(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 104247341 B
(45)授权公告日 2017.06.23

(21)申请号 201380013897.5

(22)申请日 2013.02.26

(65)同一申请的已公布的文献号
申请公布号 CN 104247341 A

(43)申请公布日 2014.12.24

(30)优先权数据
13/420,232 2012.03.14 US

(85)PCT国际申请进入国家阶段日
2014.09.12

(86)PCT国际申请的申请数据
PCT/IB2013/051525 2013.02.26

(87)PCT国际申请的公布数据
WO2013/136207 EN 2013.09.19

(73)专利权人 国际商业机器公司
地址 美国纽约

(72)发明人 B·沃克 T·A·格林菲尔德 C·巴索

(74)专利代理机构 北京市中咨律师事务所 11247
代理人 于静 张亚非

(51)Int.Cl.
H04L 12/70(2006.01)

(56)对比文件
US 2007177593 A1,2007.08.02,
CN 1574773 A,2005.02.02,
US 6331983 B1,2001.12.18,
CN 101488862 A,2009.07.22,

审查员 朱冬梅

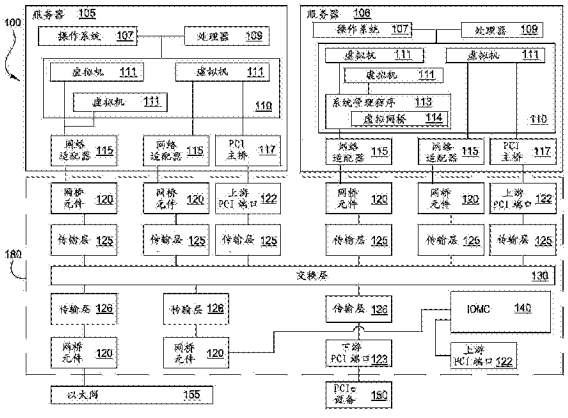
权利要求书3页 说明书27页 附图25页

(54)发明名称

分布式交换机及其多播树层次结构的动态优化方法

(57)摘要

一种分布式交换机可以包括具有一个或多个级别的代理子交换机(以及代理网桥元件)的层次结构,所述代理子交换机使得所述分布式交换机能够基于多播组的全体成员规模来扩展带宽。此外,每个代理可以根据一个或多个优化准则优化所述层次结构。例如,所述层次结构中的每个代理可以具有必需信息以便确保如果所述层次结构中的下一个代理不可用,则可以将数据路由到备用代理。可以通过跳过代理(或代理级别)进一步优化选定层次结构,以便将针对跳过的代理的数据发送到所述层次结构的较低级别中的代理。这可以更好地利用传输子交换机中的连接接口,并且消除任何不必要的代理到代理传输。



CN 104247341 B

1.一种在分布式交换机中转发多播数据帧的方法，包括：

在所述分布式交换机中的入口交换机的接收端口上接收多播数据帧；

确定层次结构的第一级别中的第一代理交换机，其中所述第一代理交换机在所述层次结构中被指定为将从所述入口交换机接收的所述多播数据帧的至少一部分转发到所述层次结构的第二级别中的第二代理交换机，其中所述入口交换机具有到所述第一代理交换机和所述第二代理交换机的直接物理连接；以及

当确定所述第一代理交换机满足至少一个优化准则时，从所述入口交换机在所述分布式交换机中转发所述部分，以便由所述第二代理交换机接收所述部分而不将所述部分转发到所述第一代理交换机。

2.根据权利要求1的方法，其中在所述分布式交换机中转发所述部分进一步包括：

将所述部分转发到所述层次结构的所述第一级别中的第三代理交换机；以及

将所述部分从所述第三代理交换机转发到所述第二代理交换机。

3.根据权利要求2的方法，其中所述优化准则包括操作中断，并且其中所述第一代理交换机不可用。

4.根据权利要求1的方法，其中在所述分布式交换机中转发所述部分进一步包括：

将所述部分从所述入口交换机直接转发到所述第二代理交换机，而不将所述部分转发到所述层次结构的所述第一级别中的任何代理交换机。

5.根据权利要求4的方法，其中所述入口交换机包括用于转发所述部分的多个连接接口，其中相对于通过将所述部分转发到所述第一级别中的所述第一代理交换机而在所述分布式交换机中转发所述部分，在所述分布式交换机中转发所述部分以便不将所述部分转发到所述第一级别中的任何代理交换机使用所述多个连接接口中的更多数量的连接接口。

6.根据权利要求4的方法，其中所述第一代理交换机通过基于与所述多播数据帧关联的多播组而在所述层次结构中被指定为将所述部分转发到所述分布式交换机中的仅一个交换机来满足所述优化准则。

7.根据权利要求1的方法，其中基于与所述多播数据帧关联的多播组，设置所述层次结构中用于在所述分布式交换机中转发所述部分的级别数。

8.根据权利要求1的方法，其中所述优化准则基于以下至少一个：所述第一代理交换机使用的连接接口数、操作中断，以及被指定为从所述第一代理交换机接收所述部分的交换机数。

9.一种计算机可读存储介质，所述计算机可读存储介质具有随其包含的计算机可读程序代码，所述计算机可读程序代码包括被配置为执行以下操作的计算机可读程序代码：

在分布式交换机中的入口交换机的接收端口上接收多播数据帧；

确定层次结构的第一级别中的第一代理交换机，其中所述第一代理交换机在所述层次结构中被指定为将从所述入口交换机接收的所述多播数据帧的至少一部分转发到所述层次结构的第二级别中的第二代理交换机，其中所述入口交换机具有到所述第一代理交换机和所述第二代理交换机的直接物理连接；以及

当确定所述第一代理交换机满足至少一个优化准则时，从所述入口交换机在所述分布式交换机中转发所述部分，以便由所述第二代理交换机接收所述部分而不将所述部分转发到所述第一代理交换机。

10.根据权利要求9的计算机可读存储介质，其中在所述分布式交换机中转发所述部分进一步包括被配置为执行以下操作的计算机可读程序代码：

将所述部分转发到所述层次结构的所述第一级别中的第三代理交换机；以及

将所述部分从所述第三代理交换机转发到所述第二代理交换机。

11.根据权利要求10的计算机可读存储介质，其中所述优化准则包括操作中断，并且其中所述第一代理交换机不可用。

12.根据权利要求9的计算机可读存储介质，其中在所述分布式交换机中转发所述部分进一步包括被配置为执行以下操作的计算机可读程序代码：

将所述部分从所述入口交换机直接转发到所述第二代理交换机，而不将所述部分转发到所述层次结构的所述第一级别中的任何代理交换机。

13.根据权利要求12的计算机可读存储介质，其中所述入口交换机包括用于转发所述部分的多个连接接口，其中相对于通过将所述部分转发到所述第一级别中的所述第一代理交换机而在所述分布式交换机中转发所述部分，在所述分布式交换机中转发所述部分以便不将所述部分转发到所述第一级别中的任何代理交换机使用所述多个连接接口中的更多数量的连接接口。

14.根据权利要求12的计算机可读存储介质，其中所述第一代理交换机通过基于与所述多播数据帧关联的多播组而在所述层次结构中被指定为将所述部分转发到所述分布式交换机中的仅一个交换机来满足所述优化准则。

15.根据权利要求9的计算机可读存储介质，其中基于与所述多播数据帧关联的多播组，设置所述层次结构中用于在所述分布式交换机中转发所述部分的级别数。

16.一种分布式交换机，包括：

所述分布式交换机的入口交换机，其接收多播数据帧并且确定层次结构的第一级别中的第一代理交换机，其中所述第一代理交换机在所述层次结构中被指定为将从所述入口交换机接收的所述多播数据帧的至少一部分转发到所述层次结构的第二级别中的第二代理交换机，

其中所述入口交换机具有到所述第一代理交换机和所述第二代理交换机的直接物理连接，并且当确定所述第一代理交换机满足至少一个优化准则时，所述入口交换机在所述分布式交换机中转发所述部分，以便由所述第二代理交换机接收所述部分而不将所述部分转发到所述第一代理交换机。

17.根据权利要求16的分布式交换机，其中在所述分布式交换机中转发所述部分进一步包括：

将所述部分转发到所述层次结构的所述第一级别中的第三代理交换机；以及

将所述部分从所述第三代理交换机转发到所述第二代理交换机。

18.根据权利要求17的分布式交换机，其中所述优化准则包括操作中断，并且其中所述第一代理交换机不可用。

19.根据权利要求16的分布式交换机，其中在所述分布式交换机中转发所述部分进一步包括：

将所述部分从所述入口交换机直接转发到所述第二代理交换机，而不将所述部分转发到所述层次结构的所述第一级别中的任何代理交换机。

20.根据权利要求19的分布式交换机，其中所述入口交换机包括用于转发所述部分的多个连接接口，其中相对于通过将所述部分转发到所述第一级别中的所述第一代理交换机而在所述分布式交换机中转发所述部分，在所述分布式交换机中转发所述部分以便不将所述部分转发到所述第一级别中的任何代理交换机使用所述多个连接接口中的更多数量的连接接口。

21.根据权利要求19的分布式交换机，其中所述第一代理交换机通过基于与所述多播数据帧关联的多播组而在所述层次结构中被指定为将所述部分转发到所述分布式交换机中的仅一个交换机来满足所述优化准则。

22.根据权利要求16的分布式交换机，其中基于与所述多播数据帧关联的多播组，设置所述层次结构中用于在所述分布式交换机中转发所述部分的级别数。

分布式交换机及其多播树层次结构的动态优化方法

背景技术

[0001] 计算机系统通常使用在公共机箱中耦合在一起的多个计算机。计算机可以是单独的服务器，它们在机箱中通过公共主干耦合。每个服务器是可插拔板，其包括至少一个处理器、板上存储器和输入/输出(I/O)接口。此外，服务器可以连接到交换机以便扩展服务器的能力。例如，交换机可以允许服务器接入其它以太网或PCIe插槽，以及允许同一或不同机箱中的服务器之间的通信。

[0002] 多播数据帧需要交换机将数据转发到多播组的所有成员。即，对于交换机接收的每一个多播数据帧，交换机都会创建数据帧的副本并且将其转发到多播组的每个成员。当组的成员增加时，交换机必须将数据帧转发到越来越多的计算机节点。

发明内容

[0003] 在此描述的实施例提供一种用于在分布式交换机中转发多播数据帧的方法和计算机程序产品。所述方法和计算机程序产品包括在所述分布式交换机中的入口交换机的接收端口上接收多播数据帧，并且确定层次结构的第一级别中的第一代理交换机。所述第一代理交换机在所述层次结构中被指定为将所述部分转发到以下至少一个：所述层次结构中的目的地交换机和第二代理交换机，并且所述目的地交换机和所述第二代理交换机都在所述层次结构的第二级别中。此外，所述层次结构增加用于在所述分布式交换机中转发所述数据帧的所述部分的可用带宽。当确定所述第一代理满足至少一个优化准则时，所述方法和计算机程序产品包括在所述分布式交换机中转发所述部分，以便由所述目的地交换机和所述第二代理交换机中的至少一个接收所述部分而不将所述部分转发到所述第一代理。

[0004] 另一个实施例提供一种分布式交换机。所述分布式交换机包括所述分布式交换机的入口交换机，其接收多播数据帧并且确定层次结构的第一级别中的第一代理交换机。所述第一代理交换机在所述层次结构中被指定为将所述多播数据帧的至少一部分转发到以下至少一个：所述层次结构中的目的地交换机和第二代理交换机，其中所述目的地交换机和所述第二代理交换机都在所述层次结构的第二级别中。此外，所述层次结构增加用于在所述分布式交换机中转发所述部分的可用带宽。当确定所述第一代理满足至少一个优化准则时，所述入口交换机在所述分布式交换机中转发所述部分，以便由所述第二级别中的所述目的地交换机和所述第二代理交换机中的至少一个接收所述部分而不将所述部分转发到所述第一代理。

附图说明

[0005] 因此，可以通过参考附图，具有获得上述方面并且可以详细理解上述方面的方式、上面简要总结的本发明实施例的更具体说明。

[0006] 但是，应该注意，附图仅示出本发明的典型实施例，因此不被视为本发明范围的限制，因为本发明可以允许其它同等有效的实施例。

[0007] 图1示出根据在此描述的一个实施例的包括分布式虚拟交换机的系统体系架构；

[0008] 图2示出根据在此描述的一个实施例的实现分布式虚拟交换机的系统的硬件表示；

[0009] 图3示出根据在此描述的一个实施例的分布式虚拟交换机；

[0010] 图4示出根据在此描述的一个实施例的能够增加带宽的图2的子交换机；

[0011] 图5A-5B示出根据在此描述的各实施例的在图4的子交换机中执行带宽增加；

[0012] 图6示出根据在此描述的一个实施例的使用以太网帧块在图4的子交换机中执行带宽增加；

[0013] 图7示出根据在此描述的一个实施例的在交换机层上传输的信元(cell)；

[0014] 图8是根据在此描述的一个实施例的带宽增加技术；

[0015] 图9是根据在此描述的一个实施例的使用分布式交换机互连的计算系统；

[0016] 图10是根据在此描述的一个实施例的用于转发多播数据帧的代理的层次结构；

[0017] 图11是根据在此描述的一个实施例的在图10中示出的层次结构的一部分的系统图；

[0018] 图12示出根据在此描述的一个实施例的在图10中示出的层次结构中的多播数据帧的一个实例路径；

[0019] 图13示出根据在此描述的一个实施例的MC组表；

[0020] 图14示出根据在此描述的一个实施例的分层数据；

[0021] 图15A-C示出根据在此描述的各实施例的用于处理操作中断的系统和技术；

[0022] 图16A-D示出根据在此描述的各实施例的用于优化层次结构的系统和技术；

[0023] 图17示出根据在此描述的一个实施例的将单播数据帧传输到主干的物理链路；

[0024] 图18示出根据在此描述的一个实施例的使用代理将多播数据帧传输到主干的物理链路；

[0025] 图19示出根据在此描述的一个实施例的将多播数据帧传输到分配给至少两个主干的目的地交换机；

[0026] 图20A-20C示出根据在此描述的各实施例的使用三种不同模式将多播数据帧传输到主干的物理链路。

具体实施方式

[0027] 对于连接到分布式交换机的计算系统(例如，服务器)，分布式虚拟交换机可以似乎是单个交换机元件。实际上，分布式交换机可以包括多个不同交换机模块，它们经由交换层互连以便每个交换机模块可以与任何其它交换机模块通信。例如，计算系统可以在物理上连接到一个交换机模块的端口，但使用交换层，能够与具有连接到WAN(例如，因特网)的端口的不同交换机模块通信。此外，每个交换机模块可以被配置为基于两个不同通信协议接受和路由数据。但是，对于计算系统，两个单独的交换机模块显示为一个交换机。

[0028] 分布式交换机可以在每个交换机模块上包括多个芯片(即，子交换机)。这些子交换机可以接收指定多个不同目的地子交换机的多播数据帧(例如，以太网帧)。接收数据帧的子交换机负责创建帧的一部分(例如帧的有效负载)的副本，并且使用分布式交换机的光纤通道网络(fabric)将该部分转发到相应的目的地子交换机。但是，子交换机可以使用多个连接接口并行传输数据帧的副本，而不是仅使用一个出口连接接口按顺序将数据帧的副

本转发到每个目的地。例如，子交换机可以具有多个Tx/Rx端口，每个端口与提供到分布式交换机中的其它子交换机的连接性的连接接口关联。接收多播数据帧的端口可以借用分配给这些其它端口的连接接口(以及关联的硬件)以便并行传输多播数据帧的副本。

[0029] 此外，可以将这些子交换机布置在分层结构中，其中选择一个或多个子交换机充当代理。将分布式交换机的子交换机组合在一起，其中将每个组分配给一个或多个代理。当子交换机接收多播数据帧时，它将分组转发到代理子交换机之一。每个代理子交换机然后可以将分组转发到另一个代理或目的地计算设备。因为代理还可以使用两个或更多连接接口并行传输分组，所以对于使用的每个代理，用于转发多播分组的带宽增加。

[0030] 此外，代理层次结构可以包括多个级别，它们形成金字塔式布置，其中较高级别代理将多播数据帧转发到较低级别代理直至到达层次结构的底部。可以定制每个多播组以便当通过分布式交换机转发多播数据时，使用这些级别中的一个或多个。此外，层次结构中的每个代理可以具有必需信息以便确保如果层次结构中的下一个代理不可用，则可以将数据路由到备用代理。

[0031] 可以通过跳过代理进一步优化选定层次结构。例如，如果层次结构的一个级别中的代理将多播数据转发到层次结构的较低级别中的仅一个代理(或目的地)子交换机，则可以跳过该代理。相反，将多播数据直接转发到较低级别中的子交换机。此外，可以优化层次结构以便确保最大程度地使用子交换机的连接接口(例如，将多播数据传输到分布式交换机中的其它子交换机的端口)。具体地说，如果跳过层次结构的某个级别将增加使用的连接接口数量，则子交换机可以将多播数据直接转发到分层级别中在跳过的级别下面的代理。

[0032] 以下参考本发明的各实施例。但是，应该理解，本发明并不限于具体描述的实施例。相反，构想以下特性和元素(无论是否与不同实施例相关)的任意组合以便实现和实施本发明。此外，尽管本发明的实施例可以实现胜过其它可能解决方案和/或现有技术的优点，但特定优点是否由给定实施例实现并不是本发明的限制。因此，以下方面、特性、实施例和优点仅是示例性的，并且不被视为所附权利要求的元素或限制，除非在权利要求(多个)中显式描述。同样，对“本发明”的引用不应该被解释为在此公开的任何发明主题的概括，并且不应该被视为所附权利要求的元素或限制，除非在权利要求(多个)中显示描述。

[0033] 所属技术领域的技术人员知道，本发明的各个方面可以实现为系统、方法或计算机程序产品。因此，本发明的各个方面可以具体实现为以下形式，即：完全的硬件实施方式、完全的软件实施方式(包括固件、驻留软件、微代码等)，或硬件和软件方面结合的实施方式，这里可以统称为“电路”、“模块”或“系统”。此外，本发明的各个方面还可以实现为在一个或多个计算机可读介质中的计算机程序产品的形式，该计算机可读介质中包含计算机可读的程序代码。

[0034] 可以采用一个或多个计算机可读介质的任意组合。计算机可读介质可以是计算机可读信号介质或者计算机可读存储介质。计算机可读存储介质例如可以是—但不限于—电、磁、光、电磁、红外线、或半导体的系统、装置或器件，或者上述的任意合适的组合。计算机可读存储介质的更具体的例子(非穷举的列表)包括：具有一个或多个导线的电连接、便携式计算机盘、硬盘、随机存取存储器(RAM)、只读存储器(ROM)、可擦式可编程只读存储器(EPROM或闪存)、光纤、便携式紧凑盘只读存储器(CD-ROM)、光存储器件、磁存储器件、或者上述的任意合适的组合。在本文件中，计算机可读存储介质可以是任何包含或存储程序的

有形介质，该程序可以被指令执行系统、装置或者器件使用或者与其结合使用。

[0035] 计算机可读的信号介质可以包括例如在基带中或者作为载波一部分传播的数据信号，其中承载了计算机可读的程序代码。这种传播的数据信号可以采用多种形式，包括—但不限于—电磁信号、光信号或上述的任意合适的组合。计算机可读的信号介质可以是计算机可读存储介质以外的任何计算机可读介质，该计算机可读介质可以发送、传播或者传输用于由指令执行系统、装置或者器件使用或者与其结合使用的程序。

[0036] 计算机可读介质上包含的程序代码可以用任何适当的介质传输，包括—但不限于—无线、有线、光缆、RF等等，或者上述的任意合适的组合。

[0037] 可以以一种或多种程序设计语言的任意组合来编写用于执行本发明的各个方面的操作的计算机程序代码，所述程序设计语言包括面向对象的程序设计语言—诸如Java、Smalltalk、C++等，还包括常规的过程式程序设计语言—诸如“C”语言或类似的程序设计语言。程序代码可以完全地在用户计算机上执行、部分地在用户计算机上执行、作为一个独立的软件包执行、部分在用户计算机上部分在远程计算机上执行、或者完全在远程计算机或服务器上执行。在涉及远程计算机的情形中，远程计算机可以通过任意种类的网络—包括局域网(LAN)或广域网(WAN)—连接到用户计算机，或者，可以连接到外部计算机(例如利用因特网服务提供商来通过因特网连接)。

[0038] 下面将参照根据本发明实施例的方法、装置(系统)和计算机程序产品的流程图和/或框图描述本发明的各个方面。应当理解，流程图和/或框图的每个方框以及流程图和/或框图中各方框的组合，都可以由计算机程序指令实现。这些计算机程序指令可以提供给通用计算机、专用计算机或其它可编程数据处理装置的处理器，从而生产出一种机器，使得这些指令在通过计算机或其它可编程数据处理装置的处理器执行时，产生了实现流程图和/或框图中的一个或多个方框中规定的功能/动作的装置。

[0039] 也可以把这些计算机程序指令存储在计算机可读介质中，这些指令使得计算机、其它可编程数据处理装置、或其它设备以特定方式工作，从而，存储在计算机可读介质中的指令就产生出包括实现流程图和/或框图中的一个或多个方框中规定的功能/动作的指令的制造品(article of manufacture)。

[0040] 也可以把计算机程序指令加载到计算机、其它可编程数据处理装置、或其它设备上，使得在计算机、其它可编程装置或其它设备上执行一系列操作步骤，以产生计算机实现的过程，从而使得在计算机或其它可编程装置上执行的指令提供实现流程图和/或框图中的一个或多个方框中规定的功能/动作的过程。

[0041] 本发明的实施例可以通过云计算基础架构提供给最终用户。云计算通常指通过网络将可扩展计算资源作为服务提供。更正式地讲，云计算可以定义为一种计算能力，这种计算能力在计算资源及其底层技术体系架构(例如，服务器、存储、网络)之间提供抽象，从而能够对共享的可配置计算资源池进行方便、按需的网络访问，可配置计算资源是能够以最小的管理成本或与服务提供者进行最少的交互就能快速部署和释放的资源。因此，云计算允许用户访问“云”中的虚拟计算资源(例如，存储、数据、应用，甚至是完整的虚拟化计算系统)，而不考虑用于提供计算资源的底层物理系统(或者那些系统的位置)。

[0042] 通常，按使用付费为用户提供云计算资源，其中仅针对实际使用的计算资源(例如用户使用的存储空间量，或者用户实例化的虚拟化系统数量)向用户收取费用。用户可以在

任何时候、从任何位置通过因特网访问位于云中的任何资源。在本发明的上下文中，用户可以访问在服务器上运行或存储的云中提供的应用或相关数据。例如，应用可以在服务器上执行，该服务器实现云中的虚拟交换机。这样做允许用户从连接到与云相连的网络(例如，因特网)的任何计算系统访问该信息。

[0043] 图1示出根据在此描述的一个实施例的包括分布式虚拟交换机的系统体系架构。第一服务器105可以包括耦合到存储器110的至少一个处理器109。处理器109可以表示一个或多个处理器(例如，微处理器)或多核处理器。存储器110可以表示包括服务器105的主存储设备的随机存取存储器(RAM)设备，以及补充级别的存储，例如高速缓冲存储器、非易失性或备用存储器(例如，可编程存储器或闪存)、只读存储器等。此外，存储器110可以被视为包括存储装置，其在物理上位于服务器105中或在耦合到服务器105的另一个计算设备上。

[0044] 服务器105可以在操作系统107的控制下操作，并且可以执行各种计算机软件应用、组件、程序、对象、模块和数据结构，例如虚拟机111。

[0045] 服务器105可以包括网络适配器115(例如，融合网络适配器)。融合网络适配器可以包括单根I/O虚拟化(SR-IOV)适配器，例如支持融合增强型以太网(CEE)的快速外围组件互连(PCIe)适配器。系统100的另一个实施例可以包括多根I/O虚拟化(MR-IOV)适配器。网络适配器115可以进一步用于实现以太网光纤通道(FCoE)协议、以太网RDMA、因特网小型计算机系统接口(iSCSI)等。一般而言，网络适配器115使用基于以太网或PCI的通信方法传输数据，并且可以耦合到一个或多个虚拟机111。此外，适配器可以促进虚拟机111之间的共享访问。尽管适配器115被示出为包括在服务器105中，但在其它实施例中，适配器可以是与服务器105分离的在物理上不同的设备。

[0046] 在一个实施例中，每个网络适配器115可以包括融合适配器虚拟网桥(未示出)，其通过协调对虚拟机111的访问来促进适配器115之间的数据传输。每个融合适配器虚拟网桥可以识别在其域(即，可寻址空间)中流动的数据。可以直接路由识别的域地址，而不在特定融合适配器虚拟网桥域之外传输数据。

[0047] 每个网络适配器115可以包括耦合到网桥元件120之一的一个或多个以太网端口。此外，为了促进PCIe通信，服务器可以具有PCI主桥117。PCI主桥117然后连接到分布式交换机180中的交换机元件上的上游PCI端口122。然后经由交换层130将数据路由到正确的下游PCI端口123，下游PCI端口123可以与上游PCI端口122位于同一或不同的交换机模块上。然后可以将数据转发到PCI设备150。

[0048] 网桥元件120可以被配置为在整个分布式虚拟交换机180内转发数据帧。例如，可以使用两个40千兆位以太网连接或一个100千兆位以太网连接来连接网络适配器115和网桥元件120。网桥元件120将网络适配器115接收的数据帧转发到交换层130。网桥元件120可以包括查找表，其存储用于转发接收的数据帧的地址数据。例如，网桥元件120可以将与接收的数据帧关联的地址数据和存储在查找表中的地址数据相比较。因此，网络适配器115不需要知道分布式交换机180的网络拓扑。

[0049] 一般而言，分布式虚拟交换机180包括多个网桥元件120，它们可以位于多个单独但互连的硬件组件上。从网络适配器115的角度来看，交换机180的操作如同单个交换机，尽管交换机180可以包括在物理上位于不同组件上的多个交换机。分布交换机180将在发生故障的情况下提供冗余。

[0050] 每个网桥元件120可以连接到一个或多个传输层模块125，这些模块将接收的数据帧转换为交换层130使用的协议。例如，传输层模块125可以将使用以太网或PCI通信方法接收的数据转换为通用数据类型(即，信元)，其经由交换层130(即，信元光纤通道网络)传输。因此，包括交换机180的交换机模块与至少两个不同通信协议兼容—例如，以太网和PCIe通信标准。即，至少一个交换机模块具有在同一交换层130上传输不同类型数据的必需逻辑。

[0051] 尽管未在图1中示出，但在一个实施例中，交换层130可以包括局部机架互连，该局部机架互连具有连接位于同一机箱和机架中的网桥元件120的专用连接，以及用于连接到其它机箱和机架中的网桥元件120的链路。

[0052] 在路由信元之后，交换层130可以与传输层模块126通信，这些模块将信元转换回对应于其相应通信协议的数据帧。网桥元件120的一部分可以促进与以太网155通信，以太网155用于接入到LAN或WAN(例如，因特网)。此外，可以将PCI数据路由到与PCIe设备150相连的下游PCI端口123。PCIe设备150可以是无源背板互连、作为附加板的扩展卡接口，或者可以由连接到交换机180的任何服务器访问的公共存储装置。

[0053] 尽管使用“上游”和“下游”描述PCI端口，但这仅用于示出一种可能的数据流。例如，在一个实施例中，下游PCI端口123可以将数据从连接到的PCIe设备150传输到上游PCI端口122。因此，PCI端口122、123均可以传输以及接收数据。

[0054] 第二服务器106可以包括连接到操作系统107的处理器109，以及包括类似于在第一服务器105中发现的一个或多个虚拟机111的存储器110。服务器106的存储器110还包括具有虚拟网桥114的系统管理程序113。系统管理程序113管理在不同虚拟机111之间共享的数据。具体地说，虚拟网桥114允许连接的虚拟机111之间的直接通信，而不需要虚拟机111使用网桥元件120或交换层130将数据传输到以通信方式耦合到系统管理程序113的其它虚拟机111。

[0055] 输入/输出管理控制器(IOMC)140(即，专用处理器)耦合到至少一个网桥元件120或上游PCI端口122，其使IOMC 140接入到交换层130。IOMC 140的一个功能可以是从管理员接收命令以便配置分布式虚拟交换机180的不同硬件元件。在一个实施例中，可以从交换层130中的单独交换网络接收这些命令。

[0056] 尽管示出一个IOMC 140，但系统100可以包括多个IOMC 140。在一个实施例中，可以将这些IOMC 140布置在层次结构中，以便选择一个IOMC 140作为主控制器，而委派其它IOMC 140作为成员控制器(或从控制器)。

[0057] 图2示出根据一个实施例的系统100的硬件级别图。服务器210和212可以在物理上位于同一机箱205中；但是，机箱205可以包括任意数量的服务器。机箱205还包括多个交换机模块250、251，这些模块包括一个或多个子交换机254(即，微芯片)。在一个实施例中，交换机模块250、251、252是硬件组件(例如，PCB板、FPGA板等)，它们在网络适配器115和网桥元件120之间提供物理支持和连接性。一般而言，交换机模块250、251、252包括连接系统200中的不同机箱205、207和服务器210、212、214的硬件，并且可以是计算系统中的单个可更换部件。

[0058] 交换机模块250、251、252(例如，机箱互连元件)包括一个或多个子交换机254和IOMC 255、256、257。子交换机254可以包括网桥元件120的逻辑或物理组—例如，每个子交换机254可以具有五个网桥元件120。每个网桥元件120可以在物理上连接到服务器210、

212。例如，网桥元件120可以将使用以太网或PCI通信协议发送的数据路由到其它网桥元件120，其它网桥元件120使用路由层连接到交换层130。但是，在一个实施例中，可以不需要网桥元件120来提供从网络适配器115到交换层130的连接性以实现PCI或PCIe通信。

[0059] 每个交换机模块250、251、252包括IOMC 255、256、257以便管理和配置系统200中的不同硬件资源。在一个实施例中，每个交换机模块250、251、252的相应IOMC可以负责配置特定交换机模块上的硬件资源。但是，因为交换机模块使用交换层130互连，所以一个交换机模块上的IOMC可以管理不同交换机模块上的硬件资源。如上面讨论的，在每个交换机模块250、251、252中，IOMC 255、256、257连接到至少一个子交换机254(或网桥元件120)，这使得每个IOMC能够在交换层130上路由命令。为了清晰起见，省略用于IOMC 256和257的这些连接。此外，交换机模块251、252可以包括多个子交换机254。

[0060] 机箱205中的虚线定义服务器210、212和交换机模块250、251之间的中平面220。即，中平面220包括在网络适配器115和子交换机254之间传输数据的数据路径(例如，导线或迹线)。

[0061] 每个网桥元件120经由路由层连接到交换层130。此外，网桥元件120还可以连接到网络适配器115或上行链路。如在此使用的，网桥元件120的上行链路端口提供扩展系统200的连接性或能力的服务。如在机箱207 中所示，一个网桥元件120包括到以太网或PCI连接器260的连接。对于以太网通信，连接器260可以使系统200接入到LAN或WAN(例如，因特网)。备选地，端口连接器260可以将系统连接到PCIe扩展槽—例如，PCIe设备150。设备150可以是其它存储装置或存储器，每个服务器210、212、214可以经由交换层130对其进行访问。有利的是，系统200用于接入到交换层130，交换层130具有与至少两种不同通信方法兼容的网络设备。

[0062] 如图所示，服务器210、212、214可以具有多个网络适配器115。如果这些适配器115中的一个发生故障，则这将提供冗余。此外，每个适配器115可以经由中平面220连接到不同交换机模块250、251、252。如图所示，服务器210的一个适配器以通信方式耦合到位于交换机模块250中的网桥元件120，而另一个适配器连接到交换机模块251中的网桥元件120。如果交换机模块250、251中的一个发生故障，则服务器210仍能够经由另一个交换模块接入交换层130。然后可以更换(例如，热插拔)发生故障的交换机模块，这导致IOMC 255、256、257和网桥元件120更新路由表和查找表以便包括新交换模块上的硬件元件。

[0063] 图3示出根据在此描述的一个实施例的虚拟交换层。系统100和200中的每个子交换机254经由网状连接模式，使用交换层130连接到彼此。即，无论是否使用子交换机254，都可以将信元(即，数据分组)路由到位于任何其它交换机模块250、251、252上的另一个其它子交换机254。这可以通过直接连接子交换机254的每个网桥元件120来实现—即，每个网桥元件120具有到每个其它网桥元件120的专用数据路径。备选地，交换层130可以使用主干-叶体系架构，其中每个子交换机254(即，叶节点)连接到至少一个主干节点。主干节点将从子交换机254接收的信元路由到正确的主干节点，然后该正确的主干节点将数据转发到正确的子交换机254。但是，本发明并不限于用于互连子交换机254的任何特定技术。

[0064] 带宽增加

[0065] 图4示出根据本发明的一个实施例的能够增加带宽的图2的子交换机。如图所示，子交换机454(即，联网元件或设备)包括五个网桥元件420和三个PCIe端口422。但是，本发

明并不限于此，并且可以包括任意数量的网桥元件、PCIe端口或者用于不同通信协议的端口。备选地，子交换机454可以仅包括网桥元件420。网桥元件420可以包含一个或多个端口421，例如先前讨论的100千兆位端口或两个40千兆位端口。此外，本发明并不限于以太网通信协议，而是可以应用于具有多播功能的任何通信方法。

[0066] 网桥元件420还包括多播(MC)复制引擎419，其执行将在端口421处接收的多播数据帧转发到目的地计算设备必需的功能。一般而言，多播数据帧包括组ID。MC复制引擎419使用组ID查找该组的不同成员。通过这种方式，MC复制引擎419确定它应该创建多播数据帧的有效负载的多少个副本，并且应向何处发送这些副本。此外，本公开还可以应用于广播数据帧。在这种情况下，接收网桥元件420将数据帧转发到连接到分布式虚拟交换机180的每个计算设备。

[0067] 每个网桥元件420和PCIe端口422与传输层(TL)425关联。TL 425将网桥元件420和PCIe端口422接收的数据从其原始格式(即，以太网或PCIe)转换为通用数据分组—即，信元。TL 425还将从交换层130接收的信元转换回其相应的通信格式，然后将数据传输到相应的网桥元件420或PCIe端口422。网桥元件420或PCIe端口422然后将转换后的数据转发到连接的计算设备。

[0068] 集成交换机路由器(ISR)450连接到传输层，并且包括连接接口455(例如，焊线、插座、端口、电缆等)以便将信元转发到分布式交换机中的其它子交换机。在一个实施例中，子交换机454与TL 425具有相同数量的接口455，然而它具有的数量可以多于或少于子交换机454上的TL425数量。在一个实施例中，将连接接口455“分配”给TL 425和网桥元件420或PCIe端口422中的一个或多个。即，如果网桥元件420或PCIe端口422接收单播数据帧，则它将使用所分配的连接接口455将数据转发到交换层130。在一个实施例中，网桥元件420之一可以借用分配给另一个网桥元件420的连接接口455(及其缓冲区)来传输多播数据帧的副本。

[0069] 尽管未示出，但ISR 450可以包括纵横式交换机(crossbar switch)，其允许同一子交换机454上的网桥元件420和PCIe端口422直接共享信息。连接接口455可以连接到纵横式交换机以便促进子交换机之间的通信。此外，ISR 450的各部分可以不位于包括子交换机454的ASIC上，但可以位于子交换机(例如，在交换机模块上)的外部。

[0070] 图5A-5B示出根据本公开的实施例在图4的子交换机中执行带宽增加。在图5A中，数据路径510示出接收的多播数据帧的有效负载采取的路径。为了清晰起见，从图中省略所有其它网桥元件、PCIe端口和TL。以太网端口421从连接到分布式交换机180的计算设备(例如，服务器105)接收多播数据帧，其可以包含多播组ID。MC复制引擎419使用组ID确定需要有效负载的多少个副本。如图所示，MC复制引擎419在一次传输中将八个有效负载副本放到ISR 450中的八个有效负载缓冲区515中。例如，子交换机454具有总线，其能够使MC复制引擎419创建多播数据帧的有效负载的一个副本，该副本被同时复制到八个有效负载缓冲区515。注意，子交换机454具有使单个网桥元件使用与其它网桥元件420或PCIe端口422关联的缓冲区的能力。因此，子交换机454上的总线控制器(例如，硬件或固件)可以阻止其它TL(TL 425B-H)接入总线，并且允许TL 425A使用共享总线将有效负载同时复制到每个有效负载缓冲区515。当然，如果需要，则这可以按顺序执行。此外，控制器或TL 425A可以确定总线接入哪个缓冲区。例如，控制器可以允许TL 425A将有效负载仅复制到缓冲区子集而不是全

部缓冲区。

[0071] 图5B示出MC复制引擎419针对不同有效负载的副本创建标头。数据路径560A-H示出MC复制引擎419针对存储在有效负载缓冲区515中的每个有效负载副本575，创建八个唯一标头580A-H。一般而言，标头580A-H提供有效负载在多播组全体成员表中指定的目的地处结束必需的路由信息。ISR 450可以将标头580A-H与有效负载副本515组合以便创建信元，然后在交换层130中转发该信元。在一个实施例中，MC复制引擎419将标头580A-H传输到相应的标头缓冲区520，每次传输一个。即，MC复制引擎419仅传输一次有效负载，但可以单独创建每个标头。因为将每个信元发送到MC组全体成员定义的不同目的地，所以信元的定制标头580A-H包含不同目的地数据。

[0072] 尽管有效负载和标头缓冲区515、520被示为单独的存储单元，但在一个实施例中，它们可以是同一存储单元的不同逻辑分区。

[0073] 将有效负载复制到ISR 450的八个缓冲区中使带宽增加八倍。即，根据一个实施例，子交换机454可以使用多达八个并行传输数据帧的接口455，而不是仅使用连接接口455之一将多播数据帧转发到多播组的所有不同目的地。此外，假设以太网端口421和连接接口455具有相同带宽(例如，100gb/s)，则子交换机可以以它接收的带宽的大约八倍带宽转发数据帧。当然，具有更多(或更少)连接接口的子交换机将相应地更改可能的带宽增加。此外，子交换机454可以被配置为使用少于总数的连接接口455。因此，带宽增加八倍是最大值，但在其它实施例中，子交换机454可以使用少于八个的连接接口455来转发多播数据帧。

[0074] 图6示出根据本公开的一个实施例的使用数据帧块在图4的子交换机中执行带宽增加。MC复制引擎619可以将有效负载分成不同的块，而不是将接收的多播数据帧的整个有效负载复制到有效负载缓冲区中。使用图5A中所述的过程，TL 625A可以将单个块675的八个副本“快照”到每个有效负载缓冲区615中。针对有效负载的每个块675重复该过程，直至将所有有效负载块675A-C加载到有效负载缓冲区615中。如在此所示的，TL 625A将接收的有效负载分成三个块—有效负载块675A-C。因此，在三次传输中，每个有效负载缓冲区615包含三个有效负载块675A-C，它们对应于接收的数据帧的完整有效负载。

[0075] MC复制引擎619针对有效负载块675A-C的每一个创建不同的标头。因此，在一个实施例中，块675A-C可以使用通过交换层130的不同路径到达同一目的地。但是，在不同块675到达相同的最终目的地之后，标头680、685和690可以包含序列号，以便与目的地关联的TL可以重新组装块以形成有效负载。将接收的有效负载分为块并且针对每个块使用单独的数据路径可以提高分布式交换机180中的数据吞吐量。

[0076] 图6示出将有效负载块675和标头680、685、690存储在ISR 650中的两个不同实施例。对于这两个实施例，假设有效负载块675A-C是同一数据帧的块并且传输到同一目的地。对于与TL 625A关联的有效负载和标头缓冲区615、620，MC复制引擎619存储标头680、685、690以便与同一地址关联的有效负载块由同一连接接口655传输。如果根据图3中所示的模式组织分布式交换机—即，每个子交换机254连接到每个其它子交换机254—则有效负载块675A-C经过同一路径到达目的地子交换机。

[0077] 相反地，MC复制引擎619可以存储标头，以便从不同连接接口655传输针对同一目的地的有效负载块675。例如，MC复制引擎619可以将与块675A关联的标头680存储在与TL 625F关联的标头缓冲区620中，但将与块675A关联的标头685存储在与TL 625G关联的标头

缓冲区620中。因此，有效负载块675A和675B都将在同一目的地处结束，但可以使用不同连接接口655并且因此使用不同通信路径传输到目的地子交换机。此外，假设ISR 650可以以接收有效负载块的任何顺序传输有效负载块675，则可以经由与TL 625F-H关联的连接接口655同时传输有效负载块675A-C。与通过同一连接接口655按顺序传输有效负载块675相比，这可以是有利的。

[0078] 图7示出根据本公开的一个实施例的在交换机层上传输的信元。信元700包括标头部分705和有效负载750。有效负载750例如可以是多播数据帧的任何部分，例如以太网帧的有效负载(或整个有效负载)的一部分。标头705包括MC组标识符710、目的地ID 715、序列号720、代理级别725和源ID 730。标头705并不限于所示部分，而是可以包括更多或更少的数据。

[0079] MC组标识符710可以是包括在接收的多播数据帧中的同一组标识符，或者与数据帧中的组标识符关联。例如，多播数据帧中的组标识符可以用作到本地表的索引以便确定MC组标识符710。当在分布式交换机180 中转发信元700时，接收子交换机能够使用MC组标识符710标识组成员。

[0080] 目的地ID 715用于通过分布式交换机180路由信元700。目的地ID 715例如可以包括子交换机ID、网桥元件号、端口号、逻辑端口号等。MC复制引擎可以将该路由信息的部分或全部放在目的地ID部分715中。

[0081] 如果如图6中所述，将多播数据帧的有效负载划分成块，则使用序列号720。在信元700到达目的地之后，指定的TL可以使用序列号720重新组合有效负载750以便生成所接收的数据帧的原始有效负载。因此，在其中未划分原始有效负载的实施例中，可以省略序列号720。

[0082] 如果接收子交换机没有足够的连接接口将多播副本传输到MC组的所有成员，则将多播副本传输到中间(即，代理)子交换机时使用代理级别725。一般而言，分布式交换机180可以使用代理的层次结构将多播数据帧传播到所有成员。代理级别725向接收子交换机指示它是层次结构中处于什么级别。这将在下面更详细地讨论。

[0083] 源ID 730与目的地ID 715一样，可以包括子交换机ID、网桥元件号、端口号、逻辑端口号等。在一个实施例中，源ID 730可以用于确保未将多播副本传输到当前正在传输信元700的同一子交换机。这可以防止循环。

[0084] 图8是根据本公开的一个实施例的带宽增加技术800。在步骤805，子交换机上的网桥元件从连接的计算设备(例如，服务器105)接收多播数据帧。通信协议可以是以太网、IP、InfiniBand，或者具有多播/广播(即，一对多)能力的任何其它通信协议。注意，InfiniBand是InfiniBand贸易协会的注册商标。

[0085] 在一个实施例中，在步骤810，网桥元件中的MC多播引擎可以将数据帧的有效负载分成不同块，但这并非必须。

[0086] 在步骤815，在一次传输中，通过借用分配给其它网桥元件或不同通信协议(即，PCIe)的TL和缓冲区资源，与接收多播数据帧的网桥元件关联的TL可以使用总线将一个有效负载块的副本快照到多个有效负载缓冲区。如图5-6中所示，在一次传输中，将八个副本同时加载到有效负载缓冲区中。在步骤820，TL可以在后续传输中将其余块传输到有效负载缓冲区中。例如，在图6中，TL 625A使用三次传输将块675A-C存储到有效负载缓冲区615中。

[0087] 在步骤825，MC复制引擎针对每个传输的块生成标头。例如，去往同一目的地的块的标头可以相同，但向接收网桥元件通知块排序的序列号除外。在一个实施例中，TL使用单独传输将定制标头放置到所借用的标头缓冲区中。注意，可以在TL将有效负载块快照到有效负载缓冲区之前、之间或之后，将标头存储在标头缓冲区中。例如，帧的有效负载的块数量可能超出有效负载缓冲区的大小，因此MC复制引擎可以将块传输到有效负载缓冲区中直至缓冲区已满，生成定制标头，并且允许ISR传输组合信元，然后再次将其余有效负载块存储在有效负载缓冲区中并且生成附加标头。

[0088] 在步骤830，ISR将来自有效负载缓冲区的有效负载块与来自标头缓冲区的其对应标头相组合，并根据目的地ID转发结果信元。在目的地TL处接收所有不同块之后，可以从多个接收的信元重构多播数据帧，并将多播数据帧的完整有效负载转发到连接到分布式交换机180的计算设备。

[0089] 在一个实施例中，ISR可能不立即逐出已转发到交换层的块。例如，子交换机上的控制器可以检测到正在使用连接接口之一传输高优先级数据，因此不能被接收多播数据帧的网桥元件借用。在这种情况下，控制器可以限制哪些有效负载缓冲区经由共享总线接收数据块。例如，子交换机可能需要将多播数据帧的副本发送到八个MC成员，但仅具有七个可用连接接口。使用七个接口将帧并行发送到七个成员之后，MC复制引擎可以生成取代一个或多个块的原始标头的一个或多个替换标头，而不是再次将块传输到有效负载缓冲区。具体地说，这些替换标头包括与在原始标头中发现的目的地不同的目的地。然后可以组合块和替换标头以便形成新信元，将该新信元转发到最后一个(即，第八个)目的地。因此，通过不立即逐出转发的块，子交换机可避免将数据块从TL重新传输到有效负载缓冲区。

[0090] 代理的层次结构

[0091] 上一部分中讨论的带宽增加可以扩展，并有利地用于继续增加MC组的带宽，这些MC组超出子交换机上的连接接口数量。即，如果图4的子交换机454需要将多播数据帧的副本发送到多于它具有连接接口的MC组成员，则子交换机仍被限于连接接口455的组合带宽(例如，8×100gb/s)来传输数据帧的有效负载。但是，使用代理子交换机(或者代理网桥元件)的层次结构允许分布式交换机继续随着MC组中的成员增加而扩展带宽。即，如果接收端口为100gb/s并且必须将多播数据帧发送到30个目的地，则分布式交换机可以使用大约为30×100gb/s的组合带宽传输多播数据帧的副本。

[0092] 图9是根据本发明的一个实施例的使用分布式交换机互连的计算系统。计算系统900包括一个或多个机架(机架1-N)，每个机架包含一个或多个机箱(机箱1-N)。为了促进可以包含在机箱1-N中的不同计算设备之间的通信，计算系统900可以使用多个子交换机1-N。具体地说，可以使用图1-2中所示的分布式交换机180来互连系统900中的多个不同计算设备。为了清晰起见，仅示出子交换机(即，包含如图4中所示的网桥元件的微芯片)。在一个实施例中，每个子交换机连接到每个其它子交换机。即，每个子交换机具有至少一条导线，该导线将其直接连接到每个其它子交换机，即使该子交换机在不同机架上。尽管如此，执行在此公开的实施例并非必需该设计。

[0093] 图10是根据本发明的一个实施例的用于转发多播数据帧的代理的层次结构。为了继续随着组成员增加而扩展带宽，计算机系统900可以建立层次结构。如图所示，针对具有136个不同子交换机的分布式交换机建立层次结构1000，其中每个子交换机具有八个连接

接口(例如,图4中所示的子交换机454)。层次结构1000分为四个级别(不包括接收多播数据帧的Rx子交换机)。分布式交换机中的所有子交换机可以分为四个组。但是,层次结构1000的级别和组成员的数量是任意的,并且例如可以取决于子交换机的总数、子交换机上的端口/连接接口的数量以及子交换机的体系架构。例如,仅具有20个子交换机的分布式交换机可能需要仅具有一个代理级别的层次结构。相反,如果每个子交换机具有可以用于并行转发分组的135个端口,则可能不需要层次结构。相反,子交换机可以通过仅使用必需数量的端口将多播数据帧并行转发到多达135个子交换机,增加用于传输多播数据的带宽。但是,使用层次结构1000可以通过允许分布式交换机容纳更多数量的子交换机而降低成本,以及增加带宽而不必使用具有更多端口的子交换机。

[0094] 示出层次结构1000以便将子交换机分配给多个代理。级别A代理—即,层次结构1000的顶级—具有四个选择的代理子交换机,或者更具体地说,可能位于也可能不位于不同子交换机上的四个代理网桥元件。为每个级别A代理分配一组子交换机。在图10中,该组由直接在包含级别A代理的方框下面的子交换机定义。即,将级别A代理1分配给子交换机0:35,将代理14分配给子交换机36:71,依此类推。因此,当接收子交换机(即,RX子交换机)接收多播数据帧时,它使用标识MC组成员的MC组表。从该信息,RX子交换机标识子交换机0:135中的哪些需要接收数据帧。如果全体成员包括组0:35中的子交换机,则RX子交换机将数据帧转发到代理1。如果0:35中的子交换机均不在MC组的全体成员中,则RX子交换机不将数据帧转发到代理1。

[0095] 假设子交换机0:35中的至少一个是MC组的成员,则当在代理1处接收分组时,可以执行类似的分析。代理1子交换机查找组全体成员,并且确定哪个级别B代理应该接收分组。为分配给级别A代理1的子交换机的子集分配级别B代理2-4。即,将代理2子交换机分配给子交换机0:11,将代理3分配给子交换机12:23,并且将代理4分配给子交换机14:35。如果组全体成员包括这三个组的每一个中的子交换机,则代理1将分组的副本转发到代理2-4。

[0096] 级别B代理还查看层次结构1000和组全体成员以便确定哪个级别C代理应该接收分组。尽管未显式示出,但将代理5分配给子交换机0:3,将代理6分配给子交换机4:7,依此类推。因此,如果子交换机1是MC组的成员,则级别C代理5将接收分组并且将其转发到子交换机1。

[0097] 在一个实施例中,从可能的目的地子交换机(即,层次结构的级别D)中选择代理子交换机。即,代理子交换机可以是子交换机0:135中的一个。此外,可以从代理被分配给的子交换机组中选择代理。例如,代理1可以是0:35中的子交换机之一,而代理5可以是组0:3中的子交换机之一,依此类推。但是,在另一个实施例中,可以从未在分配给代理的子交换机组中的子交换机来选择代理。

[0098] 备选地,代理子交换机可以不是目的地子交换机。例如,分布式交换机可以包括如下子交换机:其角色是仅用作用于转发多播业务的代理。或者,子交换机的未连接到任何计算设备的网桥元件或PCIe端口—即,多播数据帧的最终目的地—可以被选择作为代理。因此,尽管子交换机上的一个或多个网桥元件可以连接到计算设备,但是可以选择该子交换机上的未连接的网桥元件作为代理。将在后面更详细地讨论在子交换机中选择代理子交换机和代理网桥元件/TL。

[0099] 图11是根据本发明的一个实施例的在图10中示出的层次结构的一部分的系统图。

部分层次结构1100示出图10的层次结构1000的四个级别的每一个中的一个子交换机。每个子交换机可以类似于在图4-6中公开的子交换机。如图所示，RX子交换机1105在一个网桥元件420的入口端口上接收多播数据帧。使用图8中所示的过程，TL 125使用ISR 450并行转发数据帧的多达八个有效负载副本，从而相对于入口端口的带宽实现多达八倍带宽增加(假设连接接口455具有与入口端口相同的带宽)。即使带宽不同，Rx子交换机1105也相对于如下系统实现多达八倍带宽增加：该系统仅使用连接接口455之一转发多播数据帧的副本，而不是并行使用全部八个连接接口。

[0100] 配置每个子交换机1105、1110、1115和1120以便三个或四个连接接口455将数据帧的有效负载转发到代理子交换机，而保留另外四个连接接口以便将有效负载转发到位于子交换机上的网桥元件—即，本地网桥元件。使用Rx子交换机1105作为一个实例，最右侧四个连接接口455(以及关联的有效负载和标头缓冲区)专用于子交换机1105上的四个最右侧网桥元件420。因此，如果四个最右侧网桥元件420中的一个连接到作为多播数据帧目的地的计算设备，则使用最右侧连接接口455中的一个将有效负载传输到对应的网桥元件120。这可以通过ISR 450中的路由机制(例如纵横式交换机)执行。因此，ISR 450可以具有如下能力：在位于同一子交换机上的源和目的地网桥元件之间路由数据，而不使用连接到另一个子交换机的连接接口455。但是，如果Rx子交换机1105的其它四个本地网桥元件420未连接到是MC组成员的计算设备，则不使用四个最右侧连接接口455。

[0101] Rx子交换机1105使用四个最左侧连接接口455将多播数据帧的有效负载转发到多达四个级别A代理。为了简洁起见，仅示出级别A代理之一(即，子交换机1110)。为了转发数据帧的副本，MC复制引擎将使用组ID(其可以从多播数据帧的各部分以及子交换机上的接收端口配置中获得)标识组全体成员，然后它使用组全体成员确定哪个级别A代理需要数据帧的有效负载副本。使用连接接口455之一，子交换机1105的ISR 450将包含有效负载的信元传输到级别A子交换机1110。

[0102] 所属技术领域的普通技术人员应该理解，可以配置用于代理和本地网桥元件通信的连接接口455(及其关联的资源)的数量。例如，可以使用两个连接接口与四个本地网桥元件通信，这将保留六个连接接口以便与代理或目的地子交换机通信。相反，可以仅使用两个连接接口455与代理通信，而针对本地网桥元件保留六个连接接口。如果子交换机上具有七个网桥元件420而不是五个，则可以首选该配置，以便每个本地网桥元件420具有对应的连接接口455。此外，这可以影响层次结构1000，因为每个级别的代理数量减少。

[0103] 使用图8中所示的方法，级别A子交换机1110接收信元，TL 425将有效负载复制到ISR 450中的八个标头缓冲区中，并且接收网桥元件420中的MC复制引擎针对每个不同有效负载块生成标头。即，级别A子交换机1110与Rx子交换机1105执行非常类似的过程以便实现进一步带宽增加—即，通过子交换机1105和1110之间的连接传输的信元将被重新产生并且通过多达其它七个连接传输。注意，在图11中，用于接收信元的连接接口455也不用于将信元传输到代理子交换机或本地网桥元件420。但是，这并非必须。在一个实施例中，接收信元的连接接口455还可以用于转发信元，但该连接可以比其它七个接口455慢，因为它可能争用资源，这些资源用于存储和管理包含不同数据帧有效负载块的另外接收的信元。

[0104] 在一个实施例中，可以向接收子交换机通知它在层次结构中的哪个级别。即，因为子交换机例如可以是级别A和级别C代理两者，所以当子交换机将信元转发到代理时，它可

以在信元的标头中包括代理级别。如图7中所示，标头705包括代理级别725部分。在这种情况下，Rx子交换机1105中的MC复制引擎可以处于子交换机1110被用作级别A代理的代理级别725。当然，除了将代理级别信息放在标头中之外，子交换机还可以使用不同的方法。例如，Rx子交换机1105可以发送包括级别信息的特殊分组。备选地，子交换机1110可以例如使用分组ID查询主控制器或数据库，以便确定代理级别。

[0105] 借助代理信息，子交换机1110中的MC复制引擎通过查看图10中的层次结构1000，确定哪些级别B代理应该接收信元。此外，MC复制引擎可以判定多播数据帧的任何目的地是否连接到位于子交换机1110上的网桥元件420。如果是，则使用专用于本地网桥元件420的四个连接接口455，将信元转发到这些网桥元件420。

[0106] 如图所示，向至少一个级别B代理(即，子交换机1115)转发数据帧的有效负载。因此，MC复制引擎针对有效负载块生成新标头，并且将一个或多个结果分组转发到子交换机1115。

[0107] 与子交换机1105和1110一样，子交换机1115可以使用图8中的方法以便实现多达八倍的带宽增加。如图所示，级别B子交换机1115从子交换机1110接收数据分组，使用组ID号标识MC组全体成员，并且基于组全体成员，使用层次结构信息1000将多播数据帧的有效负载副本传输到级别C子交换机1120。

[0108] 级别C子交换机1120也可以执行上述带宽增加。在所示层次结构1100 中，多播数据帧在到达目的地子交换机1125之前，通过最多三个代理。当然，如果目的地计算设备连接到代理子交换机1110、1115或1120的本地网桥元件420之一，则使用专用于本地多播业务的连接接口455之一，将分组传送到本地网桥元件420。如果否，如图11中所示，经由通过所有三个代理层的后续信元路由多播数据帧的有效负载，直至有效负载到达目的地子交换机1125。

[0109] 级别C子交换机1120可以将信元路由到目的地(即，级别D)子交换机1125的连接接口455，该连接接口455与直接连接到目的地计算设备(即，服务器1130)的网桥元件420关联。具体地说，目的地子交换机1125在代理网桥元件上接收一个或多个信元。如在此所示的，这是最右侧网桥元件。但是，如果该网桥元件未连接到目的地计算设备，则它使用ISR 450将信元路由到正确的网桥元件。例如，为了节省存储器，级别C子交换机1120可能不知道目的地子交换机1125的哪些本地端口连接到目的地计算设备。相反，它仅知道代理网桥元件的位置，当它接收信元时，该代理网桥元件将数据传输到正确的本地网桥元件(即，从左侧开始的第三个网桥元件)。TL 425然后可以将一个或多个接收的信元转换回单个帧(例如，以太网帧)，该帧与在Rx子交换机1105处接收的多播数据帧具有相同的有效负载。最后，直接连接到目的地计算设备的网桥元件420使用其出口端口，将数据帧传输到目的地计算设备—即，服务器1130。

[0110] 通过这种方式，可以基于MC组中的成员数量，增加用于通过分布式交换机传输多播数据帧的有效负载的带宽。注意，层次结构随着MC组中的成员数量增加而增加带宽的能力可能不受代理数量的限制。例如，为了使带宽随着成员增加而直接增加，层次结构必须具有足够数量的代理子交换机和/或代理级别。如果代理的总数有限或者使用层次结构的太少级别，则带宽仍可以随着组成员增加而扩展，但结果带宽可能少于具有必需数量的代理的系统—即，带宽可能具有上限。这将在下一部分中更详细地讨论。

[0111] 图12示出根据本发明的一个实施例的在图10中示出的层次结构中的多播数据帧的一个实例路径。当Rx子交换机1205接收多播数据帧时，接收网桥元件使用MC组ID在MC组表1210中搜索对应的MC组全体成员1215。表1210可以位于Rx子交换机的存储器中或者分布式交换机中的其它位置。

[0112] 对于在此接收的特定多播数据帧，MC组全体成员1215包括子交换机0:35以及子交换机37。Rx子交换机1205负责确保MC组全体成员1215中的每个子交换机接收已接收的多播数据帧的有效负载副本。备选地，MC组全体成员1215可以列出要接收数据帧副本的不同计算设备，而不是列出表1210中的目的地交换机，或者接收的多播数据帧的标头可以包含目的地计算设备列表(例如，IP或MAC地址列表)。在这些情况下，分层数据1220可以包含查找表，其向子交换机1205通知哪些子交换机连接到目的地计算设备。使用该信息，子交换机然后可以标识正确的目的地子交换机。

[0113] 此外，Rx子交换机1205可以使用分层数据1220确定哪些代理应该接收副本，或者目的地计算设备是否全部连接到Rx子交换机1205，哪些本地网桥元件应该接收多播数据帧的有效负载副本。使用图10中所示的层次结构1000，将子交换机0:35分配给级别A代理1，而将子交换机37分配给级别A代理14。因此，使用至少两个连接接口，Rx子交换机1205将信元转发到在分层数据1220中标识的两个代理。

[0114] 在一个实施例中，可以针对每个子交换机具体调整层次结构。即，一个子交换机的级别A代理可以不同于另一个子交换机的级别A代理。这在不同的子交换机中分发转发分组的职责。例如，分布式交换机可以根据预定规则选择代理，这些规则例如包括可以仅将子交换机分配为最大数量的子交换机的代理，或者代理不能是同一子交换机的级别A和B代理(以便防止循环)。基于所述规则，分布式交换机可以针对每个子交换机或一组子交换机提供定制层次结构。在使用定制层次结构的分布式交换机中，标头可以包含诸如代理级别和源ID之类的信息，这能够使每个代理子交换机确定要使用哪些层次结构来转发分组。

[0115] 一旦级别A代理接收分组，使用分层数据(其可以存储在本地)，它们确定哪些级别B代理必须接收分组以便多播数据帧的有效负载到达在MC组全体成员1215中列出的所有目的地。在这种情况下，代理子交换机1使用专用于将分组传输到其它代理级别的三个连接接口，将分组转发到代理2、3和4。相反，代理14将分组转发到其三个级别B代理的仅一个—即，代理15。如前所述，在一个实施例中，代理子交换机可以包含连接到目的地计算设备的本地网桥元件。例如，代理14实际上可以是子交换机37。在这种情况下，在代理14接收分组之后，它将使用其连接接口之一将分组转发到本地网桥元件，本地网桥元件然后将分组传输到目的地计算设备。因此，不需要通过层次结构的其余部分。

[0116] 但是，在其它实施例中，分布式交换机可以包括专用代理子交换机和/或网桥元件。即，网桥元件可以专用于作为代理以便接收转发的分组，然后将这些分组的副本分发给内部连接接口。此外，专用网桥元件(或者整个子交换机)可能不连接到任何其它计算设备或外部网络，例如服务器或WAN。通过这种方式，分布式交换机确保网桥元件的硬件资源始终可用。

[0117] 假设所有级别B代理接收分组，则它们将具有新标头的有效负载转发到正确的级别C代理。在这种情况下，代理2可以将分组并行传输到代理5、6和7，代理3将分组传输到代理8、10和11，依此类推。最后，级别C代理将分组传输到目的地或发送(Tx)子交换机，或者更

具体地说，传输到Tx子交换机的网桥元件，该网桥元件具有连接到是MC组全体成员1215一部分的目的地计算设备的端口。因此，代理5将分组转发到子交换机0、1和2，代理6将分组转发到子交换机3、4和5，依此类推。但是，如果转发分组的任何一个代理是子交换机0-35—即，代理也是目的地—则级别C代理不需要将分组转发到该目的地。例如，如果将子交换机分配为充当其所分配到的同一组子交换机的代理，则可以将其从层次结构1000的第四级别中移除以便防止将两个分组发送到同一目的地。

[0118] 在一个实施例中，可以分配子交换机之一上的控制器(例如，选择作为主控制器的IOMC)以便建立一个或多个层次结构。该控制器可以持续监视分布式交换机的光纤通道网络以便确定哪些计算设备连接到不同子交换机的网桥元件。当连接更改时，控制器可以更新每个子交换机上的分层数据1220。在计算设备连接到不同子交换机(采用任何所需方式)之后，并且在分布式交换机通电之后，控制器可以检测当前配置，并且生成一个或多个层次结构。此外，如果移除或更改计算设备或子交换机，或者插入新计算设备或子交换机，则控制器可以动态检测这些更改并且基于不同配置生成新层次结构。

[0119] 在一个实施例中，控制器可以基于性能度量选择代理。例如，控制器可以使用当前未连接到计算设备的子交换机之一上的网桥元件作为代理。备选地，控制器可以针对网桥元件或子交换机监视流过网桥元件端口的网络业务、在端口中流动的特定类型业务、转发接收数据分组的响应时间等。基于该度量，控制器可以选择例如经历最少量多播业务的网桥元件作为代理。

[0120] 在一个实施例中，在选择代理网桥元件之前，控制器可以评估子交换机上的其它网桥元件或PCIe端口。如前所述，带宽增加借用与子交换机中的这些其它硬件资源关联的缓冲区和连接接口。如果这些对等资源接收或发送高优先级网络业务，则在这些子交换机上执行带宽增加可以降低高优先级网络业务的吞吐量，因为它们的已分配资源被借用以便转发重复的多播分组。因此，传输高优先级网络业务的子交换机以及这些子交换机上的网桥元件可能失去选择作为代理的资格。

[0121] 动态优化层次结构以便提供冗余并且优化性能

[0122] 除了使用上面讨论的层次结构增加分布式交换机中的可用带宽之外，还可以基于优化准则动态更改层次结构，这些准则例如包括从故障中恢复或者减少在代理之间流动的数据。

[0123] 图13示出MC组表。如图所示，MC组表1210包括多播组ID 1305、代理级别1310、主干模式1315、优化启用1320、子交换机掩码1325以及本地端口掩码1330。

[0124] MC组ID是网桥元件用于索引到表1210的号码。例如，网桥元件可以使用接收的多播数据帧的一个或多个部分获得MC组ID，该MC组ID 对应于存储在MC组表1210中的MC组ID 1305之一。网桥元件在表1210中标识正确的行之后，它可以使用子交换机掩码1325确定分布式交换机中应该接收多播数据的子交换机。

[0125] 在一个实施例中，子交换机掩码1325是位向量，其中每个位对应于交换机中的子交换机之一。位的值(即，1或0)确定对应的子交换机是否是多播数据的目的地。然而，表1210并不限于用于指定哪些子交换机是MC组成员的任何特定方法。当MC组全体成员更改时，控制器(例如，主IOMC)可以负责生成和更新每个MC组表1210中的子交换机掩码1325。

[0126] 本地端口掩码1330指定子交换机上的哪个本地网桥元件端口应该接收多播数据

帧的副本。在一个实施例中，分布式交换机中的每个子交换机与它自己的表1210关联。此外，本地端口掩码1330可以包含仅用于该子交换机上的本地端口的数据。即，表1210仅包含子交换机确定其哪个本地端口(即，网桥元件之一的本地端口)连接到目的地计算设备必需的信息。为了节省存储器，特定子交换机的表1210可能不包含分布式交换机中的任何其它子交换机的本地端口信息1330。但是，这并非必须，因为表1210可以包含分布式交换机中的一个或多个其它子交换机的本地端口信息。

[0127] 代理级别位1310设置用于每个多播组的层次结构类型。每种类型的层次结构可以基于层次结构中的代理级别数量而变化。例如，层次结构可以是一级层次结构、二级层次结构、四级层次结构等。当然，当分布式交换机中的子交换机数量增加或减少时，层次结构中的不同可能代理级别也可以增加或减少。

[0128] 一级层次结构仅使用一个级别的代理来将数据帧分发到目的地子交换机。例如，接收子交换机可以将多播数据转发到其四个本地端口以及四个或更多代理。这些代理然后负责将多播数据发送到正确的目的地计算设备。假设子交换机包含136个目的地子交换机，则该类型的层次结构(不同于四级层次结构)不会确保带宽随着目的地子交换机数量以一对一关系扩展。即，一级层次结构仍基于目的地子交换机数量增加可用带宽，但它可能是小于一对一比率。例如，代理可能必须使用同一连接接口将多播数据按顺序传输到多个不同目的地端口。但是，如果MC组仅具有几个成员(例如，少于8个子交换机)，则控制器可以设置有关一级层次结构的代理级别1310。这将平衡通过使用代理子交换机增加带宽的需要与增加的延迟，该延迟可以因以下操作产生：借用代理上的连接接口以便传输多播数据，这可以防止传输其它数据。

[0129] 二级层次结构使用两个级别的代理子交换机来传输多播数据。例如，为了实现全带宽，接收子交换机可以将多播数据传输到四个代理子交换机，每个代理子交换机然后将多播数据传输到另外四个代理。与一级层次结构一样，如果MC组的全体成员太多，则该层次结构可能不基于目的地子交换机数量以一对一关系增加带宽。取决于层次结构和MC组全体成员，可用带宽可以以小于一对一比率扩展。即，带宽可能不按照目的地子交换机数量的倍数增加。

[0130] 在另一个实例中，为了实现半带宽，接收子交换机可以将多播数据传输到十二个代理子交换机，每个代理子交换机然后将多播数据传输到另外十一个代理。因此，接收子交换机的被指定将多播数据转发到代理的连接接口之一可能必须将数据按顺序转发到三个不同代理。

[0131] 四级层次结构在图10中示出，将不会在此详细讨论。使用四级层次结构将确保即使多播数据帧是广播(即，应该将多播数据帧传输到所有子交换机)，可用带宽也随着目的地子交换机数量大约一对一扩展。即，分布式交换机中的可用带宽以近似目的地子交换机数量倍数的比率增加。

[0132] 主干模式1315和优化启用1320位将在本文档的后面讨论。

[0133] 在使用MC组ID确定要使用哪种类型层次结构以及哪些子交换机应该接收多播数据帧之后，接收子交换机可以使用分层数据1220将多播数据转发到代理或目的地子交换机。此外，子交换机可以使用本地端口掩码1330确定接收子交换机上的哪些本地网桥元件(如果有)应该接收分组。

[0134] 图14示出分层数据1220。分层数据1220包括如图10中所示的层次结构1000。使用子交换机掩码1325，子交换机确定将哪些代理分配给目的地子交换机。例如，如果接收子交换机是图10中所示的Rx子交换机，并且子交换机掩码1325列出子交换机0和36作为MC组成员，则子交换机将多播数据转发到代理1和代理14两者。

[0135] 使用代理标识寄存器1405，接收子交换机上的MC复制引擎之一确定不同代理的子交换机ID。继续上面的实例，MC复制引擎通过寄存器1405进行解析直至它发现代理1和14。关联的主子交换机ID 1410和网桥元件ID 1415(即，主子交换机上的本地网桥元件)提供路由信息，然后将该路由信息放在标头中以便将多播数据路由到代理1。因此，控制器可以很容易更改将哪个子交换机分配给层次结构中的特定代理，方法是更改与该代理关联的子交换机ID。如图所示，寄存器1405针对不同层次结构类型(即，一级、二级和四级层次结构)中的每个可能代理包含一个项。此外，每个代理可以具有备用子交换机，以防主子交换机发生故障或者被从系统中移除。在MC复制确定主子交换机不可用于充当代理之后，它使用备用子交换机ID 1420和网桥元件ID 1425(即，备用子交换机上的本地网桥元件)将多播数据路由到该代理的备用子交换机。

[0136] 在代理子交换机接收多播数据之后，它可以通过使用它自己的本地MC组表1210标识目的地子交换机来执行类似的过程，并且按照分层数据1220的指示判定是否将多播数据转发到其本地端口和/或另一个代理级别。例如，所接收的多播数据可以在其标头中包括信息，该信息标识被使用的层次结构类型(即，代理级别数量)以及当前代理级别。使用该信息，每个代理可以独立于其它代理操作。在控制器或固件针对分布式交换机中的每个子交换机填充MC组表1210和分层数据1220之后，每个代理即使在操作中断期间也可以独立操作。此外，该体系架构避免需要集中式硬件或固件必须确定路由以便传输不同多播信元。相反，每个代理子交换机具有路由多播数据的必需信息。

[0137] 图15A-C示出用于处理操作中断的系统和技术。图15A示出使用四级层次结构将多播数据转发到目的地子交换机的分布式交换机的一部分。在此，接收子交换机(子交换机6)在其网桥元件120之一的入口端口上接收多播数据帧。基于MC组全体成员和分层数据，子交换机6沿着数据路径1505将具有多播数据的多播数据转发到子交换机5(级别A代理)。子交换机5执行类似的分析，并且沿着数据路径1510将多播数据转发到级别B代理子交换机4。在使用其MC组表1210和分层数据1220的本地副本之后，子交换机4沿着数据路径1515将多播数据转发到级别C代理子交换机3，级别C代理子交换机3使用数据路径1520将多播数据转发到级别D代理子交换机2。因为子交换机2是MC组的目的地之一，所以它使用本地端口掩码1330标识正确的网桥元件，并且沿着数据路径1525将多播数据传输到网桥元件。尽管未示出，但本地网桥元件使用本地端口将结果数据帧传输到目的地计算设备。

[0138] 注意，数据路径1505-1520穿过交换机光纤通道网络，而数据路径1525可以是子交换机2中的本地传输。

[0139] 图15B示出与图15A中所示相同的系统，但操作中断除外。在这种情况下，子交换机4暂时不可用、被移除或发生故障。备选地，数据路径1510可能被断开连接或切断。在任何一种情况下，中断阻止子交换机5将多播数据转发到子交换机4。对于这种情况，分布式交换机可以包括硬件通知系统，在它检测到子交换机不可用之后，将广播消息传输到分布式交换机中的所有子交换机。基于该通知，当传输到级别B代理时，子交换机5使用图14中所示的备

用子交换机ID 1420和网桥元件ID 1425。在这种情况下，子交换机ID 1420是子交换机1。在子交换机1沿着数据路径1530接收多播数据之后，它确定其是哪个代理级别，并且基于分层数据，使用数据路径1535将多播数据转发到子交换机3(即，级别C代理)。

[0140] 备选地，因为每个代理子交换机可以独立操作，所以每个子交换机的分层数据1220可以不同。即，子交换机1可以将多播数据转发到不同级别C代理。只要将该不同代理和代理子交换机3分配给层次结构中的同一组目的地子交换机(即，包括子交换机2的组)，则多播数据将到达其正确目的地。如果子交换机3和4布置在分布式交换机中，则可以首选在使用备用级别B代理时使用不同的级别C代理，以便如果一个子交换机发生故障，则其它子交换机也可能不可用。控制器可能先前配置子交换机1的分层数据1220以便使用不同于子交换机3的级别C代理—即，不同主交换机ID 1410和网桥元件ID 1415，而不是要求子交换机1尝试将多播数据转发到子交换机3。

[0141] 最后，控制器可以更新子交换机5的层次结构数据1220，以便提供不同主子交换机以替换子交换机4。

[0142] 图15C示出用于处理中断的技术。在步骤1550，子交换机接收多播数据。所述数据可以是来自连接的计算设备的接收的多播数据帧，或者是从分布式交换机中的另一个子交换机接收的多播数据。

[0143] 在步骤1555，接收子交换机上的接收网桥元件例如使用图13的子交换机掩码1325，确定MC组全体成员。在标识目的地子交换机之后，网桥元件可以确定要使用什么类型的层次结构转发多播数据。此外，多播数据可以向接收网桥元件通知层次结构的当前级别—即，子交换机是否从较高级别代理接收分组—因此接收子交换机知道要引用层次结构的哪个部分。

[0144] 在步骤1560，网桥元件将组全体成员与层次结构(例如，树结构)相比较，以便确定哪些代理应该接收多播数据。即，如果网桥元件确定它从较高级别代理接收多播数据，则它评估较低级别代理以便确定其中哪些代理应该接收多播数据的副本。

[0145] 使用图10作为参考，假设接收子交换机是代理3。该代理然后确定MC组的哪些成员也在分配给它的子交换机组(即，子交换机12:23)中。如果子交换机12:23是MC组成员，则在代理3下面的每个级别C代理(即，代理8:10)接收多播数据。但是，如果仅子交换机12:16在MC组中，则仅代理8和9从代理3接收多播数据。

[0146] 在标识较低级别代理之后，接收子交换机使用代理标识寄存器1405确定将多播数据路由到标识的代理所需的位置信息。具体地说，位置信息可以包括主子交换机ID 1410和网桥元件ID 1415。

[0147] 在转发多播数据之前，在步骤1570，子交换机可以判定它是否接收到指示预定代理(或多个代理)不可用的通知。但是，本公开并不限于判定网络一部分是否经历中断的任何特定方法。例如，子交换机可以首先传输多播数据而不判定目的地子交换机是否可用。但是，如果子交换机随后确定未接收多播数据(例如，未接收到确认信号)，则它可以推断具有系统中断。

[0148] 如果主代理执行功能，则在步骤1575，子交换机将多播数据传输到在子交换机ID 1410和网桥元件ID 1415中列出的代理子交换机和网桥元件。

[0149] 如果未执行功能，则在步骤1580，子交换机将多播数据传输到在备用交换机ID

1420和网桥元件ID 1425中列出的备用代理子交换机和网桥元件。

[0150] 图16A-D示出用于优化层次结构的系统和技术。图16A类似于图12，只是更改MC组全体成员1615以便包括子交换机0:25和子交换机37。基于图10中所示的层次结构1000，为了将多播数据传送到MC组的每个子交换机成员，Rx子交换机1605将多播数据转发到代理1和14两者。这些代理子交换机转而将多播数据转发到适当的级别B代理，依此类推。但是，这可以导致将多播数据不必要地传输到一个或多个代理。

[0151] 图16B示出图16A中所示的子交换机优化层次结构的结果。具体地说，图16B中的多播数据的传播路径避免到代理的不必要中间传输。如图所示，沿着数据路径1650将多播数据直接转发到子交换机37，而不是将多播数据按顺序从代理14转发到代理15，然后转发到代理18。因为代理可以用于使用两个连接接口并行传输多播数据以便增加带宽，所以当它们仅使用一个连接接口时，可以通过跳过代理改进延迟。因此，可以跳过将多播数据传输到仅一个目的地的每个代理。

[0152] 数据路径1655示出优化层次结构的另一个位置。当评估将多播数据转发到哪些级别A代理时，Rx子交换机1605可以使用层次结构数据1220确定因为代理1需要将数据转发到多个代理，所以不能跳过代理1。因此，Rx子交换机1605将多播数据转发到代理1。但是，因为代理4仅将数据转发到一个代理(即，代理11)，所以代理1可以跳过代理4并且将多播数据直接传输到代理11。此外，代理1可以确定不能跳过代理11，因为代理11负责将多播数据传送到两个目的地交换机—即，子交换机24和25。因为每个代理独立操作并且可以访问至少在当前级别下面的分层级别的分层数据，所以代理仍可以根据目的地子交换机数量而增加可用带宽，以及避免将多播数据传输到仅使用一个连接接口转发多播数据的代理而产生的某种不必要延迟。

[0153] 在一个实施例中，用于优化不同子交换机的能力是可能的，因为每个子交换机或者至少每个代理子交换机包含在其当前级别下面的层次结构级别的层次结构数据。因此，子交换机能够使用层次结构的树结构确定每个较低级别代理必须将多播数据转发到多少个子交换机。

[0154] 图16C示出可以执行的另一个优化。具体地说，图16C通过标识未使用的连接接口，优化图16A中所示的系统。如图11中所示，Rx子交换机1105可以分配四个连接接口455以便将多播数据转发到四个其它子交换机。基于该示出的布置，在图16A和16B中，Rx子交换机1605仅使用四个分配的连接接口中的两个将多播数据转发到其它子交换机。相比之下，图16C中的Rx子交换机1605使用全部四个连接接口，从而避免将多播数据传输到代理1。

[0155] 在Rx子交换机1605标识代理之后，它确定可用连接接口的总数。子交换机1605然后判定所标识的代理之一是否将多播数据传输到小于或等于可用连接接口加上为该代理分配的连接接口的数量的多个子交换机。在此，Rx子交换机1605具有两个可用连接接口加上被分配为将数据传输到代理1的连接接口。因为代理1将多播数据转发到仅三个其它子交换机(代理2、3和11)，所以Rx子交换机1605可以实际上将多播数据直接转发到这三个子交换机。

[0156] 类似于图16B中所示的优化，图16C中所示的优化也跳过仅传输到一个其它子交换机的代理。即，因为Rx子交换机1605具有分配给代理14的一个连接接口，并且基于上面表达的关系，代理14仅将多播数据传输到一个其它子交换机(即，代理15)，所以子交换机1605确

定它可以直接传输到代理15。当向代理15和18应用类似的分析时，它导致将多播数据从Rx子交换机1605直接传输到子交换机37。

[0157] 在另一个实施例中，子交换机跳过分层级别(例如，级别A)，前提是该级别将数据转发到下一个分层级别(例如，级别B)中的四个或更少的全部子交换机。当应用于图16A时，代理1和4仅将数据传输到下一个分层级别(级别B)中的四个全部子交换机。因此，可以跳过这些代理，并且Rx子交换机1605的四个连接接口可以将多播数据直接传输到级别B代理。将该优化与图16B中所示的优化(即，如果代理传输到仅一个其它子交换机，则跳过该代理)相组合将导致图16C中所示的优化层次结构。

[0158] 尽管未在图16A-C中示出，但当确定跳过较低级别代理时，在一个实施例中，较高级别代理可以考虑较低级别代理是否也是目的地子交换机。例如，如果代理4是子交换机24，则可能不跳过该代理，因为代理4必须将多播数据转发到其本地端口之一以及子交换机25。但是，在这种情形中，可以跳过代理11，因为它将多播数据仅转发到子交换机25(由于子交换机24(即，代理4)已经接收数据)。较高级别代理可以通过参考代理标识寄存器1405，确定哪些代理也是目的地子交换机(即，哪些代理具有耦合到目的地计算设备的本地端口/网桥元件)。

[0159] 此外，尽管图16A-B示出在四级层次结构中跳过代理，但是该相同过程也可以用于使用代理子交换机的任何类型的层次结构。

[0160] 图16D是用于优化层次结构遍历的技术1600。在步骤1655，代理或接收子交换机接收多播数据，并且在步骤1660，标识应被转发多播数据的代理。具体地说，子交换机可以使用MC组表1210的本地副本标识MC组成员，并且基于层次结构数据1220，确定将这些组成员分配给哪些代理。

[0161] 在步骤1665，子交换机评估是否可以跳过标识的代理。如上面公开的，此确定可以基于代理是否将多播数据传输到仅一个其它子交换机、标识的代理之一是否将多播数据传输到小于或等于可用连接接口加上为该代理分配的连接接口的数量的多个子交换机，或者标识的代理是否将数据转发到少于接收交换机已分配连接接口的全部子交换机。此外，子交换机可以考虑代理是否也是将多播数据转发到本地网桥元件端口的目的地子交换机。

[0162] 如果不能跳过代理，则在步骤1670，子交换机将多播数据转发到标识的代理。

[0163] 但是，如果可以跳过代理，则在步骤1675，子交换机可以通过将多播数据直接转发到低于标识的代理的层次结构级别中的子交换机来跳过代理。

[0164] 控制器可以通过针对在MC组表1210中列出的每个MC组更改优化启用位1320的值，启用和禁用该优化。即，不同的优化方法可以提供不同的优点。例如，确保在每个子交换机上使用最大数量的连接接口可以减少代理之间的交换机业务，但也可以防止与子交换机上的不同网桥元件关联的其它交换机业务使用连接接口。因此，当设置优化启用位1320时，控制器或系统管理员可以考虑这些优缺点。

[0165] 到聚合链路的多播帧传送

[0166] 链路聚合控制协议由IEEE 802.3ad标准定义。具体地说，链路聚合(也称为主干化或链路捆绑)是将多个物理链路绑定为一个聚合(逻辑)链路或主干的过程(在本公开中，“主干”和“聚合链路”可以交换使用)。采用这样的方式跨链路发送业务：组成两个端节点之间的流的帧始终采取相同路径。这通常通过以下操作实现：对帧标头的选定字段执行散列

运算以便选择要使用的物理链路。这样做可以平衡跨一组物理链路的业务并且避免帧在给定流中的错序。

[0167] 图17示出在分布式交换机中将单播数据帧传输到主干的一个物理链路。具体地说，源子交换机1705通过子交换机在其网桥元件120之一处的入口端口，从开始节点(例如，服务器、在计算设备上运行的应用等)接收单播数据帧。在经由主干1720将单播数据转发到连接到分布式交换机的端节点(例如，交换机、服务器、应用等)之前，源子交换机1705可以使用链路聚合控制协议确定在路由单播数据时使用三个物理链路1725_{1-3}中的哪一个。该过程在此称为“链路选择”。如标准定义的，源子交换机1705使用单播数据帧的标头中的信息(例如，目的地和源MAC地址和/或EtherType)选择物理链路1725_{1-3}之一。因此，如果接收在标头中具有相同MAC地址和/或EtherType的另一个单播数据帧，则该帧也使用与前一个单播数据帧相同的数据路径到达端节点。

[0168] 源子交换机1705接收单播数据帧，并且基于包含在标头中的信息执行链路选择。例如，可以配置链路选择，即使两个单播数据帧的标头包含相同的源和目的地MAC地址但包含不同的EtherType(例如，IPv4与IPv6)，源子交换机1705也使用不同的物理链路1725_{1-3}转发分组。换句话说，标头字段用作散列键并且与主干配置信息相比较，以便选择主干1720的物理链路1725。通过这种方式，链路选择可以使用同一主干跨不同物理链路1725_{1-3}分散业务。此外，因为相同的标头字段导致选择相同的物理链路1725，所以维持数据业务的顺序。

[0169] 在图17所示的实例中，源子交换机1705选择物理链路1725_2作为适当的物理链路。数据路径1715示出子交换机6将单播数据转发到目的地子交换机1710(即，子交换机3)，目的地子交换机1710然后经由物理链路1725_2将单播数据传输到端节点。

[0170] 图18示出使用代理将多播数据帧传输到主干的物理链路。子交换机6在入口端口上接收多播数据帧，而不是接收单播数据帧。子交换机6可以解析多播数据帧，并且标识创建散列键必需的标头字段。但是，在一个实施例中，子交换机6可能无法确定要使用主干1720的哪个物理链路(或端口)，因为子交换机6未存储任何其它子交换机的本地端口信息。如前面讨论的，分布式交换机中的子交换机可以包含分布式交换机中的所有不同子交换机的子交换机掩码信息，但可能未存储不同子交换机的本地端口信息。接收子交换机可能不知道哪些本地端口是主干1720的一部分，因此不能将多播数据帧发送到具有由散列键定义的正确端口的子交换机。

[0171] 因为包括子交换机的半导体芯片上的区域有限，所以存储每个子交换机上的不同主干的本地端口信息也许不可能。如图13所示，每个MC组的本地端口信息可以需要40个位。在具有数百个子交换机和数百个MC组(每个组可以启用不同本地端口)的系统中，存储所有子交换机的本地端口信息的存储要求不可行。相反，分布式交换机可以延迟选择要使用主干的哪个端口。即，不同于接收多播数据帧的子交换机的子交换机可以执行链路选择。

[0172] 在子交换机6接收多播数据帧之后，它可以生成散列键并且将该键放在它创建的每个信元的标头中，以便将多播数据转发到分布式交换机中的其它子交换机。如数据路径1805、1810和1815所示，子交换机6将包含多播数据的信元转发到层次结构定义的多个代理。当代理接收多播数据时，它们可以使用其本地MC组表判定其本地端口之一是否应接收多播数据的副本。例如，在多播数据到达子交换机4之后，它将使用表并且确定其网桥元件120之一确实与针对MC组启用的本地端口关联。但是，因为该本地端口是主干1720的一部

分，所以分析在此并未结束。

[0173] 子交换机4判定是否应使用其本地端口在主干1720中传输该特定多播数据。为了进行此判定，子交换机4通过将信元标头中的散列键与主干配置数据相比较，执行链路选择。如果该过程产生与子交换机4上的本地端口匹配的端口ID，则子交换机4使用物理连接1725_3将多播数据传输到端节点。但是，如数据路径1820和1825所示，子交换机4确定不应使用其本地端口，并且继续基于层次结构转发多播数据。当在子交换机3上接收多播数据时，它还可以基于散列键确定正确的端口。因为得到的主干端口ID与子交换机3的本地端口匹配，所以它使用数据路径1830和1835，利用连接1725_2将多播数据帧转发到端节点。还将沿着数据路径1835转发具有相同散列键的所有后续多播数据帧。当然，不同散列键可以导致使用物理连接1725_1或1725_3传输多播数据，而不是使用物理连接1725_2。

[0174] 通过这种方式，在不同于接收子交换机的子交换机处进行链路选择(即，确定哪个端口和对应的物理连接将转发多播数据帧)。

[0175] 图19示出将多播数据帧传输到分配给至少两个主干的目的地交换机。MC组全体成员可以包括任意数量的主干或聚合链路。在此，子交换机0:2的端口1950组成聚合链路1，而子交换机36、71和73的端口1950组成聚合链路2。与同一聚合链路关联的子交换机可以位于不同机箱或机架上。例如，子交换机71可以与子交换机73在物理上位于不同的机架上。

[0176] 在一个实施例中，Rx子交换机1905使用MC组表1910中的MC组全体成员1915确定用于多播数据的目的地子交换机(即，子交换机0:2、36、71和73)。使用分层数据1220，Rx子交换机标识应该为哪些级别A代理提供多播数据的副本，以便数据到达目的地子交换机。使用图10作为实例层次结构，图19示出多播数据通过层次结构的级别A-D的传播。当然，在其它实施例中，可以优化层次结构以便跳过一个或多个代理。

[0177] 尽管MC组全体成员1915指定与聚合链路1和2关联的所有子交换机接收多播数据，但是聚合链路控制协议规定可以选择每个聚合链路的仅一个端口1950来传输多播数据。此外，针对具有相同相关标头部分的任何随后接收的多播数据帧，必须使用同一链路。因此，对于MC组的每个接收的多播数据帧，仅子交换机0:2之一在聚合链路1中传输多播数据帧，并且仅子交换机36、71、73之一在聚合链路2中传输多播数据帧。

[0178] 可以将聚合链路表1960存储在分布式交换机中的每个子交换机上。聚合链路表1960可以包括主干配置信息，将该信息与散列键比较时，在主干中标识特定端口作为选定端口。具体地说，网桥元件使用主干ID索引到聚合链路表1960以便标识特定主干。表1960列出分布式交换机中是主干一部分的每个端口。在分布式交换机中标识与特定端口关联的所有端口(或者物理连接)之后，网桥元件使用散列键标识主干中的一个端口作为选定端口。因此，每个散列键唯一地标识主干中的仅一个端口(即，选定端口)(尽管多个散列键可以映射到同一端口)。在子交换机从聚合链路表1960中标识选定端口ID之后，它可以将该ID与其本地端口ID相比较并且判定它们是否匹配。如果匹配，则子交换机为该本地端口提供多播数据的副本，然后将该副本经由聚合链路转发到端节点。

[0179] 图20A-C示出用于在实现层次结构的分布式交换机中传输多播数据的三个实施例。但是，本公开并不限于这些实施例。

[0180] 每个MC组的主干模式位1315可以用于指示子交换机在接收属于MC组的多播数据时使用三个实施例中的哪一个。

[0181] 实施例1

[0182] 图20A示出将多播数据帧传输到MC组的每个成员。在该实施例中，Rx子交换机2002使用MC组表2004确定目的地子交换机2010和2016均是多播数据帧所属的MC组的成员。假设简单的一级层次结构，Rx子交换机2002通过其ISR 450将多播数据转发到级别A代理子交换机2006。在该实例中，将目的地子交换机2010和2016均分配给级别A代理2006。代理子交换机2006使用两个连接接口455将多播数据转发到目的地子交换机2010和2016。虚线示出多播数据在通过分布式交换机传播时的路径。

[0183] 每个目的地子交换机具有两个本地端口2015，它们连接到主干2022的相应物理链路2014_{1-2}。即，相应目的地子交换机2010、2016的MC组表2012、2018中的本地端口掩码指示针对与多播数据关联的MC组启用端口2015。

[0184] 但是，两个本地端口与同一主干2022关联，如实线2024所示。因此，在目的地子交换机2010、2016将多播数据传输到端口2015之前，两个子交换机2010、2016可以基于接收的散列键执行链路选择，以便判定它们的本地端口2015是否是正确的本地端口。

[0185] 对于目的地子交换机2016，接收网桥元件120(即，最左侧网桥元件)使用MC组表2018中的本地端口掩码判定是否针对与多播数据关联的MC组启用其本地端口之一。因为启用本地端口2015(即，该本地端口是用于将多播数据的副本传输到端节点的候选者)，所以网桥元件120可以确定与本地端口2015关联的主干ID。可以将该信息存储在子交换机2016上标识属于主干2022的所有端口的寄存器中。网桥元件120使用主干ID索引到链路聚合表2020以便标识正确的主干及其关联的端口。然后使用接收的信元标头中的散列键标识选定端口。如在此所示，特定散列键的选定端口不是目的地子交换机2016的本地端口2015。因此，接收网桥元件可以忽略多播数据(即，丢弃包含多播数据的分组/信元)。

[0186] 在目的地子交换机2010的接收网桥元件120(即，最右侧网桥元件)接收多播数据之后，它使用MC组表2012中的本地端口掩码判定是否针对与多播数据关联的MC组启用其本地端口之一。因为启用本地端口2015，所以接收网桥元件120可以使用主干寄存器判定本地端口2015是否属于主干。在这种情况下，本地端口2015是主干2022的一部分。使用散列键和主干2022的主干ID，网桥元件120散列到链路聚合表2014以便判定得到的选定端口是否与本地端口2015具有相同的ID。在这种情况下，端口ID匹配。

[0187] 因此，接收网桥元件120将多播数据转发到与本地端口2015关联的网桥元件120(即，最左侧网桥元件120)。虚线2030示出网桥元件120使用物理链路2024_1将多播数据帧从目的地子交换机2010转发到聚合链路的端节点。如果Rx子交换机2002接收具有相同散列键的另一个多播数据帧，则多播数据将遵循与虚线所示相同的路径(即，未从目的地子交换机2016传输多播数据帧)。但是，不同散列键可以导致目的地子交换机2016沿着主干2022传输多播数据帧，而目的地子交换机2010忽略多播数据。

[0188] 使用该过程，延迟链路选择直至多播数据到达目的地子交换机，该子交换机包含标识正确本地端口以便在主干上通信时使用所需的本地端口信息。

[0189] 目的地子交换机2010可以与目的地子交换机2016并行执行相同的过程。即，两个目的地子交换机独立地执行链路选择。因此，两个子交换机2010、2016可以同时执行链路选择，但这并非必须。

[0190] 尽管未示出，但该过程还可以应用于具有与单个主干关联的两个或更多端口的目

的地子交换机。例如，如果目的地子交换机2010具有与主干2022关联的两个启用的端口，则这些启用的本地端口中的仅一个是选定端口。因此，仅选定端口传输多播数据，而另一个启用的端口则不传输。

[0191] 尽管在没有选定端口的目的地子交换机中忽略多播数据，但如果MC组包含多个较小聚合链路(相对于具有一个或两个较大聚合链路的MC组)，则该实施例可以是优选的。对于包括大量较小聚合链路(例如，多于十个)的MC组，即使目的地子交换机没有选定端口，它也可能针对是不同聚合链路的选定端口的另一个本地端口(或者针对不是任何聚合链路的一部分的端口)需要多播数据。相比之下，如果MC组的所有目的地端口均是单个聚合链路的一部分，则将多播数据传输到所有子交换机可能效率低下，因为除了一个子交换机以外的所有子交换机都将忽略该多播数据。因此，可以基于与MC组关联的聚合链路的数量和大小，设置MC组的主干模式位1315。

[0192] 在一个实施例中，散列键可能不与信元一起传输；相反，目的地子交换机2010、2016中的每一个可以生成散列键。即，在子交换机之间传输的信元可以包含多播数据帧标头中用于生成散列键的必需信息。

[0193] 实施例2

[0194] 图20B示出按照主干将多播数据帧传输到MC组的仅一个成员。具体地说，当建立MC组表时，主控制器(即，主IOMC)可以确保启用MC组全体成员2036中的每个聚合链路的仅一个端口。例如，子交换机掩码2038包括至少三个主干(主干1、2和3)，其中对于每个主干，仅列出具有启用的端口的子交换机作为多播数据的目的地。每个主干的具有启用的端口的目的地子交换机在此称为指定子交换机。

[0195] 相比之下，图20A中所示的实施例包括主干，其中启用相应子交换机上的至少两个端口。Rx子交换机将多播数据转发到主干中具有启用的端口的每个子交换机，尽管该多播数据可能被忽略。

[0196] Rx子交换机2032使用MC组表2034的子交换机掩码2038确定MC组全体成员2036。控制器先前已配置子交换机掩码2038，以便针对每个主干存在仅一个指定子交换机。使用代理子交换机2040，将多播数据转发到所有目的地和指定子交换机。尽管代理子交换机2040将多播数据路由到至少三个指定子交换机(以及任意数量的目的地子交换机)，但为了清晰起见，仅示出一个指定子交换机。具体地说，该图示出针对主干1传输多播数据。

[0197] 指定子交换机2042执行链路选择，以便基于散列键确定正确的本地端口。因为具有三个与主干1关联的子交换机，所以指定子交换机2042确定这些子交换机中的哪个包含用于多播数据的正确选定端口。使用上述过程，接收网桥元件120(即，最右侧网桥元件)查询其本地端口掩码并且确定它是主干的指定子交换机—即，它是主干中具有启用的端口的唯一子交换机。接收网桥元件120然后可以查询主干寄存器以便确定主干ID。使用该ID和散列键，网桥元件可以在链路聚合表2044中标识选定端口。如果指定子交换机2042上的启用的端口2043与选定端口相同，则使用启用的端口2043将多播数据帧转发到端节点。这标记为选项1。

[0198] 备选地，指定子交换机2042标识与主干1关联的哪个子交换机具有选定端口并且将多播数据转发到该子交换机。例如，指定子交换机2042可以在链路聚合表2044中具有附加部分，该部分标识主干1中的所有其它端口的位置数据和ID。基于该信息，指定子交换机

2042确定这些端口ID中的哪个与选定端口匹配。例如，如果选定端口是子交换机2046上的端口2047，则指定子交换机2042将多播数据传输到子交换机2046(选项2)。但是，如果选定端口是子交换机2048上的端口2049，则将多播数据转发到该子交换机(选项3)。

[0199] 与主干中的多个目的地子交换机可以执行链路选择的实施例1相比，在此，每个主干中的仅一个子交换机执行链路选择。但是，如果选定端口不是指定子交换机上的启用的本地端口，则实施例2可以相对于实施例1添加其它跳跃，因为指定子交换机将多播数据传输到包含选定端口的子交换机(如选项2和3所示)。

[0200] 注意，子交换机2046和2048可以称为“目的地”子交换机，尽管Rx子交换机(以及代理子交换机2040)的子交换机掩码2038不知道子交换机2046、2048上的本地端口是传输多播数据的候选者。即，控制器向Rx子交换机2032和代理子交换机2040隐藏该信息，以便当所有三个目的地子交换机2043、2046、2048中的仅一个具有被选择用于传输数据的端口时，防止将多播数据发送到所有这三个目的地子交换机。

[0201] 实施例3

[0202] 图20C示出按照主干将多播数据帧传输到MC组的仅一个成员。第三实施例与实施例1和2的主要区别是不执行链路选择。相反，在Rx子交换机2052曾经接收多播数据帧之前，控制器可以选择用于传输多播数据帧的“选定端口”。

[0203] 与实施例2中一样，每个MC组的每个主干仅启用一个端口。因此，在子交换机掩码2054中仅将具有该启用的端口的子交换机标记为目的地子交换机(即，指定子交换机2058)。使用代理子交换机2056，将多播数据转发到指定子交换机2058。但是，指定子交换机2058不执行任何链路选择。相反，接收网桥元件120(最右侧网桥元件120)使用本地端口掩码确定与启用的端口(即，端口2062)关联的网桥元件120。接收网桥元件120将多播数据传输到启用的端口，启用的端口然后经由主干1的物理连接2060将数据帧传输到端节点。因此，当控制器填充MC组表并且针对特定MC组按照主干启用仅一个端口时，控制器选择选定端口。所启用的端口是选定端口。

[0204] 有利的是，与实施例1相比，将数据传输到不包含选定端口的目的地子交换机时，实施例3避免必须忽略多播数据。此外，与实施例2中不同，实施例3不需要其它跳跃以便从指定子交换机转到具有选定端口的目的地子交换机。但是，实施例3未从链路选择的负载平衡方面受益。即，如果随后接收同一MC组中的多播数据帧，但该多播数据帧具有完全不同的源或目的地MAC地址和/或EtherType，则Rx子交换机2053仍将多播数据传输到指定子交换机2058，指定子交换机2058使用同一端口2062将后续数据帧传输到端节点。在实施例1和2中，不同散列键可以导致使用不同端口。但是，如果MC业务是工作负载的一小部分，则可以优选实施例3，因为它不向交换机光纤通道网络中注入其它业务。

[0205] 代理层次结构可以包括多个级别，它们形成金字塔式布置，其中较高级别代理将多播数据帧转发到较低级别代理直至到达层次结构的底部。可以定制每个多播组以便当通过分布式交换机转发多播数据时，使用这些级别中的一个或多个。此外，层次结构中的每个代理可以具有必需信息以便确保如果层次结构中的下一个代理不可用，则可以将数据路由到备用代理。

[0206] 可以通过跳过代理进一步优化选定层次结构。例如，如果层次结构的一个级别中的代理将多播数据转发到层次结构的较低级别中的仅一个代理(或目的地)子交换机，则可

以跳过该发送代理。实际上，将多播数据直接转发到较低级别中的子交换机。此外，可以优化层次结构以便确保最大程度地使用子交换机的连接接口（例如，将多播数据传输到分布式交换机中的其它子交换机的端口）。具体地说，如果跳过层次结构的某个级别将增加使用的连接接口数量，则子交换机可以将多播数据直接转发到分层级别中在跳过的级别下面的代理。

[0207] 附图中的流程图和框图显示了根据本发明的不同实施例的系统、方法和计算机程序产品的可能实现的体系架构、功能和操作。在这点上，流程图或框图中的每个方框可以代表一个模块、程序段或代码的一部分，所述模块、程序段或代码的一部分包含一个或多个用于实现规定的逻辑功能的可执行指令。也应当注意，在有些作为替换的实现中，方框中所标注的功能可以以不同于附图中所标注的顺序发生。例如，两个连续的方框实际上可以基本并行地执行，它们有时也可以按相反的顺序执行，这依所涉及的功能而定。也要注意的是，框图和/或流程图中的每个方框、以及框图和/或流程图中的方框的组合，可以用执行规定的功能或动作的专用的基于硬件的系统来实现，或者可以用专用硬件与计算机指令的组合来实现。

[0208] 尽管以上所述涉及本发明的各实施例，但可以构想本发明的其它和进一步实施例而不偏离本发明的基本范围，并且本发明的范围由下面的权利要求确定。

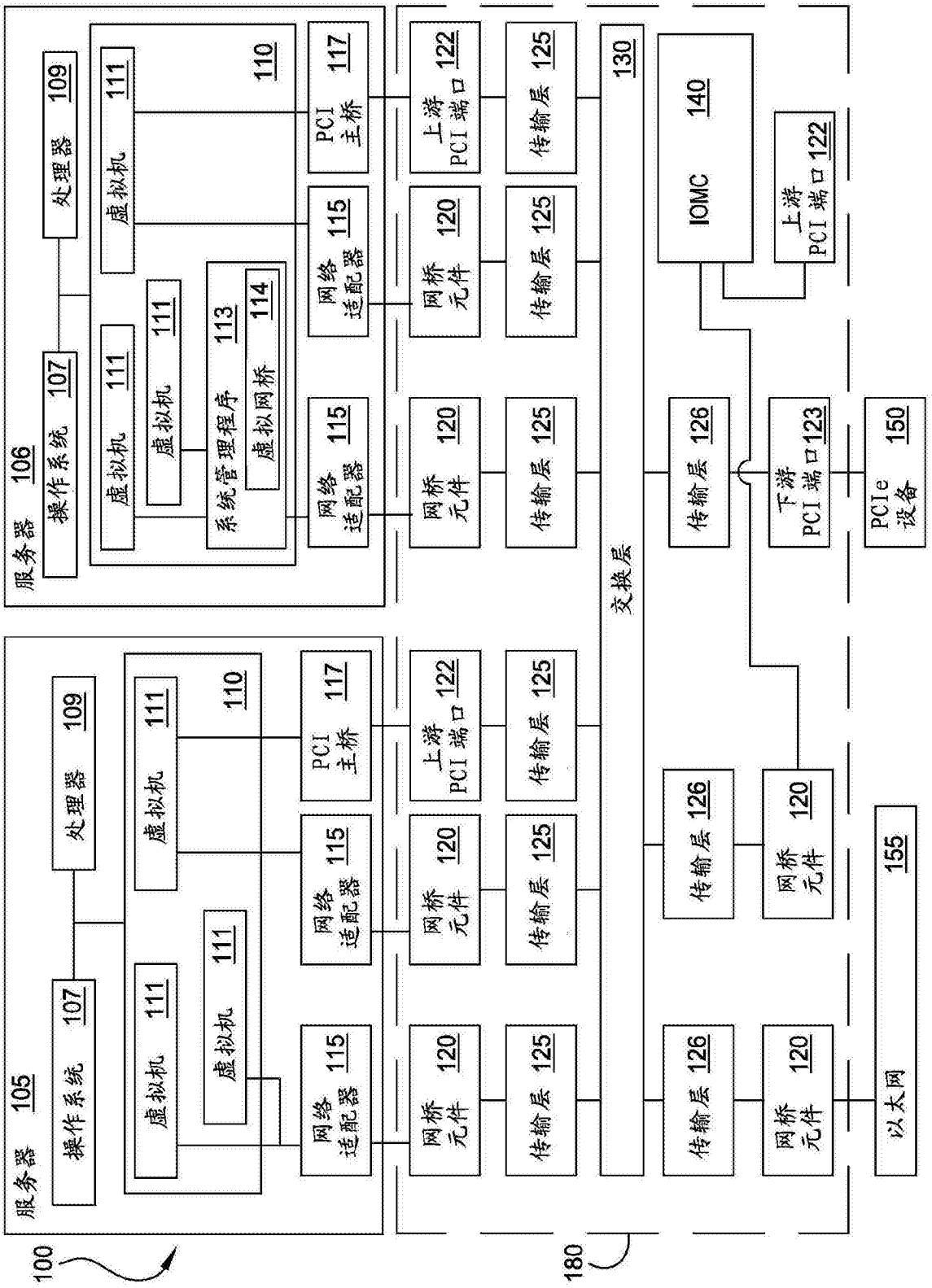

服务器 105
操作系统 107
处理器 109
虚拟机 111
虚拟机 111
虚拟机 111
110
网络适配器 115
网桥元件 120
传输层 125
PCI主桥 117
上游PCI端口 122
传输层 125
交换层 130
传输层 126
下游PCI端口 123
PCIe设备 150
传输层 126
网桥元件 120
以太网 155
服务器 106
操作系统 107
处理器 109
虚拟机 111
虚拟机 111
虚拟机 111
系统管理程序 113
虚拟网桥 114
110
网络适配器 115
网桥元件 120
传输层 125
PCI主桥 117
上游PCI端口 122
传输层 125
IOMC 140
上游PCI端口 122
100
180

图1

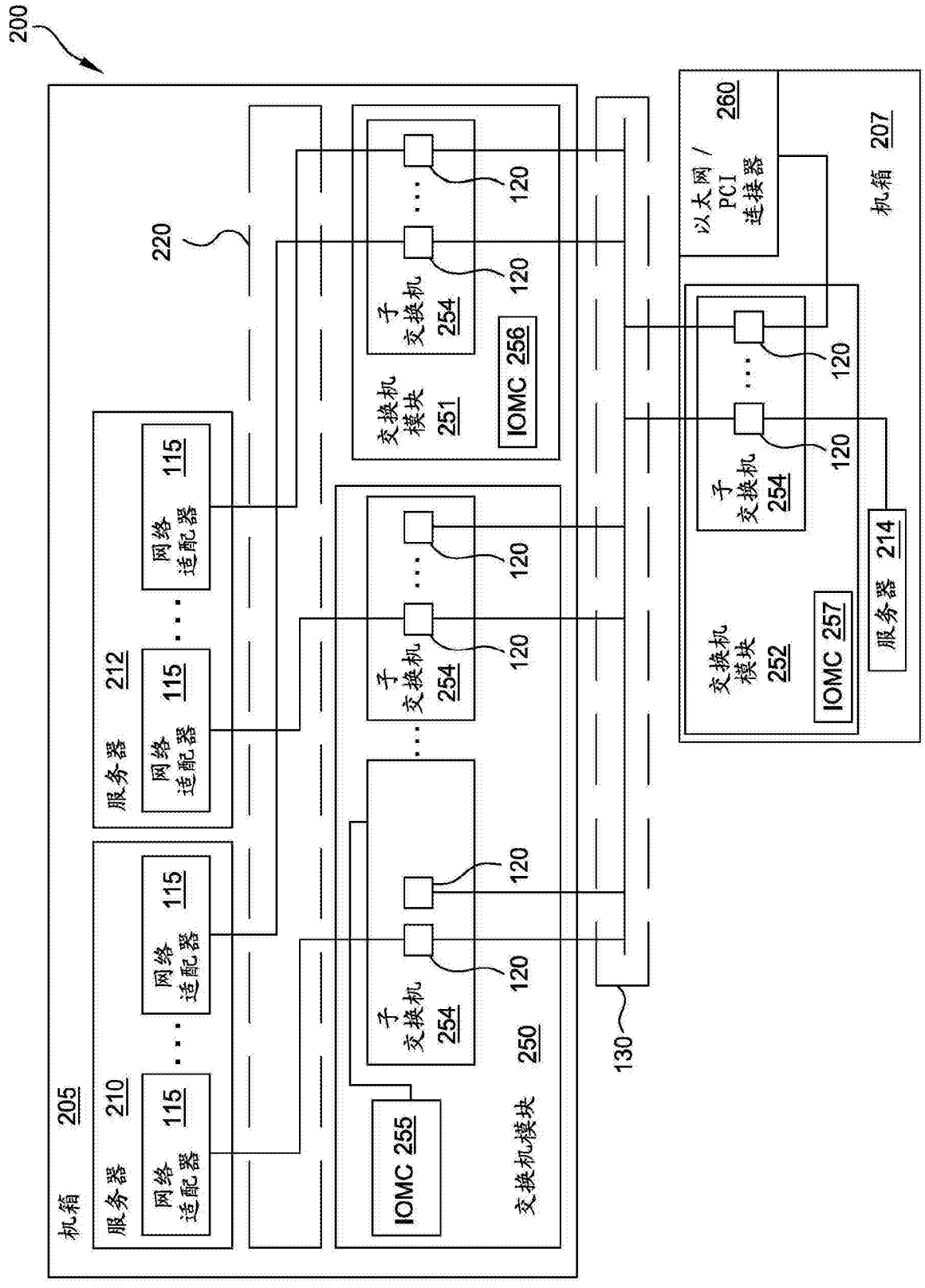

200
机箱 205
服务器 210
网络适配器 115
服务器 212
网络适配器 115
220
交换机模块 250
IOMC 255
子交换机 254
120
交换机模块 251
IOMC 256
130
以太网 / PCI 连接器 260
机箱 207
交换机模块 252
IOMC 257
服务器 214

图2

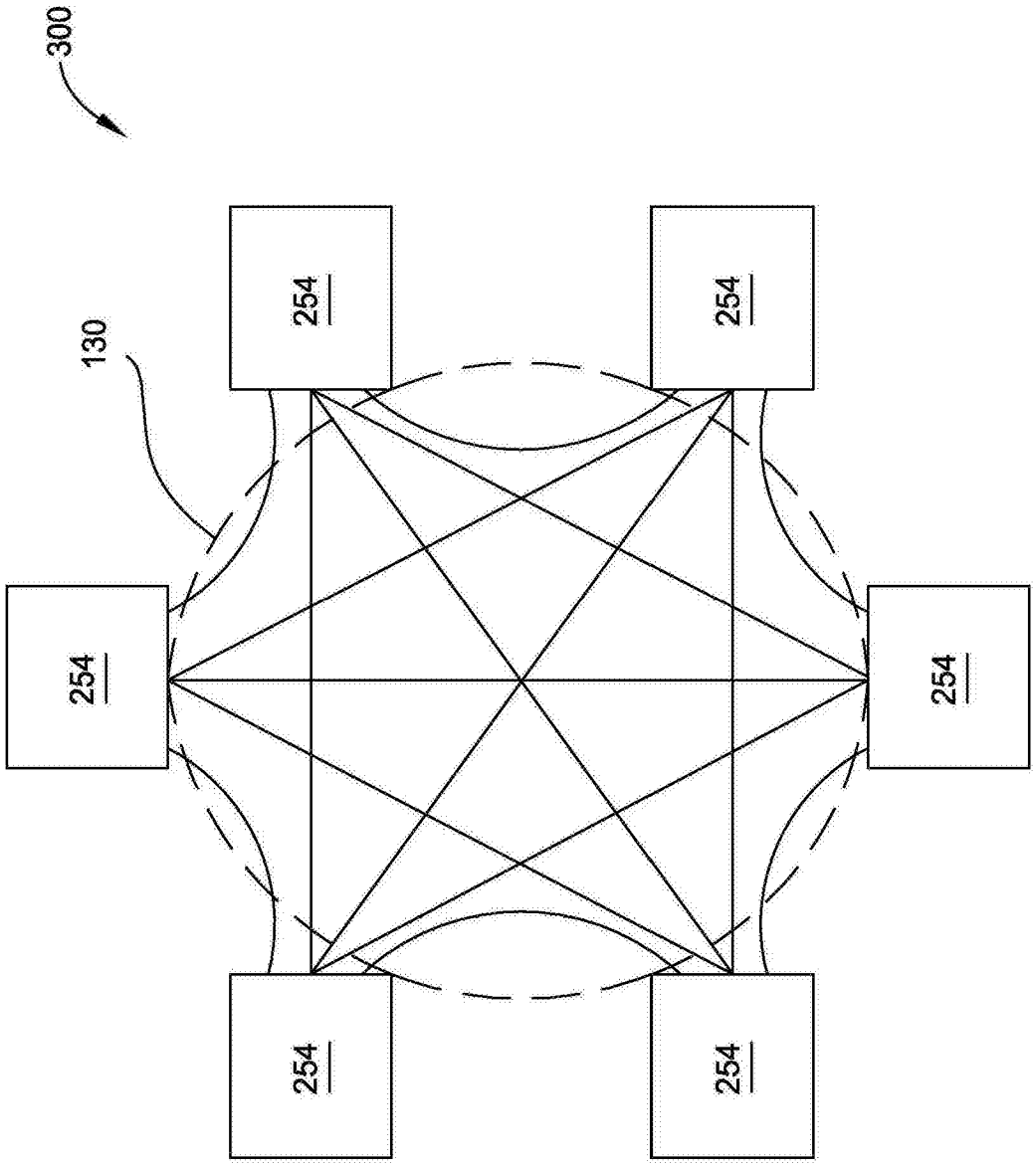


图3

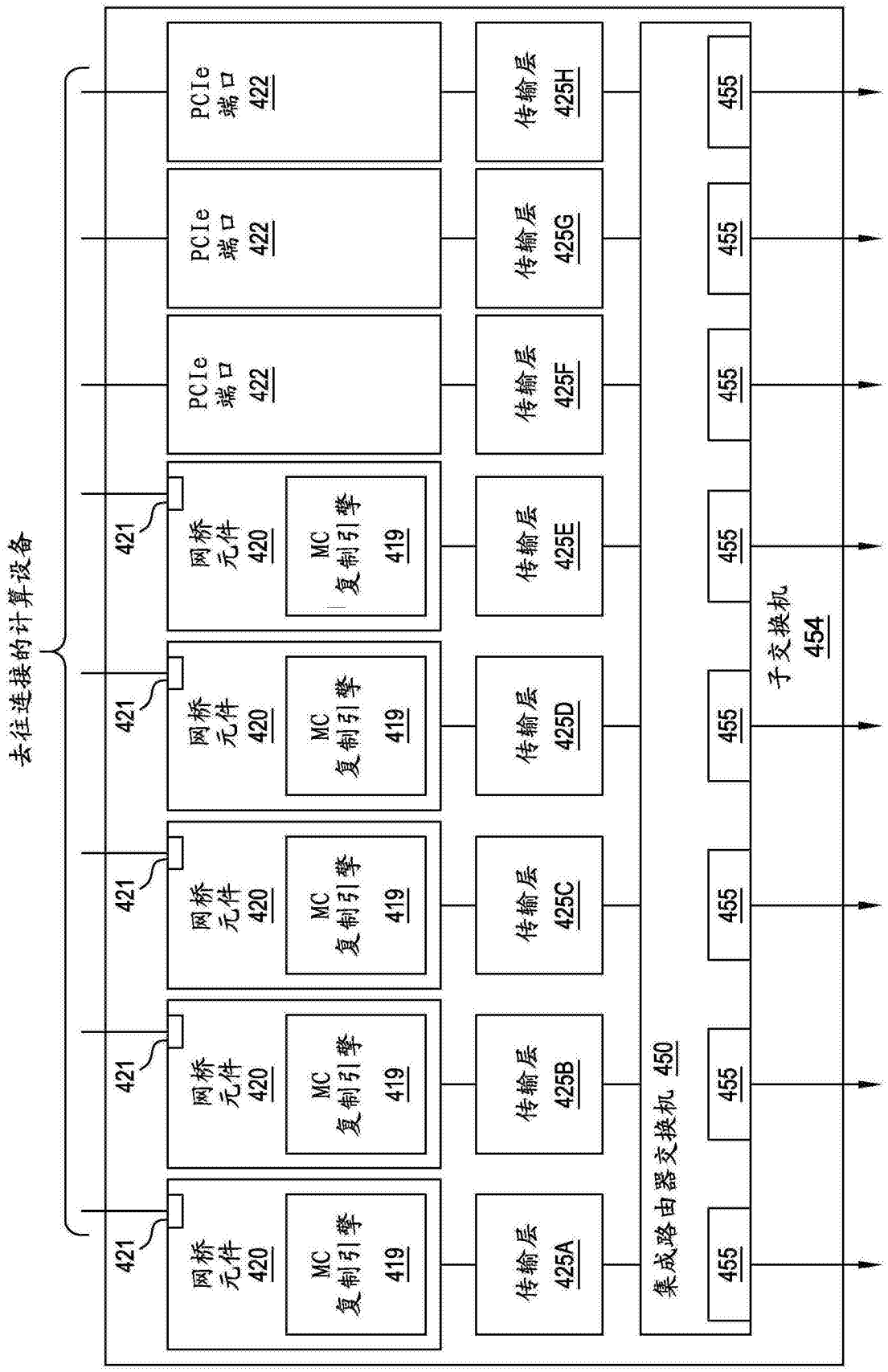

去往连接的计算设备
421
网桥元件 420
MC复制引擎 419
传输层 425A
传输层 425B
传输层 425C
传输层 425D
传输层 425E
传输层 425F
传输层 425G
传输层 425H
PCIe端口 422
集成路由器交换机 450
455
子交换机 454
去往交换层 130

图4

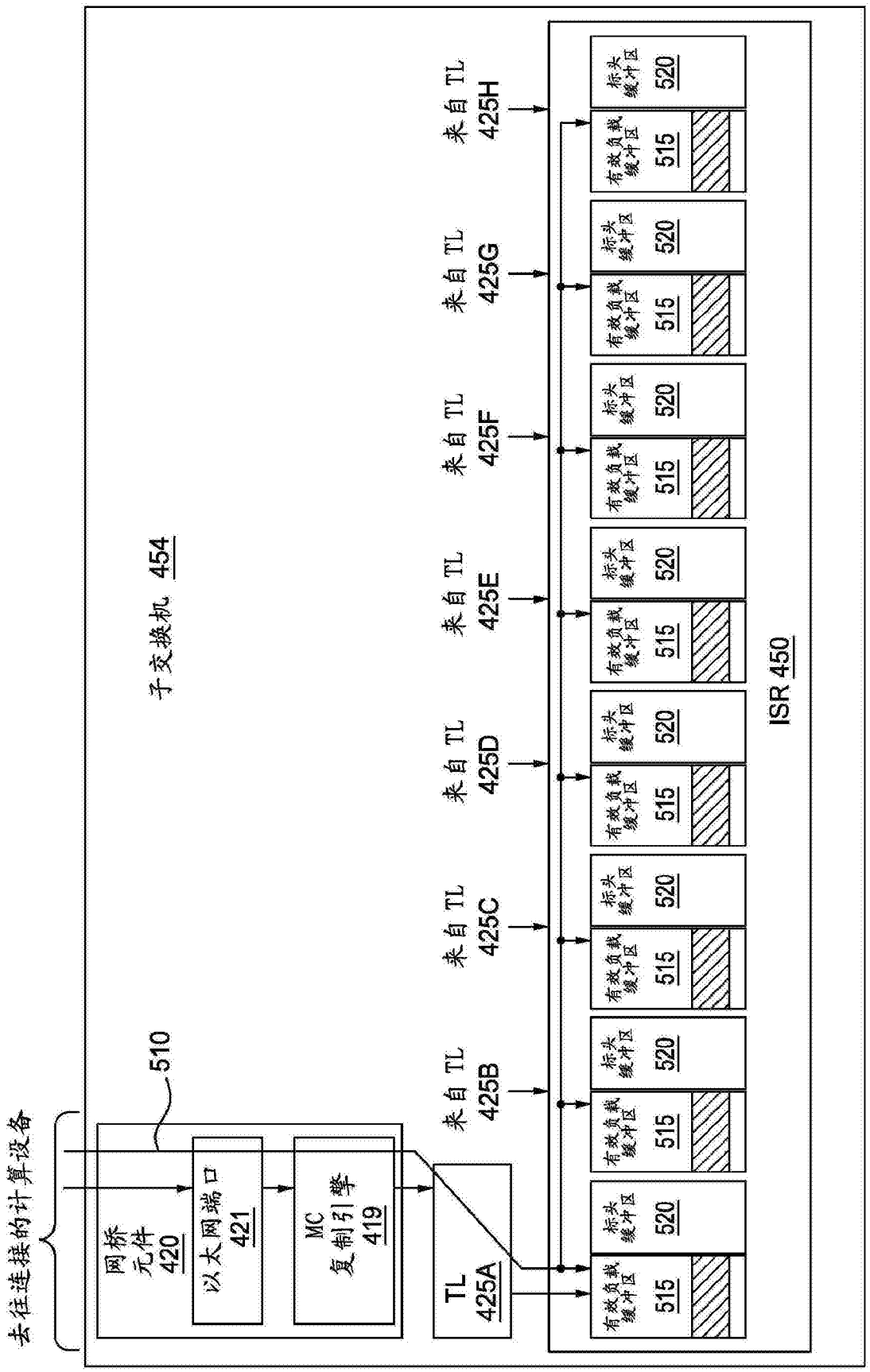

去往连接的计算设备
510
网桥元件 420
以太网端口 421
MC 复制引擎 419
TL 425A
子交换机 454
来自 TL 425B
来自 TL 425C
来自 TL 425D
来自 TL 425E
来自 TL 425F
来自 TL 425G
来自 TL 425H
ISR 450
有效负载缓冲区 515
标头缓冲区 520

图5A

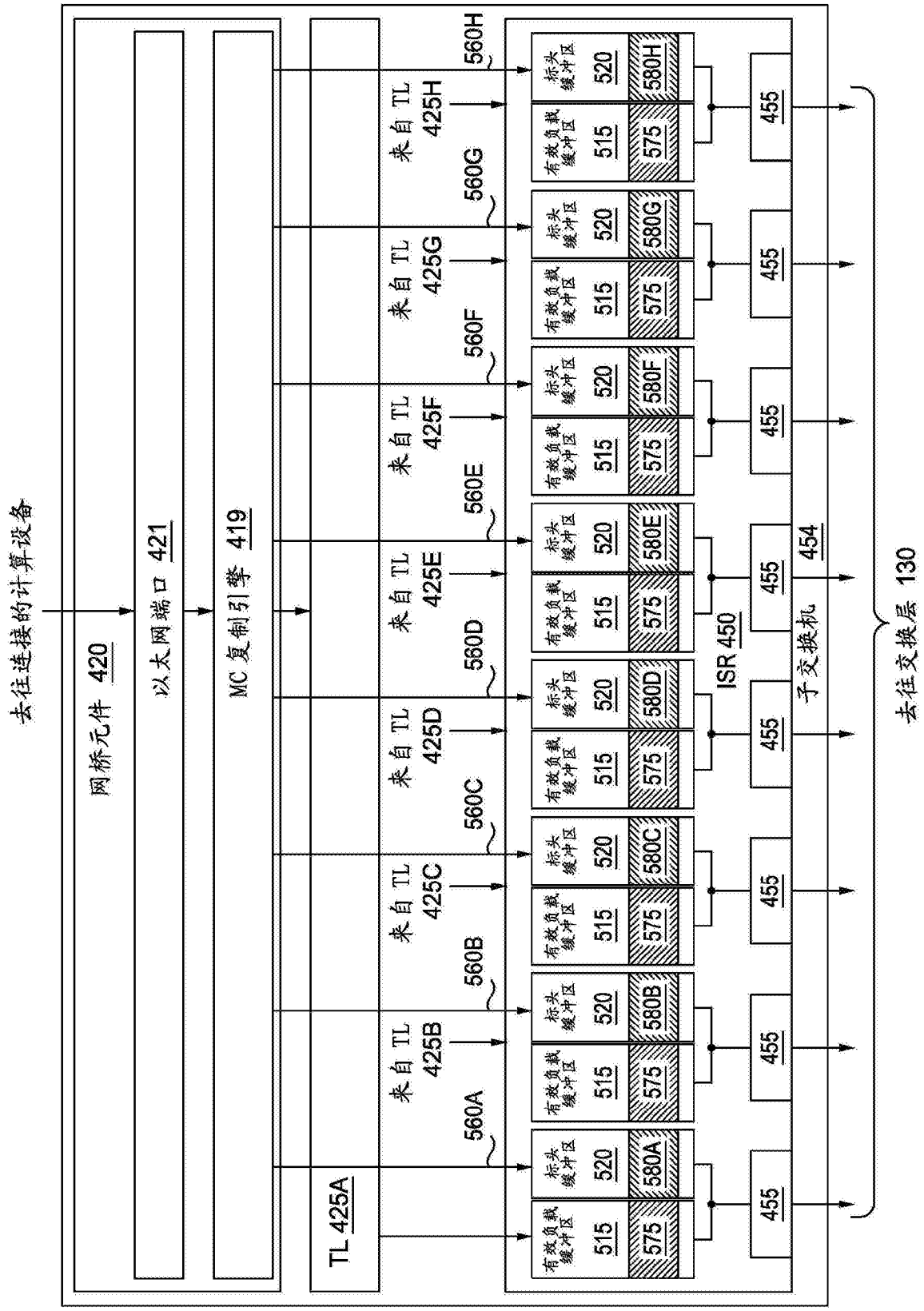

去往连接的计算设备
网桥元件 420
以太网端口 421
MC 复制引擎 419
TL 425A
来自 TL 425B
来自 TL 425C
来自 TL 425D
来自 TL 425E
来自 TL 425F
来自 TL 425G
来自 TL 425H
560A
560B
560C
560D
560E
560F
560G
560H
有效负载缓冲区 515
标头缓冲区 520
575
580A
580B
580C
580D
580E
580F
580G
580H
ISR 450
子交换机 454
455
去往交换层 130

图5B

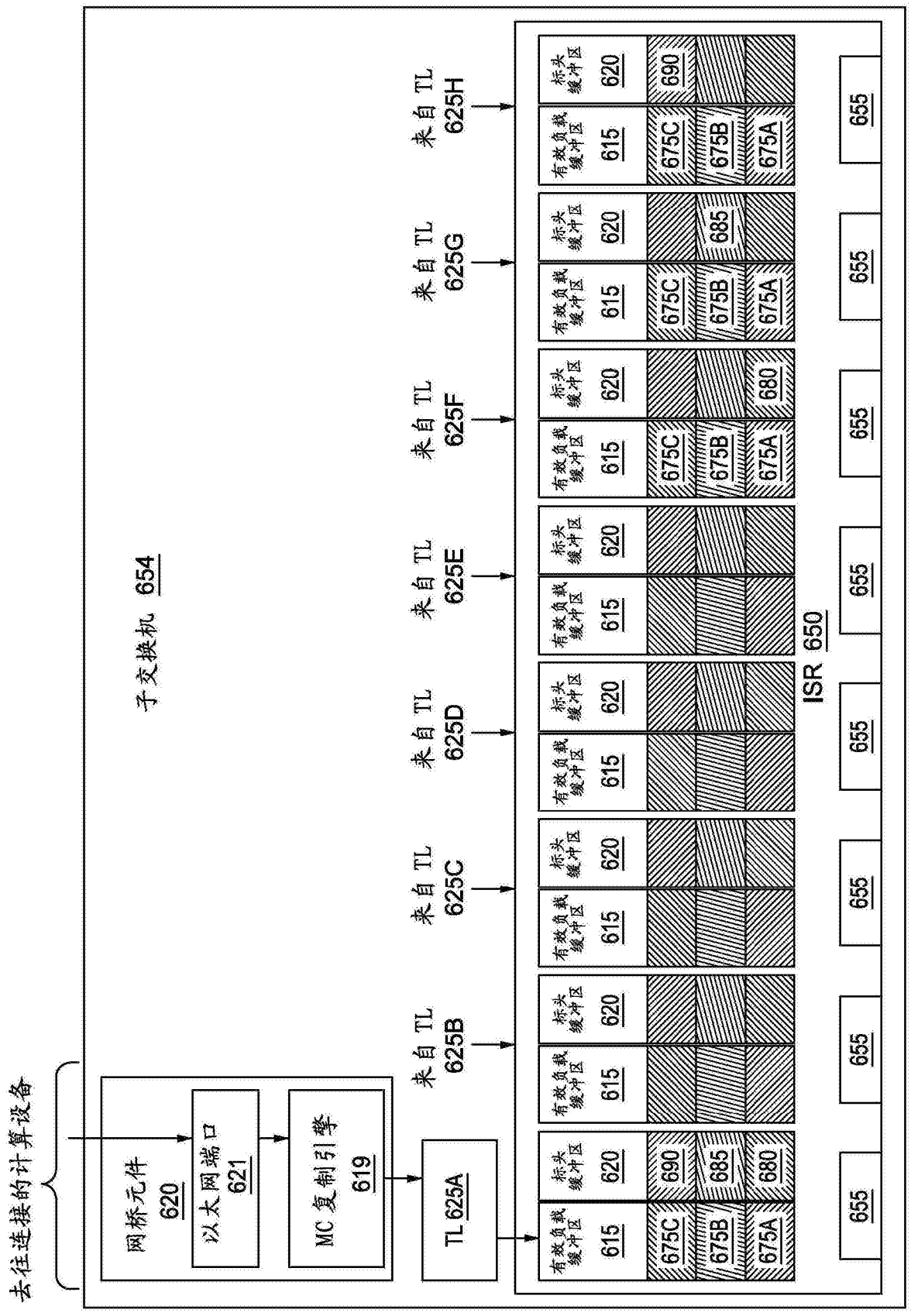

去往连接的计算设备
网桥元件 620
以太网端口 621
MC 复制引擎 619
TL 625A
子交换机 654
来自 TL 625B
来自 TL 625C
来自 TL 625D
来自 TL 625E
来自 TL 625F
来自 TL 625G
来自 TL 625H
ISR 650
有效负载缓冲区 615
标头缓冲区 620
675C
675B
675A
690
685
680
655

图6

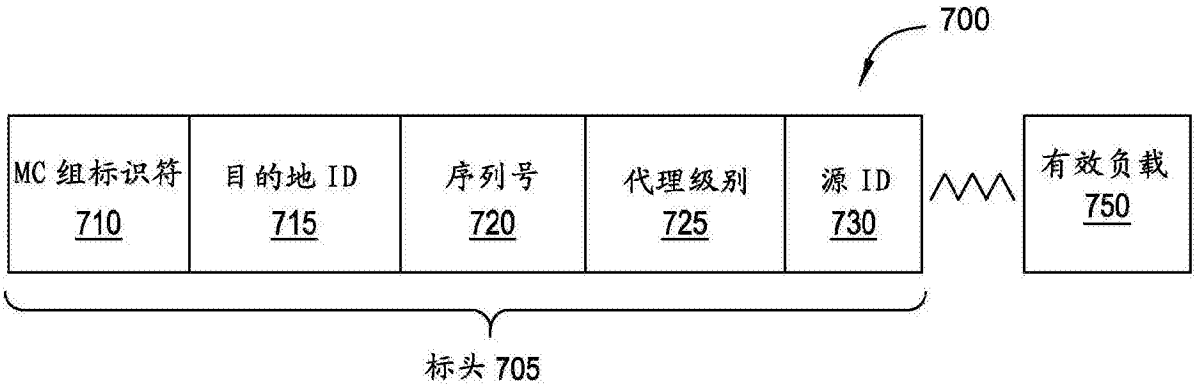

700
MC 组标识符
710
目的地 ID
715
序列号
720
代理级别
725
源 ID
730
有效负载
750
标头 705

图7

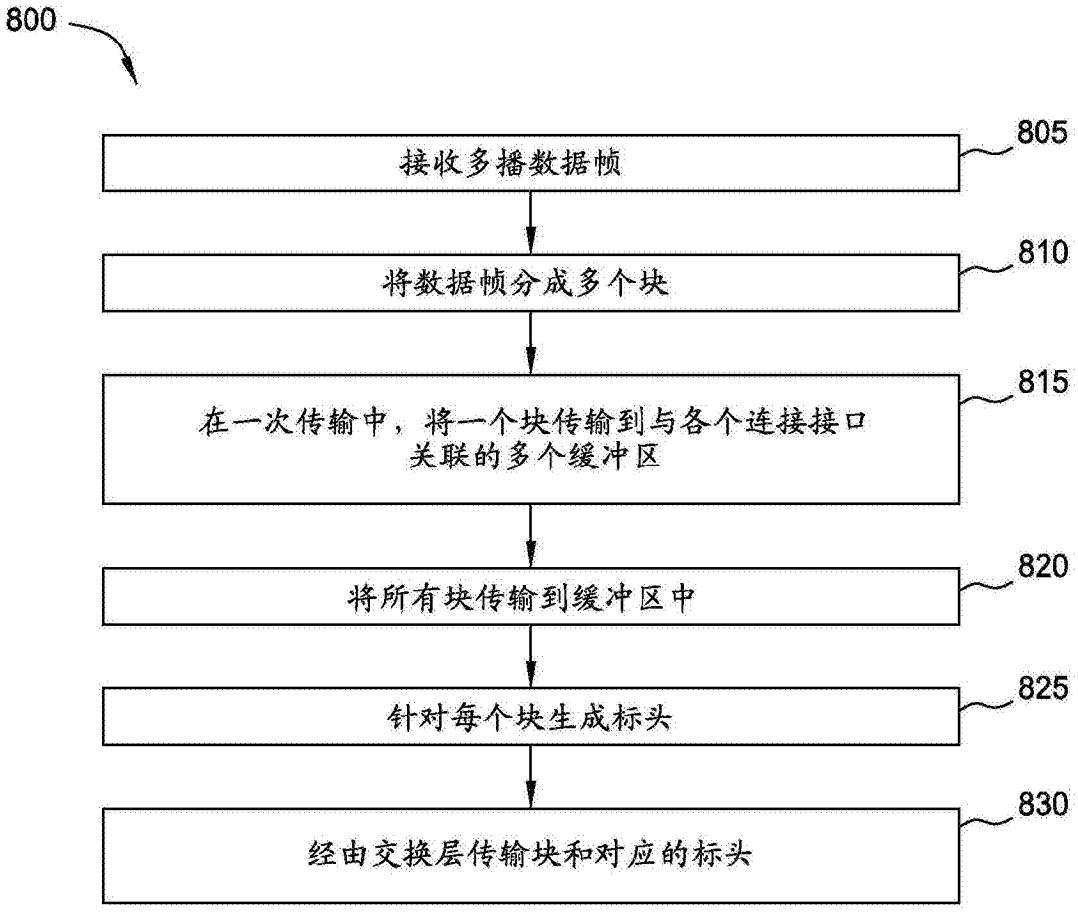

800
接收多播数据帧
805
将数据帧分成多个块
810
在一次传输中，将一个块传输到与各个连接接口关联的多个缓冲区
815
将所有块传输到缓冲区中
820
针对每个块生成标头
825
经由交换层传输块和对应的标头
830

图8

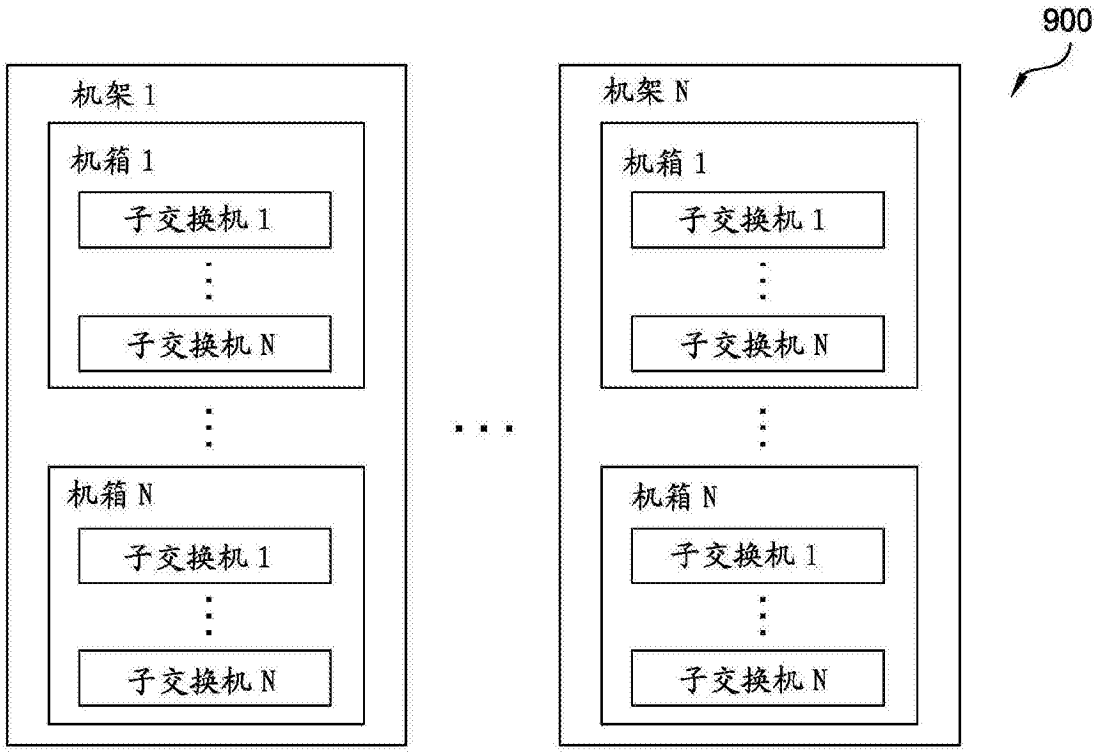

900
机架 1
机箱 1
子交换机 1
子交换机 N
机箱 N
子交换机 1
子交换机 N
机架 N
机箱 1
子交换机 1
子交换机 N
机箱 N
子交换机 1
子交换机 N

图9

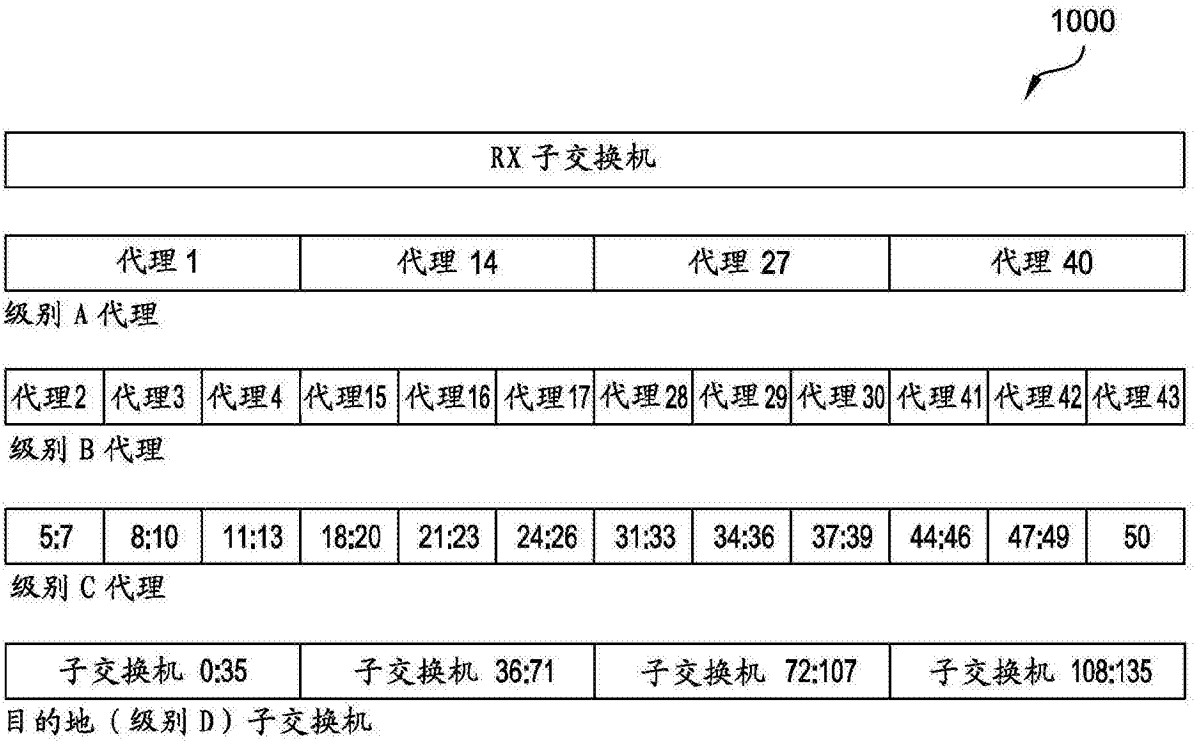

1000
RX 子交换机
代理 1
代理 14
代理 27
代理 40
级别 A 代理
代理2 代理3 代理4 代理15 代理16 代理17 代理28 代理29 代理30 代理41 代理42 代理43
级别 B 代理
5:7 8:10 11:13 18:20 21:23 24:26 31:33 34:36 37:39 44:46 47:49 50
级别 C 代理
子交换机 0:35
子交换机 36:71
子交换机 72:107
子交换机 108:135
目的地（级别 D）子交换机

图10

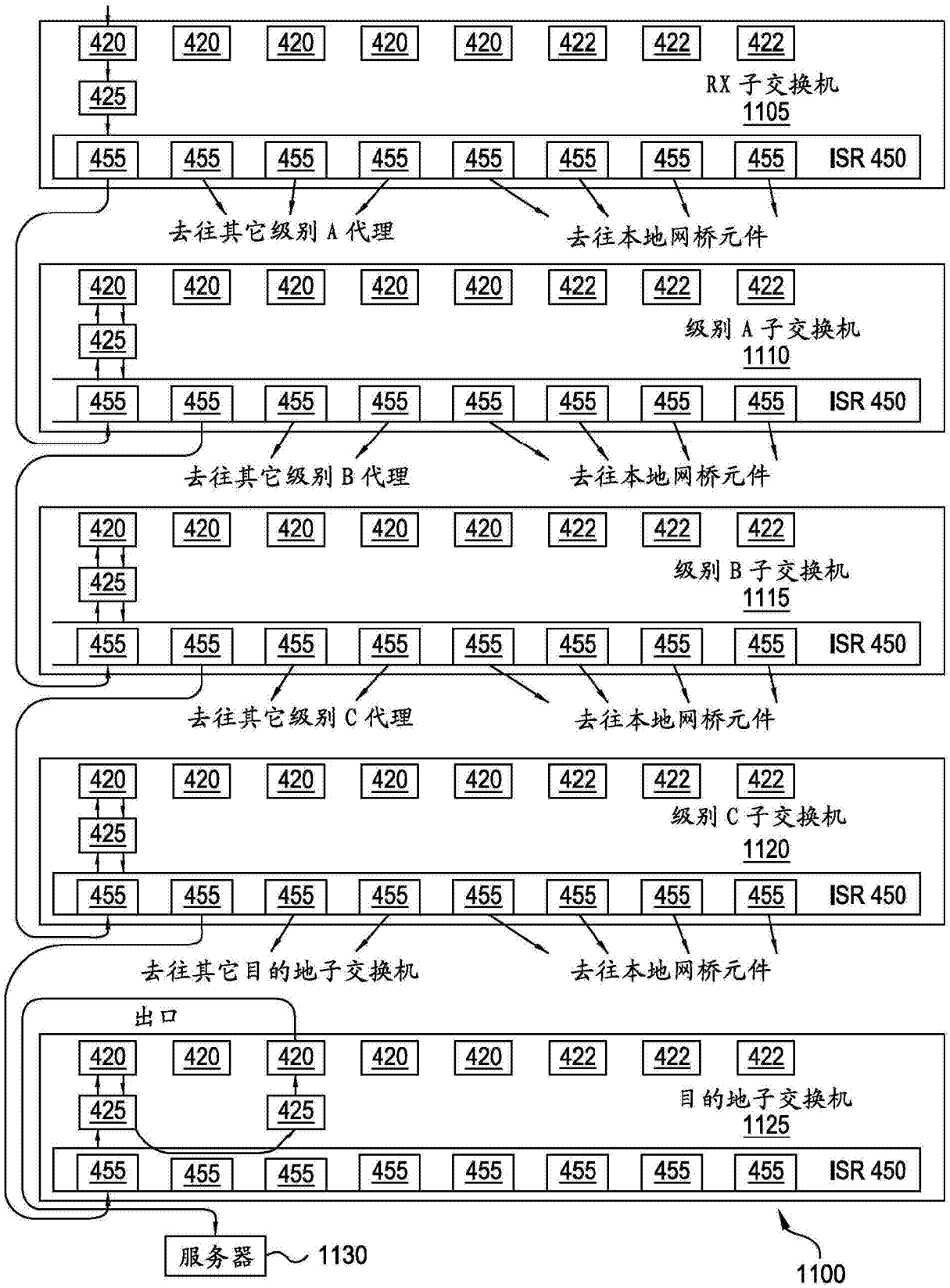

420
425
455
422
RX 子交换机
1105
ISR 450
去往其它级别 A 代理
去往本地网桥元件
级别 A 子交换机
1110
去往其它级别 B 代理
级别 B 子交换机
1115
去往其它级别 C 代理
级别 C 子交换机
1120
去往其它目的地子交换机
出口
目的地子交换机
1125
服务器
1130
1100

图11

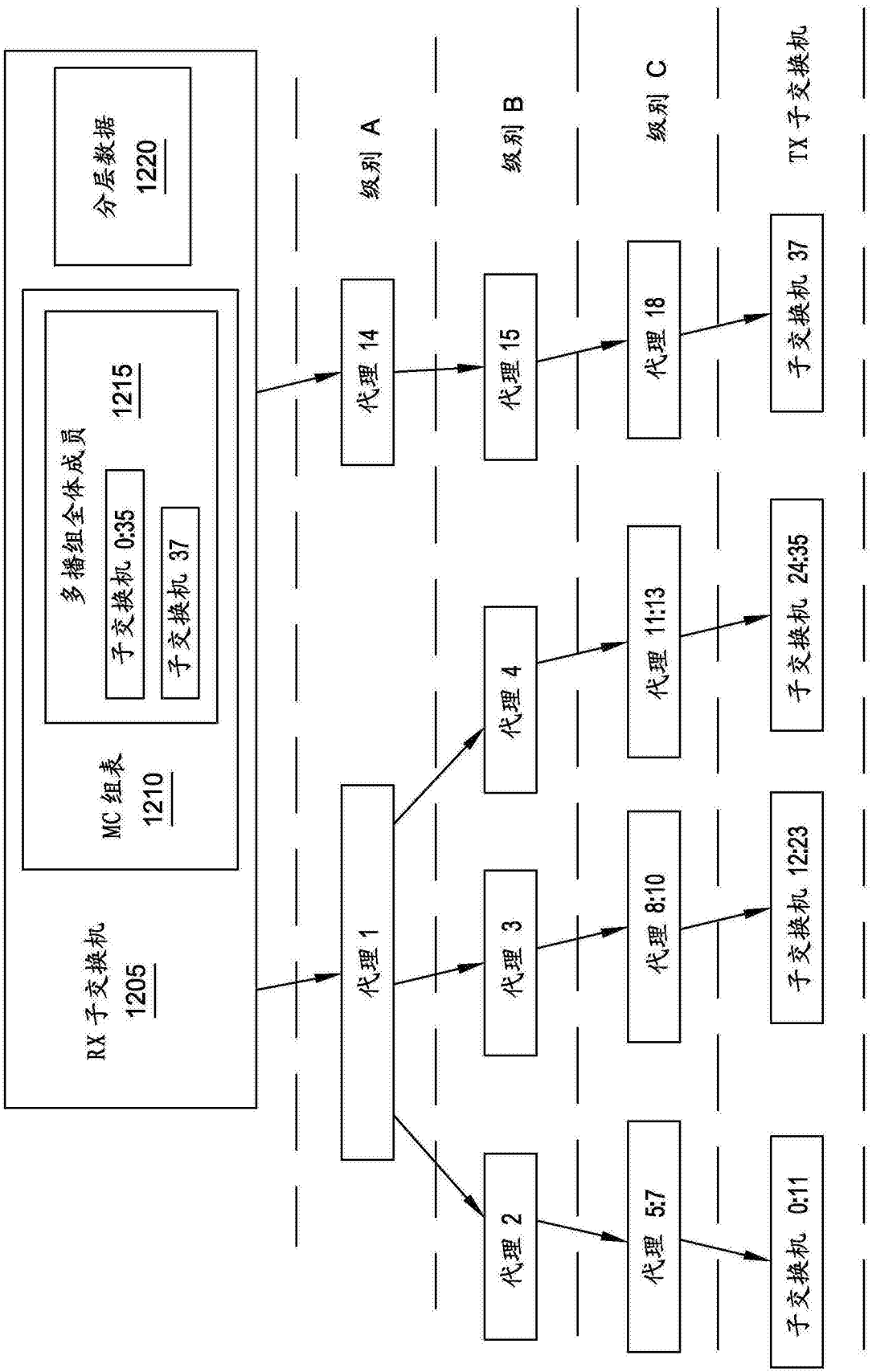

RX 子交换机
1205
MC 组表
1210
多播组全体成员
1215
子交换机 0:35
子交换机 37
分层数据
1220
级别 A
级别 B
级别 C
TX 子交换机
代理 1
代理 2
代理 3
代理 4
代理 5:7
代理 8:10
代理 11:13
代理 14
代理 15
代理 18
子交换机 0:11
子交换机 12:23
子交换机 24:35
子交换机 37

图12

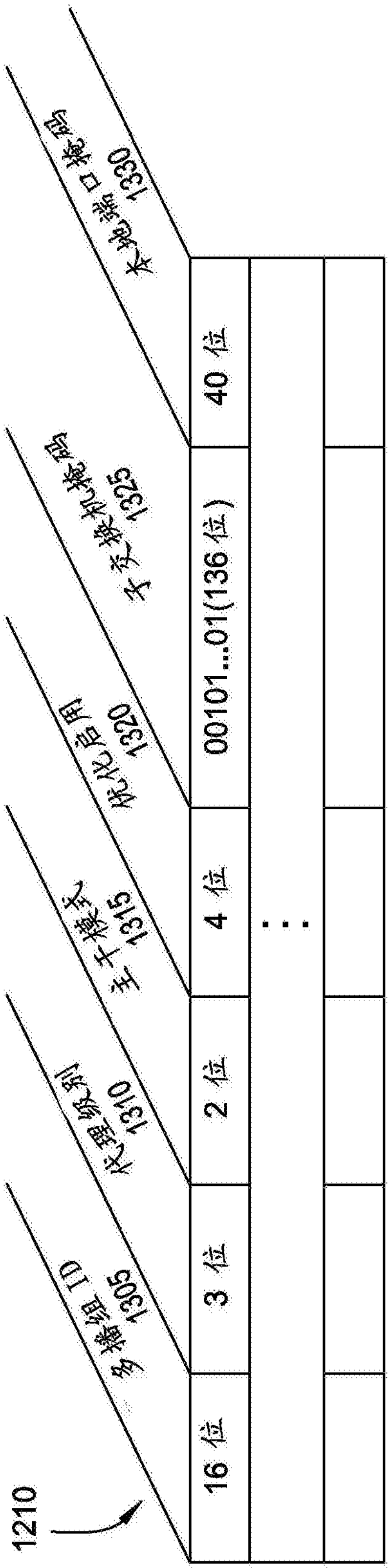

1210
多播组 ID
1305
代理级别
1310
主干模式
1315
优化启用
1320
子交换机掩码
1325
本地端口掩码
1330
16 位
3 位
2 位
4 位
00101...01(136 位)
40 位
...

图13

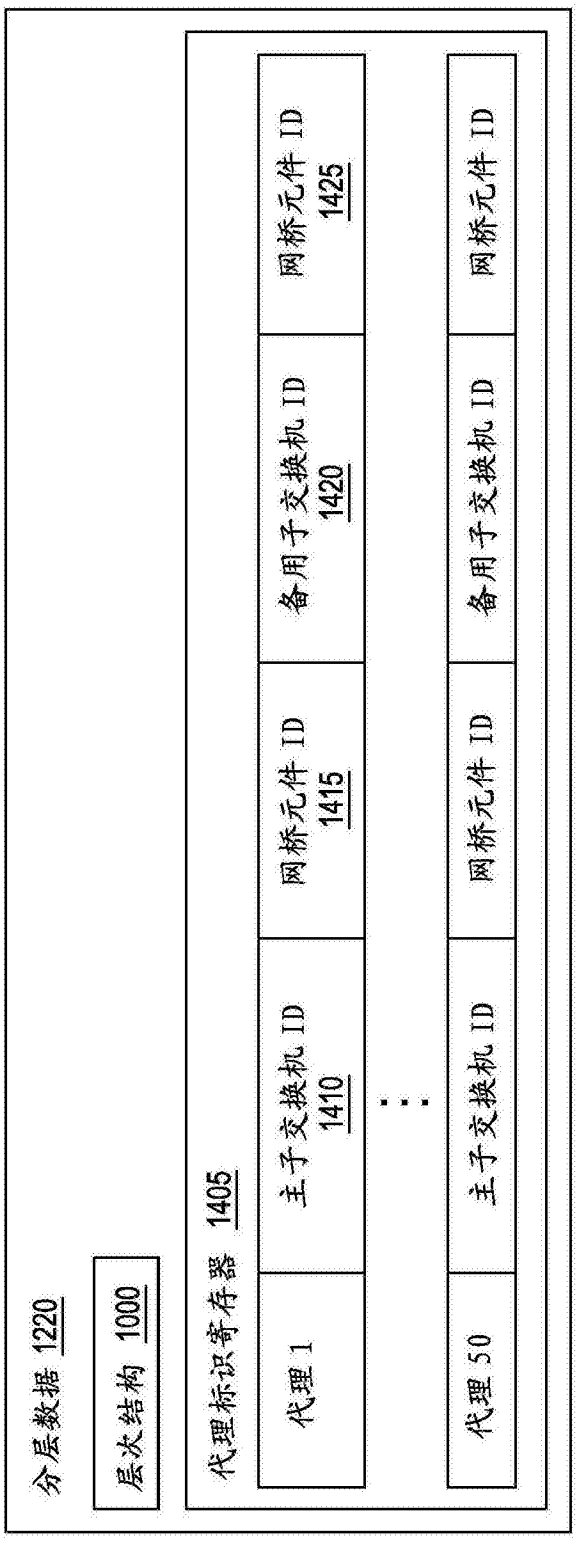

分层数据 1220
层次结构 1000
代理标识寄存器 1405
代理 1
主子交换机 ID 1410
网桥元件 ID 1415
备用子交换机 ID 1420
网桥元件 ID 1425
代理 50
主子交换机 ID
网桥元件 ID
备用子交换机 ID
网桥元件 ID

图14

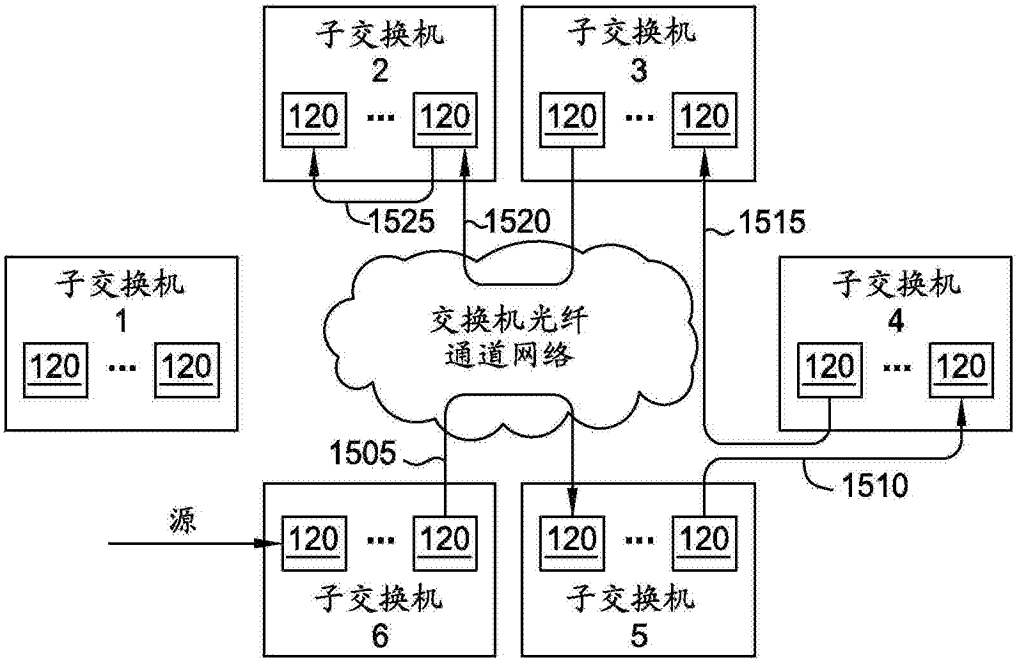

子交换机
2
120
120
子交换机
3
120
120
1525
1520
1515
子交换机
1
120
120
交换机光纤
通道网络
子交换机
4
120
120
1505
1510
源
120
120
子交换机
6
120
120
子交换机
5

图15A

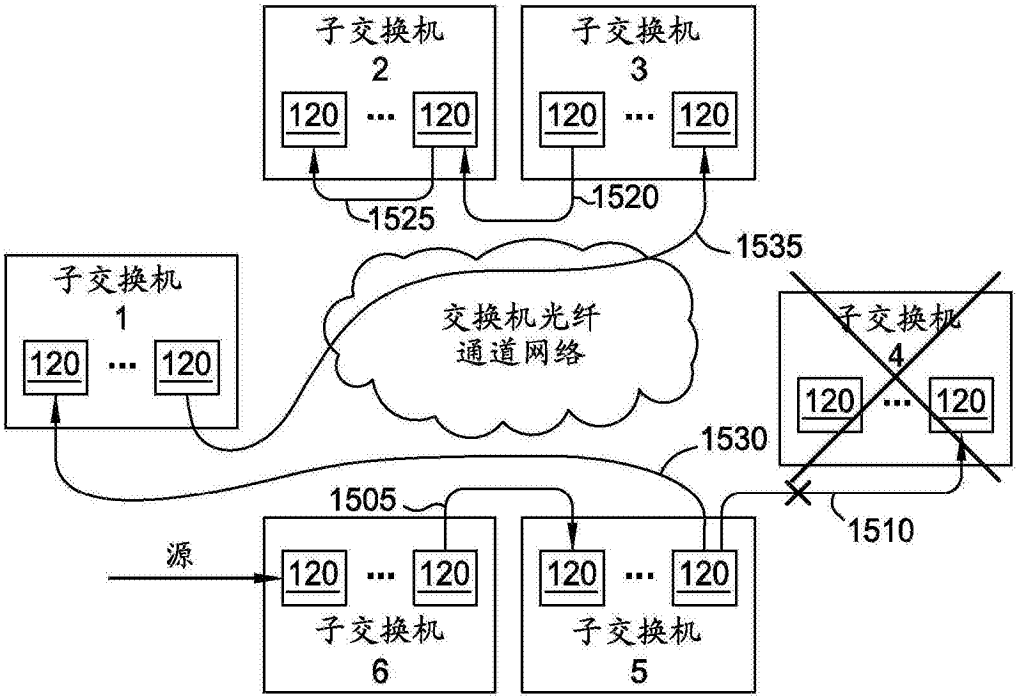

子交换机
2
120
120
子交换机
3
120
120
1525
1520
1535
子交换机
1
120
120
交换机光纤
通道网络
子交换机
4
120
120
1530
1505
1510
源
120
120
子交换机
6
120
120
子交换机
5

图15B

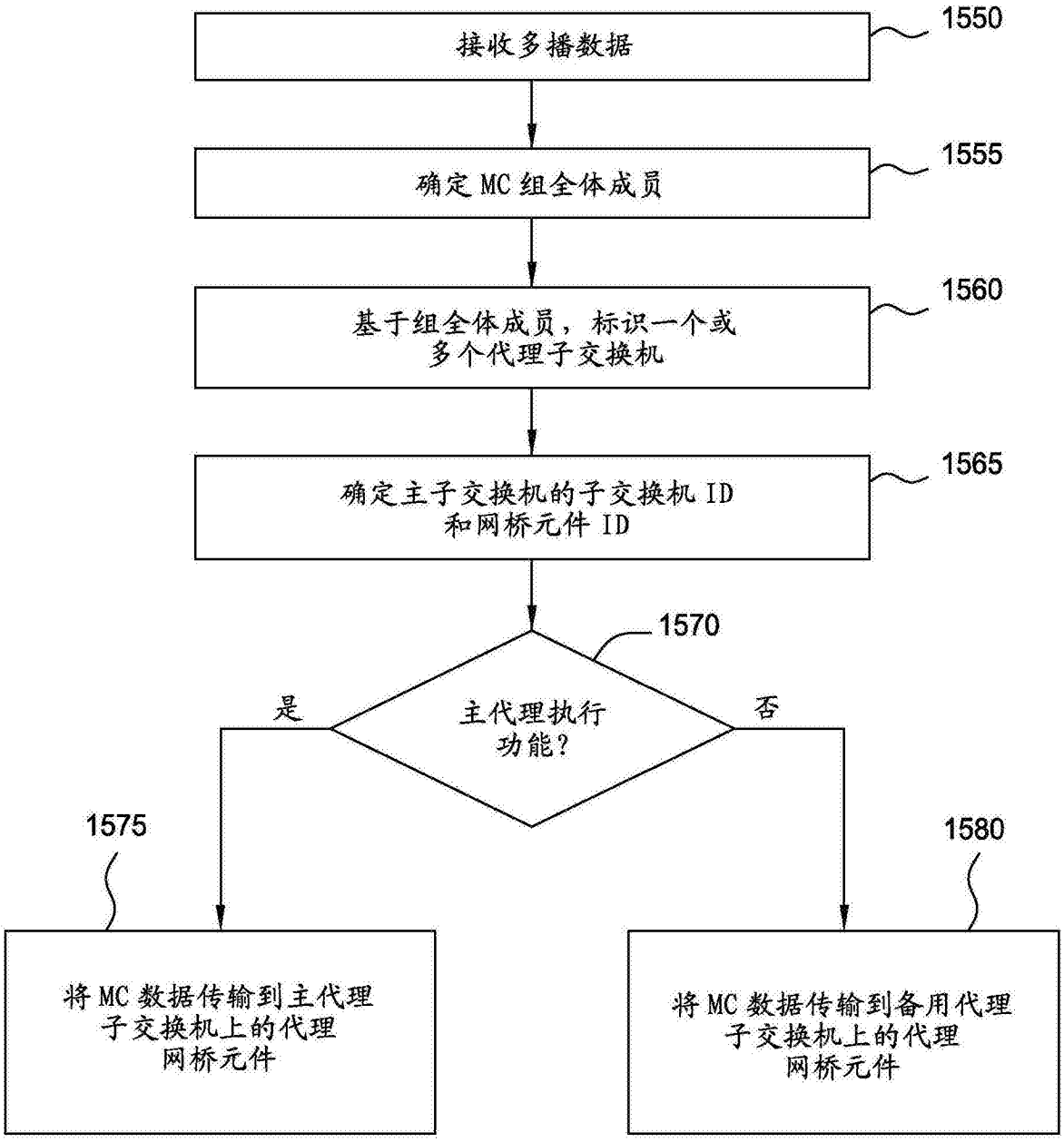

接收多播数据
1550
确定 MC 组全体成员
1555
基于组全体成员，标识一个或
多个代理子交换机
1560
确定主子交换机的子交换机 ID
和网桥元件 ID
1565
1570
主代理执行
功能?
是
否
1575
将 MC 数据传输到主代理
子交换机上的代理
网桥元件
1580
将 MC 数据传输到备用代理
子交换机上的代理
网桥元件

图15C

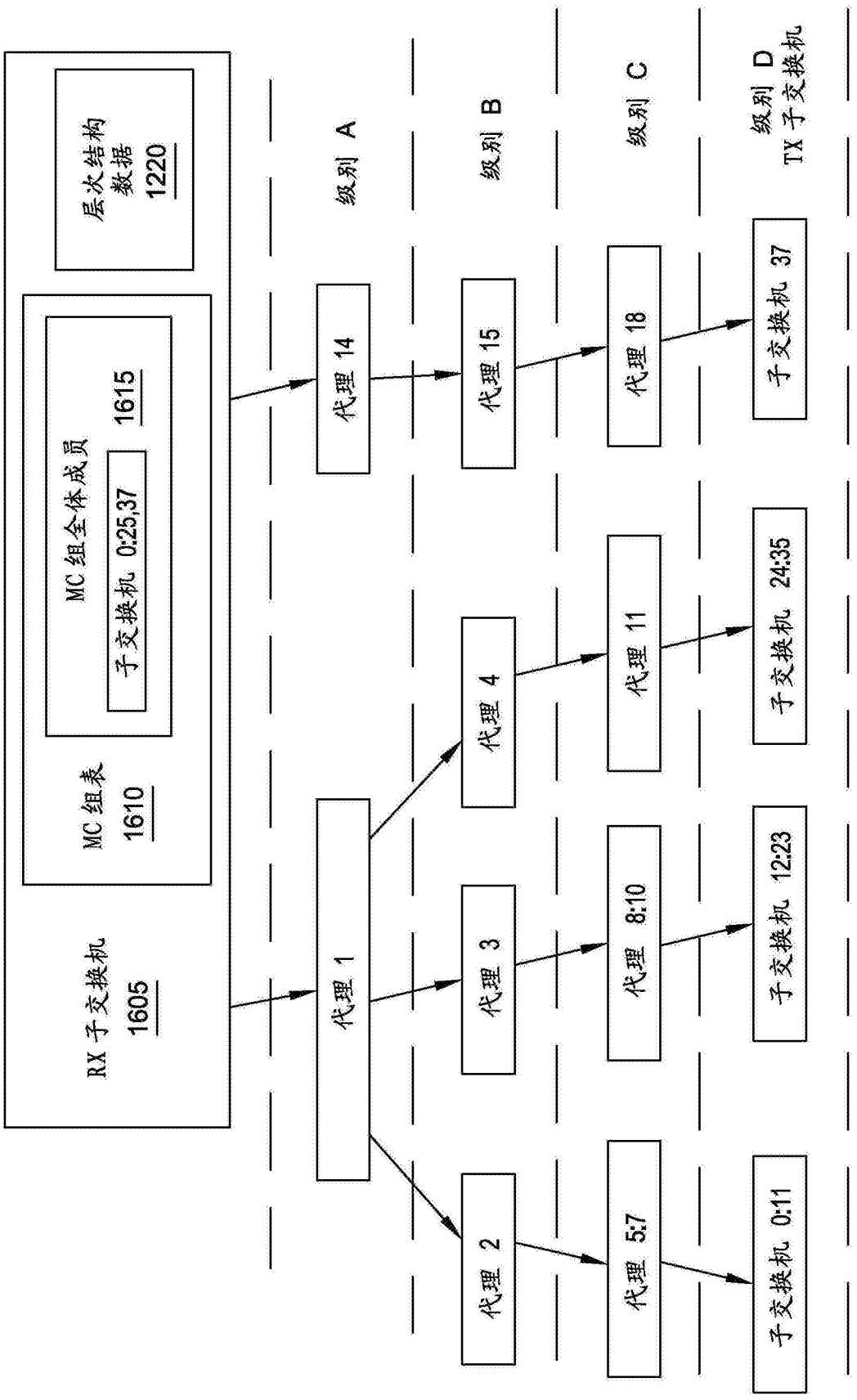

RX 子交换机
1605
MC 组表
1610
MC 组全体成员
子交换机 0;25;37
1615
层次结构
数据
1220
级别 A
级别 B
级别 C
级别 D
TX 子交换机
代理 1
代理 2
代理 3
代理 4
代理 5;7
代理 8;10
代理 11
代理 14
代理 15
代理 18
子交换机 0:11
子交换机 12;23
子交换机 24;35
子交换机 37

图16A

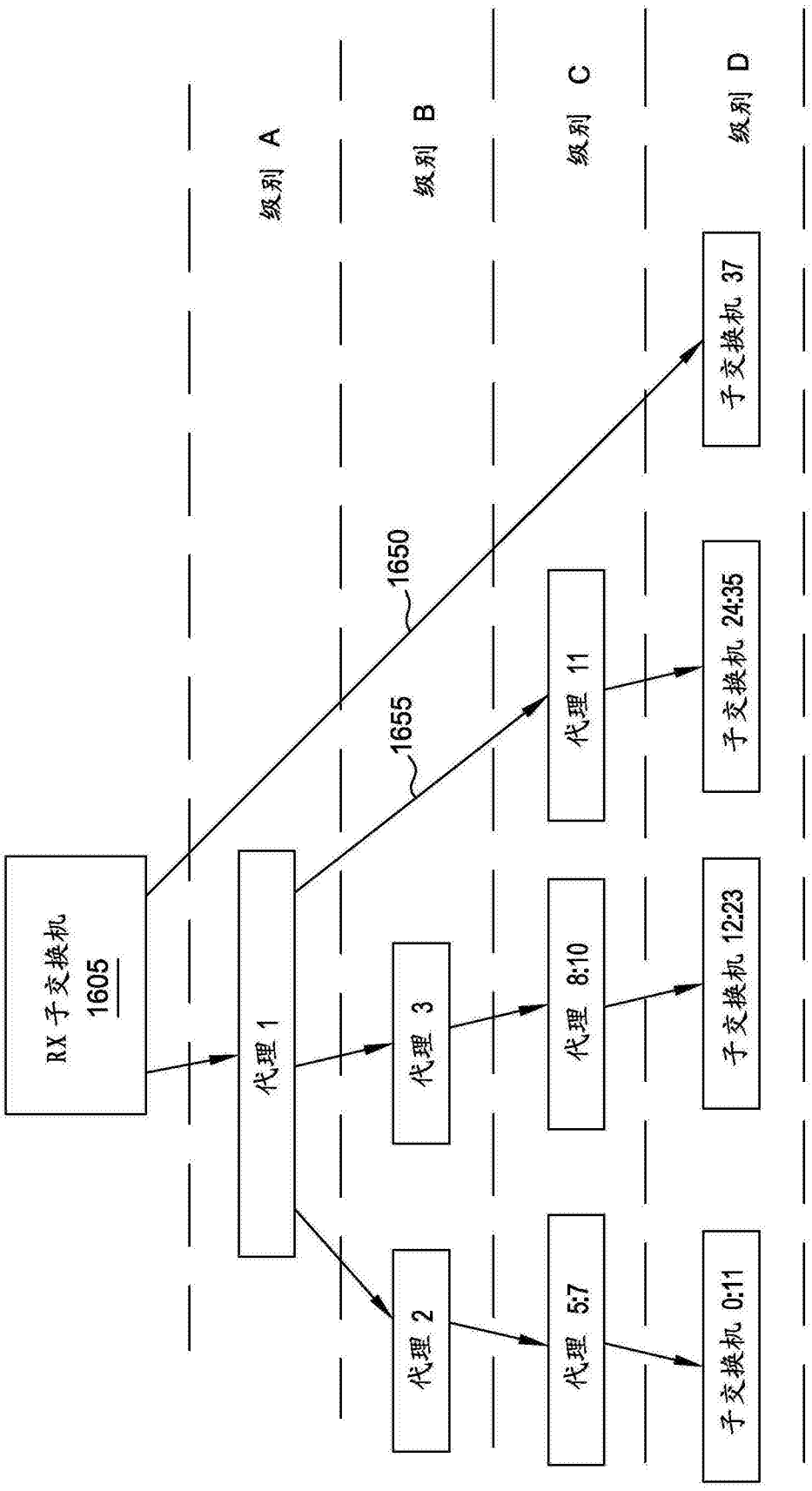

级别 A
级别 B
级别 C
级别 D
RX 子交换机
1605
代理 1
代理 2
代理 3
代理 5:7
代理 8:10
代理 11
子交换机 0:11
子交换机 12:23
子交换机 24:35
子交换机 37
1650
1655

图16B

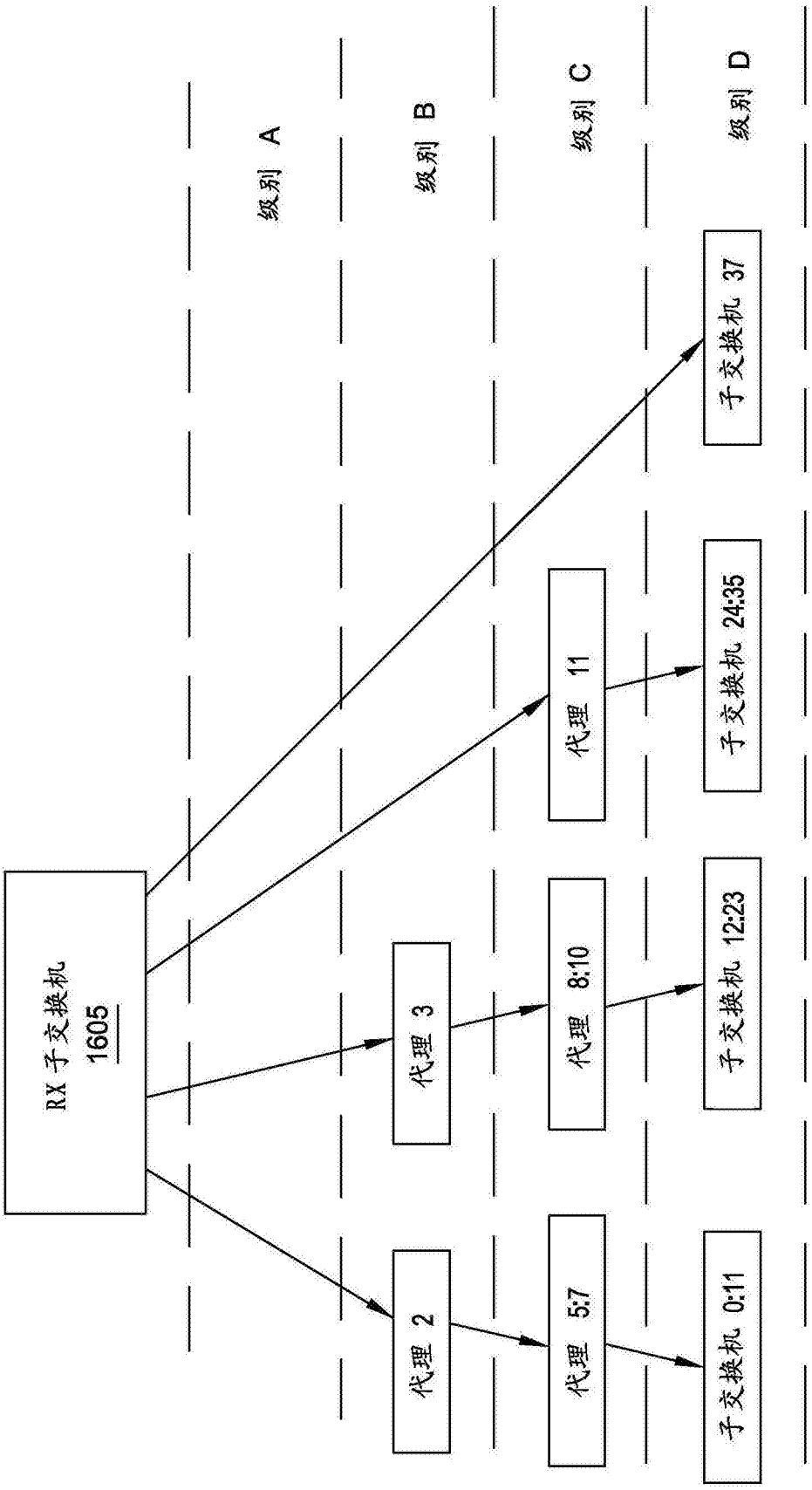

级别 A
级别 B
级别 C
级别 D
RX 子交换机
1605
代理 2
代理 3
代理 5:7
代理 8:10
代理 11
子交换机 0:11
子交换机 12:23
子交换机 24:35
子交换机 37

图16C

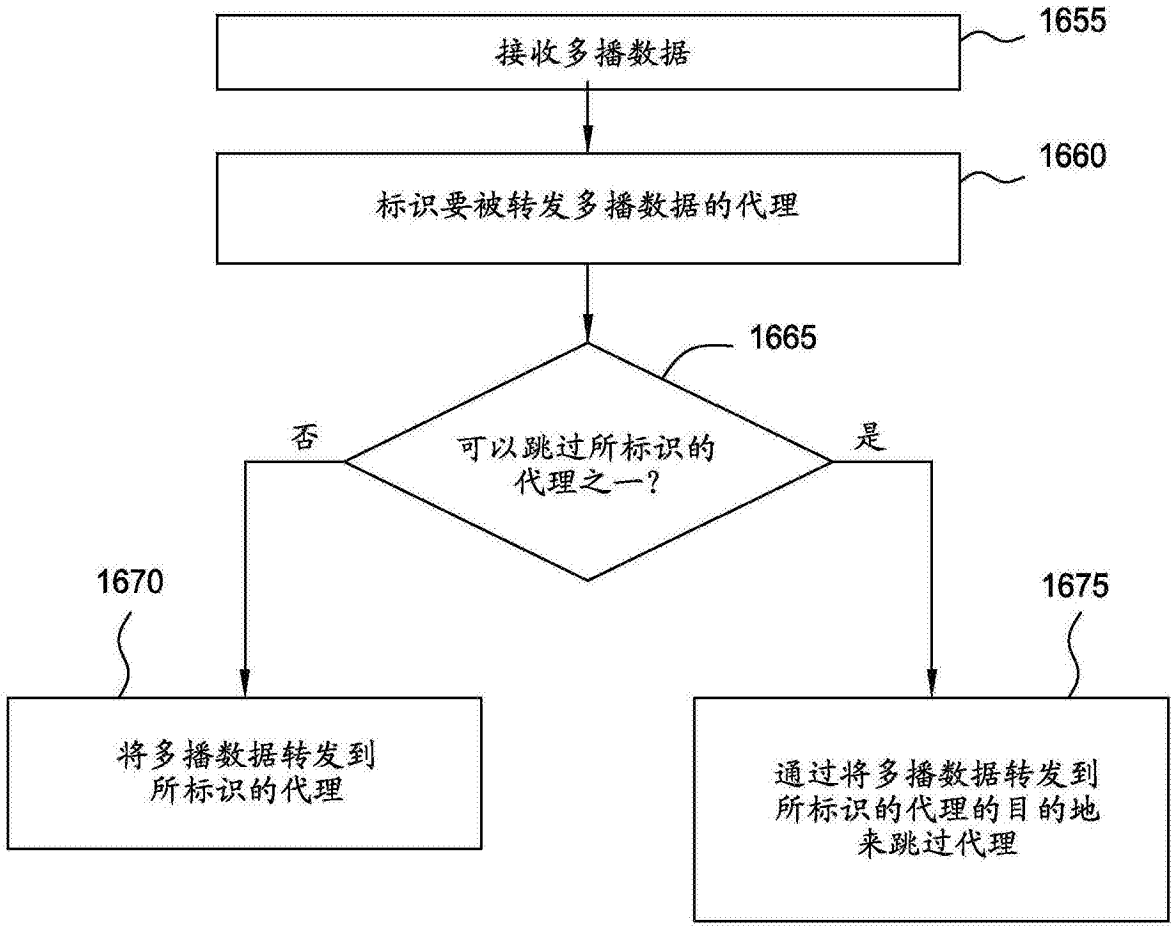

接收多播数据
1655
标识要被转发多播数据的代理
1660
1665
可以跳过所标识的代理之一?
否
是
1670
1675
将多播数据转发到所标识的代理
通过将多播数据转发到所标识的代理的目的地来跳过代理

图16D

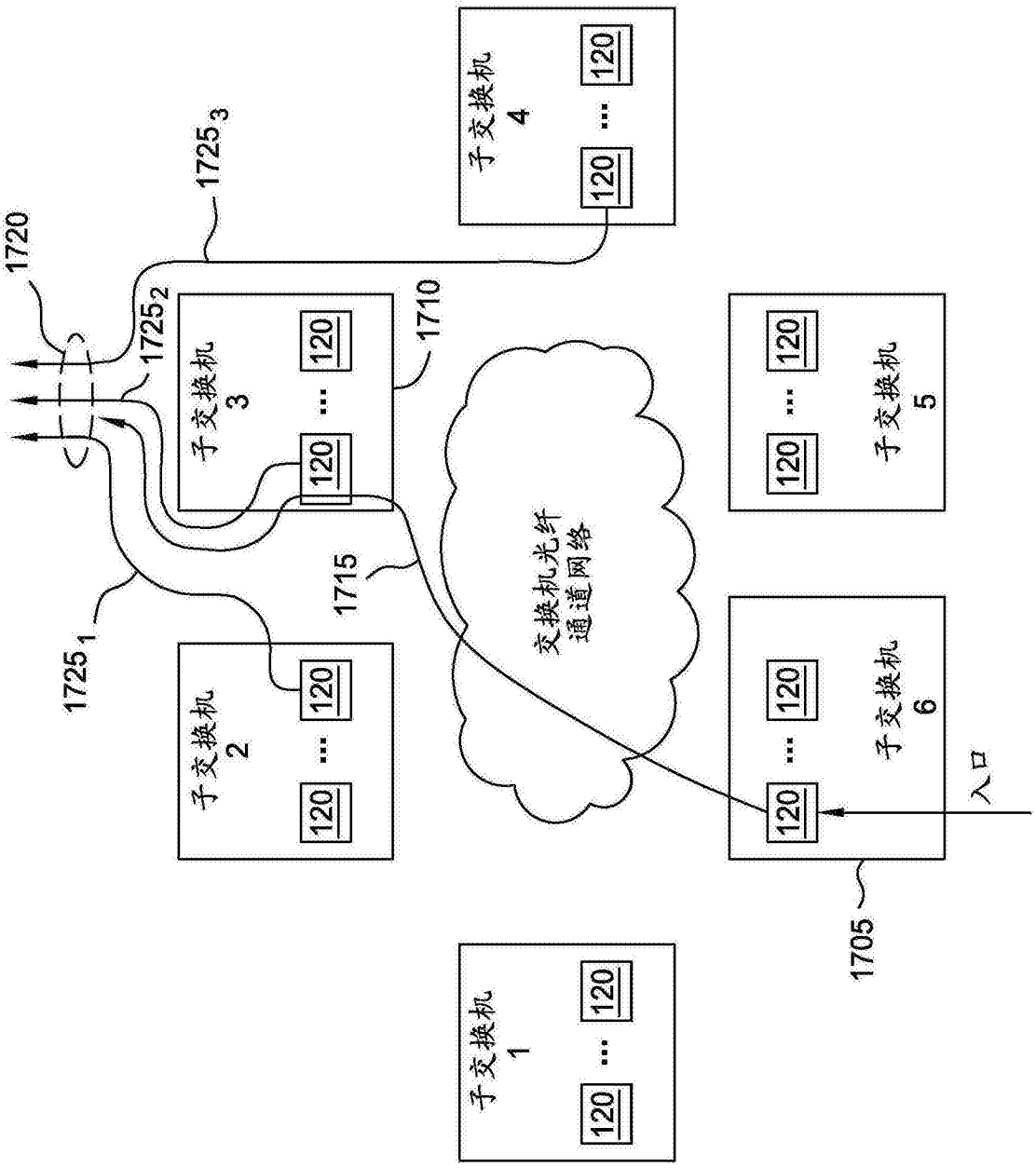

1720
1725 1
1725 2
1725 3
1710
1715
1705
入口
交换机光纤
通道网络
子交换机
1
子交换机
2
子交换机
3
子交换机
4
子交换机
5
子交换机
6
120

图17

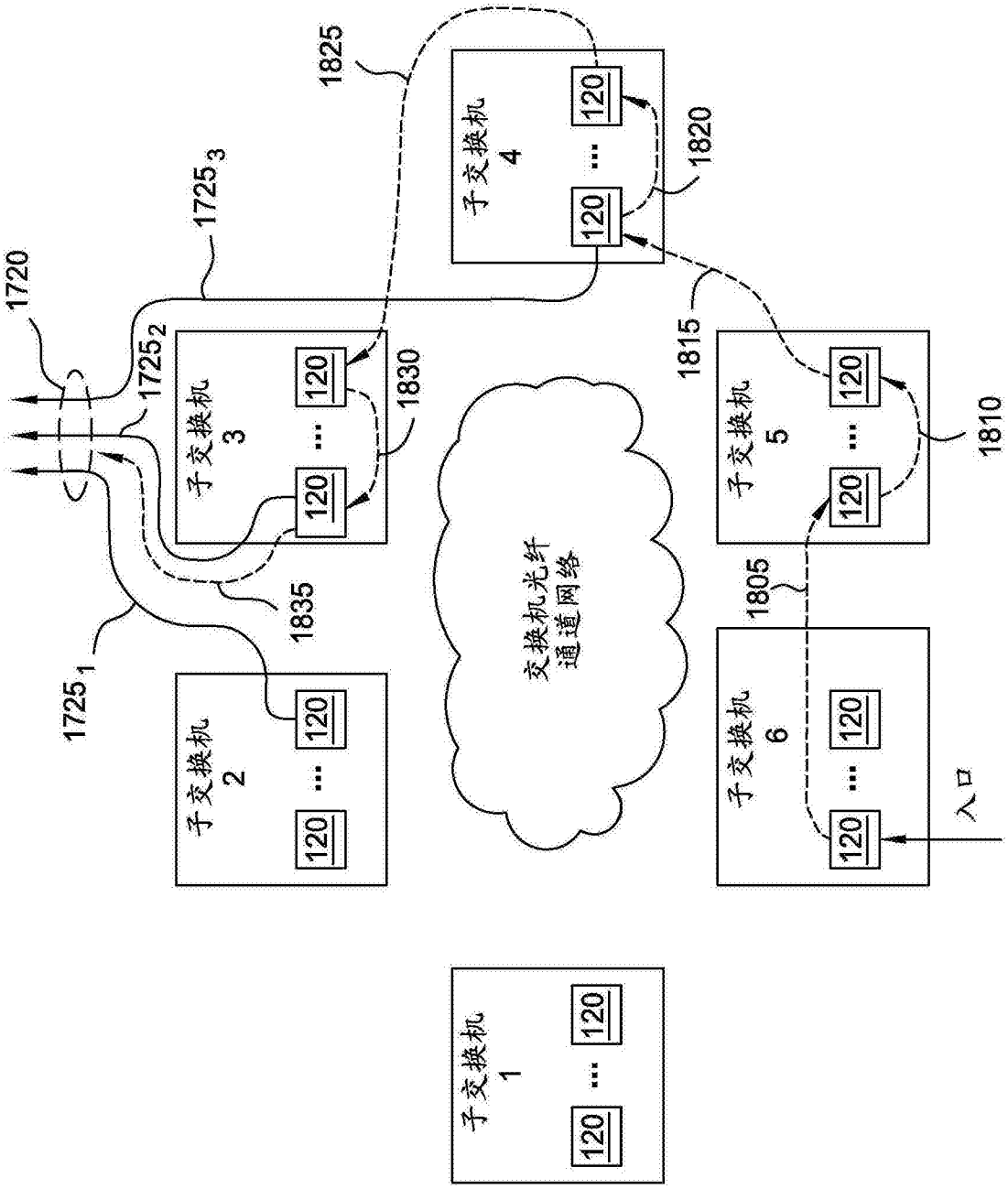

1825
1725 3
1720
1725 2
1725 1
1835
1830
1820
1815
1810
1805
子交换机 1
子交换机 2
子交换机 3
子交换机 4
子交换机 5
子交换机 6
120
交换机光纤通道网络
入口

图18

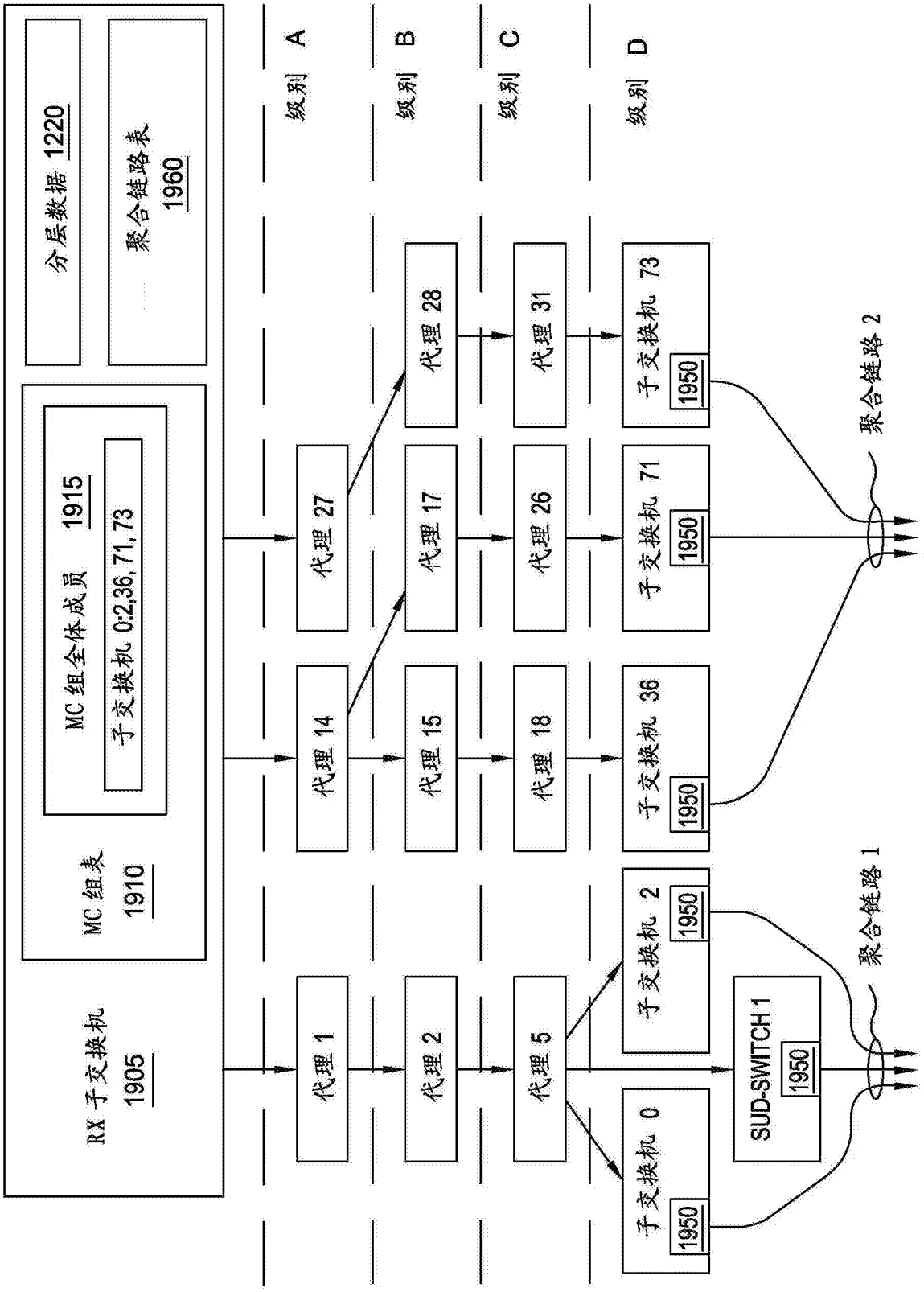

RX 子交换机 1905
MC 组表 1910
MC 组全体成员 1915
子交换机 0:2,36,71,73
分层数据 1220
聚合链路表 1960
级别 A
级别 B
级别 C
级别 D
代理 1
代理 2
代理 5
代理 14
代理 15
代理 18
代理 27
代理 17
代理 26
代理 28
代理 31
子交换机 0 1950
子交换机 2 1950
SUD-SWITCH 1 1950
子交换机 36 1950
子交换机 71 1950
子交换机 73 1950
聚合链路 1
聚合链路 2

图19

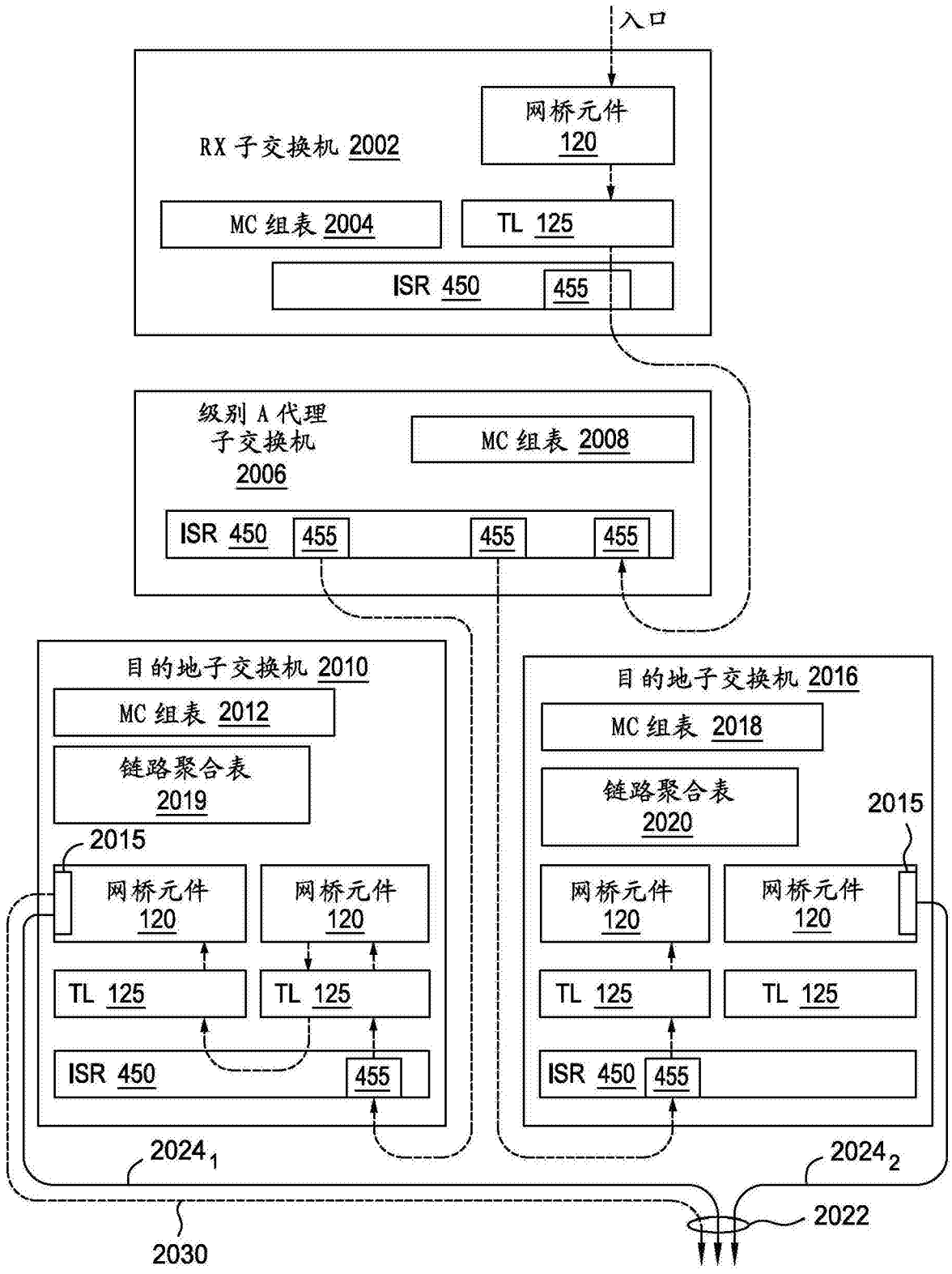

入口
网桥元件
120
RX 子交换机 2002
MC 组表 2004
TL 125
ISR 450
455
级别 A 代理
子交换机
2006
MC 组表 2008
ISR 450
455
455
455
目的地子交换机 2010
MC 组表 2012
链路聚合表
2019
2015
网桥元件
120
网桥元件
120
TL 125
TL 125
ISR 450
455
目的地子交换机 2016
MC 组表 2018
链路聚合表
2020
2015
网桥元件
120
网桥元件
120
TL 125
TL 125
ISR 450
455
2024 1
2024 2
2022
2030

图20A

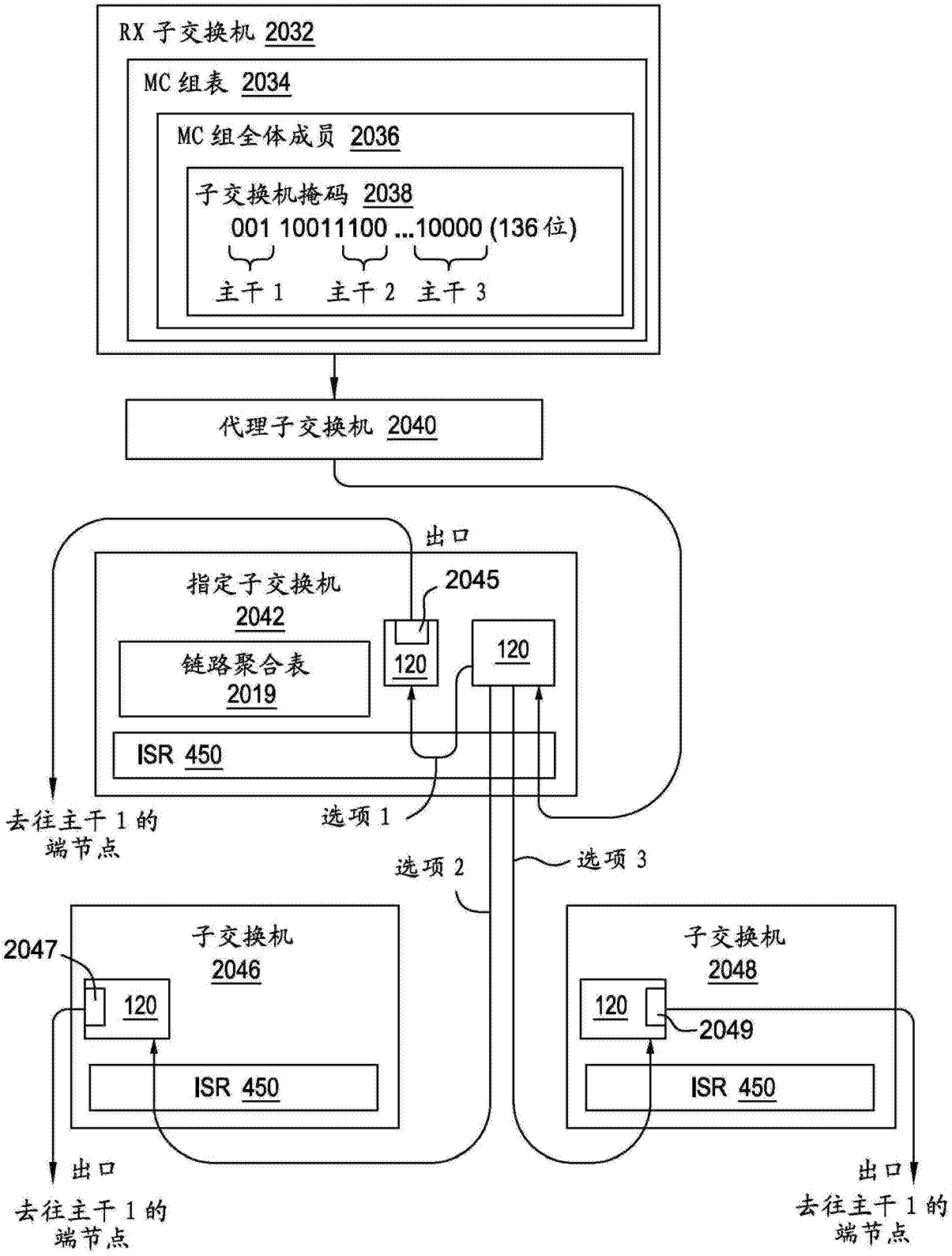

RX 子交换机 2032
MC 组表 2034
MC 组全体成员 2036
子交换机掩码 2038
001 10011100 ...10000 (136 位)
主干 1
主干 2
主干 3
代理子交换机 2040
出口
指定子交换机
2042
2045
链路聚合表
2019
120
120
ISR 450
去往主干 1 的
端节点
选项 1
选项 2
选项 3
2047
子交换机
2046
120
ISR 450
子交换机
2048
120
2049
ISR 450
出口
去往主干 1 的
端节点
出口
去往主干 1 的
端节点

图20B

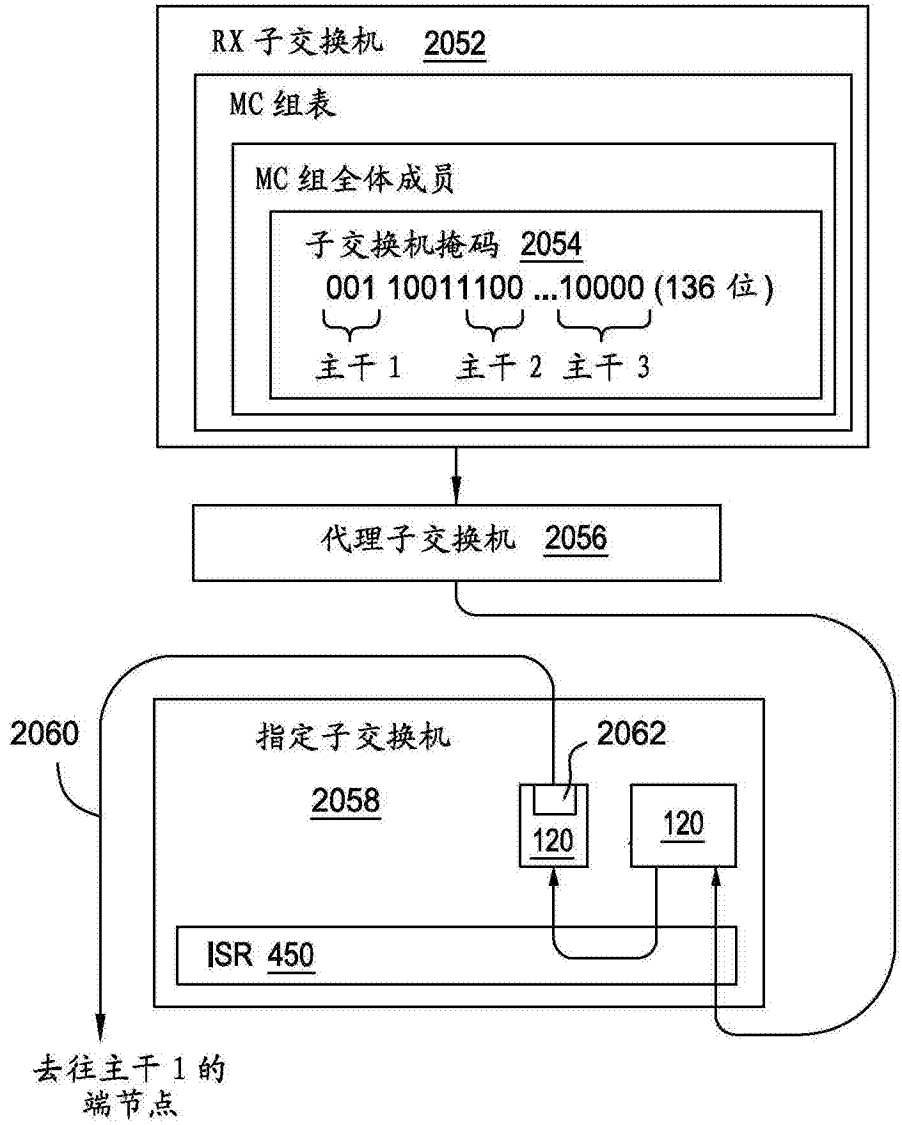

RX 子交换机 2052
MC 组表
MC 组全体成员
子交换机掩码 2054
001 10011100 ...10000 (136 位)
主干 1
主干 2
主干 3
代理子交换机 2056
2060
指定子交换机
2058
2062
120
120
ISR 450
去往主干 1 的
端节点

图20C