LID (x)
VF2 1034
LID (y)
VM2
1035
管理程序 1031

图10

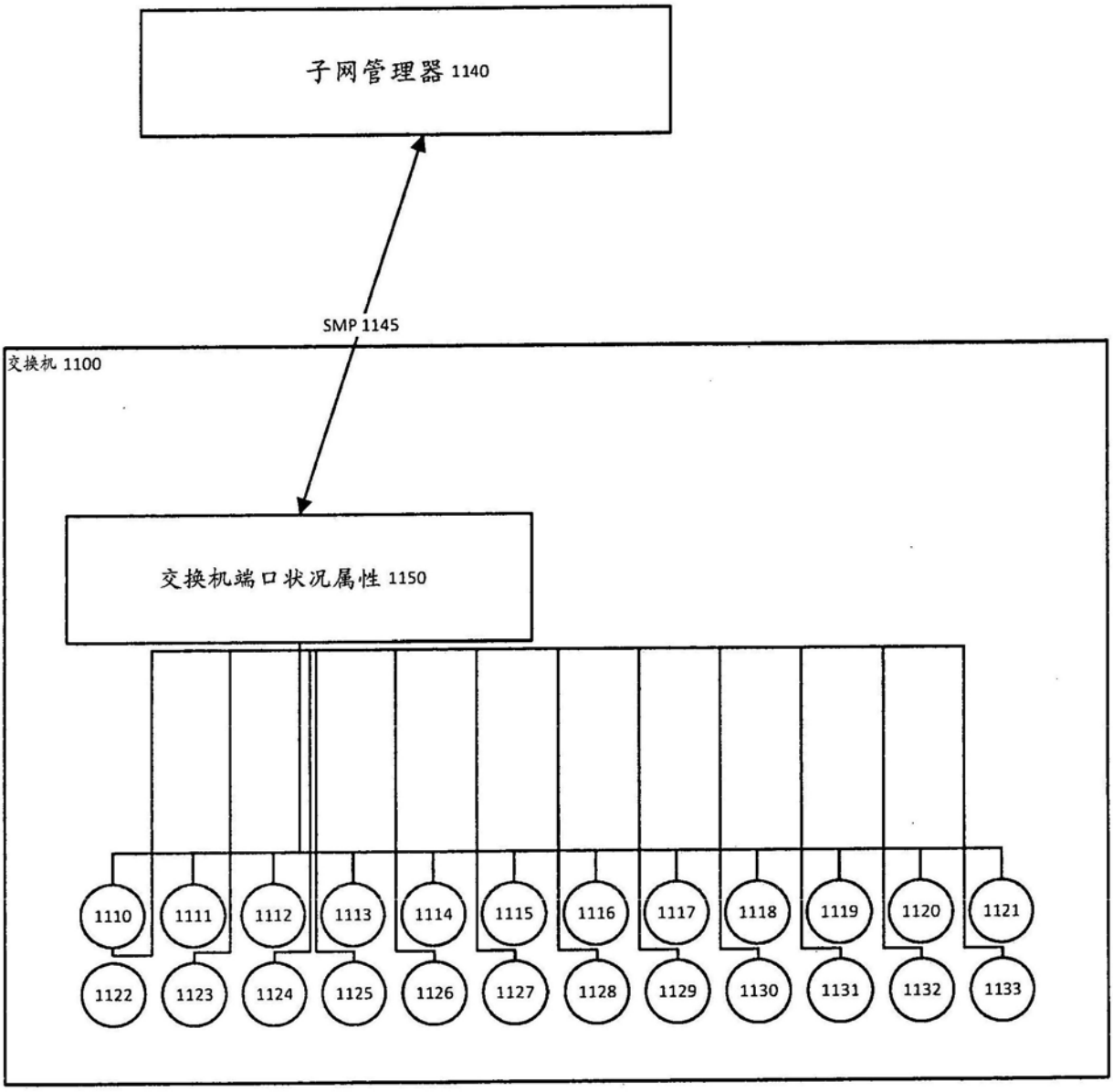

子网管理器 1140
SMP 1145
交换机 1100
交换机端口状况属性 1150
1110
1111
1112
1113
1114
1115
1116
1117
1118
1119
1120
1121
1122
1123
1124
1125
1126
1127
1128
1129
1130
1131
1132
1133

图11

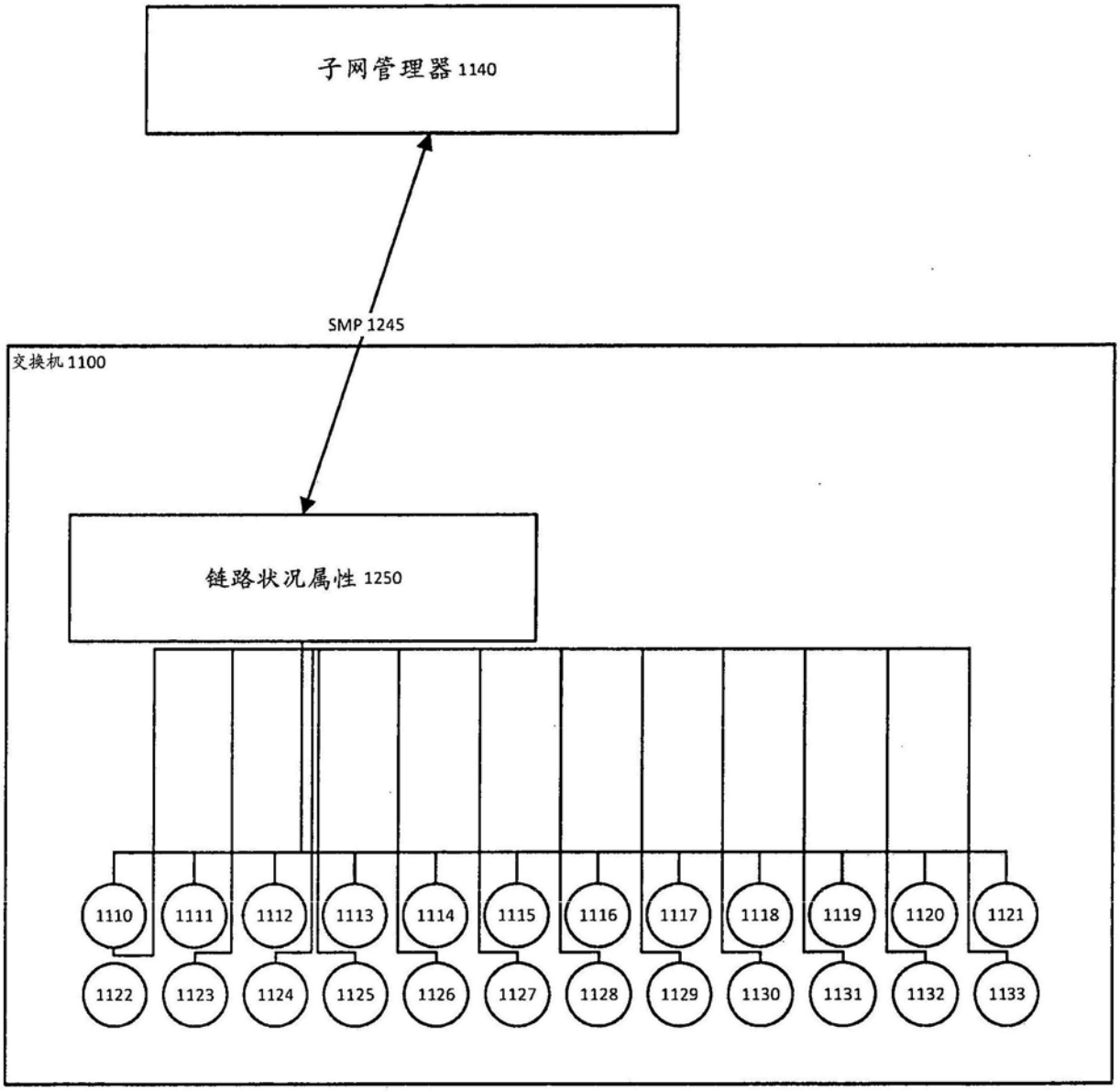

子网管理器 1140
SMP 1245
交换机 1100
链路状况属性 1250
1110
1111
1112
1113
1114
1115
1116
1117
1118
1119
1120
1121
1122
1123
1124
1125
1126
1127
1128
1129
1130
1131
1132
1133

图12

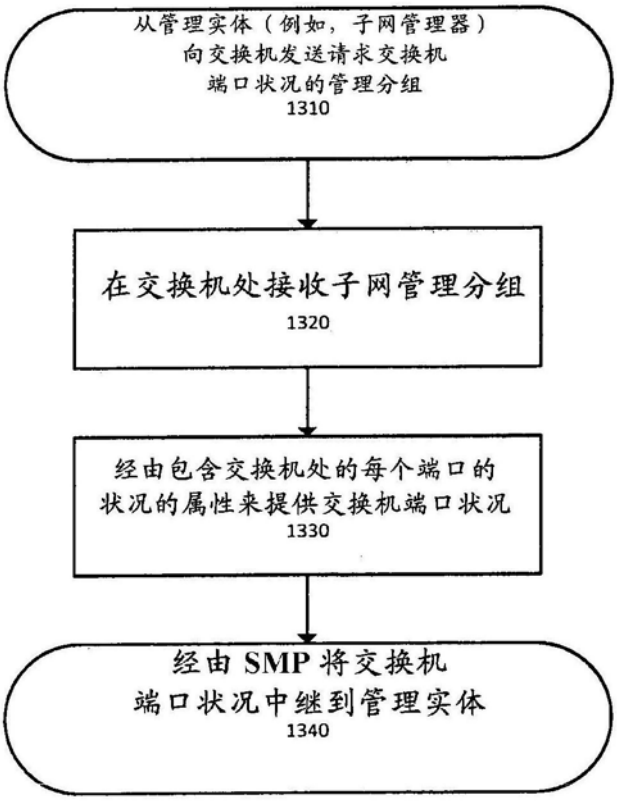

从管理实体（例如，子网管理器）
向交换机发送请求交换机
端口状况的管理分组
1310
在交换机处接收子网管理分组
1320
经由包含交换机处的每个端口的
状况的属性来提供交换机端口状况
1330
经由 SMP 将交换机
端口状况中继到管理实体
1340

图13

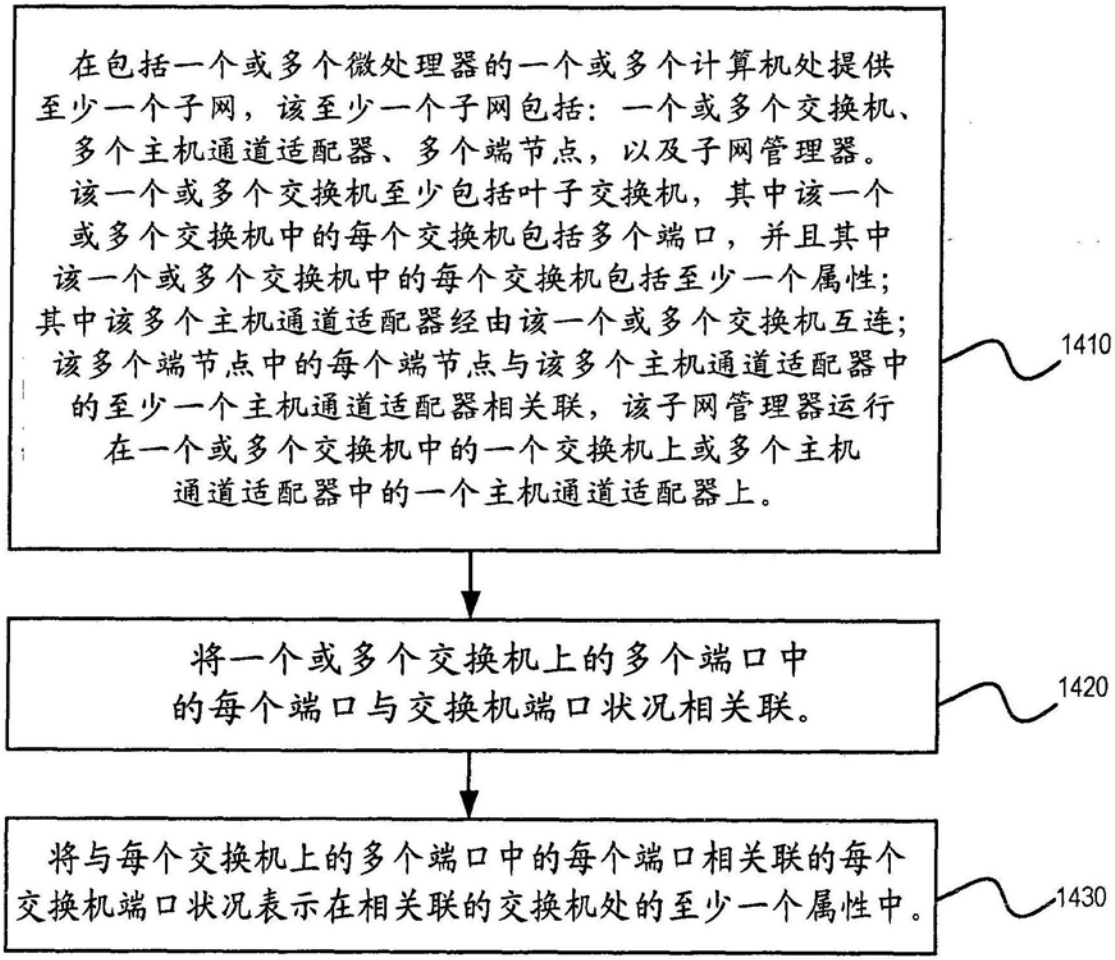

在包括一个或多个微处理器的一个或多个计算机处提供至少一个子网，该至少一个子网包括：一个或多个交换机、多个主机通道适配器、多个端节点，以及子网管理器。该一个或多个交换机至少包括叶子交换机，其中该一个或多个交换机中的每个交换机包括多个端口，并且其中该一个或多个交换机中的每个交换机包括至少一个属性；其中该多个主机通道适配器经由该一个或多个交换机互连；该多个端节点中的每个端节点与该多个主机通道适配器中的至少一个主机通道适配器相关联，该子网管理器运行在一个或多个交换机中的一个交换机上或多个主机通道适配器中的一个主机通道适配器上。
1410
将一个或多个交换机上的多个端口中的每个端口与交换机端口状况相关联。
1420
将与每个交换机上的多个端口中的每个端口相关联的每个交换机端口状况表示在相关联的交换机处的至少一个属性中。
1430

图14

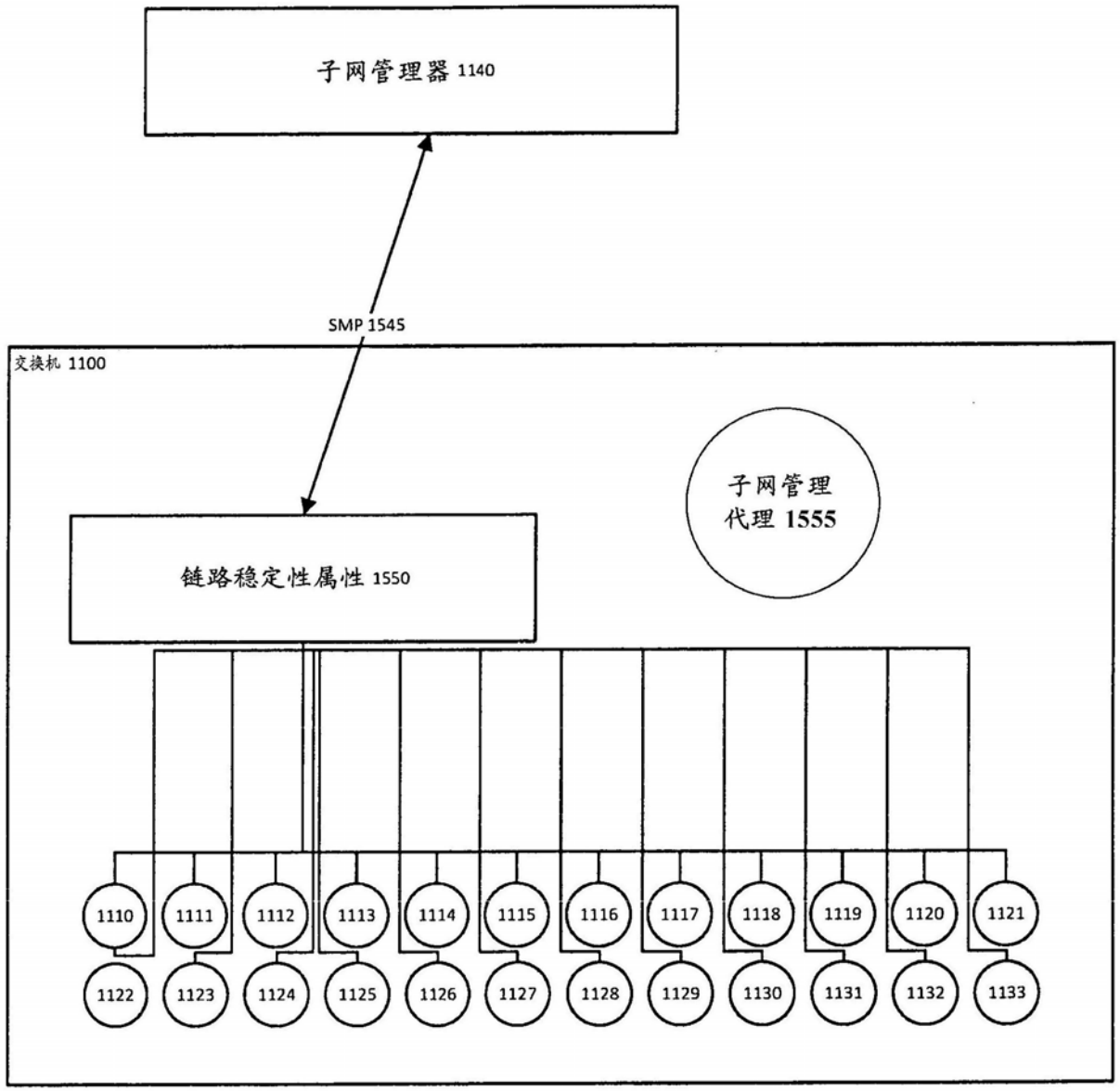

子网管理器 1140
SMP 1545
交换机 1100
子网管理
代理 1555
链路稳定性属性 1550
1110
1111
1112
1113
1114
1115
1116
1117
1118
1119
1120
1121
1122
1123
1124
1125
1126
1127
1128
1129
1130
1131
1132
1133

图15

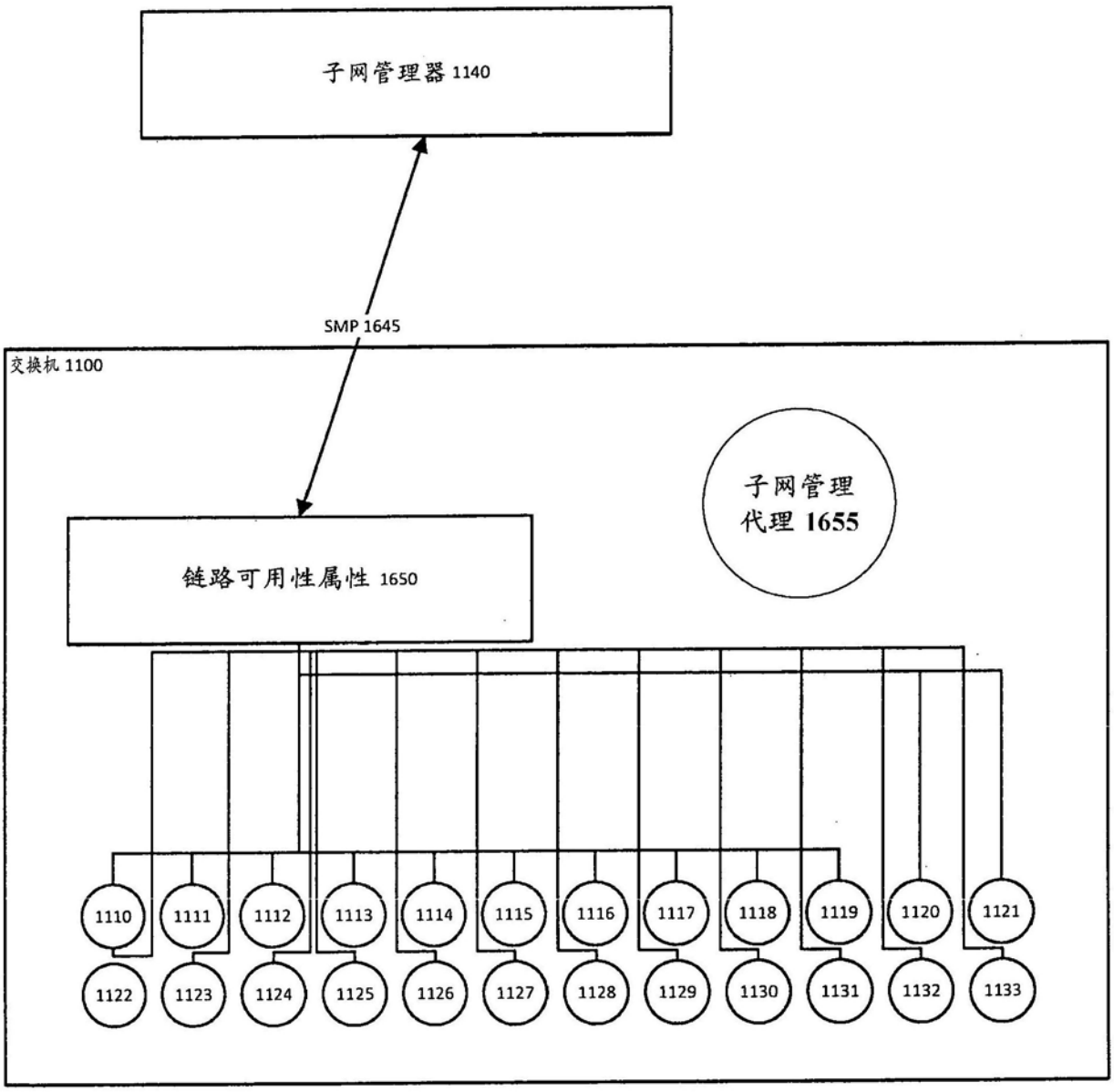

子网管理器 1140
SMP 1645
交换机 1100
子网管理
代理 1655
链路可用性属性 1650
1110
1111
1112
1113
1114
1115
1116
1117
1118
1119
1120
1121
1122
1123
1124
1125
1126
1127
1128
1129
1130
1131
1132
1133

图16

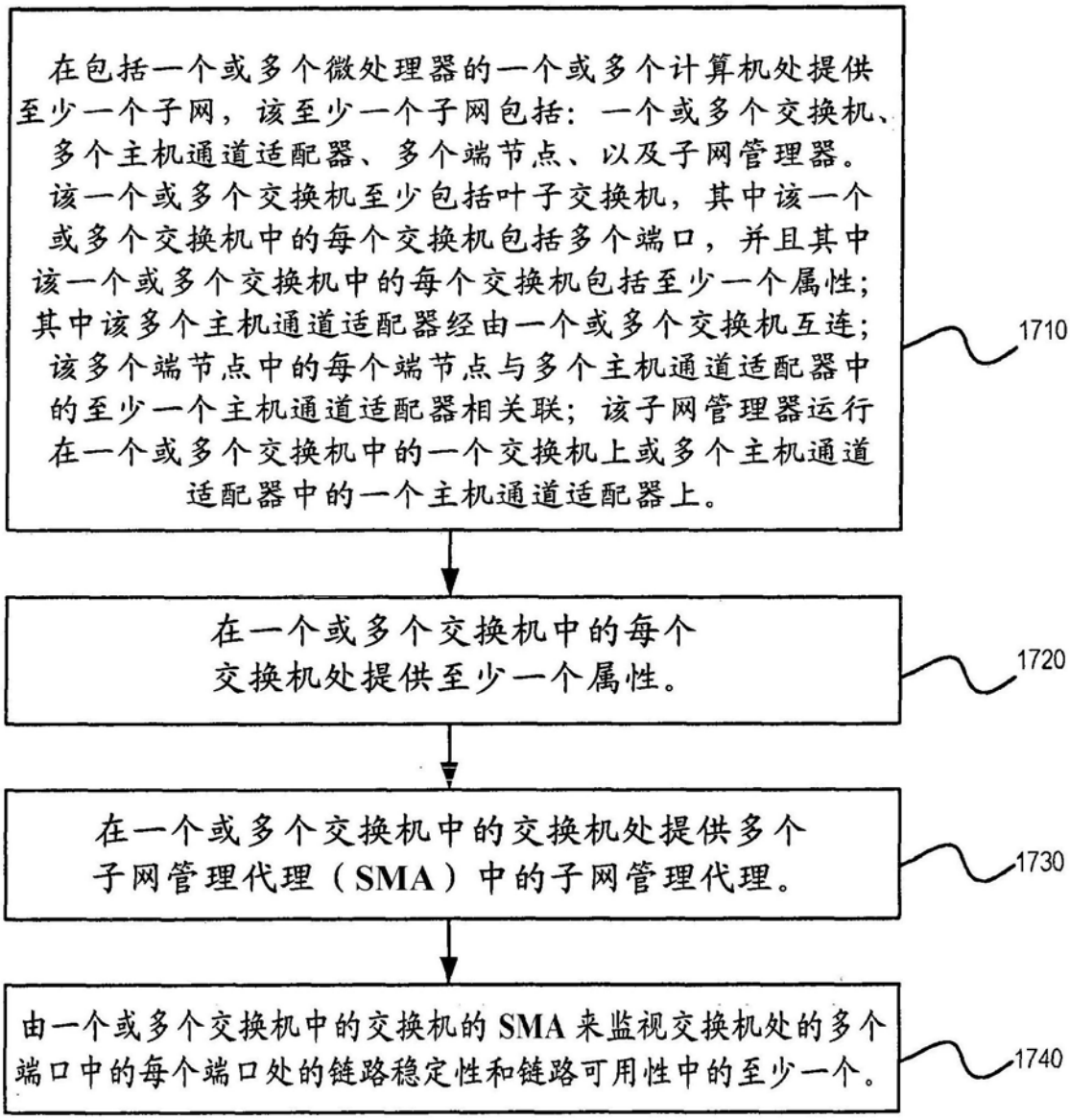

在包括一个或多个微处理器的一个或多个计算机处提供至少一个子网，该至少一个子网包括：一个或多个交换机、多个主机通道适配器、多个端节点、以及子网管理器。该一个或多个交换机至少包括叶子交换机，其中该一个或多个交换机中的每个交换机包括多个端口，并且其中该一个或多个交换机中的每个交换机包括至少一个属性；其中该多个主机通道适配器经由一个或多个交换机互连；该多个端节点中的每个端节点与多个主机通道适配器中的至少一个主机通道适配器相关联；该子网管理器运行在一个或多个交换机中的一个交换机上或多个主机通道适配器中的一个主机通道适配器上。
1710
在一个或多个交换机中的每个交换机处提供至少一个属性。
1720
在一个或多个交换机中的交换机处提供多个子网管理代理（SMA）中的子网管理代理。
1730
由一个或多个交换机中的交换机的SMA来监视交换机处的多个端口中的每个端口处的链路稳定性和链路可用性中的至少一个。
1740

图17