(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 105340230 A
(43)申请公布日 2016.02.17

(21)申请号 201480034586.1
(22)申请日 2014.06.16
(30)优先权数据
13/920,604 2013.06.18 US
(85)PCT国际申请进入国家阶段日
2015.12.17
(86)PCT国际申请的申请数据
PCT/US2014/042521 2014.06.16
(87)PCT国际申请的公布数据
WO2014/204850 EN 2014.12.24
(71)申请人 阿尔卡特朗讯公司
地址 法国布洛涅-比扬古
(72)发明人 P·V·纳卢尔
(74)专利代理机构 北京市中咨律师事务所 11247
代理人 李昱炜 杨晓光
(51)Int.Cl.
H04L 12/931(2006.01)
H04L 12/751(2006.01)
H04L 12/703(2006.01)
H04L 12/713(2006.01)

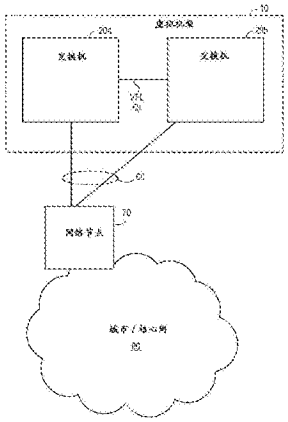
权利要求书2页 说明书8页 附图7页

(54)发明名称

虚拟机架拓扑管理

(57)摘要

通过与VC-ISIS协议相关的定制协议数据单元(PDU)的协议扩展部分内在虚拟机架内的交换机之间传达虚拟机架拓扑信息,实现了虚拟机架内的拓扑管理。虚拟机架拓扑信息可被用于(例如)检测虚拟机架的拓扑,并在虚拟机架的每一个交换机内生成最短路径树以用于虚拟机架内的交换机之间的路由。



CN 105340230 A

1.一种在包括至少两个交换机的虚拟机架内的交换机，该交换机包括：

耦合至虚拟组织链路(VFL)的多个VFL端口，其中，所述VFL将所述虚拟机架内的至少两个交换机的中的每一个相互连接，从而使所述至少两个交换机能够作为一个逻辑交换机操作；

处理器，其用于：

经由所述VFL端口之一从虚拟机架内的额外交换机接收定制协议数据单元(PDU)，

从所述PDU内的协议扩展部分提取出虚拟机架拓扑信息，以及

采用所述虚拟机架拓扑信息检测所述虚拟机架的拓扑结构，并生成供所述虚拟机架内的路由所使用的最短路径树。

2.根据权利要求1所述的交换机，其中，所述定制PDU是定制Hello PDU和定制链路状态PDU中的一个，所述协议扩展部分是所述定制PDU内的类型长度值字段。

3.根据权利要求2所述的交换机，其中，所述定制Hello PDU包括所述类型长度值字段中的一个内的亚秒Hello间隔，其指示从所述额外交换机发送的相继定制Hello PDU之间的亚秒时间间隔。

4.根据权利要求1所述的交换机，其中，所述VFL端口被耦合至各自的VFL链路，每一链路将所述交换机耦合至所述至少两个交换机中的其他交换机中的一个。

5.根据权利要求1所述的交换机，其中，所述处理器还采用所述定制PDU内的所述虚拟机架拓扑信息以及从所述虚拟机架内的其他交换机接收的额外定制PDU内的额外虚拟机架拓扑信息来检测所述虚拟机架内的所述至少两个交换机中的每一个。

6.根据权利要求1所述的交换机，其中，所述处理器还采用所述虚拟机架拓扑信息来从所述至少两个交换机中推选用于所述虚拟机架的主交换机。

7.根据权利要求1所述的交换机，还包括：

维护转发表的存储器；并且

其中，所述处理器将所述最短路径树编程到所述转发表内。

8.根据权利要求1所述的交换机，其中，所述处理器还采用所述虚拟机架拓扑信息来检测所述虚拟机架内的所述至少两个交换机中的一个或多个的故障，并根据检测到的故障生成新的最短路径树。

9.一种非暂态存储设备，其具有有形地体现于其上并且可从其访问可由至少一个处理器解释的指令集，所述指令集被配置为使处理器执行操作以用于：

在交换机中的接收交换机处经由多个虚拟组织链路(VFL)端口之一从虚拟机架内的交换机接收定制协议数据单元(PDU)，所述多个VFL端口被耦合至VFL，其中，所述VFL将所述虚拟机架内的交换机中的每一个相互连接，从而使得所述交换机能够作为一个逻辑交换机操作；

从所述定制PDU内的协议扩展部分提取虚拟机架拓扑信息；以及

采用所述虚拟机架拓扑信息来检测所述虚拟机架的拓扑结构，并生成供所述虚拟机架内的交换机之间的路由所使用的最短路径树。

10.一种用于虚拟机架内的拓扑管理的方法，包括：

在交换机中的接收交换机处经由多个虚拟组织链路(VFL)端口之一从虚拟机架内的交换机接收定制协议数据单元(PDU)，所述多个VFL端口被耦合至VFL，其中，所述VFL将所

述虚拟机架内的交换机中的每一个相互连接，从而使得所述交换机能够作为一个逻辑交换机操作；

从所述定制 PDU 内的协议扩展部分提取虚拟机架拓扑信息；以及

采用所述虚拟机架拓扑信息来检测所述虚拟机架的拓扑结构，并生成供所述虚拟机架内的交换机之间的路由所使用的最短路径树。

虚拟机架拓扑管理

技术领域

[0001] 本发明总体上涉及数据网，尤其涉及虚拟机架架构。

背景技术

[0002] 数据网络允许许多不同的计算设备，例如，个人电脑、IP 电话设备或服务器相互通信和 / 或与附接至网络的各种其他网络单元或远程服务器通信。例如，数据网络可以包括但不限于城市以太网或企业以太网，其支持包括（例如）IP 语音 (VoIP)、数据和视频应用在内的多个应用。这样的网络按常规包括很多用于通过网络路由流量的互连节点，一般将其称为交换机或路由器。

[0003] 往往在各种节点在网络的特定区域内的位置的基础上对它们进行区分，通常根据网络的规模用两个或三个“级”或“层”来表征所述位置。就常规而言，三级网络由边缘层、汇聚层和核心层构成（而两级网络则仅由边缘层和核心层构成）。数据网络的边缘层包括边缘（又称为接入）网络，其通常提供从诸如局域网的企业网或本地网与城市或核心网的连接。边缘 / 接入层是网络的入口点，即，客户网络名义上附接至该层，驻留在边缘层内的交换机被称为边缘节点。不同类型的边缘网络包括数字用户线路、光纤同轴电缆混合网 (hybrid fiber coax)(HFC)、光纤到户和企业网，例如，校园网和数据中心网。边缘节点可以为附接设备执行（例如）L2 交换功能。边缘节点通常被连接至一个或多个企业交换机和 / 或客户网络中的终端设备，并且还连接至终接来自多个边缘节点的接入链路的汇聚层。驻留在汇聚层内的交换机被称为汇聚交换机。汇聚交换机可以执行（例如）经由汇聚链路从边缘节点接收的流量的 L2 交换和 L3 路由。汇聚层被连接至城市或核心网络层，其执行对从（三级网络中的）汇聚交换机接收的或者从（二级网络中的）边缘节点接收的流量执行 L3 或 IP 路由。可以认识到，在网络的每一递增层中的节点通常具有更大的容量和更快的吞吐量。

[0004] 数据网所面临的关键困难之一是需要网络弹性，即，即使可能出现部件故障、链路故障等也能够保持高可用性的能力，这一点对于提供令人满意的网络性能至关重要。网络弹性可以部分地通过拓扑冗余，即，通过提供冗余节点（和冗余节点内的冗余部件）和节点之间的多重物理通路，以及部分地通过 L2/L3 协议利用所述冗余从而在发生故障时收敛到备选路径上以用于交换 / 路由通过网络的业务流来实现。

[0005] 在一种已知的解决方案中，虚拟机架被用于提供冗余，其同时还提供提高的吞吐量和带宽。在虚拟机架内，两个或更多物理以太网交换机可被耦合到一起，从而借助于统一控制面及配置文件形成作为单个交换机 / 路由器操作的单逻辑形式因子。通常通过建立和维护同步转发表以及在运行于交换机上的对应 / 对等应用之间交换控制信息来由虚拟机架主交换机管理路由和交换引擎冗余。因而，邻居发现、最佳转发路径、故障探测和恢复必须在虚拟机架内全部得到支持。

附图说明

[0006] 图 1 示出了虚拟机架的实施例的示意性方框图；

[0007] 图 2 示出了虚拟机架的另一实施例的示意性方框图；

[0008] 图 3 示出了虚拟机架内的交换机的实施例的示意性方框图；

[0009] 图 4 和图 5 示出了定制 Hello 协议数据单元 (PDU) 的示范性格式的实施例；

[0010] 图 6 和图 7 是定制链路状态 PDU 的示范性格式的实施例；

[0011] 图 8 示出了用于虚拟机架内的拓扑管理的方法的实施例的示范性流程图。

具体实施方式

[0012] 图 1 示出了根据本发明的虚拟机架 10 的实施例。虚拟机架 10 包括两个或更多以太网交换机 20a、20b，它们一起形成了一个逻辑交换机。虚拟机架 10 具有外部节点用以向虚拟机架 10 转发流量的媒体访问控制 (MAC) 地址和 Internet(IP) 协议地址。虚拟机架 10 内的每一交换机 20a、20b 还分配有用于在交换机 20a、20b 之间进行路由的唯一标识符（即，用于驻留在交换机上的软件部件之间的通信的 IP 地址或其他内部标识符）。

[0013] 经由虚拟组织链路 (VFL)50 将以太网交换机 20a、20b 耦合到一起。VFL50 为交换机 20a、20b 之间的信息交换提供连接，所述信息交换与流量转发、MAC 寻址、多播流、地址解析协议 (ARP) 表、层 2 控制协议（例如，生成树、以太网环路防止、逻辑链路探测协议）、路由协议（例如，RIP、OSPF、BGP) 以及将虚拟机架 10 连接至其他上游 / 下游节点的链路的状态有关。在每一个交换机 20a、20b 内维护用于外部节点的 MAC 地址 / 转发表，从而能够使交换机 20 之间的桥接或路由包能够抵达外部目的地设备。例如，在包将从虚拟机架 10 内的一个交换机（例如，交换机 20a) 被路由至另一交换机（例如，交换机 20b) 以用于传输给外部目的地设备时，将预先计划的报头加到所述包上，该报头包括源交换机 20a 的标识符和目的地交换机 20b 的标识符。

[0014] 交换机 20a、20b 是单独的物理交换机，每一个都可作为独立的交换机操作。可以将交换机 20a、20b 一起包装到单个物理机架内或者包装到两个或更多单独的物理机架内。根据机架配置，交换机 20a 和 20b 可以处于同一地理区域内，例如，中央局或数据中心，或者可以处于单独的地理位置上，例如，不同的建筑物或城市，以提供地理多样性。

[0015] 此外，虚拟机架 10 内的交换机 20a 和 20b 可以是汇聚交换机、边缘交换机或企业交换机。在交换机 20a、20b 是企业交换机的实施例中，将虚拟机架 10 向下游连接至局域网 (LAN) 内的一个或多个终端设备，向上游连接至一个或多个边缘交换机。在交换机 20a、20b 是边缘交换机的实施例中，将虚拟机架 10 向下游连接至一个或多个企业交换机 LAN 和 / 或本地网内的终端设备，向上游连接至一个或多个汇聚交换机或网络节点，例如，城市 / 核心网 80 内的网络交换机和 / 或路由器。在交换机 20a、20b 是汇聚交换机的实施例中，将虚拟机架 10 向下游连接至一个或多个边缘交换机，向上游连接至城市 / 核心网 80 内的一个或多个网络节点。在图 1 中，虚拟机架 10 代表耦合至城市 / 核心网 80 内的一个或多个网络节点 70 的边缘或汇聚交换机。

[0016] 在实施例中，虚拟机架 10 和网络节点 70 之间的连接是由多机架链路集群 (MC-LAC)60 形成的，在所述集群中，两个或更多物理链路将网络节点 70 与虚拟机架 10 内的交换机 20a、20b 中的两者或更多连接，如 2011 年 1 月 20 日提交的发明名称为“System and Method for Multi-Chassis Link Aggregation”的美国专利申请 13/010169 中所述，

通过引用将其并入本文。例如，如图 1 所示，各个外部物理链路将交换机 20a、20b 的每一个连接至城市 / 核心网 80 内的网络节点 70 以形成 MC-LAG 60。在示范性实施例中，虚拟机架 10 和 / 或网络节点 70 可以采用负载均衡技术分配跨越 MC-LAG 60 的所有可用链路的流量。例如，对于通过 MC-LAG 60 传输的每一个包而言，基于涉及在源及目的地 Internet 协议 (IP) 或者媒体存取控制 (MAC) 地址信息上运行的散列函数的负载均衡算法来选择物理链路之一。

[0017] 在另一实施例中，可以采用标准链路集群 (LAG) 或者其他干线或链路将交换机 20a、20b 连接至上游和 / 或下游节点。应当理解，文中采用的词语“LAG”是指采用链路集合控制协议 (LACP) 的在两个节点之间的多个物理链路的捆绑，以在其间形成单逻辑信道，所述链路集合控制协议 (LACP) 在 2008 年 11 月 3 日发布的 IEEE 802.1AX-IEEE 802.3ad 中定义。

[0018] 不管虚拟机架 10 与上流和 / 或下游节点之间的连接的类型如何，交换机 20a、20b 的操作对于上游和下游节点而言都是透明的，上游和下游节点将它们作为一个逻辑设备 (虚拟机架 10) 来对待。因此，上流和下游节点能够活跃地向虚拟机架 10 转发流量，同时交换机 20a、20b 之间的 MAC 地址表和其他转发信息的同步由 VFL 50 上的 L2 包流和控制消息发送驱动。

[0019] 在一个实施例中，交换机 20a 和 20b 在活跃 / 不活跃环境内操作，其中，并非所有的外部链路都同时活跃地转发流量 (即，一个交换机上的外部链路是活跃地，而其他交换机上的外部链路则保持不活跃或者“待命”)。在这一实施例中，可以采用生成树协议 (STP) 使备用通路脱离待命模式进入活跃状态，从而在活跃链路发生故障时重新建立连接。在另一实施例中，交换机 20a 和 20b 在活跃 / 活跃环境内操作，其中，所有外部连接都同时活跃 (即，所有交换机上的外部链路都是活跃的)。在这一实施例中，STP 可以不必在网络拓扑结构的一些或所有部分中运行以用于环路防止 (例如，可以在 VFL 50 之上以及在将虚拟机架 10 连接至网络 80 内的上游 / 核心交换机的链路之上仍然采用 STP)。

[0020] 根据不同的实施例，虚拟机架 10 内的交换机 20a、20b 采用唯一的虚拟机架中间系统到中间系统协议 (VC-ISIS) 以用于 VFL 50 上的通信。VC-ISIS 协议使得交换机 20a 和 20b 能够交换用于机架 10 内的拓扑管理的系统特异性信息 (下文称为拓扑信息)。例如，可以将拓扑信息用于虚拟机架 10 内的邻居探测、拓扑公告、最佳转发路径确定 (最短路径桥接)、故障探测和故障恢复。作为例子而非限制，所述拓扑信息可以包括虚拟机架应用、交换机硬件和系统软件性能参数所特有的信息。

[0021] VC-ISIS 协议定义了各种可以添加到现有的 ISIS 协议上的定制协议扩展部分，如 ISo/IEC 10589:2002 中所定义的，并且将其称为 VC-ISIS，以区别于现有 ISIS 实现的协议 (例如，用于 IP 的 ISIS 和 ISIS-SPB)。VC-ISIS 协议的定制协议扩展部分携带着拓扑信息，其能够以 (例如) 类型 / 长度 / 值 (TLV) 字段形式实施，以定义 VC-ISIS 协议的定制协议数据单元 (PDU)，例如，定制 Hello PDU 和 / 或定制链路状态 PDU。定制 PUD 有助于虚拟机架 10 内的交换机 20a 和 20b 的自动探测和配置以及以虚拟机架 10 内的每一个交换机 20a 和 20b 为根的最短路径桥接树的生成。此外，交换机 20a 和 20b 之间的定制 PDU 的周期性交换能够为故障恢复过程中的最短路径树的重新计算获得快速的最佳收敛时间。例如，在一个实施例中，按照亚秒时间间隔在交换机 20a 和 20b 之间交换定制 Hello PDU，从而议定

快速的故障探测。所交换的拓扑信息也可被采用以促进虚拟机架内的主交换机的推选。

[0022] 所述 VC-ISIS 协议还是规模可调的，因而不管虚拟机架 10 内的交换机 20a 和 20b 的数量有多少，都能够就转发路径收敛提供鲁棒的、可接受的性能。例如，如图 2 所示，采用一起形成了 VFL 50 的各 VFL 链路 40 将六个交换机 20a-20f 按照网状拓扑结构耦合到一起。可以认识到，随着虚拟机架 10 内的交换机 20a 和 20b 的数量的提高，拓扑探测和最短路径桥接变得越来越关键。VC-ISIS 协议（具有协议扩展部分）能够实现并优化任何类型的虚拟机架拓扑结构中的拓扑管理。因而，VC-ISIS 协议提供了有效率并且可靠的用于虚拟机架拓扑管理的控制面功能，而无需用户 / 管理员方面的外来干预或可视性。

[0023] 图 3 示出了虚拟机架内的交换机 20 的示范性实施例。交换机 20 包括一个或多个虚拟组织链路 (VFL) 端口 30a-30c 以及一个或多个外部端口 35a-35c。VFL 端口 30a-30c 提供与形成 VFL 的链路的连接。外部端口 35a-35c 提供与通往外部上游和 / 或下游节点的链路的连接。外部端口 35a-35c 中的一个或多个可以包括用于 MC-LAG 物理链路、LAG 或其他干线组、固定链路等的成员端口。VFL 端口 30a-30c 和外部端口 35a-35c 可以具有相同的物理接口类型，例如铜端口 (CAT-5E/CAT-6)、多模光纤端口 (SX) 或单模光纤端口 (LX)。在另一实施例中，VFL 端口 30a-30c 和外部端口 35a-35c 可以具有一个或多个不同的物理接口类型。

[0024] 交换机 20 还包括处理器 22、机架管理模块 (CMM)23（其既可以包括主用 CMM，又可以包括备用 CMM）、虚拟机架 (VC) 拓扑引擎 24、交换组织 25 和非暂态存储设备 26。VC 拓扑引擎 24 包括可由处理器 38 解释和运行的算法（或指令集），从而使处理器 38 执行用于交换机 20 内的虚拟机架拓扑管理的操作，例如，邻居发现、拓扑公告、最短路径桥接、故障探测和恢复。此外，CMM 23 还包括可由处理器 38 解释和运行的算法（或指令集），从而使处理器执行用于管理交换桥（机架）的操作。可以将所述 VC 拓扑引擎 24 和 CMM 23 存储到（例如）非暂态存储设备 26 内或者交换机 20 的另一非暂态存储设备内。

[0025] 一般将文中采用的“处理器”一词理解为驱动通用计算机的设备。作为例子而非限制，“处理器”38 可以包括微处理器、微控制器、中央处理单元 (CPU)、现场可编程门阵列 (FPGA)、专用集成电路 (ASIC) 或任何其他处理设备中的一个或多个。此外，一般应当将文中采用的“非暂态存储设备”一词理解为包括用于存储供通用计算机采用的数据和 / 或程序的设备。作为例子而非限制，“非暂态存储设备”39 可以包括数据存储器、随机存取存储器 (RAM)、只读存储器 (ROM)、闪速存储器、光盘、ZIPTM 驱动器、磁带驱动器、数据库或其他类型的存储设备或存储介质中的一个或多个。

[0026] VC 拓扑引擎 24 与 CMM 23 结合使得虚拟机架内的其他交换机的探测以及交换机 20 与虚拟机架内的其他交换机之间的路由的最短路径树 (SPT)28 的生成自动化。CMM 23 与虚拟机架内的其他交换机 (VC 交换机）上的 CMM 建立逻辑进程间通信 (IPC)，从而与其他 VC 交换机交换 VC-ISIS 协议数据单元 300。VC 拓扑引擎 24 生成传输给其他 VC 交换机的 VC-ISIS PDU 300，并对经由 VFL 端口 30a-30c 中的一个或多个从其他 VC 交换机接收的 VC-ISIS PUD 300 进行处理。应当指出，VC 拓扑引擎 24 可以作为 CMM 23 的部分运行或者独立于 CMM 23（与之并列）运行。

[0027] VC 拓扑引擎 24 从接收到的 VC-ISIS PDU 300 提取拓扑信息 320，并通过汇聚从每一个 VC 交换机接收的拓扑信息 320 建立虚拟机架的拓扑表示（地图）。这一地图指示

（例如）每一个 VFL 端口 30a-30c 能够到达的 VC 交换机。基于该地图以及其他拓扑信息 320（例如，链路状态信息），VC 拓扑引擎 24 建立 SPT 28，其指示至虚拟机架内的具体交换机的最低成本（最短）路径以用于转发流量。例如，VC 拓扑引擎 24 能够计算下一跳信息以及等成本路径集，从而建立可以用于（例如）负载均衡的相邻性集 (adjacency set)。

[0028] CMM 23 利用 SPT 28 连同接收到的 VC-ISIS PDU 300 以及其他 PDU（即在虚拟机架之外生成的 PDU）更新在存储设备 26 内维护的 MAC/HDI 转发表 27。MAC/HDI 转发表 27 包括 MAC 地址条目的列表，所述 MAC 地址条目例如其他 VC 交换机、交换机 20 和其他 VC 交换机内的软件、交换机 20 和其他 VC 交换机内的硬件以及外部（上游或下游）设备的 MAC 地址。MAC 地址条目包括在对包进行桥接或路由使之抵达具有相关 MAC 地址的设备的过程中采用的相关硬件设备信息 (HDI)。目的地硬件设备信息包括例如与目的地 MAC 地址相关的交换机 20 或另一 VC 交换机的端口标识符。MAC/HDI 转发表 27 可以包括一个或多个表格，例如，源干线图、干线位图表、干线组表、VLAN 映射表等。此外，VC 拓扑引擎 24 可以将 SPT 28 编程到 MAC/HDI 转发表 27 内。

[0029] 在示范性操作中，VC 拓扑引擎 24 可由处理器 22 执行，从而经由 VFL 端口 30a-30c 中的一个或多个与虚拟机架内的其他交换机通信，以运行发现虚拟机架内的每一个交换机以及每一个交换机的各种属性（例如，每一个交换机的标识符、交换机的 MAC 地址、交换机优先级、与每一个交换机相关的 VLAN 等）以生成 SPT 28 的拓扑发现过程。例如，VC 拓扑引擎 24 能够对在一个或多个 VFL 端口 30a-30c 上从虚拟机架内的一个或多个其他交换机接收到的 VC-ISIS 协议数据单元 (PDU) 进行处理，其方式为从 PDU 300 提取拓扑信息 320 并由拓扑信息 320 生成 SPT 28。CMM 23 还可由处理器 22 执行，从而基于 SPT 28 和各种接收到的 PDU（例如，VC-ISIS PDU 300 和 / 或外部 PDU）创建或更新 MAC/HDI 转发表 27。例如，VC 拓扑引擎 24 能够向 CMM 23 提供从接收到的 VC-ISIS PDU 300 中提取的拓扑信息 320 以用于更新 MAC/HDI 转发表 27。之后，处理器 22 能够采用 SPT 28 和 / 或 MAC/HDI 转发表 27 经由交换组织 25 将输入流量（抵达 VFL 端口 30a-30c 或外部端口 35a-35c 的）交换至交换机 20 上的其他端口 30a-30c 或 35a-35c。

[0030] 在另一实施例中，VC 拓扑引擎 24 与 CMM 23 结合进一步自动化在交换机 20 和 / 或虚拟机架内的其他交换机初始化时和 / 或在交换机 20 和 / 或虚拟机架内的其他交换机的恢复过程期间交换机 20 的配置。例如，VC 拓扑引擎 24 能够发送和 / 或接收与安装在交换机和 / 或其他 VC 交换机上的软件的一个或多个软件许可证有关的许可证信息供 CMM 23 或其他软件模块使用，以确保对于安装在 VC 交换机上的所有软件而言虚拟机架都具有适当的软件许可证。

[0031] 作为另一范例，VC 拓扑引擎 24 可以发送和 / 或接收某些供 CMM 23 或其他软件模块在推选虚拟机架内的主交换机的过程中使用的拓扑信息。主交换机是虚拟机架以及在虚拟机架内的交换机上运行的所有应用的管理（配置和监控）的中心点。例如，主交换机可以负责控制交换机之间的负载分配、交换 / 路由和通信以及管理虚拟机架内的冗余。

[0032] 在示范性操作中，VC 拓扑引擎 24 可由处理器 22 执行以支持基于每一交换机的交换机属性和 / 或其他推选标准来推选主交换机的主机推选过程。例如，在示范性操作中，在虚拟机架初始化时，在虚拟机架内启动主机推选过程，从而从虚拟机架内的所有交换机中推选出主机交换机。主机推选过程可以利用（例如）虚拟机架主机推选算法（作为 CMM 23

的部分或者独立于 CMM 23 运行），该算法考虑各种推选标准，例如，哪一交换机具有最低标识符 /MAC 地址、哪一交换机具有最高优先级、哪一交换机具有最长正常运行时间和 / 或其他标准来推选主交换机。例如可以由 VC 拓扑引擎 24 采用从接收到的 PDU 300 提取的拓扑信息 320 来提供所述推选标准。

[0033] 在一个实施例中，所接收到的 VC-ISIS PDU 300 是定制 Hello PDU。此外，交换机 20 还可以生成定制 Hello PDU，并将所述 Hello PDU 发送出所有的 VFL 端口 30a-30c。如上所述，交换机 20 能够对接收到的定制 Hello PDU 320 进行处理，以发现邻居并建立相邻性。例如，共享公共 VFL 链路的虚拟机架内的交换机将变成 VC-ISIS 邻居——如果它们的定制 HelloPDU 含有满足形成相邻性的标准的某一信息。基于相邻性信息，交换机可以建立拓扑图，继而生成最短路径树 (SPT)28。

[0034] 如上文进一步讨论的，定制 Hello PDU 300 包括虚拟机架特有的拓扑信息 320。例如，定制 Hello PDU 300 可以包括对虚拟机架独一无二的交换机标识符以及该交换机的 MAC 地址。此外，在示范性实施例中，定制 Hello PDU 300 还可以包括亚秒 Hello 时间间隔值，其能够使定制 Hello PDU 300 按照亚秒时间间隔交换（生成并发送至每一交换机）定制 Hello PDU 300。通过更加频繁地交换定制 Hello PDU 300，VC 拓扑引擎 24 能够更加快速地检测到虚拟机架内的故障。例如，如果将 Hello 时间间隔设为 100 毫秒，那么在 VC 拓扑引擎 24 未在预定的多个 100 毫秒间隔内在链路上或者从具体的 VC 交换机接收到 Hello PDU 的情况下，VC 拓扑引擎 24 能够确定在 VFL 链路上或者 VC 交换机上发生了故障（即，在已经经过了两个 Hello 时间间隔而未从特定的交换机接收到 Hello PDU 的情况下，VC 拓扑引擎将确定在特定交换机上发生了故障）。

[0035] 因此，VC 拓扑引擎 24 能够维护计时器或其他工具以监测在特定 VFL 端口 30a-30c 上和 / 或从特定 VC 交换机接收到的相继 Hello PDU 之间的持续时长。VC 拓扑引擎 24 还可以包括一组指令，从而基于相继接收到的 Hello PDU 之间经过的时间长度判断是否发生了故障。此外，VC 拓扑引擎 24 能够维护一组指令，从而根据探测到故障使 VC 拓扑引擎 24 恢复。例如，所述指令集可以指示 VC 拓扑引擎 24 重建虚拟机架的拓扑图，并基于新的拓扑图更新 SPT 28。

[0036] 在另一实施例中，所接收到的 VC-ISIS PDU 300 是定制链路状态 PDU。此外，交换机 20 还可以生成定制链路状态 PDU，并将所述 Hello PDU 发送出所有的 VFL 端口 30a-30c。如上所述，定制链路状态 PDU 300 包括对虚拟机架特定的拓扑信息 320。例如，定制链路状态 PDU 300 可以包括许可证信息、交换机优先级、交换机正常运行时间以及虚拟机架内的主交换机的标识符。应当理解定制 Hello 和 / 或链路状态 PDU 中包含的具有 VC 特异性的拓扑信息 320 可以发生变化，实施例不限于任何特定类型的 VC 拓扑信息。此外，应当理解，可以将具有 VC 特异性的拓扑信息 320 仅包含到定制 Hello PDU 或者仅包含到链路状态 PDU 内（而非两种类型的 PDU 均包含）。

[0037] 图 4 示出了 VC-ISIS（定制）Hello PDU 400 的示范性格式。VC-ISIS Hello PDU 400 包括标准 Hello 字段 405（例如，域内路由协议鉴别器、长度指示符等）连同类型 / 长度 / 值 (TLV) 字段 410。在 TLV 字段 410 内能够包含各种 VC 特异的拓扑信息 420。作为例子而非限制，如图 5 所示，这样的拓扑信息 420 可以包括操作机架标识符 421（即，虚拟机架内的交换机的唯一标识符）连同交换机的 MAC 地址 422。此外，拓扑信息 420 可以包括机

架角色 423(即,该交换机是主机还是从机)、机架类型 424、指定网络接口 (NI) 插槽标识符 425(即,连接至其他 VC 交换机的 NI 卡标识符)以及操作控制 VLAN 426(即,所述交换机负责的具体 VLAN(VLAN 标签 ID))。所述拓扑信息 420 还可以包括亚秒 Hello 间隔 427,如上文所述。

[0038] 图 6 示出了 VC-ISIS(定制)链路状态 PDU 500 的示范性格式。VC-ISIS 链路状态 PDU 500 包括标准链路状态字段 505(例如,域内路由协议鉴别器、长度指示符等)连同类型 / 长度 / 值 (TLV) 字段 510。在 TLV 字段 510 内可以包括各种 VC 特异的拓扑信息 520。作为例子而非限制,如图 7 所示,这样的拓扑信息 520 可以包括主用 CMM 标识符 521 和次级 CMM 标识符 522、正常运行时间(即,交换机已经工作的持续时长)、许可证配置 524(即,与安装在该交换机和 / 或虚拟机架内的其他交换机上的软件的软件许可证有关的信息)、所配置的机架优先级 525(即,虚拟机架内的交换机的优先级)以及机架群标识符 526(即虚拟机架身份)。此外,拓扑信息 520 还可以包括候选主机的机架标识符 527(即,虚拟机架内的备用主交换机的唯一标识符)、候选主机的 MAC 地址 528、主机的机架标识符 529(即,虚拟机架内的主交换机的唯一标识符)以及主机的 MAC 地址 530。

[0039] 图 8 示出了用于虚拟机架内的拓扑管理的方法 800 的实施例的示范性流程图。所述方法开始于 810,其中,在虚拟机架内的交换机的 VFL 链路上接收 VC-ISIS 协议数据单元 (PDU)。在 820,交换机从 VC-ISIS PDU 中提取虚拟机架拓扑信息。交换机在 830 中从所述虚拟机架拓扑信息探测虚拟机架的拓扑结构,并在 840 中生成最短路径树。

[0040] 如这里可能使用的,术语“基本”和“大致”为其对应术语和 / 或相关的术语提供了行业接受的公差。这样的行业容许公差从不到百分之一到百分之五十不等,其对应于但不限于部件值、集成电路工艺变化、温度变化、起落时间和 / 或热噪声。这样的物件之间的相对性从百分之几的差异到重大差异不等。文中还可能采用词语“耦合至”和 / 或“耦合”,它们包括物件之间的直接耦合和 / 或物件之间的经由居间物件的间接耦合(例如,物件包括但不限于部件、元件、电路和 / 或模块),其中,对于间接耦合而言,居间物件不修改信号信息但是可以调整其电流水平、电压水平和 / 或功率水平。文中还可能采用推断耦合(即,根据推断某一元件被耦合至另一元件),其按照与“耦合至”相同的方式包括两个物品之间的直接和间接耦合。文中可以采用“可操作”一词指示物件包括处理模块、数据、输入、输出等当中的一个或多个以执行所描述的或者必需的对应功能中的一个或多个,并且还可以包括与一个或多个其他物件的推断耦合以执行所描述的或者必需的对应功能。文中还可以采用词语“连接至”和 / 或“连接”或“相互连接”,其包括节点 / 设备之间的直接连接或链路和 / 或节点 / 设备之间的经由居间物件(例如,物件包括但不限于部件、元件、电路、模块、节点、设备等)的间接连接。文中还可能采用推断连接(即,根据推断某一元件被连接至另一元件),其按照与“连接至”相同的方式包括两个物品之间的直接和间接连接。

[0041] 上文还借助于说明指定功能的执行及其关系的方法步骤描述了实施例。为了描述的方便起见,这些功能构建块和方法步骤的边界和顺序是随意定义的。替代边界和顺序可被定义,只要适当地执行指定功能和关系即可。因而,任何这样的替代边界或顺序都处于所要求保护的本发明的精神和范围内。类似地,文中可能还随意地定义了流程图块以说明某些重要功能。就所使用的范围而言,可以另行定义仍然执行某些重要功能的流程图块边界和顺序。因而,这样的功能构建块和流程图块的替代定义和替代顺序处于所要求保护的本

发明的精神和范围内。本领域技术人员还将认识到文中的功能构建块和其他举例说明块、模块和部件可以由所示出的或者通过一个或多个运行适当软件等的分立部件、网络、系统、数据库或其组合来实施。

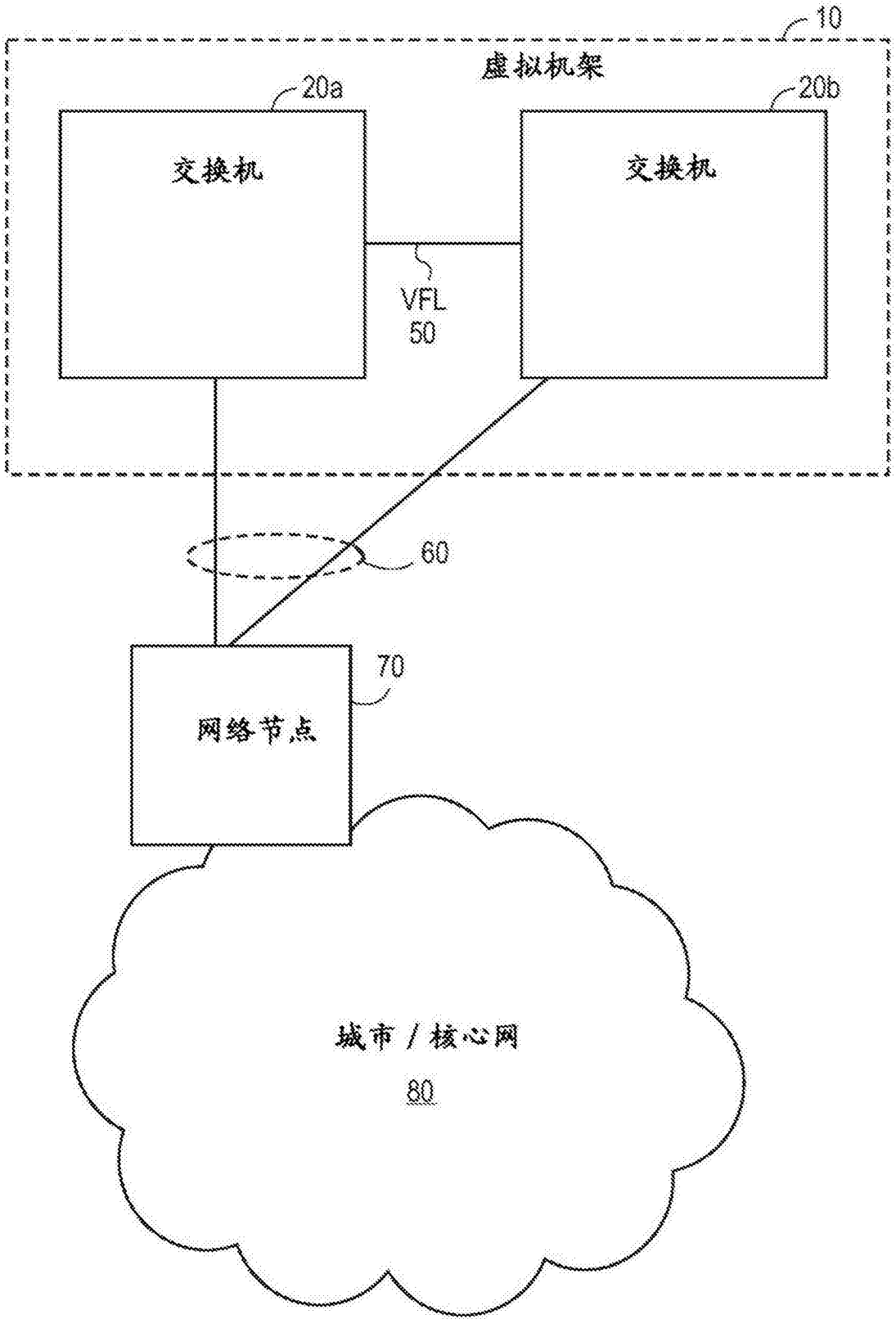


图 1

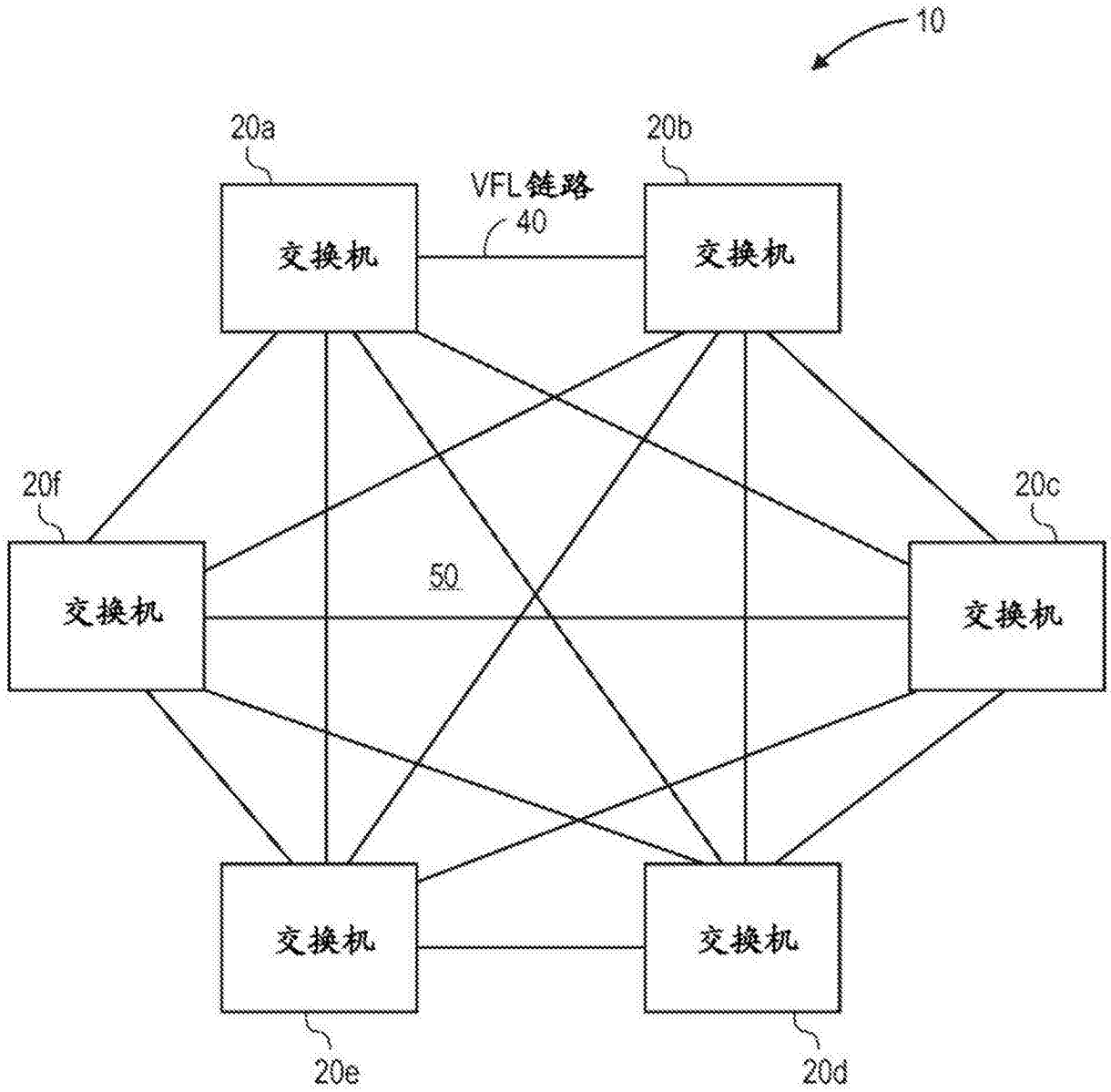


图 2

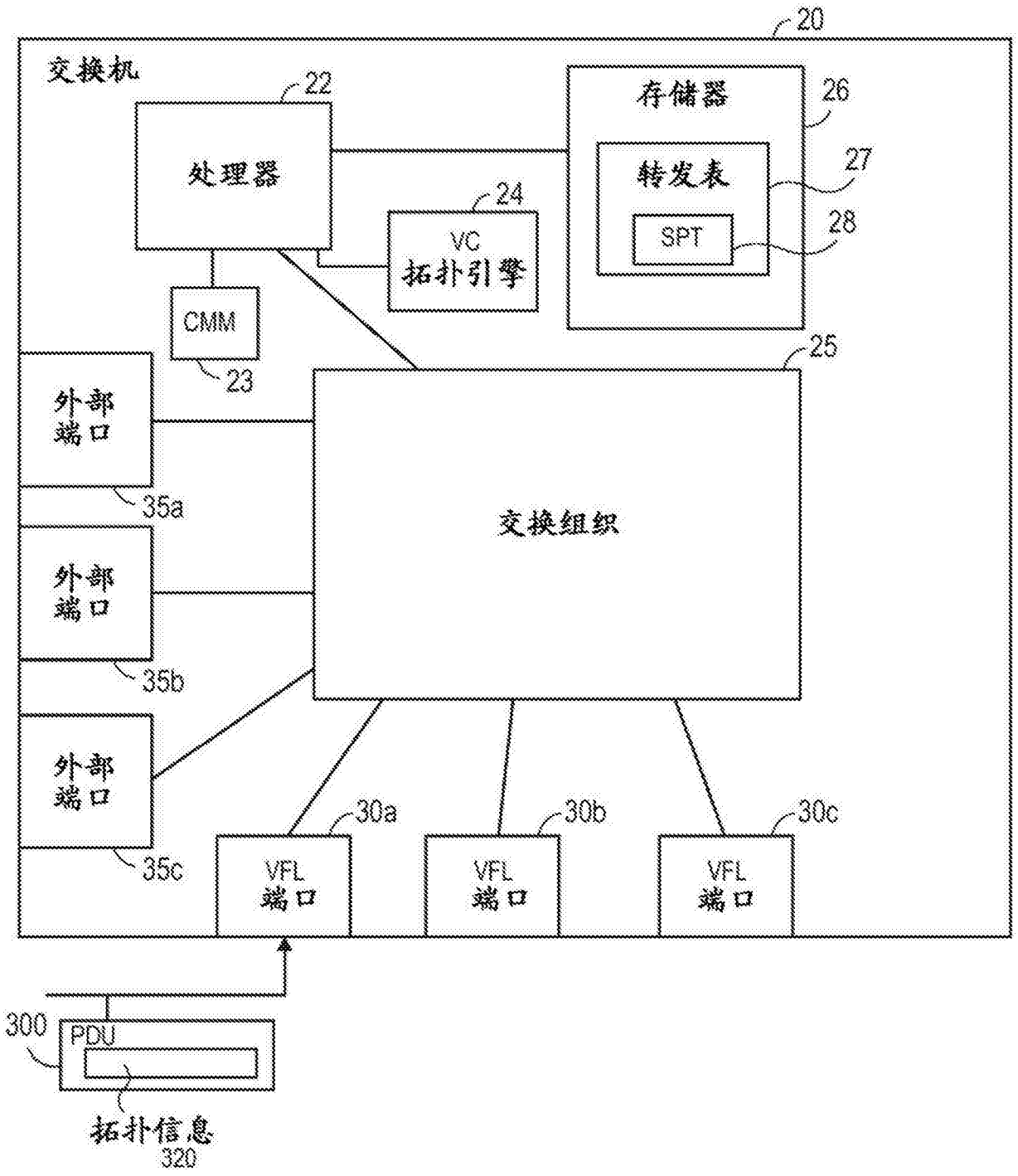

20
交换机
22
处理器
存储器
26
转发表
27
SPT
28
24
VC
拓扑引擎
CMM
23
25
外部端口
35a
交换组织
外部端口
35b
外部端口
35c
30a
30b
30c
VFL端口
VFL端口
VFL端口
300
PDU
拓扑信息
320

图 3

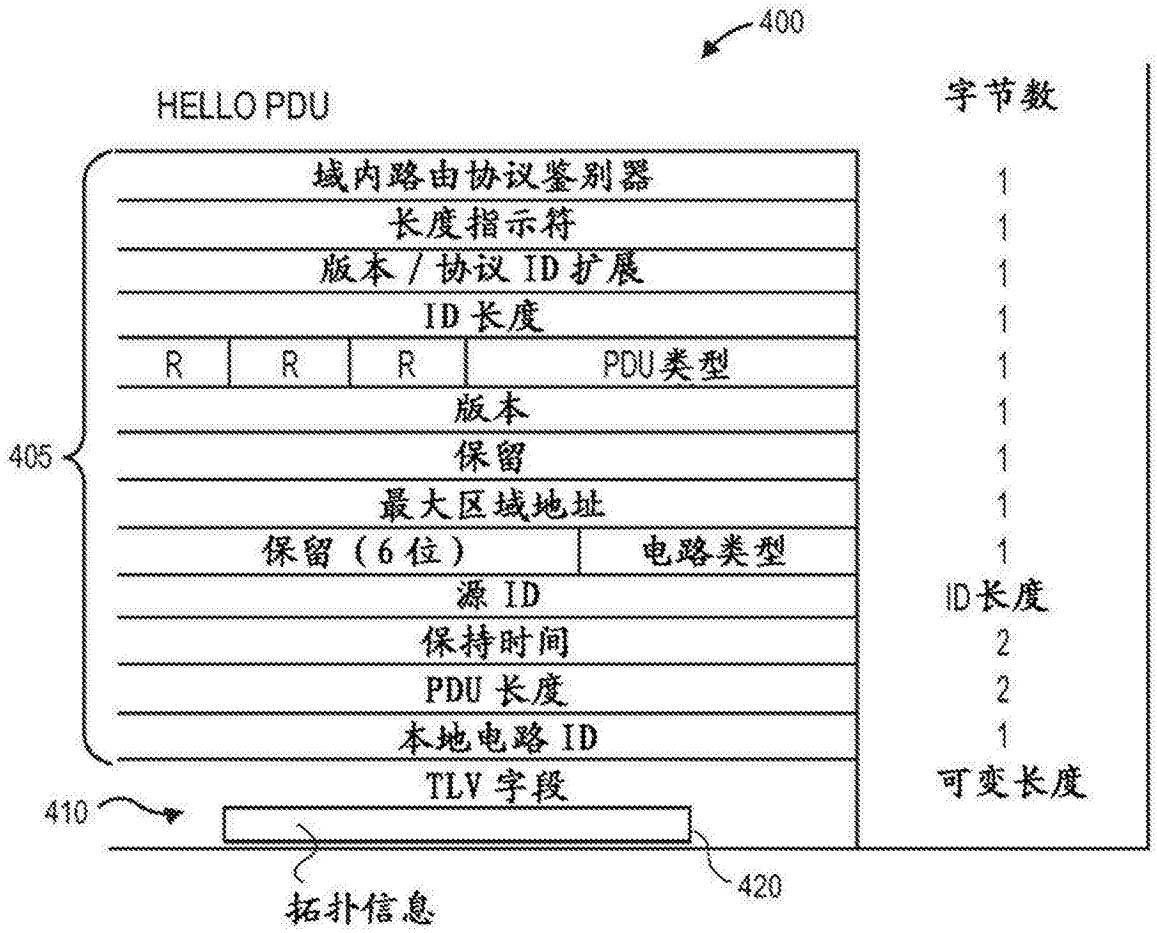

400
HELLO PDU
字节数
域内路由协议鉴别器
1
长度指示符
1
版本 / 协议 ID 扩展
1
ID 长度
1
R
R
R
PDU类型
1
版本
1
保留
1
最大区域地址
1
保留（6 位）
电路类型
1
源 ID
ID长度
保持时间
2
PDU 长度
2
本地电路 ID
1
TLV 字段
可变长度
405
410
420
拓扑信息

图 4

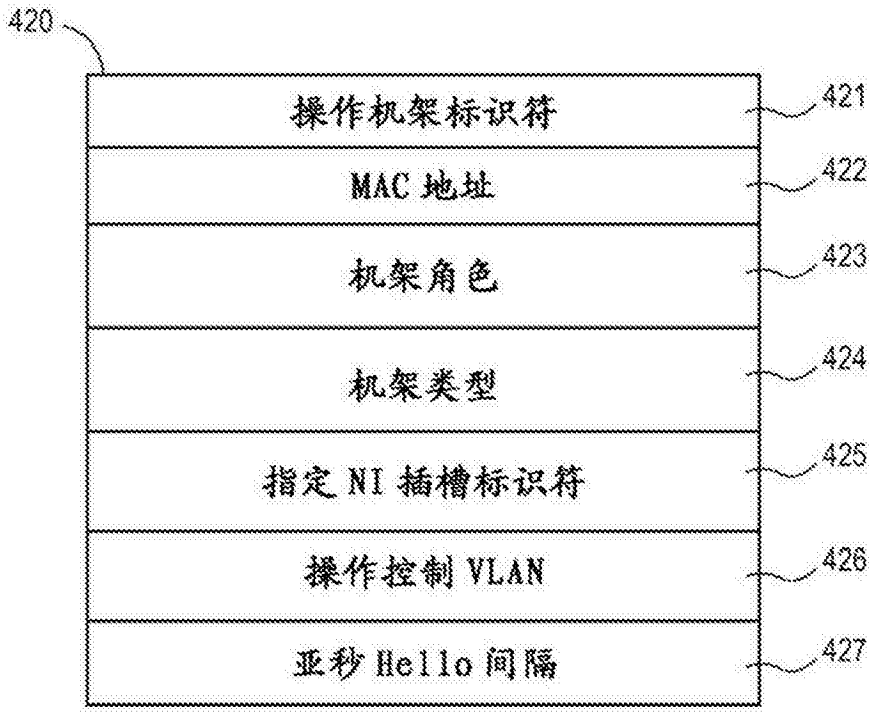

420
操作机架标识符
421
MAC 地址
422
机架角色
423
机架类型
424
指定 NI 插槽标识符
425
操作控制 VLAN
426
亚秒 Hello 间隔
427

图 5

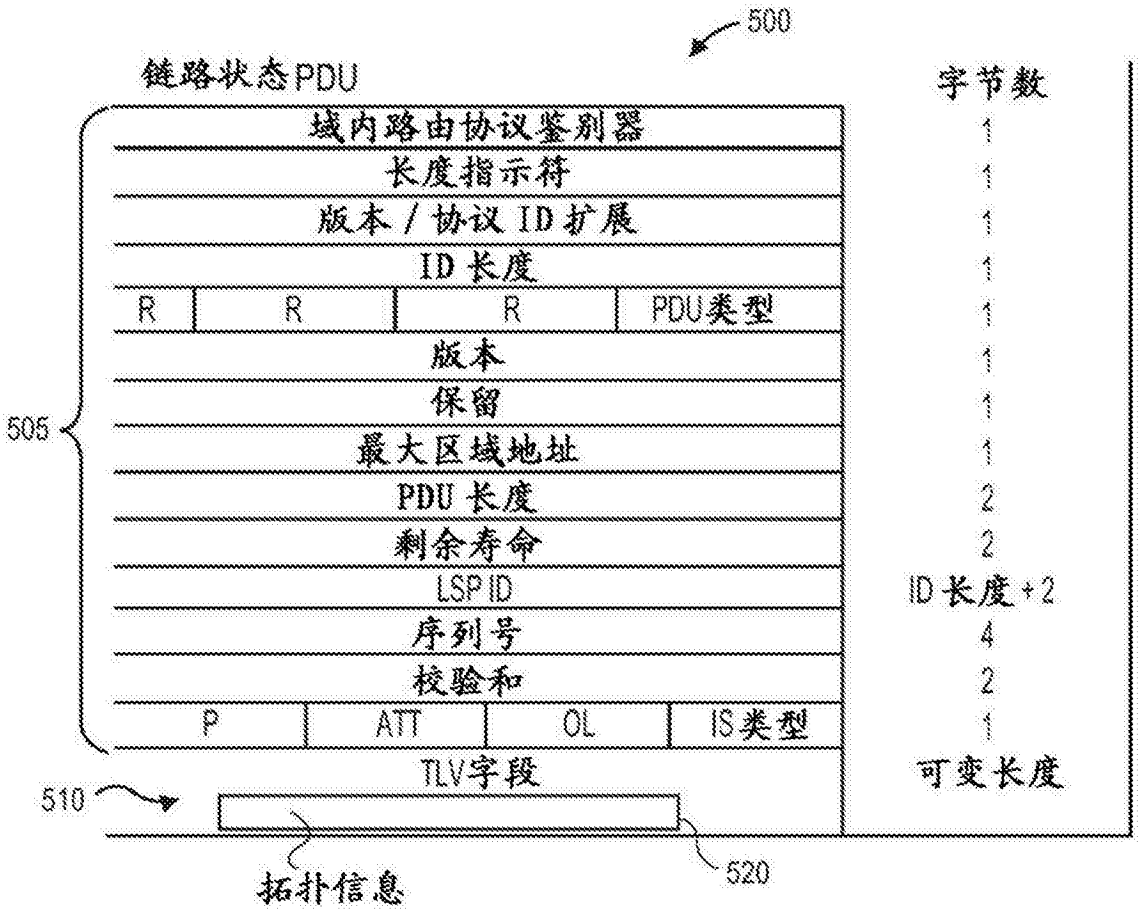

500
链路状态PDU
字节数
域内路由协议鉴别器
1
长度指示符
1
版本/协议ID扩展
1
ID长度
1
R
R
R
PDU类型
1
版本
1
保留
1
505
最大区域地址
1
PDU长度
2
剩余寿命
2
LSP ID
ID长度+2
序列号
4
校验和
2
P
ATT
OL
IS类型
1
TLV字段
可变长度
510
拓扑信息
520

图 6


520
主用 CMM 标识符
521
次级 CMM 标识符
522
正常运行时间
523
许可证配置
524
所配置的机架优先级
525
机架群标识符
526
候选主机的机架标识符
527
候选主机的 MAC 地址
528
主机的机架标识符
529
主机的 MAC 地址
530

图 7

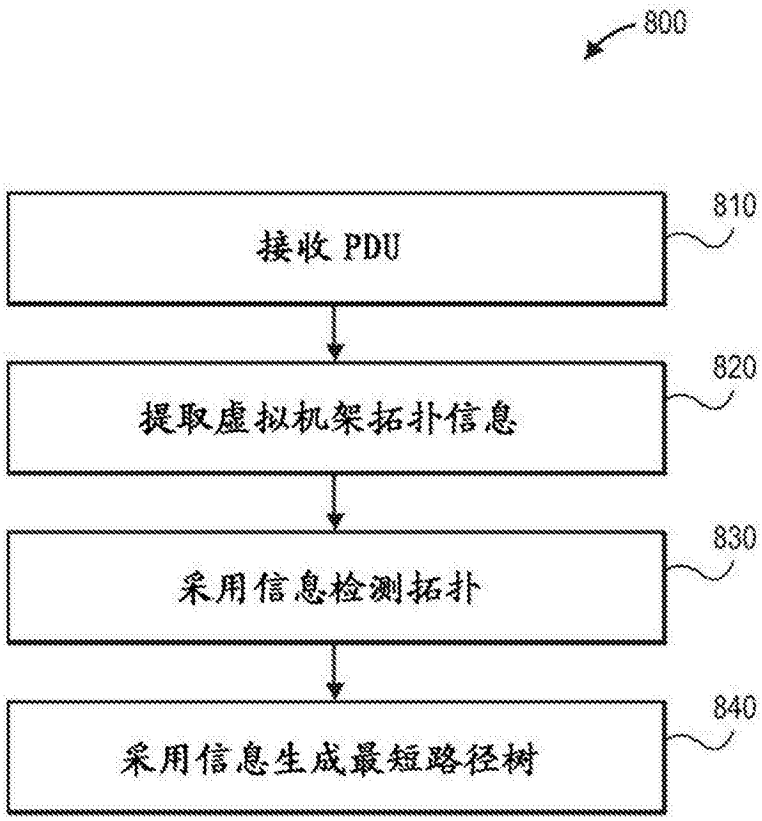

800
接收 PDU
810
提取虚拟机架拓扑信息
820
采用信息检测拓扑
830
采用信息生成最短路径树
840

图 8