(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 1663194 B
(45)授权公告日 2010.06.23

(21)申请号 03807560.1

(22)申请日 2003.03.25

(30)优先权数据
10/114,568 2002.04.01 US

(85)PCT申请进入国家阶段日
2004.09.29

(86)PCT申请的申请数据
PCT/US2003/009328 2003.03.25

(87)PCT申请的公布数据
WO2003/085910 EN 2003.10.16

(73)专利权人 思科技术公司
地址 美国加利福尼亚州

(72)发明人 毛里利奥·科梅托 斯科特·S·李

(74)专利代理机构 北京东方亿思知识产权代理有限责任公司 11258
代理人 王怡

(51)Int.Cl.
H04L 12/56(2006.01)

(56)对比文件
EP 0772121 A1,1997.05.07, 全文.
CN 1332546 A,2002.01.23, 全文.
Ezio Valdevit.Fabric shortest Path First Version 2.2000,2-4,16.
Network Working Group.Multiprotocol Label Switching Architecture.2001,20-24.
Brocade.Optimizing the performance and managementof2Gbit/secSAN fabrics with ISL trunking.2002,1-2.
Venkat Rangan.RE:FCIP/iFCP:Guarantee IN-order delivery for FC N/NL_ports.2001,1.

审查员 张江波

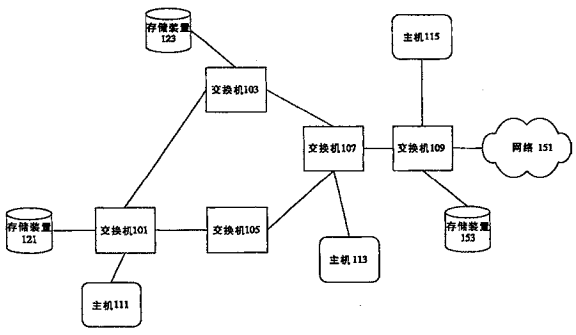
权利要求书 2 页 说明书 10 页 附图 12 页

(54)发明名称

用于在光纤信道结构中选择性地传递帧的方法和系统

(57)摘要

本发明提供了用于改进光纤信道帧传递的方法和装置。所提供的技术通过智能地延迟或丢弃所选择的光纤信道帧,以获得帧的良序传递。所提供的其他技术通过使用标签交换和帧标签来获得良序传递。所述各种技术可应用于链路状态或信道改变等场合中。



CN 1663194 B

1. 一种方法，用于在光纤信道结构中选择性地传递帧，该方法包括：

在具有路由表的光纤信道交换机处检测光纤信道结构链路改变；

生成与拓扑版本号相关联的更新路由表，其中生成更新路由表包括确定对应于所述更新路由表中每个条目的下一跳、进入标签和目的地；

在光纤信道交换机处接收帧，所述帧具有对应于所述更新路由表中第一条目的第一目的地和第一标签，其中所述第一标签对应于进入标签；

确定所述光纤信道交换机是否已接收到与所述更新路由表中的所述第一条目对应的第一外发标签，其中所接收到的第一外发标签具有与所述更新路由表相同的拓扑版本号；以及

如果确定光纤信道交换机尚未接收到所述第一外发标签，则丢弃所述帧。

2. 如权利要求 1 所述的方法，其中所述第一标签等同于与所述第一条目相关联的第一进入标签。

3. 如权利要求 1 所述的方法，还包括：

如果确定所述光纤信道交换机已经接收到所述第一外发标签，则将所述帧转发到所述下一跳。

4. 如权利要求 3 所述的方法，其中所述路由表与特定的虚拟存储区域网络相关联。

5. 如权利要求 4 所述的方法，其中所述拓扑版本号是使用所述虚拟存储区域网络中每个光纤信道交换机的标称号来导出的。

6. 如权利要求 5 所述的方法，其中所述第一标签和第一外发标签是 MPLS 标签。

7. 如权利要求 1-6 中任意之一所述的方法，还包括：

接收第一链路状态控制消息，该消息具有所述第一外发标签，并具有与所述路由表的拓扑版本号相对应的拓扑版本号。

8. 如权利要求 7 所述的方法，还包括：

向所述光纤信道结构中的其他光纤信道交换机公告所述第一标签。

9. 一种系统，用于在光纤信道结构中选择性地传递帧，该系统包括：

用于在具有路由表的光纤信道交换机处检测光纤信道结构链路改变的装置；

用于生成与拓扑版本号相关联的更新路由表的装置，其中生成更新路由表包括确定对应于所述更新路由表中每个条目的下一跳、进入标签和目的地；

用于在光纤信道交换机处接收帧的装置，所述帧具有对应于所述更新路由表中第一条目的第一目的地和第一标签，其中所述第一标签对应于进入标签；

用于确定所述光纤信道交换机是否已接收到与所述更新路由表中的所述第一条目对应的第一外发标签的装置，其中所接收到的第一外发标签具有与所述更新路由表相同的拓扑版本号；以及

用于如果确定光纤信道交换机尚未接收到所述第一外发标签，则丢弃所述帧的装置。

10. 如权利要求 9 所述的系统，其中所述第一标签等同于与所述第一条目相关联的第一进入标签。

11. 如权利要求 9 所述的系统，还包括：

用于如果确定所述光纤信道交换机已经接收到所述第一外发标签，则将所述帧转发到所述下一跳的装置。

12. 如权利要求 11 所述的系统，其中所述路由表与特定的虚拟存储区域网络相关联。

13. 如权利要求 12 所述的系统，其中所述拓扑版本号是使用所述虚拟存储区域网络中每个光纤信道交换机的标称号来导出的。

14. 如权利要求 13 所述的系统，其中所述第一标签和第一外发标签是 MPLS 标签。

15. 如权利要求 9-14 中任意之一所述的系统，还包括：

用于接收第一链路状态控制消息的装置，该消息具有所述第一外发标签，并具有与所述路由表的拓扑版本号相对应的拓扑版本号。

16. 如权利要求 15 所述的系统，还包括：

用于向所述光纤信道结构中的其他光纤信道交换机公告所述第一标签的装置。

用于在光纤信道结构中选择性地传递帧的方法和系统

[0001] 相关申请的交叉引用

[0002] 本申请与Scott S.Lee和Dinesh G.Dutt同时递交的美国专利申请No.10/114,394(律师案卷号No.ANDIP009)相关,该申请题为“LabelSwitching In Fibre Channel Networks”,作为整体在此通过引用而被包含,用于各种目的。

技术领域

[0003] 本发明涉及光纤信道网络。更具体地说,本发明涉及用于在链路状态发生改变或信道状态发生改变的情况下,在光纤信道网络中提供光纤信道帧的良序(in order)传递的方法和装置。

背景技术

[0004] 许多传统网络协议允许分组序列的乱序(out of order)传递。基于TCP/IP的网络中的网络节点可以接收乱序分组集合,并在接收后对分组进行重排序。如果分组是沿着不同的路径到达目的地的,则它们经常是乱序到达的。

[0005] 然而,一些光纤信道设备例如磁盘、磁盘阵列或其他存储机构不能处理乱序帧。链路和信道状态改变是一些可能会在光纤信道结构中引起帧的乱序传递的情况。两个光纤信道实体间被视为单个链路的多个链路在此被称为一个信道。现有网络中的一些机制要求在链路状态改变时冲刷(flush)网络中的所有帧。当网络中的路径和路由改变时,冲刷所有的帧可以防止乱序传递。即使指向关联目的地的帧的路径没有改变时,也会冲刷所有的帧。然而,不管是显示地还是隐式地冲刷所有的帧,都会严重扰乱网络的运行,因为过多的帧被丢弃,并且网络的运行至少会暂时被停顿。

[0006] 因此,希望可以提供一些方法和装置,以用于改进光纤信道帧传递,并且尤其是在链路状态和信道改变时提供良序传递。

发明内容

[0007] 本发明提供了一些方法和装置来改进光纤信道帧传递。提供了用于通过智能地延迟或丢弃所选择的光纤信道帧,来获得良序帧传递的技术。其他用于获得良序传递的技术是通过使用标签交换和帧标签来提供的。所述各种技术可在链路或信道状态改变等情况期间应用。

[0008] 根据各个实施例,提供了一种方法,用于在光纤信道结构中选择性地传递光纤信道帧。在光纤信道实体处标识一组下一跳(next hop)。所述下一跳组被用于基于目的地标识符来转发帧。检测光纤信道结构链路改变。标识一组更新的下一跳。所述更新的下一跳组被用于在考虑到光纤信道结构链路改变的同时,基于目的地标识符来转发在光纤信道实体处接收的帧。将所述下一跳组与所述更新的下一跳组进行比较。如果确定所述下一跳组与所述更新的下一跳组不同,则在一段预定的时间内防止朝向所述更新的下一跳组的帧转发。

[0009] 根据另一个实施例，提供了一种方法，用于在光纤信道结构中选择性地传递帧。标识一个用于将帧从第一光纤信道实体转发到第二光纤信道实体的信道。该信道包括多个将所述第一光纤信道实体连接到所述第二光纤信道实体的链路。在第一光纤信道实体处检测所述信道的改变。标识一个用于将帧从第一光纤信道实体转发到第二光纤信道实体的更新信道。该更新信道与所述信道不同。阻塞所接收的用于在所述更新信道上转发，但尚未被置入与所述更新信道相关联的输出队列中的帧。

[0010] 在另一个实施例中，提供了一种方法，用于在光纤信道结构中选择性地传递帧。在光纤信道交换机处检测光纤信道结构链路改变。生成与拓扑版本号相关联的更新路由表。“生成更新路由表”包括确定对应于所述更新路由表中每个条目的下一跳、进入标签(incoming label)和目的地。在光纤信道交换机处接收到帧。所述帧包括对应于所述更新路由表中第一条目的第一目的地和第一标签。确定光纤信道交换机是否已接收到与所述更新路由表中的所述第一条目对应的第一外发标签(outgoing label)，其中所接收到的第一外发标签具有与所述更新路由表相同的拓扑版本号。如果确定光纤信道交换机尚未接收到所述第一外发标签，则丢弃所述帧。

[0011] 在下面对本发明的说明以及附图中，将更详细地描述本发明的上述及其他特征和优点，所述说明和附图例示性地说明了本发明的原理。

附图说明

[0012] 参考下面的描述并结合附图可最佳地理解本发明，所述附图说明了本发明的具体实施例。

[0013] 图 1 示意性地表示了可以使用本发明的技术的网络。

[0014] 图 2 示意性地表示了经受光纤信道结构链路改变的光纤信道结构。

[0015] 图 3A 示意性地表示了一个路由表，其示出了一组下一跳。

[0016] 图 3B 示意性地表示了一个路由表，其示出了一组更新的下一跳。

[0017] 图 4 示意性地示出了虚拟输出队列。

[0018] 图 5 的处理流程图示出了光纤信道帧的阻塞。

[0019] 图 6 的处理流程图示出了光纤信道帧的丢弃。

[0020] 图 7 示意性地表示了信道处可能的重排序。

[0021] 图 8A 示意性地表示了转发信道表。

[0022] 图 8B 示意性地表示了更新的转发信道表。

[0023] 图 9 的处理流程图示出了信道改变时光纤信道帧的转发。

[0024] 图 10 示意性地表示了标签交换路由器。

[0025] 图 11 示意性地表示了链路改变期间的标签交换路由器。

[0026] 图 12 示意性地表示了外标签(out label)被部分解析期间的标签交换路由器。

[0027] 图 13 的处理流程图示出了使用输入和外发标签，用于获得良序传递的技术。

具体实施方式

[0028] 下面将详细描述本发明的一些具体实施例，其中包括了发明人认为是实施本发明的最佳模式。这些具体实施例的示例被示出在附图中。尽管本发明是结合具体实施例来描

述的，但是应当理解到这并非是将本发明局限于所描述的实施例上。相反，如果一些替换方案、修改和等同物可以被包含在本发明如所附权利要求所定义的精神和范围之内，就应覆盖这些替换方案、修改和等同物。

[0029] 本发明的方法和装置提供了光纤信道帧的良序传递。根据各个实施例，若干网络状况可能会导致帧到光纤信道设备的乱序传递。本发明的技术提供了对某些光纤信道帧的延迟、阻塞、丢弃和/或标签化，以将帧良序传递给光纤信道设备。在一个实施例中，将要遍历新路径的帧被阻塞，以允许沿着旧路径行进的帧或者先到达目的地，或者从网络中被丢弃。

[0030] 图1示意性地表示了一种网络示例，其可以使用本发明的技术。图1示出了使用光纤信道实现的存储区域网络。交换机101耦合到交换机103和105，还耦合到主机111和存储装置121。在一个实施例中，主机111是服务器或客户端系统，而存储装置121是任何存储子系统，例如单个磁盘或独立磁盘冗余阵列(RAID)。交换机105耦合到交换机107。交换机107连接到主机113，而交换机103连接到存储装置123。交换机109连接到主机115、交换机107、存储装置153和外部网络151，外部网络151可能使用也可能不使用光纤信道。为了使主机111访问网络151，可以使用经过交换机105的路径。应注意，任何包括处理器、存储器和到光纤信道结构的连接的装置都可以是光纤信道交换机。

[0031] 用于在光纤信道网络中将交换机彼此连接的端口在此被称为非F端口，而用于将交换机连接到主机的端口在此被称为F端口。在一个示例中，非F端口被用于将交换机105连接到交换机107，而F端口被用来将交换机107连接到主机113。类似地，FL端口被用来将交换机103连接到存储装置123。F端口和FL端口等端口在此被称为边缘端口。其他端口被称为非边缘端口。

[0032] 根据各个实施例，从主机111发送到网络151或存储设备153的帧包括诸如交换标识符、序列和序列号等参数。交换标识符可以提供与该帧属于哪个交换有关的信息。序列可以提供与该帧属于所述交换的哪个部分有关的信息，而序列号可以提供与多个帧应如何被排序有关的信息。序列号可用来实现光纤信道帧的良序传递。

[0033] 一些光纤信道设备，例如某些存储磁盘和磁盘阵列，需要以帧被发送的顺序来接收帧。传统网络例如TCP/IP网络没有这一需求，因为TCP/IP网络一般具有在接收时对帧进行重新排序的机制。如果在光纤信道网络中依次发送具有序列号191、192和193的帧，则接收所述帧的光纤信道设备可能期望所述帧具有与它们被发送时相同的顺序。光纤信道设备可能不能够处理帧的乱序接收。

[0034] 在静态光纤信道网络中，帧一般是以它们的发送顺序被接收的。然而，一些情况可能会导致光纤信道帧的乱序传递。尤其是链路状态改变可能会导致乱序传递。

[0035] 图2示意性地表示了经受光纤信道结构链路改变的光纤信道结构。图2示出了链路改变的一个示例，其可能会导致光纤信道帧的乱序传递。在交换机103和交换机107之间引入了一个使用非边缘端口的新链路。随着交换机103和交换机107之间新链路的引入，可以生成新版本的路由表。可以使用多种路由表生成算法，例如光纤信道最短路径优先(FSPF)。从主机111开始，经过交换机101、105和107而到达交换机109的流量现在可以经过交换机101、103和107到达交换机109。在新链路引入之前，可用于将帧从交换机101发送到存储设备153的下一跳组是交换机105。

[0036] 可用于将帧从一个光纤信道实体发送到另一个光纤信道实体的一组相邻光纤信道实体在此被称为一组下一跳。在链路改变之后，可以在交换机中更新下一跳组。在一个示例中，在链路状态改变后，用于将帧从交换机 101 发送到网络 151 的下一跳组从只有交换机 105 改变为交换机 103 和交换机 105 二者。可用于将帧从源发送到目的地，并且在链路改变或生成更新的路由表后被更新的一组相邻实体在此被称为一组更新的下一跳。应注意，下一跳组可包括一个或多个相邻节点。在一个示例中，下一跳组是单个相邻实体。在另一个示例中，下一跳组包括多个相邻实体。

[0037] 具有更新的下一跳组可能会导致光纤信道帧的乱序传递。在一个示例中，交换机 101 处在一个序列中发送的早先的帧可能会经过交换机 105，而相同序列中后来的帧可能会经过交换机 103。多种网络状况可能会使得后来的帧在早先的帧经过交换机 105 到达交换机 109 之前，就经过交换机 103 到达交换机 109。在一个示例中，早先的帧由于交换机 105 处的拥塞而在交换机 105 处变慢，而后来的帧由于交换机 103 和交换机 107 之间的新的高带宽链路而快速通过交换机 103。存储设备 153 先接收到从主机 111 后发送的帧，而后接收先发送的帧，它可能不能够处理这些乱序的帧。

[0038] 图 3A 和 3B 图示了两个路由表，其示出了与交换机 101 处的下一跳组和更新的下一跳组有关的信息。图 3A 示出了用于在交换机 101 处接收的帧、以及尚未在交换机 103 和交换机 107 之间建立链路的网络的下一跳组。当在交换机 101 处接收到帧时，示出该帧目的地的标识符可用于引用路由表中的一个条目。在一个示例中，帧的目的地是交换机 107，可引用条目 309 来确定下一跳组是交换机 105。如果确定所接收的帧的目的地是交换机 101，则可使用该路由表来将帧传递给与交换机 101 相关联的处理器。路由表也可用来丢弃帧。在一个示例中，可将一个值例如空值置入下一跳组中，并可在参考路由表后丢弃具有与空值相关联的目的地的帧。在此，将命令丢弃具有特定目的地的帧的条目称为邻接丢弃(adjacencydrop)。

[0039] 在一个实施例中，为交换机所属的每个虚拟存储区域网络 (VSAN) 都提供了路由表。应注意，一个光纤信道交换机可以是许多不同 VSAN 的一部分，可以为该交换机与之相关联的每个 VSAN 都提供路由表。

[0040] 在增加了连接交换机 103 和交换机 107 的链路后，更新路由表。图 3B 示意性地表示了一个路由表，其示出了更新的下一跳组。根据各个实施例，基于条目 329，其目的地被设置为交换机 107 的帧可沿着交换机 103 或交换机 105 转发。在稳定的拓扑中，一个特定流或交换中的所有的帧都遵循相同的路径。

[0041] 应注意，路由表可允许两个路径，或者它可以选择一个最佳路径。如果所选择的最佳路径所具有的下一跳是交换机 105，则更新的下一跳组与图 3A 所示链路改变前的下一跳组相同。如果为目的地是交换机 107 的帧选择的路径是交换机 103，则更新的下一跳组与链路改变前路由表中的下一跳组不同。确定更新的下一跳组与初始的下一跳组是否不同，这对于决定是否阻塞或丢弃特定帧是有帮助的。在一个示例中，如果即使在光纤信道结构中的链路改变后，特定序列的帧的路径仍保持不变，则帧不会被阻塞或丢弃。如果该序列的帧的路径仍然相同，则所述帧会被良序传递到目的地。如果该序列的帧的路径已改变，则存在帧被乱序传递的风险。

[0042] 可能会影响光纤信道帧的传递顺序的机制之一是光纤信道交换机内的队列。首先

从主机发送的帧可能停留在与交换机 105 相关联的队列中，而后来从主机发送的帧可能被快速的传过交换机 103。图 4 示意性地示出了根据各个实施例，可与光纤信道交换机相关联的队列。虽然下面将描述一种特定类型的队列，但是应注意，可使用与各种输入和输出端口相关联的多种不同输入和输出队列来实现本发明的技术。

[0043] 交换机 401 连接到外部节点 451、453、455 和 457。交换机 401 包括与每个交换端口相关联的共享存储器的缓冲区 403。缓冲区 403 与外部节点 451 相关联。为了清楚起见，与外部节点 453、455 和 457 相关联的缓冲区未被示出。缓冲区 403 可保存将发送到外部节点 453、455 和 457 的流量，并可保存指向外部节点 451 的回送流量。

[0044] 在典型的实现方式中，将发送到各个外部节点的帧都被置入相同的缓冲区 403 中。因此，当交换机 401 接收到将要发送到特定节点例如外部节点 453 的大量帧时，与外部节点 453 相关联的帧可使用整个缓冲区 403。根据各个实施例，存储在缓冲区 403 中的帧可通过帧描述符队列 411-447 中的指针来引用。每个帧描述符可包含一个指针或引用，其标识了该帧存储在缓冲区 403 中的何处。指向共享缓冲区的指针或引用在此被称为描述符。描述符还可标识其他信息，例如帧优先级。

[0045] 在一个示例中，仲裁器 405 使用轮转方法来选择帧。在第一轮中，选择将发送到外部节点 453 的帧。在第二轮中，选择将发送到外部节点 455 的帧，等等。更具体地说，仲裁器 405 可首先选择将发送到外部节点 453 并与描述符 411 相关联的高优先级帧，然后选择将发送到外部节点 455 并与描述符 421 相关联的高优先级帧，然后选择将发送到外部节点 457 并与描述符 431 相关联的高优先级帧，等等。应注意，本领域的普通技术人员将会认识到，可以使用多种技术来选择帧。

[0046] 具有基于目的地而被分配的缓冲区的排队系统在此被称为虚拟输出队列 (VOQ)。VOQ 在 Tamir Y.，Frazier G.：“High performance multi-queuebuffers for VLSI communications switches”，Proc.Of 15th Ann.Symp.OnComp.Arch.，pp.343-354，June 1988 中有进一步的描述，该文献的内容通过引用而被整体包含，用于所有目的。标识了两个节点之间具有特定特性的流量的抽象概念在此被称为流 (flow)。在一个示例中，通过源标识符、目的地标识符、优先级、等级和交换标识符来引用流。也可使用其他特性。然而应注意，也可只通过源和目的地标识符来引用流。

[0047] 根据各个实施例，特定流中的帧可能因为缓冲区 403 满而被阻塞。如果对于同一序列中的帧提供了另一个路由，则后来的帧可能会比停留在发生拥塞的交换机中的帧更快地穿过光纤信道结构。在一种实现方式中，后来的帧可沿着另一个路由被阻塞，以使得早先的帧先到达目的地。先发送的帧在此称为在先帧，而后来由源发送的帧在此称为在后帧。可使用若干机制来阻塞在后帧。在一个实施例中，仲裁器 405 可以仅仅是不选择将序列的帧发送到外部节点。在另一个实施例中，在后帧可能根本不被排队以获得发送调度，直到一段时间过去。

[0048] 可以若干方式来确定所述时间段。根据各个实施例，光纤信道交换机被配置成在帧被丢弃前，将该帧缓冲不超过光纤信道交换机延迟的时间。然而应注意，在光纤信道交换机延迟过去之后，一些交换机可能不会丢弃帧。在一个实施例中，如果帧被保存在交换机中的缓冲区中超过延迟时间段，则使帧从虚拟输出队列中出列并丢弃。光纤信道交换机延迟可依赖于交换速度和网络拥塞。帧在被丢弃前停留在光纤信道交换机中的时间量在此被称

为光纤信道交换机延迟。

[0049] 帧在被丢弃前停留在光纤信道网络中的时间量在此被称为光纤信道结构漏延迟或网络漏延迟。根据各个实施例，将光纤信道交换机延迟乘以帧遍历光纤信道结构所花费的最大跳数，从而计算出光纤信道结构漏延迟或网络漏延迟。可使用很多种技术来确定光纤信道交换机延迟和光纤信道结构漏延迟。

[0050] 图 5 的处理流程图示出了来自所连接的主机或磁盘的帧的转发。在 501，检测到光纤信道结构链路改变。如上所述，光纤信道结构链路改变可能会导致帧序列到光纤信道设备的乱序传递。根据各个实施例，可基于链路更新消息的接收或链路更新消息的发送来检测光纤信道结构链路改变。在 503，基于新的链路状态信息，为每个目的地生成更新的下一跳组。可以使用诸如光纤信道最短路径 (FSPF) 之类的算法来生成更新的下一跳组。利用与网络拓扑有关的新信息，交换机能更好地确定什么是最佳路径以将帧发送到特定目的地。在 505，为每个目的地确定下一跳组是否等于更新的下一跳组。如果在 507 确定所有的下一跳组都与对应的更新的下一跳组相等，则不采取任何动作。

[0051] 例如，如果将帧发送到目的地的初始路径是经过交换机 103 和 107，而发送帧到目的地的新路径也是经过交换机 103 和 107，则不需要任何动作。然而，如果将帧发送到目的地的更新路径是 105 和 107，则更新的下一跳组与初始的下一跳组不等同。可以应用阻塞和丢弃机制。在 509，对于其下一跳组不同于对应的更新的下一跳组的每个目的地，阻塞用于发送具有所述目的地的帧的队列或虚拟输出队列。根据各个实施例，阻塞虚拟输出队列可能需要不将帧发送到下一跳。在 511，使用更新的下一跳组来更新路由表。然后在 513，在结构漏延迟或结构漏时间段 (drain period) 期间阻塞其下一跳组有改变的队列。

[0052] 结构漏延迟使得有时间将仍停留在拥塞网络交换机中的在先分组或者被传递，或者从网络中丢弃。与每个虚拟输出队列相关联的可能是不同的结构漏时间段，或者单一的结构漏时间段可适用于所有的虚拟输出队列。在结构漏时间段过去后，在 515 可解除对被阻塞的队列的阻塞。现在可发送被阻塞队列中的在后帧，因为在先帧或者已被丢弃，或者已被传递到目的地。可以将在后帧发送到目的地，而不会存在早先发送的帧仍在网络中等待乱序到达目的地的风险。

[0053] 图 5 中描述的技术可以以逐一 VSAN、逐一目的地的方式应用于光纤信道结构中的任何端口。然而，根据各个实施例，图 5 中描述的技术被应用于将交换机连接到主机或磁盘的边缘端口。用于边缘端口的技术集中于阻塞流量，以避免注入可能会被丢弃的流量。

[0054] 图 6 的处理流程图示出了光纤信道交换机之间的帧转发。在 601，检测到光纤信道结构链路改变。在 603，为每个目的地计算更新的下一跳组。在 605，对于每个目的地，确定下一跳组是否等于更新的下一跳组。如果在 607 所有的下一跳组都等于更新的下一跳组，则不需要采取任何动作。如果不是所有的下一跳组都等于对应的更新的下一跳组，则在 609，丢弃具有与改变了的下一跳组相关联的目的地的帧。

[0055] 在 611 可以更新路由表，并且在 613，光纤信道交换机在与虚拟输出队列相关联的对应的结构漏时间段期间等待。对于特定队列，在结构漏时间段过去之后，本来将被置于与将结构漏时间段相关联的队列中的帧现在被转发，而不是被丢弃。

[0056] 如上所述，本发明的技术可应用于在检测到链路改变后，使得在光纤信道结构中实现光纤信道帧的良序传递。一般地，当增加新节点或新链路，或者从光纤信道拓扑中减掉

旧的节点或旧链路时，就存在光纤信道帧的乱序传递的风险。

[0057] 然而，网络拓扑的改变并不是可能触发乱序传递的唯一事件。两个交换机之间的信道的改变也可导致乱序帧传递。

[0058] 图 7 示意性地表示了可能会导致乱序传递的信道改变。交换机 107 和 109 原来可通过冗余链路 701、703 和 705 而互连，这些链路形成信道 709。在交换机 107 和 109 之间传输的流量可基于诸如公平性和负载平衡等因素而在信道 709 中的不同链路之间分布。在另一个示例中，在发送和接收交换机处可使用相同的哈希函数来确定下一次访问哪个链路。使用相同的哈希函数可提供良序传递。

[0059] 当增加新链路 707 以形成更新的信道 711 时，现在有更多的链路可用于交换机 107 和交换机 109 之间的传输。然而应注意，网络拓扑没有发生改变。交换机 107 和交换机 109 仍连接在光纤信道结构中。根据各个实施例，当信道改变时，没有必要改变或更新路由表。路由表可能只是指示帧将从交换机 107 转发到交换机 109。从交换机 107 到交换机 109 的转发没有被新链路 707 的加入所影响，其中新链路 707 对初始信道 709 的加入形成了更新信道 711。

[0060] 然而，向信道加入链路也可能会引发光纤信道帧的乱序传递。使用上述示例，其中具有序列号 1-6 的帧在链路 701、703 或 705 上传输，具有序列号 7 和 8 的帧沿着链路 707 传输。链路 707 可能没有发生拥塞，或者具有更高的带宽，使得具有序列号 7 和 8 的帧在帧 1-6 之前到达交换机 109。本发明的技术提供了光纤信道帧的阻塞和丢弃，以实现到交换机 109 下游的交换机和节点的良序传递。阻塞在此也被称为延迟。

[0061] 如上所述，当信道改变时，路由表没有必要改变，因为交换机 107 和交换机 109 之间仍存有链路。用于确定信道中哪些链路可用于传输帧的逻辑和机制在此被称为转发信道表。应注意，转发信道表可能会改变。图 8A 示出了交换机 107 的转发信道表。对于下一跳 109，条目 803 表明链路 701、703 和 705 可用于将帧从交换机 107 发送到下一跳 109。

[0062] 图 8B 示出了信道改变后交换机 107 的转发信道表。在信道发生改变后，条目 813 提供的信息表明链路 701、703、705 和 707 可用于将帧从交换机 107 发送到下一跳 109。

[0063] 图 9 的处理流程图示出了根据各个实施例，信道改变期间光纤信道帧的转发。在 901，交换机检测到信道的改变。一个或更多链路的加入或删除可以是信道改变。在 903，丢弃基于路由表下一跳而被定向到所述信道的流量。在一个实施例中，丢弃要通过所述信道传输但尚未排队的流量，以使转发信道表中的改变可以适应所述信道改变。然而，已经排队的流量仍然保留。如上所述，已经排队的流量可能驻留在与特定流相关联的虚拟输出队列中。在 905，标记并阻塞用于沿着所述信道中的链路发送的队列。在一个示例中，“标记并阻塞所述队列”包括使得所有已经在所述队列中的帧被发送，同时在交换机漏延迟时间段期间阻塞任何新的帧。

[0064] 然后在 907，可以更新转发信道表，以在信道中增加和 / 或去除链路。在 909，先前在 903 被丢弃的朝向所述信道的流量现在可以被允许进入队列。然后在 911 交换机可以在修正的光纤信道交换机漏延迟期间等待。根据各个实施例，修正的光纤信道交换机漏延迟比标准的光纤信道交换机漏延迟要长，以允许所有已经在队列中的帧被发送而不是被丢弃。在 913，与信道中的链路相关联的队列被释放，队列中的所有的帧或者被发送，或者如果太旧就被丢弃，并且现在可以良序传递新的帧序列。

[0065] 虽然根据各个实施例可以使用帧的延迟等机制，但是本发明的技术还考虑到了使用标签来提供光纤信道帧的良序传递。使用标签的许多原因之一是标签提供了访问路由表中条目的一种快速机制。与查看目的地地址不同，目的地标识符可以是“内标签”(in label)或进入标签，其可用来迅速引用路由表条目。还可因为多种其他原因而使用标签。在光纤信道网络中使用标签在由发明人 Scott S.Lee 和 Dinesh G.Dutt 同时递交的美国申请 No.10/114,394 中进行了描述，该申请题为“Label Switching In FibreChannel Networks”(律师案卷号 No.ANDIP009)，其整体通过引用而被包含，以用于所有目的。用于实现标签交换的体系结构的一个示例是 RFC3031 中描述的多协议标签交换 (MPLS)，其整体也通过引用而被包含以用于所有目的。

[0066] 图 10 示意性地表示了包括标签交换路由器的光纤信道网络，所述标签交换路由器可以基于与帧相关联的标签来转发分组。除了包含目的地地址之外，帧还包括内标签作为目的地标识符，其在此还被称为进入标签，其使得交换机可快速访问路由表中的条目。例如，标签交换路由器 1004 可以接收目的地为 2，内标签为 420 的帧。应注意，在图 10-12 中，值为 1-5 的目的地或下一跳表示交换机 1001-1005。标签交换路由器 1004 可以访问它的路由表 1014，以识别出下一跳是标签交换路由器 1002，而“外标签”应该是 220。外标签在此还被称为外发标签。根据各个实施例，标签交换路由器 1004 将对应于路由表中所述内标签的帧标签值 420 替换为对应于路由表 1014 中所述外标签的帧标签 220。

[0067] 通过替换标签值，标签交换路由器 1004 向下一跳路由器 1002 提供了标签信息，以使得标签交换路由器 1002 可类似地快速访问路由表条目。应注意，虽然可以提供标签交换来快速访问路由表中的条目，但是标签交换可因为多种原因而被使用。本发明的技术表明可以使用标签来良序传递帧。

[0068] 当标签交换路由器 1002 接收到来自标签交换路由器 1004 的帧时，该标签交换路由器使用标签 220 来访问路由表 1012 中的条目。使用内标签 220，标签交换路由器 1002 识别出该帧已到达最终跳交换机，不再需要被转发到另一个交换机。然后可将该帧转发到最终目的地，它可能是主机或磁盘。

[0069] 多种技术可用来生成具有标签的路由表。在一个实施例中，根据 FSPF 协议，在接收到链路状态更新分组时生成路由表。路由表可被周期性地生成，或在识别出有链路状态改变时生成。根据各个实施例，新生成的路由表与一个标称号 (incarnation number) 相关联。光纤信道结构中所有标称号的组合在此被称为拓扑版本号。在一个实施例中，每次在交换机处生成新的路由表时，都将标称号加 1。根据各个实施例，光纤信道网络中的每个标签交换路由器都不仅生成朝向每个目的地的新转发路由，而且每个标签交换路由器都还生成不同于先前的内标签组的新的内标签。

[0070] 图 11 示意性地示出了链路状态改变期间的标签交换路由器。在此，标签交换路由器 1004 和标签交换路由器 1001 之间不再有链路。在整个光纤信道结构中传播链路状态更新分组或链路状态记录。为了从标签交换路由器 1004 发送到标签交换路由器 1001，标签交换路由器 1004 不再能直接转发到标签交换路由器 1001。相反，标签交换路由器 1004 将帧发送到标签交换路由器 1002 或标签交换路由器 1003。这些新的路由被反映在路由表 1114 中。标签交换路由器 1004 生成具有新的下一跳组的新路由表 1114，并生成新的内标签 411 来替换旧内标签 410。

[0071] 根据各个实施例，标签交换路由器 1004 发送标签交换控制消息到光纤信道结构中的其他标签交换路由器，以去除过期的内标签 410。发送到光纤信道结构中的其他标签交换路由器的标签交换控制消息包含拓扑版本号。在一个实施例中，其他标签交换路由器验证标签交换控制消息的版本号与所述路由表的拓扑版本号是否匹配。如果标签交换控制消息的拓扑版本号与特定标签交换路由器中的路由表的拓扑版本号不匹配，或者如果控制消息包含较旧的版本号，则丢弃该标签交换控制消息。

[0072] 如果标签交换控制消息的拓扑版本号比特定标签交换路由器中的路由表的拓扑版本号更新，则可存储该标签交换控制消息并在以后有更新版本的路由表版本时使用。

[0073] 如果拓扑版本号匹配，则接收到标签交换控制消息的标签交换路由器可去除过期的标签。例如，如果标签交换路由器 1004 生成新的内标签 411 来替换旧的内标签 410，并将撤换过期标签的标签交换控制消息发送到标签交换路由器 1002，则标签交换路由器 1002 将去除与目的地 ID 1 和下一跳 4 相关联的外标签 410。根据各个实施例，用于去除过期标签的标签交换控制消息在此被称为标签撤换消息。

[0074] 在外标签未被解析时，丢弃光纤信道帧以防止光纤信道帧的乱序传递。在图 11 和 12 中，“？”用于表明一个标签未被解析。例如，假设标签交换路由器 1005 向标签交换路由器 1001 发送了帧 1 和 2 以最终到达标签交换路由器 1004。在标签交换路由器 1001 和标签交换路由器 1004 之间的链路故障之后，标签交换路由器 1005 向标签交换路由器 1002 发送了帧 3 和 4 以最终到达标签交换路由器 1004。如果不丢弃帧 1 和 2，则它们可能会在帧 3 和 4 之后到达标签交换路由器 1004。然而，由于标签交换路由器 1001 处用于标签交换路由器 1004 的外标签在链路故障之后未被解析，因此帧 1 和 2 被丢弃，而帧 3 和 4 可以良序到达标签交换路由器 1004。也就是说，由于路由表 1111 中与目的地 4 相关联的外标签 440 不再是准确的，因此可以丢弃帧 1 和 2。

[0075] 为了解析外标签，标签交换路由器 1004 向其他标签交换路由器例如标签交换路由器 1002 公告包括内标签 411 在内的新的内标签。用于增加新标签的标签交换控制消息在此被称为标签映射消息。在一个示例中，标签交换路由器 1004 发送具有一个标签的标签映射消息到标签交换路由器 1002。标签交换路由器 1002 处接收的标签映射消息可以指示标签交换路由器 1002 使用新的内标签 411 作为与下一跳 4 和目的地 1 相关联的新的外标签。

[0076] 图 12 示意性地示出了外标签尚未被完全解析时，光纤信道结构中的标签交换路由器。标签交换路由器 1004 已生成了路由表 1214，将一组旧的内标签 410、420、430、440 和 450 分别替换为更新的内标签栈 411、421、431、441 和 451。标签交换路由器 1004 还解析了它的外标签，并将外标签 110、110、220、330、250 和 350 分别替换为新的外标签 211、311、221、331、251 和 351。根据各个实施例，在标签交换路由器 1004 已经从其他所有路由器接收到具有与路由表 1214 的版本号对应的版本号的标签映射消息后，解析路由器 1004 处的外标签。

[0077] 然而，标签交换路由器 1002 尚未完全解析路由表 1212 中的外标签。虽然路由表 1212 包含了一组新的内标签，但是外标签组只是被部分解析了。具体地说，虽然标签交换路由器 1002 具有对应于下一跳 4 的新的外标签，但是它不具有对应于下一跳 5 的新的外标签。这可能是因为收到了来自标签交换路由器 1004 的标签映射消息，但没有来自标签交换

路由器 1005 的标签映射消息。如果标签交换路由器 1002 接收到要发送到下一跳 5 的帧，则丢弃该帧，因为与下一跳 5 相关联的外标签未被解析。

[0078] 随着链路状态更新消息和标签交换控制消息传递通过整个光纤信道结构，以与图 12 中的标签交换路由器 1004 和标签交换路由器 1001 解析外标签相同的方式，每个交换机最终都可解析外标签。

[0079] 图 13 的处理流程图示出了使用标签的帧传递。在 1301，检测到光纤信道结构链路改变。在 1303，基于链路状态信息来计算对于每个目的地的更新的下一跳组。根据各个实施例，在 1305 标识拓扑版本号。拓扑版本号可以是光纤信道结构中所有交换机的标称号的组合。在一个实施例中，拓扑版本号包括附接在一起的所有版本号。在另一个实施例中，拓扑版本号是各个标称号的校验和。在另一个实施例中，拓扑版本号是对网络中的交换机有意义的唯一号码。在 1307，交换机可以发送标签交换控制消息来撤换以前公告的标签。在 1309，每个标签交换路由器可以生成新的内标签或进入标签，并将所述新的内标签或进入标签公告给结构中其他交换机。

[0080] 应注意，本发明的技术不一定需要以任何特定的顺序来执行。例如，在一个实施例中，标签交换路由器可在撤换其他标签交换路由器处的外标签的同时生成并公告内标签。

[0081] 在 1311，丢弃具有旧标签的帧。在外标签未被解析时也丢弃帧。例如，如果与下一跳 6 相关联的外标签未被解析，则丢弃被配置成发送到下一跳 6 的帧。在 1313，使用来自其他标签映射消息的信息或来自其他标签路由器的公告的信息，解析外标签。在 1315，当解析了外标签时，不再丢弃帧。

[0082] 虽然已参考具体实施例来具体示出并描述了本发明，但是本领域内的技术人员将会理解到可以对所公开的实施例的形式和细节作出改变，而不会偏离本发明的精神或范围。例如，本发明的实施例可利用多种网络协议和体系结构来实施。可在多种不同时刻发送平抑消息 (quench message) 等指令。因此，本发明应被解释成包含所有落在本发明的真正精神和范围内的所有变体和等同物。

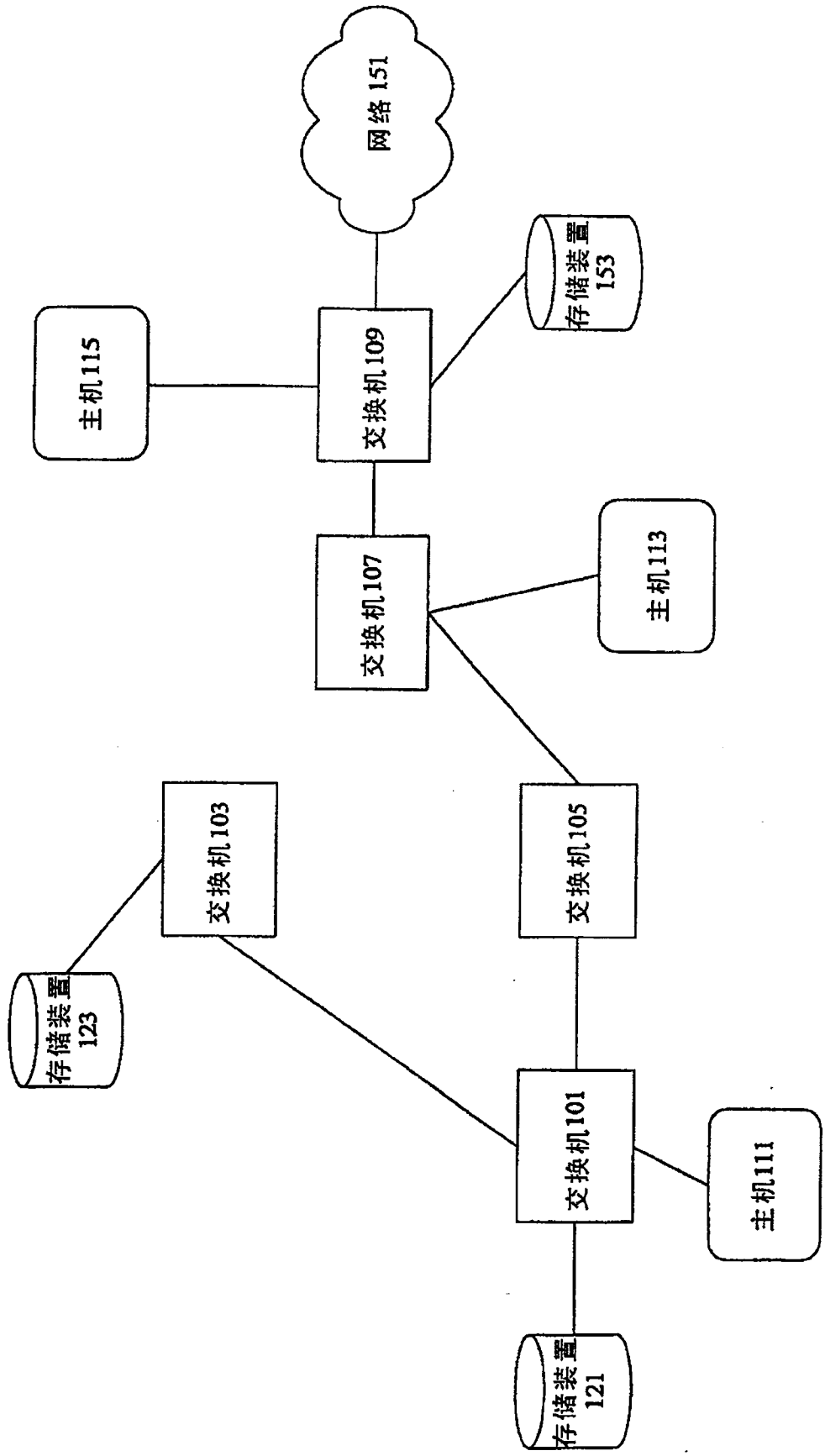

网络 151
存储装置 153
主机 115
交换机 109
交换机 107
主机 113
交换机 103
交换机 105
存储装置 123
交换机 101
主机 111
存储装置 121

图 1

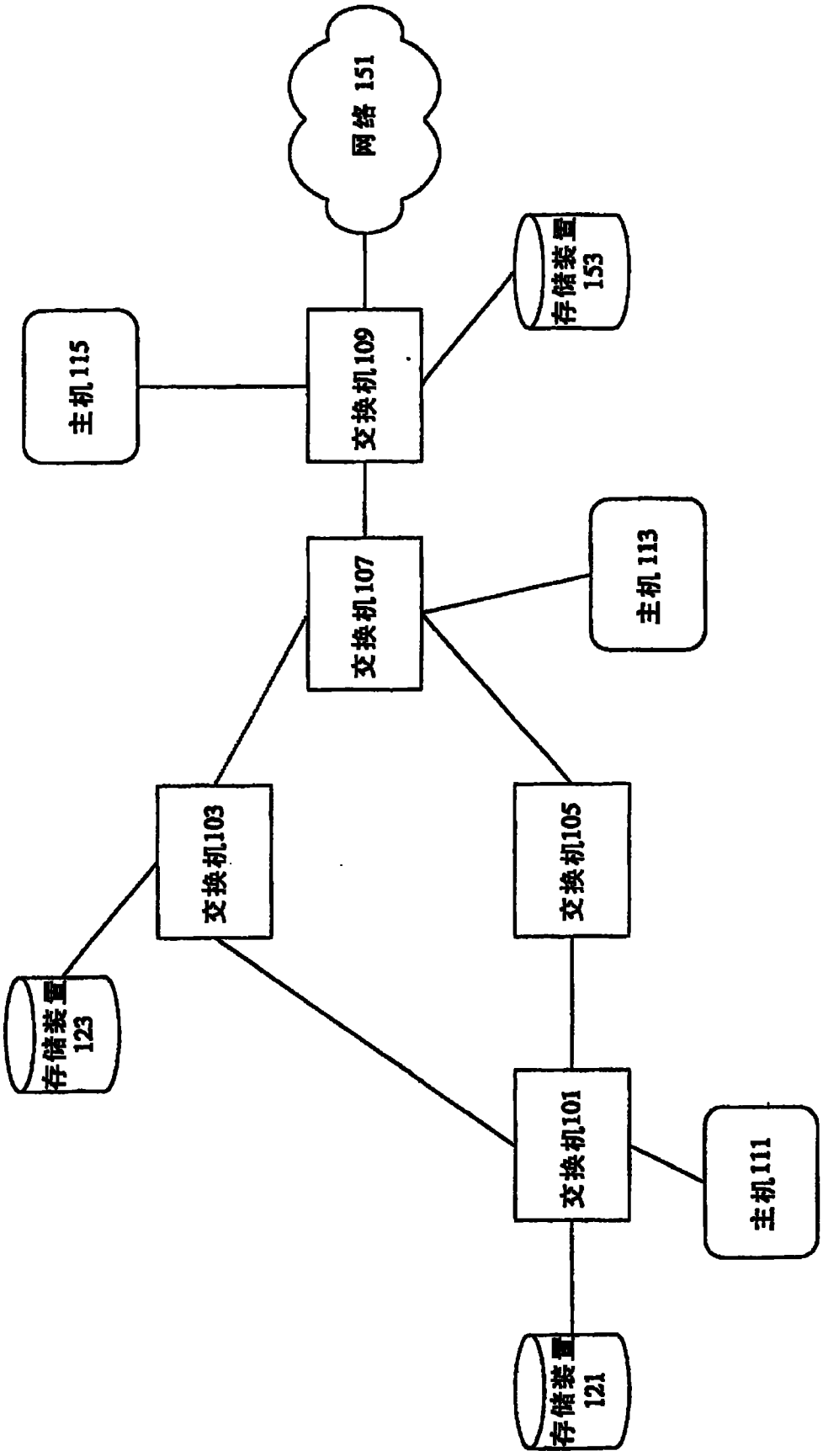

网络 151
存储装置 153
主机 115
交换机 109
交换机 107
主机 113
交换机 103
交换机 105
存储装置 123
交换机 101
主机 111
存储装置 121

图 2

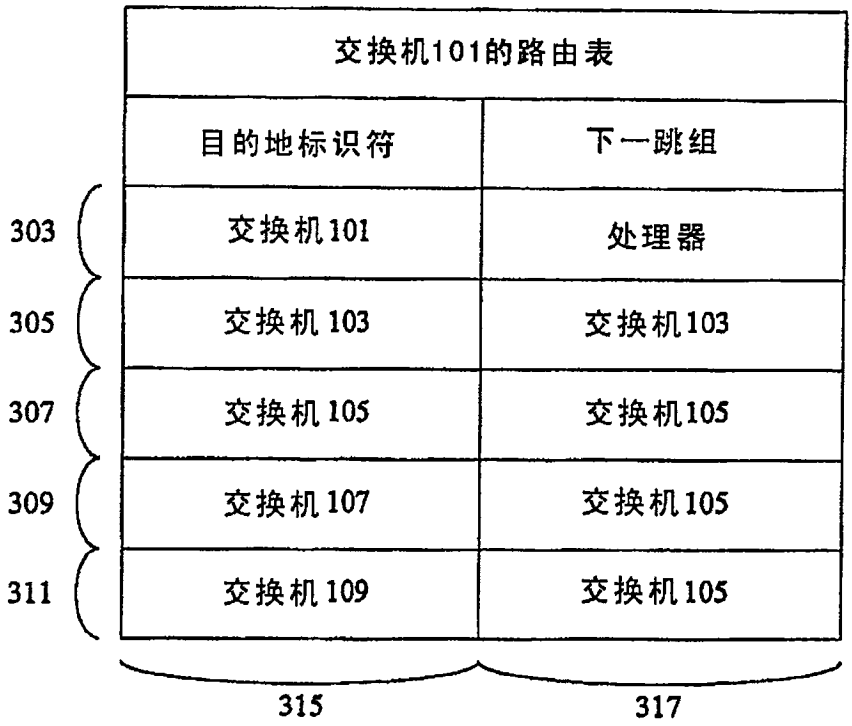

交换机101的路由表
目的地标识符
下一跳组
303
交换机101
处理器
305
交换机 103
交换机103
307
交换机 105
交换机105
309
交换机 107
交换机105
311
交换机 109
交换机105
315
317

图 3A

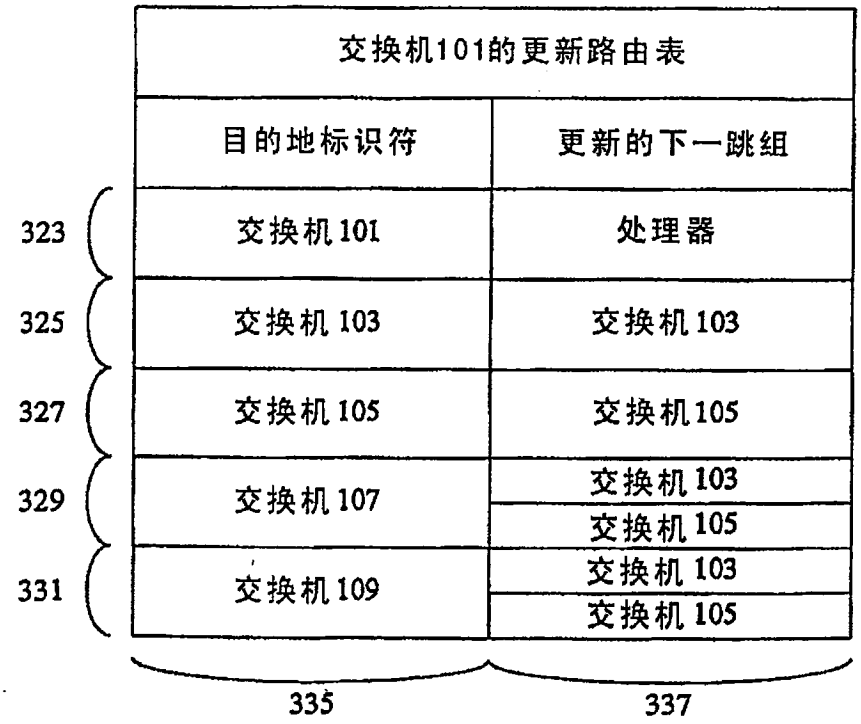

交换机101的更新路由表
目的地标识符
更新的下一跳组
323
交换机101
处理器
325
交换机 103
交换机103
327
交换机 105
交换机105
329
交换机 107
交换机 103
交换机 105
331
交换机 109
交换机 103
交换机 105
335
337

图 3B

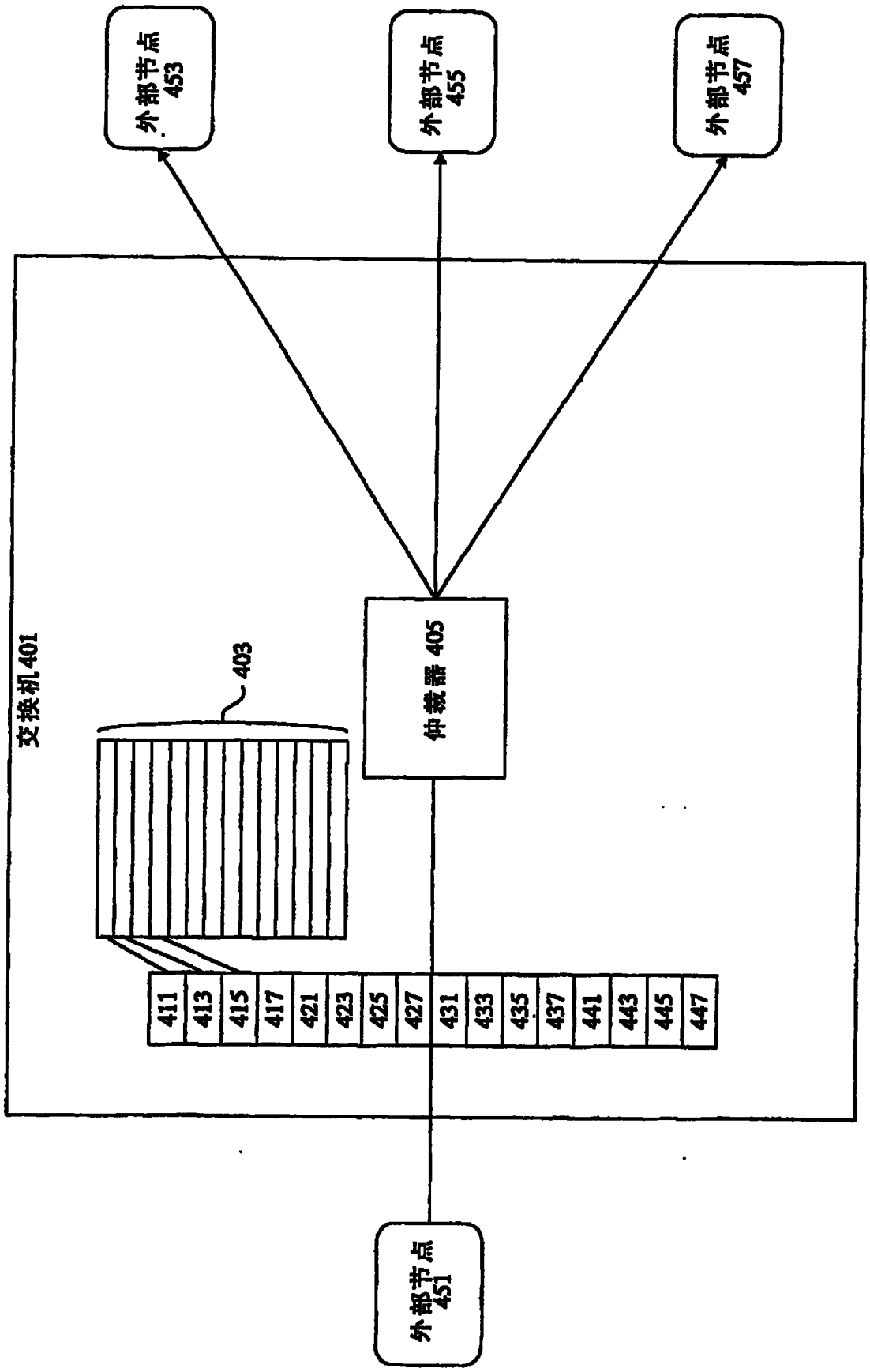

外部节点 453
外部节点 455
外部节点 457
交换机 401
403
仲裁器 405
411
413
415
417
421
423
425
427
431
433
435
437
441
443
445
447
外部节点 451

图 4

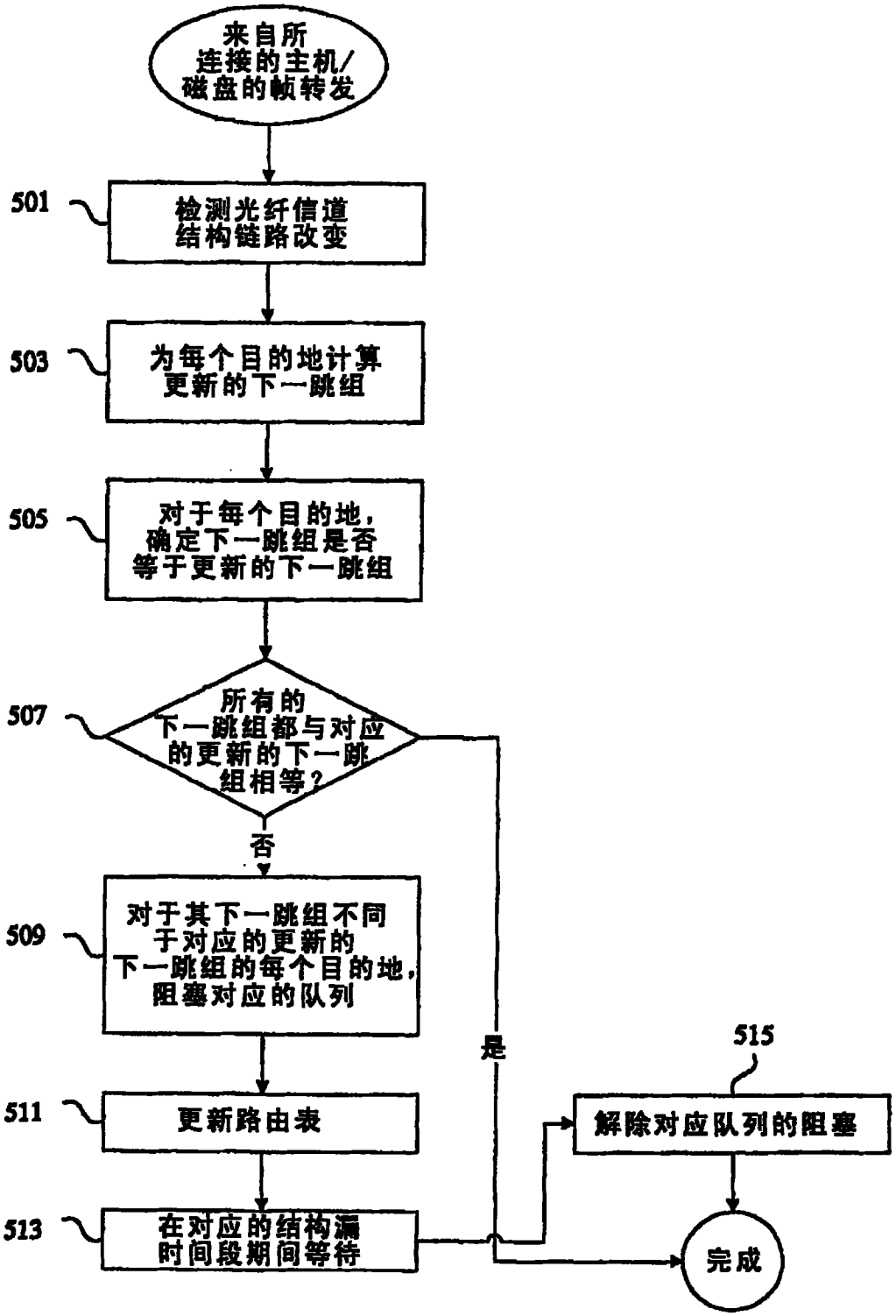

来自所连接的主机/磁盘的帧转发
501
检测光纤信道结构链路改变
503
为每个目的地计算更新的下一跳组
505
对于每个目的地，确定下一跳组是否等于更新的下一跳组
507
所有的下一跳组都与对应的更新的下一跳组相等？
否
是
509
对于其下一跳组不同于对应的更新的下一跳组的每个目的地，阻塞对应的队列
511
更新路由表
513
在对应的结构漏时间段期间等待
515
解除对应队列的阻塞
完成

图 5

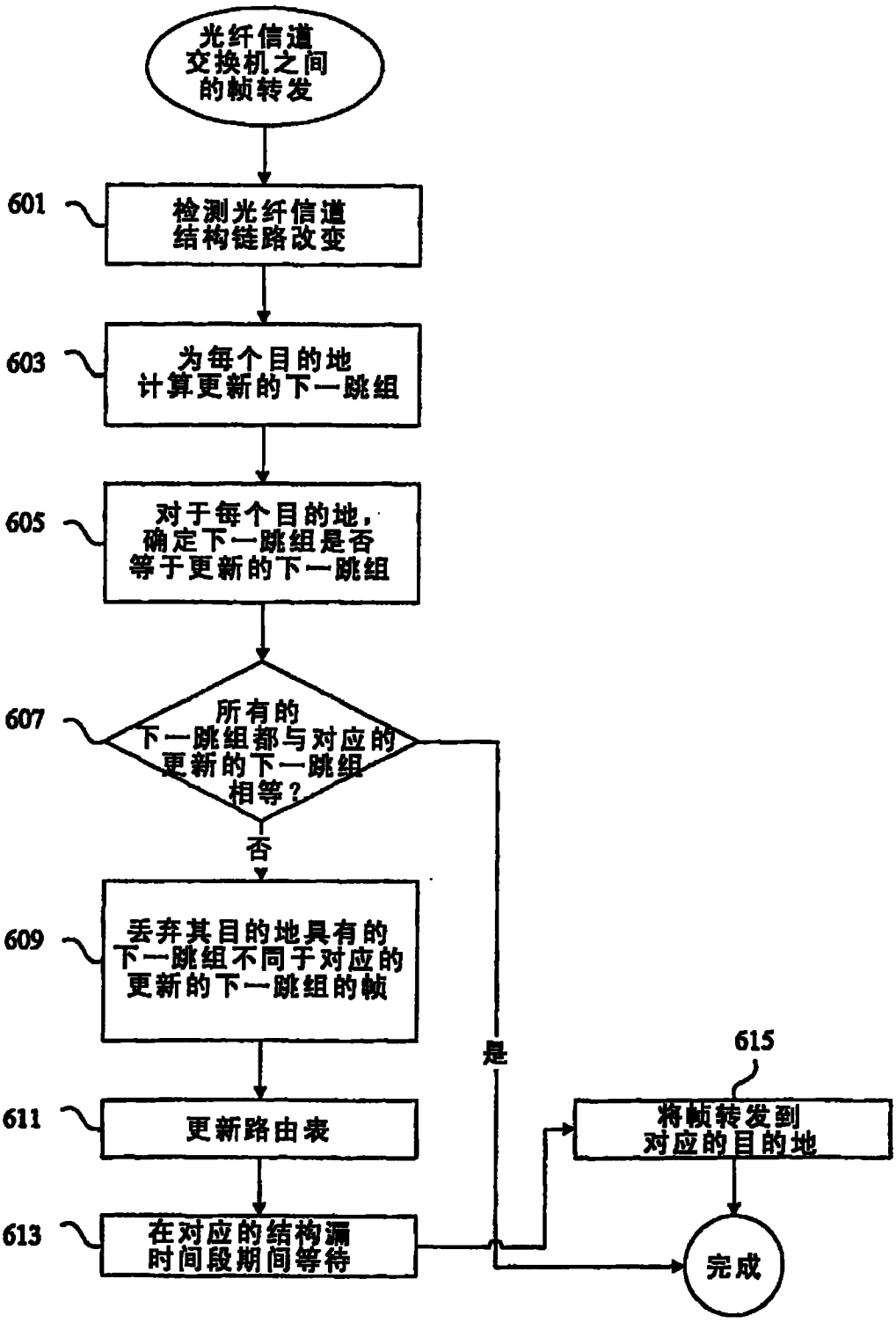

光纤信道
交换机之间
的帧转发
601
检测光纤信道
结构链路改变
603
为每个目的地
计算更新的下一跳组
605
对于每个目的地，
确定下一跳组是否
等于更新的下一跳组
607
所有的
下一跳组都与对应的
更新的下一跳组
相等？
否
是
609
丢弃其目的地具有的
下一跳组不同于对应的
更新的下一跳组的帧
611
更新路由表
613
在对应的结构漏
时间段期间等待
615
将帧转发到
对应的目的地
完成

图 6

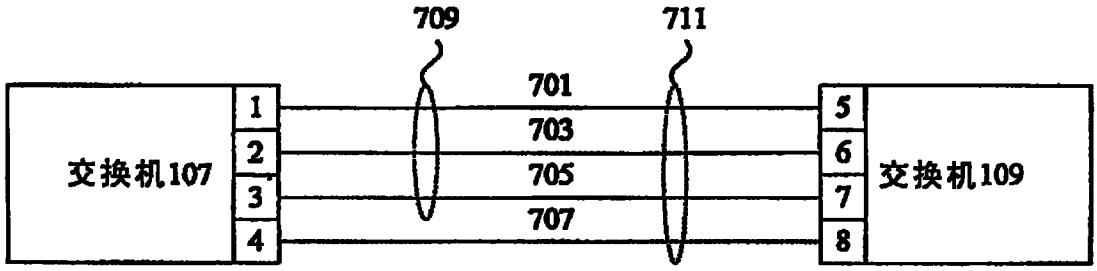


图 7

交换机107的转发信道表	
下一跳	链路
109	链路701, 703, 和 705
...	...

803 805 807 809

图 8A

交换机107的更新转发信道表	
下一跳	链路
109	链路701, 703, 705, 和707
...	...

813 815 817 819

图 8B

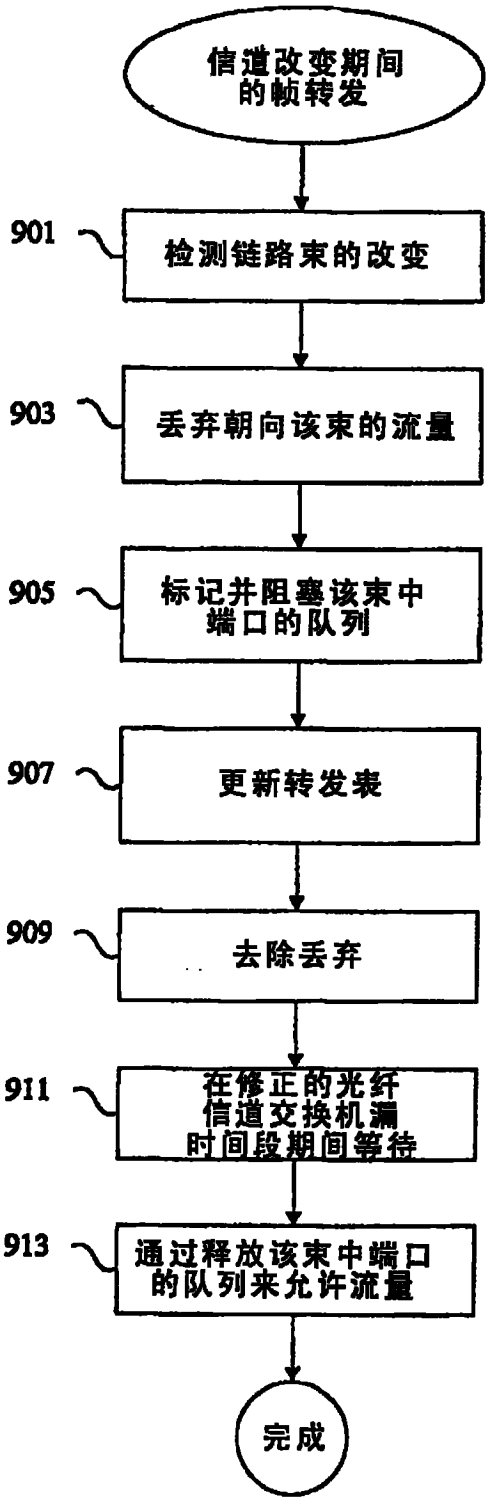

信道改变期间
的帧转发
901
检测链路束的改变
903
丢弃朝向该束的流量
905
标记并阻塞该束中
端口的队列
907
更新转发表
909
去除丢弃
911
在修正的光纤
信道交换机漏
时间段期间等待
913
通过释放该束中端口
的队列来允许流量
完成

图 9

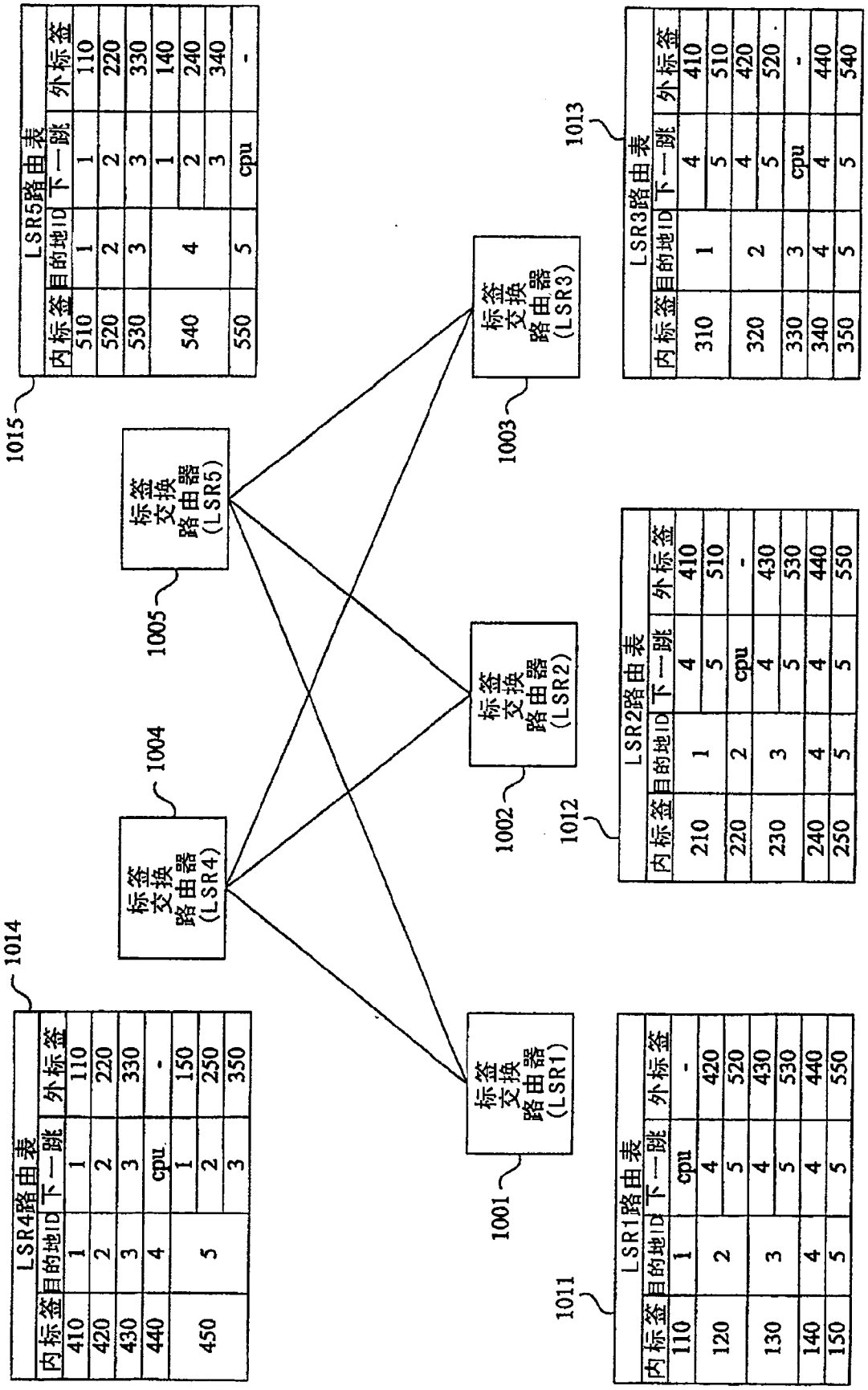

LSR5路由表
内标签 目的地ID 下一跳 外标签
510 1 1 110
520 2 2 220
530 3 3 330
540 4 1 140
2 240
3 340
550 5 cpu -
1015
LSR3路由表
内标签 目的地ID 下一跳 外标签
310 1 4 410
5 510
320 2 4 420
5 520
330 3 cpu -
340 4 4 440
350 5 5 540
1013
LSR4路由表
内标签 目的地ID 下一跳 外标签
410 1 1 110
420 2 2 220
430 3 3 330
440 4 cpu -
450 5 1 150
2 250
3 350
1014
LSR2路由表
内标签 目的地ID 下一跳 外标签
210 1 4 410
5 510
220 2 cpu -
230 3 4 430
5 530
240 4 4 440
250 5 5 550
1012
LSR1路由表
内标签 目的地ID 下一跳 外标签
110 1 cpu -
120 2 4 420
5 520
130 3 4 430
5 530
140 4 4 440
150 5 5 550
1011
标签交换路由器(LSR5)
1005
标签交换路由器(LSR3)
1003
标签交换路由器(LSR4)
1004
标签交换路由器(LSR2)
1002
标签交换路由器(LSR1)
1001

图 10

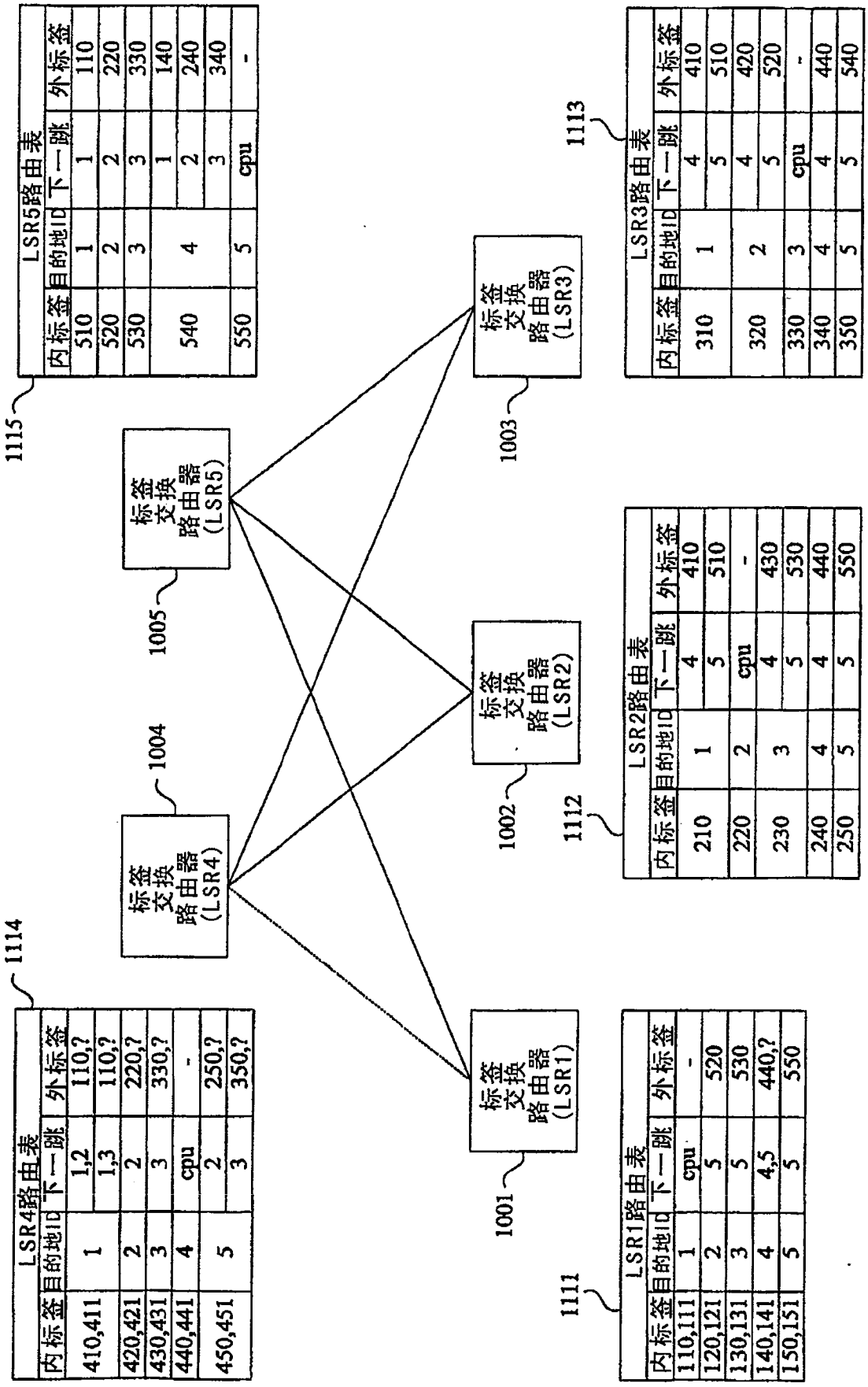

LSR5路由表
内标签 目的地ID 下一跳 外标签
510 1 1 110
520 2 2 220
530 3 3 330
540 4 1 140
2 240
3 340
550 5 cpu -
1115
标签交换路由器(LSR5)
1005
标签交换路由器(LSR3)
1003
LSR3路由表
内标签 目的地ID 下一跳 外标签
310 1 4 410
5 510
320 2 4 420
5 520
330 3 cpu -
340 4 4 440
350 5 5 540
1113
1004
标签交换路由器(LSR4)
1002
标签交换路由器(LSR2)
1112
LSR2路由表
内标签 目的地ID 下一跳 外标签
210 1 4 410
5 510
220 2 cpu -
230 3 4 430
5 530
240 4 4 440
250 5 5 550
1114
LSR4路由表
内标签 目的地ID 下一跳 外标签
410,411 1 1,2 110,?
1,3 110,?
420,421 2 2 220,?
430,431 3 3 330,?
440,441 4 cpu -
450,451 5 2 250,?
3 350,?
1001
标签交换路由器(LSR1)
1111
LSR1路由表
内标签 目的地ID 下一跳 外标签
110,111 1 cpu -
120,121 2 5 520
130,131 3 5 530
140,141 4 4,5 440,?
150,151 5 5 550

图 11

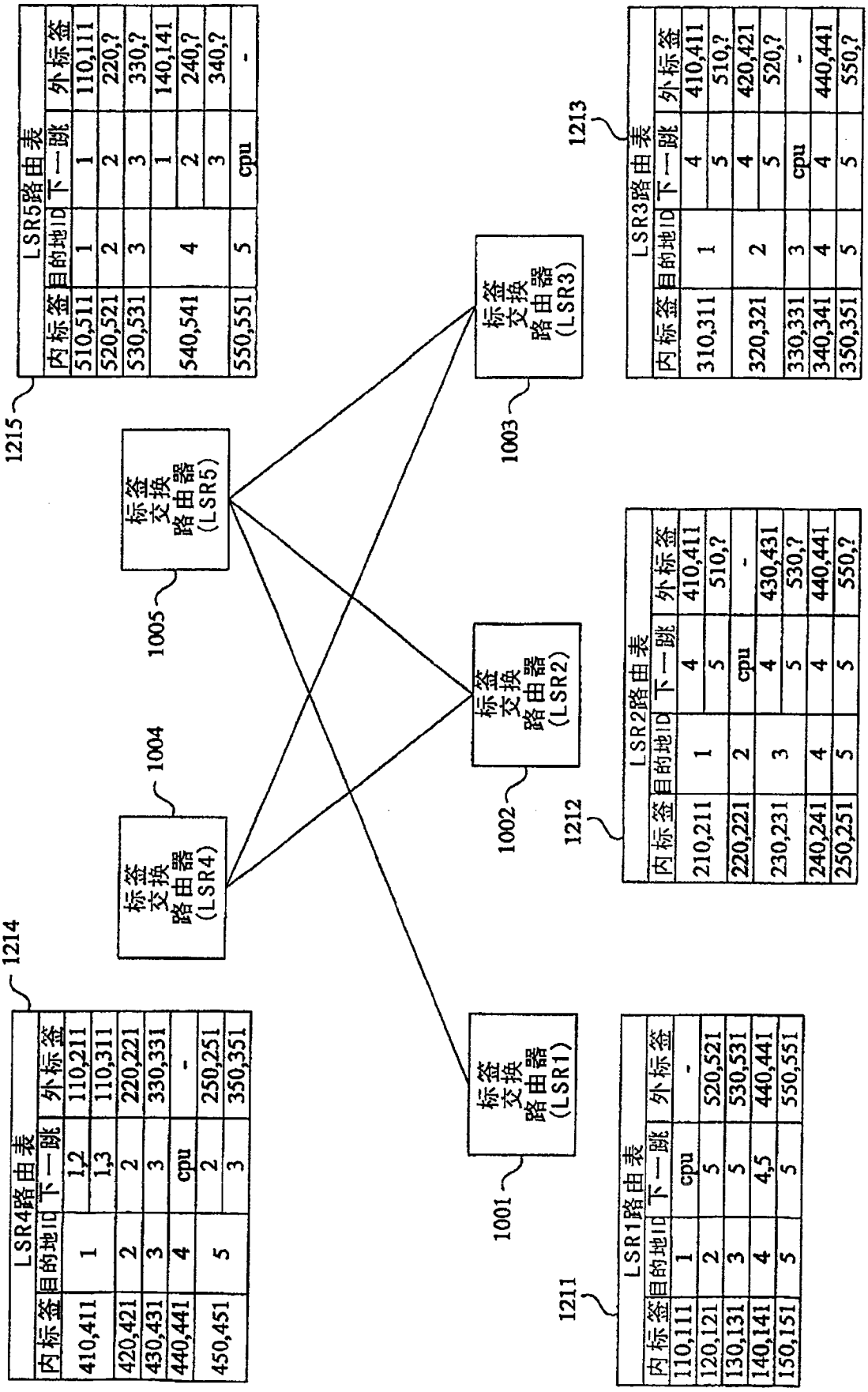

LSR5路由表
内标签 目的地ID 下一跳 外标签
510,511 1 1 110,111
520,521 2 2 220,?
530,531 3 3 330,?
540,541 4 1 140,141
2 240,?
3 340,?
550,551 5 cpu -
1215
标签交换路由器(LSR5)
1005
标签交换路由器(LSR4)
1004
1214
LSR4路由表
内标签 目的地ID 下一跳 外标签
410,411 1 1,2 110,211
1,3 110,311
420,421 2 2 220,221
430,431 3 3 330,331
440,441 4 cpu -
450,451 5 2 250,251
3 350,351
标签交换路由器(LSR3)
1003
1213
LSR3路由表
内标签 目的地ID 下一跳 外标签
310,311 1 4 410,411
5 510,?
320,321 2 4 420,421
5 520,?
330,331 3 cpu -
340,341 4 4 440,441
350,351 5 5 550,?
标签交换路由器(LSR2)
1002
1212
LSR2路由表
内标签 目的地ID 下一跳 外标签
210,211 1 4 410,411
5 510,?
220,221 2 cpu -
230,231 3 4 430,431
5 530,?
240,241 4 4 440,441
250,251 5 5 550,?
标签交换路由器(LSR1)
1001
1211
LSR1路由表
内标签 目的地ID 下一跳 外标签
110,111 1 cpu -
120,121 2 5 520,521
130,131 3 5 530,531
140,141 4 4,5 440,441
150,151 5 5 550,551

图 12

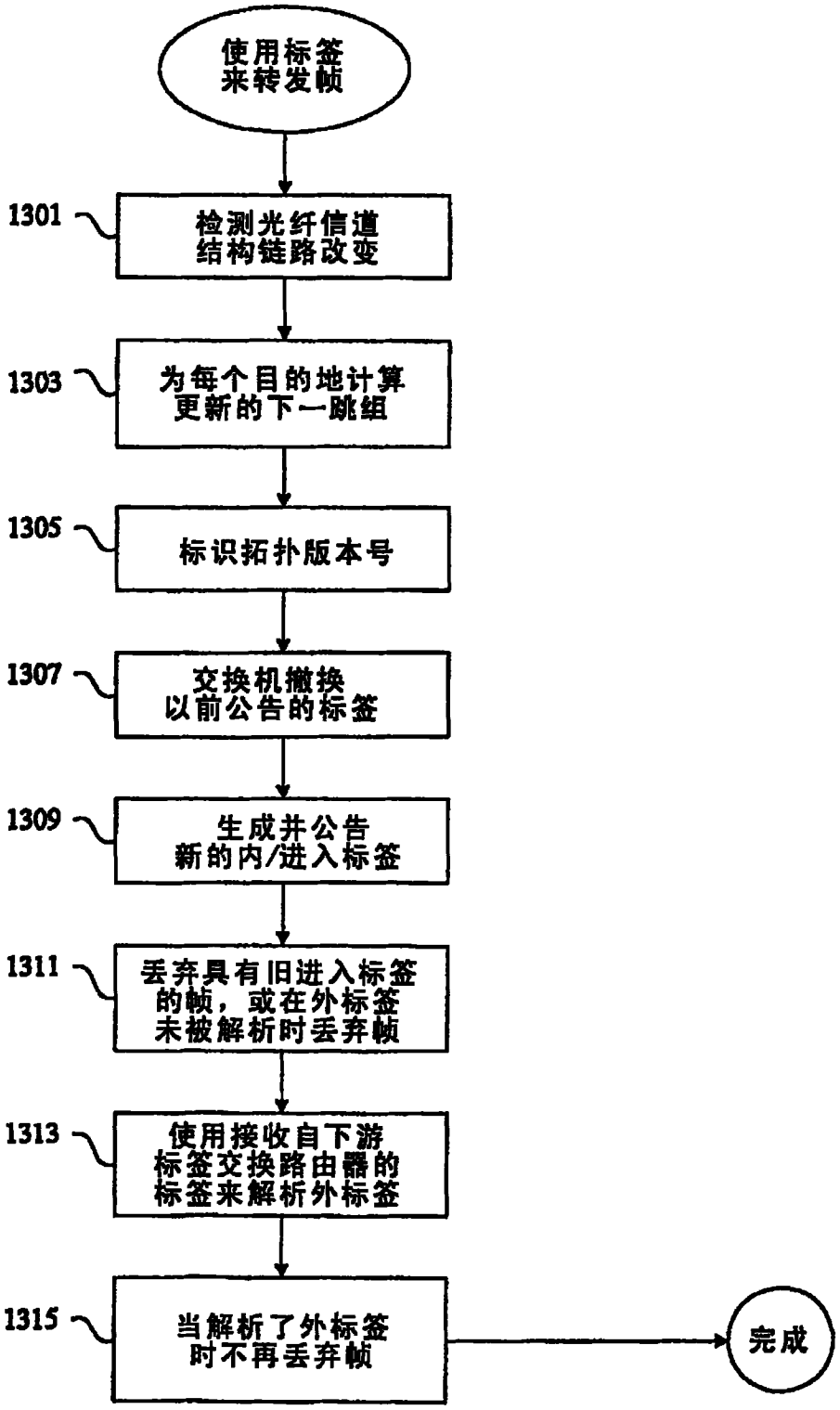

使用标签
来转发帧
1301
检测光纤信道
结构链路改变
1303
为每个目的地计算
更新的下一跳组
1305
标识拓扑版本号
1307
交换机撤换
以前公告的标签
1309
生成并公告
新的内/进入标签
1311
丢弃具有旧进入标签
的帧，或在外标签
未被解析时丢弃帧
1313
使用接收自下游
标签交换路由器的
标签来解析外标签
1315
当解析了外标签
时不再丢弃帧
完成

图 13