



(12) 发明专利申请

(10) 申请公布号 CN 103098426 A

(43) 申请公布日 2013.05.08

(21) 申请号 201180045232.3

(74) 专利代理机构 中国专利代理(香港)有限公司 72001

(22) 申请日 2011.09.16

代理人 汤春龙 朱海煜

(30) 优先权数据

12/886,439 2010.09.20 US

(51) Int. Cl.

H04L 12/723(2013.01)

(85) PCT申请进入国家阶段日

2013.03.20

(86) PCT申请的申请数据

PCT/IB2011/054070 2011.09.16

(87) PCT申请的公布数据

W02012/038870 EN 2012.03.29

(71) 申请人 瑞典爱立信有限公司

地址 瑞典斯德哥尔摩

(72) 发明人 S. 基尼 P. 德索扎

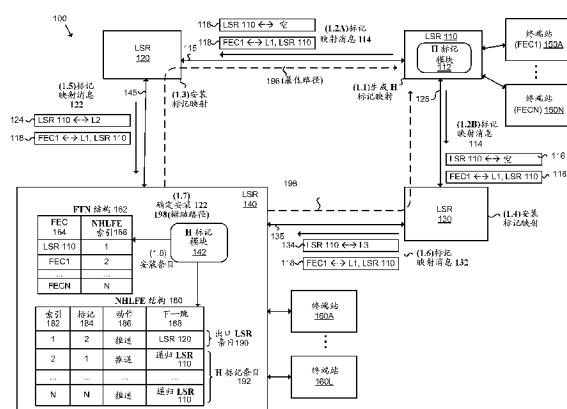
权利要求书4页 说明书14页 附图11页

(54) 发明名称

使用分层标记栈改进 LDP 收敛的方法和设备

(57) 摘要

描述了用于改进 MPLS(多协议标记交换)网络中 LDP(标记分发协议)收敛时间的方法。建立分层 LSP 以传输属于附连到出口 LSR 的 FEC 的分组。分层 LSP 包含出口 LSRLSP, 其对于附连到出口 LSR 的每一个 FEC 而言是公用的并形成从入口 LSR 通过一个或多个中间 LSR 到出口 LSR 的路径。当标记交换送往附连到出口 LSR 的 FEC 的分组时使用出口 LSRLSP。分层 LSP 还包含每个 FEC 的独特的 FECLSP, 其由出口 LSR 用于标识分组并将分组转发到那个 FEC。响应于改变入口 LSR 到达出口 LSR 的下一跳的拓扑改变, 入口 LSR 修改转发结构中的条目以改变出口 LSRLSP 的下一跳, 并且基本上不修改 FECLSP 的任何转发结构条目。通过减少在拓扑改变之后的转发结构修改减少了 LDP 收敛时间。



1. 一种在 MPLS(多协议标记交换) 网络中充当 LSR(标记交换路由器) 的第一网络单元中用于改进 LDP(标记分发协议) 收敛时间的方法, 所述方法包括如下步骤:

为分别属于多个终端站的多个 FEC(转发等效类) 建立分层 LSP(标记交换路径), 其中每个 FEC 的所述分层 LSP 包含:

出口 LSR LSP, 所述出口 LSR LSP 对于每一个所述 FEC 而言是公用的, 并形成到第二网络单元的路径, 并且当在所述 MPLS 网络中标记交换分组时使用, 所述第二网络单元充当所述多个 FEC 的出口 LSR; 以及

独特的 FEC LSP, 所述独特的 FEC LSP 由所述第二网络单元用于标识分组并将分组转发到所述 FEC; 以及

响应于改变所述第一网络单元针对所述出口 LSR LSP 的下一跳的拓扑改变, 修改转发结构以至少改变针对所述出口 LSR LSP 的所述下一跳, 而基本上没有修改所述 FEC LSP 的任何转发结构条目;

由此, 通过减少在所述拓扑改变之后的转发结构修改改进了 LDP 收敛时间。

2. 如权利要求 1 所述的方法, 其中, 所述第一网络单元充当入口 LSR, 并且其中, 建立所述分层 LSP 的所述步骤还包含如下步骤:

从充当中间 LSR 的第三网络单元接收多个标记映射消息, 所述多个标记映射消息包含:

第一出口 LSR 标记映射消息, 其指示映射到所述第二网络单元的 IP 地址的标记, 其中所述标记的值由所述第三网络单元分配, 以及

用于所述多个 FEC 中每个 FEC 的独特的分层标记映射消息, 每一个所述分层标记映射消息指示映射到那个 FEC 的标记由所述出口 LSR 始发, 并且还指示始发所述标记的所述出口 LSR 的身份; 以及

基于所述出口 LSR 标记映射消息在转发结构中安装出口 LSR 标记条目, 使得指示的标记将被推送到用于送往所述多个 FEC 中任一个 FEC 的出局分组的标记栈上, 并且那些分组将被传送到所述第三网络单元; 以及

对于每个独特的分层标记映射消息, 基于那个分层标记映射消息在所述转发结构中安装分层标记条目, 使得指示的标记将被推送到用于送往那个 FEC 的出局分组的所述标记栈上, 并且所述出口 LSR 标记条目应该被访问。

3. 如权利要求 2 所述的方法, 还包括如下步骤:

从第四网络单元接收多个标记映射消息, 所述第四网络单元充当提供到所述第二网络单元的备选下一跳的中间 LSR, 所述多个标记映射消息包含:

第二出口 LSR 标记映射消息, 所述第二出口 LSR 标记映射消息指示映射到所述第二网络单元的 IP 地址的标记, 其中, 所述标记的值由所述第四网络单元分配, 以及

用于所述多个 FEC 的独特的分层标记映射消息;

其中, 所述拓扑改变将所述第一网络单元针对所述出口 LSR LSP 的所述下一跳从所述第三网络单元改变到所述第四网络单元; 以及

其中, 修改所述转发结构以至少改变针对所述出口 LSR LSP 的所述下一跳的所述步骤包含: 改变所述转发结构中的所述出口 LSR 标记条目, 使得由所述第四网络单元分配的所述标记将被推送到送往所述多个 FEC 中任一个 FEC 的所述出局分组上, 并且那些分组将被

传送到所述第四网络单元。

4. 如权利要求 2 所述的方法,还包括如下步骤:

接收送往所述多个终端站中第一终端站的未标记分组;

基于所述分组中的所述目的地 IP 地址来确定所述第一终端站的所述 FEC;

基于所述 FEC 来访问所述分层标记条目;

将所访问的分层标记条目中指示的标记推送到用于所述分组的标记栈上;

访问由所述所访问的分层标记条目所指示的所述出口 LSR 标记条目;

将在所述所访问的出口 LSR 标记条目中指示的所述标记推送到用于所述分组的所述标记栈上;以及

将标记的分组传送到所述出口 LSR 标记条目中标识的下一跳。

5. 如权利要求 1 所述的方法,其中,所述拓扑改变是如下项其中之一:链路故障、节点故障、度量改变和具有到所述第二网络单元的更优化路径的新路由。

6. 一种在 MPLS(多协议标记交换)网络中充当出口 LSR(标记交换路由器)的第一网络单元中用于始发标记映射消息以从充当入口 LSR 的第二网络单元通过充当中间 LSR 的一个或多个第三网络单元的集合到所述第一网络单元建立分层 LSP(标记交换路径)的方法,所述方法包括如下步骤:

生成出口 LSR 标记映射消息,所述出口 LSR 标记映射消息包含映射到所述第一网络单元的 IP 地址的标记;

对于分别属于其中所述第一网络单元是出口的多个终端站的多个 FEC(转发等效类)中的每个 FEC,为那个 FEC 生成分层标记映射消息,所述分层标记映射消息指示由所述第一网络单元为那个 FEC 始发的标记并指示所述第一网络单元始发了那个标记;

将所述出口 LSR 标记映射消息和所述分层标记映射消息传送到所述第一网络单元的一个或多个对等体,以允许为所述多个 FEC 中的每个 FEC 建立所述分层 LSP,所述分层 LSP 包含外部 LSP 和特定于那个 FEC 的内部 LSP,所述外部 LSP 定义到达所述第一网络单元的路径,所述内部 LSP 由所述第一网络单元用于标识分组并将分组转发到那个 FEC;

由此,所述分层 LSP 允许所述第二网络单元通过改变与所述外部 LSP 相关联的一个或多个转发条目来对影响到所述第一网络单元的可到达性的拓扑改变作出反应,而基本上没有修改与任何所述内部 LSP 相关联的任何转发条目。

7. 如权利要求 6 所述的方法,其中,包含在所述出口 LSR 标记映射消息中的所述标记是空标记,并且其中,所述第一网络单元的 IP 地址是所述第一网络单元的环回 IP 地址。

8. 如权利要求 6 所述的方法,其中,响应于从所述第二网络单元接收到标记请求消息生成至少一个所述分层标记映射消息。

9. 如权利要求 6 所述的方法,其中,所述分层标记映射消息包含分层标记 TLV(类型长度值),所述分层标记 TLV 包含:

通用标记子 TLV,所述通用标记子 TLV 包含由所述第一网络单元为那个 FEC 始发的标记,以及

FEC 子 TLV,所述 FEC 子 TLV 包含所述第一网络单元的 IP 地址。

10. 如权利要求 6 所述的方法,其中,至少一个所述 FEC 的至少一个所述分层标记映射消息包含那个 FEC 的度量类型和度量值。

11. 一种网络单元,其充当 MPLS(多协议标记交换)网络中的入口 LSR(标记交换路由器),以使用分层 LSP(标记交换路径)在所述 MPLS 网络上传输属于终端站的分组,所述网络单元包括:

控制卡,所述控制卡包含分层标记模块,所述分层标记模块可操作以:

处理从出口 LSR 为属于所述终端站的多个 FEC(转发等效类)始发的分层标记映射消息和从所述出口 LSR 始发的出口 LSR 标记映射消息,所述分层标记映射消息各包含由所述出口 LSR 为 FEC 始发的标记的映射并指示所述出口 LSR 始发了那个标记,并且所述出口 LSR 标记映射消息各包含映射到所述出口 LSR 的 IP 地址的标记,使得为所述多个 FEC 中的每个 FEC 建立分层 LSP,其中每个 FEC 的所述分层 LSP 包含:

基于所述出口 LSR 标记映射消息的出口 LSR LSP,所述出口 LSR LSP 对于每一个所述 FEC 而言是公用的并提供到所述出口 LSR 的下一跳;以及

独特的 FEC LSP,所述独特的 FEC LSP 承载信息以标识分组并将分组转发到所述 FEC;

将表示所述分层 LSP 的一个或多个转发结构条目下载到所述网络单元的一个或多个线卡;以及

响应于改变到所述出口 LSR 的所述下一跳的拓扑改变,修改所述出口 LSR LSP 的一个或多个转发结构条目并将所述一个或多个转发结构条目下载到所述一个或多个线卡以改变针对所述出口 LSR 的所述下一跳,而基本上没有修改和下载所述 FEC LSP 的任何转发结构条目;

由此,通过减少在所述拓扑改变之后的转发结构修改来改进 LDP 收敛时间。

12. 如权利要求 11 所述的网络单元,其中,所述分层标记模块还可操作以处理分层标记映射消息,所述分层标记映射消息指示与所述多个 FEC 相关联的度量值,以确定当对于单个 FEC 接收到由不同出口 LSR 始发的多个分层标记映射消息时安装哪个分层标记映射消息。

13. 如权利要求 11 所述的网络单元,其中,所述分层标记模块要基于接收的出口 LSR 标记映射消息来将出口 LSR 标记条目下载到所述一个或多个线卡上的转发结构,其指示标记以推送到标记栈上,以便到达所述出口 LSR。

14. 如权利要求 13 所述的网络单元,其中,每一个所述分层标记映射消息的处理都包含所述分层标记模块基于那个分层标记映射消息将分层标记条目下载到所述一个或多个线卡上的转发结构,使得指示的标记将被推送到用于送往在所述分层标记映射消息中指示的所述 FEC 的出局分组的标记栈上,并且所述出口 LSR 标记条目需要被访问以将到达所述出口 LSR 的所述标记推送到所述标记栈上。

15. 如权利要求 14 所述的网络单元,还包括:

一个或多个线卡,所述一个或多个线卡包含一个或多个分组处理实体以使用分层 LSP 来转发送往所述终端站的分组,其包括对于送往那些终端站之一的每个接收的未标记分组执行如下操作:

基于所述分组的所述目的地 IP 地址来确定那个终端站的所述 FEC;

基于那个 FEC 来访问分层标记条目;

将所访问的分层标记条目中指示的所述标记推送到用于所述分组的标记栈上;

访问由所述所访问的分层标记条目所指示的出口 LSR 标记条目;

将在所述所访问的出口 LSR 标记条目中指示的所述标记推送到用于所述分组的所述标记栈上 ; 以及

将标记的分组传送到所述所访问的出口 LSR 标记条目中标识的下一跳。

16. 一种第一网络单元, 所述第一网络单元在 MPLS(多协议标记交换) 网络中充当用于分别属于多个终端站的多个 FEC(转发等效类) 的出口 LSR(标记交换路由器), 并始发标记映射消息以从充当所述 FEC 的入口 LSR 的第二网络单元通过充当中间 LSR 的一个或多个第三网络单元的集合建立分层 LSP(标记交换路径), 所述第一网络单元包括 :

控制卡, 所述控制卡包含分层标记模块, 所述分层标记模块可操作以 :

生成出口 LSR 标记映射消息, 所述出口 LSR 标记映射消息包含映射到所述第一网络单元的 IP 地址的标记,

对于所述多个 FEC 中的每个 FEC, 生成那个 FEC 的分层标记映射消息, 所述分层标记映射消息指示由所述第一网络单元为那个 FEC 始发的标记并指示所述第一网络单元始发了那个标记 ; 以及

使得将生成的出口 LSR 标记映射消息和所述分层标记映射消息传送到第三网络单元的集合以通过所述 MPLS 网络传播, 从而允许为每一个所述 FEC 建立分层 LSP, 所述分层 LSP 包含外部 LSP 和特定于那个 FEC 的内部 LSP, 所述外部 LSP 定义到达所述第一网络单元的 IP 地址的路径, 所述内部 LSP 由所述第一网络单元用于标识分组并将分组转发到那个 FEC ;

由此, 所述分层 LSP 允许所述第二网络单元通过改变与所述外部 LSP 相关联的一个或多个转发条目对影响到所述第一网络单元的可到达性的拓扑改变作出反应, 而基本上没有修改与任何所述内部 LSP 相关联的任何转发条目。

17. 如权利要求 16 所述的网络单元, 其中, 包含在所述出口 LSR 标记映射消息中的所述标记是空标记, 并且其中, 所述第一网络单元的 IP 地址是所述第一网络单元的环回 IP 地址。

18. 如权利要求 16 所述的网络单元, 其中, 所述分层标记模块可操作以响应于从所述第二网络单元接收到标记请求消息而生成至少一个所述分层标记映射消息。

19. 如权利要求 16 所述的网络单元, 其中, 每个分层标记映射消息要包含分层标记 TLV(类型长度值), 所述分层标记 TLV 包含 :

通用标记子 TLV, 所述通用标记子 TLV 包含由所述第一网络单元为那个 FEC 始发的所述标记, 以及

FEC 子 TLV, 所述 FEC 子 TLV 包含所述第一网络单元的 IP 地址。

20. 如权利要求 16 所述的网络单元, 其中, 所述分层标记模块还在所述分层标记映射消息中包含所述 FEC 的度量类型和度量值。

使用分层标记栈改进 LDP 收敛的方法和设备

技术领域

[0001] 本发明的实施例涉及连网领域；并且更具体地说，涉及使用分层标记堆积来改进 LDP 收敛。

背景技术

[0002] (在 2007 年 10 月的请求注释 (RFC) 5036 中描述的) 标记分发协议 (LDP) 用于广告转发等效类 (FEC) 到标记的映射。IP(因特网协议) 前缀 FEC 用于沿路由的路径设立标记交换路径 (LSP)。LDP 使用路由表中的路由来广告 IP 前缀 FEC 的标记映射。随着网络中 FEC 的数量增大，标记的数量也对应地增大。例如，在各运行 LDP 的多个标记交换路由器 (LSP) 的标记交换网络中，出口 LSR 为每个独特的出口下一跳分配非空标记。如果每个前缀都具有独特的下一跳，则每个前缀将必须分配独特的标记。在具有许多订户 (例如可能数千到数百万) 的无线和 / 或有线订户终接情形中，分配的标记数量很大。

[0003] 不是前缀的出口的 LSR 使用如下其中一项技术可知前缀 (连同其相关联的标记) 与那个前缀的出口 LSR 之间的关联性：使用 IP 路由表以分配 FEC 的标记；运行链路状态协议 (例如在 RFC 2328 (1998 年 4 月) 中描述的 OSPF (开放最短路径优先)、在 RFC 1142 (1990 年 2 月) 中描述的 IS-IS (中间系统到中间系统))，或运行附加协议 (例如 RFC 4271 (2006 年 1 月) 中描述的 BGP (边界网关协议))。

[0004] 然而，以上技术在一些情况下可能不是可能的或优选的。例如，在许多情况下，存在如下需要：边缘 LSR 运行诸如静态或 RIP (路由信息协议) 等简单路由协议，其具有用于冗余的 BFD (双向转发检测)。例如，网络提供商可能需要在无线订户终接情形下是基站网络单元的入口 LSR，以运行比较简单的非链路状态路由协议。这些协议未给出 FEC 与出口 LSR 之间的关联性。

[0005] 此外，IGP 的收敛时间可能比较高。例如，如果存在大量前缀，并且以高速率向路由表添加和删除它们 (例如移动订户在站之间移动可引起比较频繁地修改路由表)，则 IGP 的收敛时间比较高。例如，无线网络 (例如 4G 或 LTE 网络) 中的典型边缘 LSR 可支持数十万或百万的订户 (每一个订户在 IGP 中都可具有独特的前缀)。IGP 也有多个抑制 (dampen) 机制，它们可增大收敛时间 (例如 LSA (链路状态广告) 生成延迟 (例如通过使用在 RFC 2328 中描述的 MinLSInterval)、LSA 调步定时器 (例如使用在 RFC 2328 中描述的重传定时器 (RxmtInterval) 和 SPF (最短路径优先) 抑制定时器)。而且，当 LSDB (链路状态数据库) 很大时，SPF 在典型 SPF 运行期间执行许多次存储器存取，并且当将路由下载到公用存储器 (store) (例如 RIB (路由信息库)) 时，也需要相当大量的处理和 / 或存储器存取。这些全都增大了 IGP 的收敛时间。

[0006] 运行诸如 BGP 等附加协议不是优选的，原因在于它增加了开销 (无论是在资本支出 (CapEx) 上还是运营支出 (OpEx) 上)。例如，在资本支出上的开销包含开发和 / 或支持另一协议所需的开发资源，所述另一协议包括诸如可缩放性、高可用性和 / 或冗余等特征；运行附加协议所需的额外 CPU 和 / 或存储器，其随着冗余而增大；全网格连接 (网络中每对

LSR 之间一个连接) 的需要增大了计算资源 (CPU 循环、存储器等) 量以随着网络中 LSR 数量的增大而增长; 并且如果避免了全网格, 则它需要单独的路由 - 反射器 (RR), 其是单独类型的网络单元。在运营支出上的开销包括在网络设计上引入复杂性, 原因在于 BGP 必须在每一个 LSR 上被配置; 如果使用 RR(其通常是单独的网络单元), 则它需要由运营商维护; 并且配置和维护 BGP 协议需要专业且昂贵的人员的专业知识。

发明内容

[0007] 本文描述了通过使用分层 LSP(标记交换路径)来改进 LDP(标记分发协议)收敛时间。在一个实施例中, 在 MPLS(多协议标记交换)网络中为分别属于多个终端站的多个 FEC(转发等效类)建立分层 LSP。每个 FEC 的分层 LSP 包含: 出口 LSR(标记交换路由器) LSP, 其对于每一个 FEC 而言是公用的并且形成这些 FEC 的出口 LSR 的路径, 并且当在 MPLS 网络中标记交换分组时使用; 并且还包含独特的 FEC LSP, 其由出口用于标识分组并将分组转发到所述 FEC; 响应于改变第一网络单元针对所述出口 LSR LSP 的下一跳的拓扑改变, 修改转发结构以至少改变所述出口 LSR LSP 的下一跳, 而基本上没有修改所述 FEC LSP 的任何转发结构条目。通过减少在拓扑改变之后的转发结构修改改进了 LDP 收敛时间。

[0008] 在一个实施例中, 在 MPLS 网络中充当出口 LSR 的网络单元执行如下操作。网络单元生成包含映射到第一网络单元的 IP 地址的标记的出口 LSR 标记映射消息。对于属于其中网络单元是出口的终端站的每个 FEC, 它为那个 FEC 生成分层标记映射消息, 其指示由网络单元为那个 FEC 始发的标记, 并指示网络单元始发了那个标记。所述出口标记映射消息和所述分层标记映射消息被传送到网络单元的对等体, 以允许为每个 FEC 建立分层 LSP, 所述分层 LSP 包含定义到达网络单元的路径的外部 LSP 和特定于那个 FEC 的内部 LSP, 其由网络单元用于标识分组并将分组转发到所述 FEC。所述分层 LSP 允许入口 LSR 通过改变与外部 LSP 相关联的一个或多个转发条目对影响到出口 LSR 的可到达性的拓扑改变作出反应, 而基本上没有修改与任何内部 LSP 相关联的任何转发条目, 由此改进了 LDP 收敛时间。

[0009] 在一个实施例中, 充当入口 LSR 的网络单元包含控制卡, 该控制卡包含分层标记模块。分层标记模块可操作以处理从出口 LSR 为属于终端站的多个 FEC 始发的分层标记映射消息和从出口 LSR 始发的出口 LSR 标记映射消息, 所述分层标记映射消息各包含由出口 LSR 为 FEC 始发的标记的映射并指示出口 LSR 始发了那个标记, 并且出口 LSR 标记映射消息各包含映射到出口 LSR 的 IP 地址的标记, 使得为多个 FEC 中的每个 FEC 建立分层 LSP。每个 FEC 的分层 LSP 包含: 基于出口 LSR 标记映射消息的出口 LSR LSP, 其对于每一个 FEC 都是公用的, 并提供到出口 LSR 的下一跳; 以及独特的 FEC LSP, 其承载信息以标识分组并将分组转发到该 FEC。分层标记模块还可操作以将表示所述分层 LSP 的一个或多个转发结构条目下载到所述网络单元的一个或多个线卡。响应于改变到出口 LSR 的下一跳的拓扑改变, 分层标记模块可操作以修改出口 LSR LSP 的一个或多个转发结构条目并将所述一个或多个转发结构条目下载到一个或多个线卡以改变针对出口 LSR 的下一跳, 而基本上没有修改和下载 FEC LSP 的任何转发结构条目。通过减少在拓扑改变之后的转发结构修改改进了 LDP 收敛时间。

[0010] 在一个实施例中, 充当多个 FEC 的出口 LSR 的网络单元包含控制卡, 该控制卡包含分层标记模块, 分层标记模块可操作以生成包含映射到第一网络单元的 IP 地址的标记的

出口 LSR 标记映射消息。对于每一个 FEC，分层标记模块可操作以为那个 FEC 生成分层标记映射消息，其指示由网络单元为那个 FEC 始发的标记，并指示网络单元始发了那个标记，并且还可操作以使生成的出口 LSR 标记映射消息和分层标记映射消息被传送到充当中间或中转 LSR 的一个或多个网络单元以允许为每一个 FEC 建立分层 LSP，所述分层 LSP 包含定义到达充当出口 LSR 的网络单元的 IP 地址的路径的外部 LSP 和特定于那个 FEC 的内部 LSP，内部 LSP 由充当出口 LSR 的网络单元用于标识分组并将分组转发到那个 FEC。分层 LSP 允许入口 LSR 通过改变与外部 LSP 相关联的一个或多个转发条目对影响到出口 LSR 的可到达性的拓扑改变作出反应，而基本上没有修改与任何内部 LSP 相关联的任何转发条目，由此改进了 LDP 收敛。

附图说明

[0011] 通过参考用于例证本发明实施例的如下描述和附图可最好地理解本发明。在附图中：

图 1 是根据一个实施例例证分层标记映射消息的分发和分层标记的管理的数据流程图；

图 2 是根据一个实施例更详细例证充当 LSR 的示范网络单元的框图；

图 3 例证了根据一个实施例在标记映射消息中使用的分层标记映射的示范消息格式；

图 4 例证了根据一个实施例在标记映射消息中使用的度量 TLV 的示范消息格式；

图 5 例证了根据一个实施例的标记映射消息的示范消息格式；

图 6 例证了根据一个实施例的标记请求消息的示范消息格式；

图 7 是例证根据一个实施例入口 LSR 安装标记映射消息使得当传输属于具体 FEC 的分组时将使用 LSP 分层的示范操作的流程图；

图 8 例证了根据一个实施例使用 LSP 分层的示范分组流；

图 9 例证了根据一个实施例当接收到送往远程 FEC 的分组时由入口 LSR 执行的示范操作；

图 10 是例证根据一个实施例当拓扑改变影响出口 LSR 的可到达性时执行的示范操作的数据流程图；以及

图 11 例证了根据一个实施例在图 10 中例证的拓扑改变之后使用 LSP 分层的示范分组流。

具体实施方式

[0012] 在以下描述中，阐述了许多特定细节。然而，要理解，本发明实施例可以在没有这些特定细节的情况下实施。在其它实例中，众所周知的电路、结构和技术未详细示出，以免模糊了对此描述的理解。本领域技术人员用所包含的描述将能够实现适当的功能性，而无需过多实验。

[0013] 在说明书中提到“一个实施例”、“一实施例”、“示例实施例”等指示所描述的实施例可包含具体特征、结构或特性，但每个实施例可能不一定都包含该具体特征、结构或特性。此外，这种短语不一定是指同一实施例。另外，当结合实施例描述具体特征、结构或特性时，认为它在本领域技术人员的知识范围内，以结合其它实施例来实现这种特征、结构或

特性,而不管是否明确描述了。

[0014] 在以下说明书和权利要求书中,可使用术语“耦合”和“连接”,连同它们的派生词。应该理解,这些术语不打算作为彼此的同义词。“耦合”用于指示两个或更多单元彼此协同操作或交互作用,它们可以直接或者可以不直接物理接触或电接触。“连接”用于指示在彼此耦合的两个或更多单元之间建立通信。

[0015] 本文所用的网络单元(例如路由器、交换机、桥、基站等)是以通信方式互连网络上其它设备(例如其它网络单元、终端站等)的连网设备件,其包含硬件和软件。订户终端站(例如服务器、工作站、膝上型电脑、掌上电脑、移动电话、智能电话、多媒体电话、通过因特网协议的语音(VOIP)电话、便携式媒体播放器、GPS单元、游戏系统、机顶盒等)访问通过因特网提供的内容/服务和/或在叠加于因特网上的虚拟私用网络(VPN)上提供的内容/服务。所述内容和/或服务通常由属于服务或内容供应商的一个或多个终端站(例如服务器终端站)或参与对等服务的终端站提供,并且可包含公用网页(免费内容、店面、搜索服务等)、私用网页(例如提供电子邮件服务的用户名/密码访问的网页等)、通过VPN的公司网等。通常,订户终端站(例如通过(有线或无线)耦合到接入网的客户驻地设备)耦合到边缘网络单元,边缘网络单元(例如通过一个或多个核心网络单元)耦合到其它边缘网络单元,其它边缘网络单元耦合到其它终端站(例如服务器终端站)。

[0016] 描述了用于使用分层标记交换路径(LSP)来改进LDP收敛的方法和设备。在一个实施例中,充当FEC的出口LSR的网络单元除了广告与出口LSR的IP地址(例如属于出口LSR并具有来自其它LSR的路径的环回(loopback)地址或其它地址)相关联的标记(例如空标记)(其属于分层LSP的外部LSP(本文中有时称为出口LSR LSP))(本文中称为出口LSR标记映射)之外,还广告由那个出口LSR为那个FEC始发的标记(其属于对应分层LSP的内部LSP(本文中有时称为FEC LSP))(本文称为分层标记映射)。通过这么做,使用LDP来分发FEC到其对应出口LSR的映射及其标记映射(该标记由那个出口LSR始发)。使用这些标记映射,LSP分层用于传输属于那个FEC的分组。到那个FEC的出口LSR的路径在分层中较低,在其上遂穿在分层中较高的LSP(特定于那个FEC的)。

[0017] 响应于导致下一跳改变的到出口LSR的路径改变,仅对应于在分层中较低的路径的下一跳需要在数据平面中被重新编程,这改进了LDP的收敛时间,并可减少在链路故障或节点故障期间的业务损耗持续时间。此外,仅到出口LSR的标记映射必须在诸如链路状态IGP等路由协议中承载,由此缩小了在路由协议中承载的信息大小,从而引起更快的路由协议收敛。

[0018] 充当入口LSR的网络单元和充当中转LSR的网络单元安装属于外部LSP的标记映射(出口LSR标记映射),标记映射当从出口LSR穿过到入口LSR时可被修改。例如,中转LSR将属于外部LSP的标记的值对换到它们自己的标记空间的值(保持与出口LSR的地址的关联性)。充当分层LSP的入口LSR的网络单元在其NHLFE(下一跳标记转发条目)结构中安装FEC的分层标记映射,使得当将分组转发到那个FEC时使用分层LSP。例如,在一个实施例中,充当入口LSR的网络单元在其NHLFE中安装对应于外部LSP的标记映射的条目(本文中称为出口LSR标记映射条目),并安装包含对应于外部LSP的递归下一跳到标记映射的分层标记映射的条目(本文中称为分层标记映射条目)。当转发送往这些FEC之一的分组时,入口LSR在其FTN中查找该FEC以确定对应的NHLFE结构。所得到的NHLFE指示

要推送到标记栈上的 FEC 标记（由出口 LSR 为那个 FEC 始发的标记），并且包含到出口 LSR 的递归下一跳标记映射。入口 LSR 将 FEC 标记推送到栈上，并访问对应于出口 LSR 标记映射的 NHLFE，其包含推送到栈上的标记以及朝向出口 LSR 的下一跳。入口 LSR 将那个标记推送到标记栈上，并将标记的分组传送到在标记映射中标识的到达出口 LSR 的下一跳。除了出口 LSR（可能还有执行倒数第二跳弹出的倒数第二中间 LSR）之外，分层标记的分组在网络中基于外部标记进行标记交换。

[0019] 响应于影响出口 LSR 的可到达性的拓扑改变（例如链路故障、节点故障、度量改变、存在新路由等），并假设存在到出口 LSR 的不同路由（尽管它可能是次优路由），仅需要改变属于外部 LSP 的 NHLFE 中的标记条目（出口 LSR 标记映射条目），而不是附连到出口 LSR 的每个 FEC 的每个条目。因而，代替需要修改这些 FEC 的每一个条目，仅需要修改那些 FEC 的对应于到达出口 LSR 的条目。这改进了 LDP 的收敛时间，并可减少在链路故障或节点故障期间的业务损耗持续时间。

[0020] 图 1 是根据一个实施例例证分层标记映射消息的分发和分层标记的管理的数据流程图。网络 100 包含 LSR（标记交换路由器）110、120、130 和 140，它们是同一 MPLS 域的一部分。LSR 110、120、130 和 140 中的每个都在网络单元上实现。LSR 140 通过链路 145 与 LSR 120 耦合，并通过链路 135 与 LSR 130 耦合。LSR 110 通过链路 115 与 LSR 120 耦合，并且还通过链路 125 与 LSR 130 耦合。应该理解，所例证的 LSR 数量是示范性的，原因在于网络中可能存在更多或更少的 LSR。

[0021] 图 2 根据一个实施例更详细例证了充当 LSR 的示范网络单元。网络单元 200 包含控制平面 210 和数据平面 250（有时称为转发平面或媒体平面）。控制平面 210 确定如何路由数据（例如分组）（例如所述数据的下一跳和所述数据的出局端口），而数据平面 250 负责转发该数据。控制平面 210 包含 IGP（内部网关协议）模块 215 和 LDP（标记分发模块）220。IGP 模块 215 可运行诸如 OSPF（开放最短路径优先）或 IS-IS（中间系统对中间系统）等链路状态协议，或运行诸如 RIP（路由信息协议）等其它协议。IGP 模块 215 与其它网络单元通信，以交换路由并基于一个或多个路由度量来选择那些路由。选择的 IGP 路由被存储在 RIB（路由信息库）225 中。IGP 模块 215 也能使未被选择且存储在 RIB 225 中的路由条目存储在本地 RIB（例如 IGP 本地 RIB）中。

[0022] LDP 模块 220 与其对等体（LDP 对等体）交换标记映射信息。例如，LDP 模块 220 可生成标记映射消息，以及从其对等体接收标记映射消息。LDP 模块 220 依赖于由 IGP 模块 215 提供给 RIB 225 的基础路由信息，以便转发标记分组。LDP 模块 220 分配标记，并将与转发标记分组相关的其它信息（例如 NHLFE 信息、ILM（入局标记映射）信息、FTN 信息）存储在 MPLS 信息库 230 中。LDP 模块 220 包含分层标记模块 222，该分层标记模块 222 将 LDP 模块 220 的功能性扩展成在标记映射始发和标记管理期间支持分层标记，将在本文后面对此进行更详细描述。

[0023] 控制平面 210 基于 RIB 225 和 MPLS 信息库 230 用路由信息来对数据平面 250 进行编程。具体地说，来自 RIB 225 的某信息被编程到 FIB（转发信息库）255，并且来自 MPLS 信息库 230 的某信息被编程到 ILM 结构 260、NHLFE 结构 265 和 FTN 结构 270。

[0024] 在一个实施例中，网络单元 200 包含一个或多个线卡（line card）（有时称为转发卡）的集合和一个或多个控制卡的集合。线卡和控制卡的集合通过一个或多个机构（例如

耦合这些线卡的第一全网格和耦合所有这些卡的第二全网格)耦合在一起。线卡集合通常构成数据平面，并且可各存储 FIB 255、ILM 260、NHLFE 265 和 FTN 270，它们将用在转发分组时。具体地说，FTN 270 用于转发未标记(例如它们是在入口 LSR 从 MPLS 域外部接收的)但在转发前要标记的分组。ILM 260 用于转发标记的分组。控制卡通常运行包含 IGP 模块 215、LDP 模块 220 的路由协议，并存储 RIB 225 和 MPLS 信息库 230。

[0025] 将参考 LSR 110(其充当分别具有 FEC 1-N 的终端站 150A-N 的出口 LSR)、LSR 140(其充当入口 LSR)以及 LSR 120 和 130(其充当中间(中转)LSR)来描述图 1 和随后附图。因而，将相对于作为分组的目的地来描述终端站 150A-N(但是应该理解，终端站 150A-N 也可以是分组的源)。因而，为了到达终端站 150A-N，分组必须穿过 LSR 110。应该理解，在 LSR 110 与终端站 150A-N 之间可能存在其它网络单元和/或设备(例如一个或多个接入网单元)。FEC 1-N 中的每个都是标识 LDP LSP 上分组传送目的地的标识符(例如 IP 地址前缀、主机地址、用于传输伪线的伪线 ID(PWID))。如图 1 中所描绘的，终端站 150A-N 中的每个都与不同 FEC 相关联。为了简单起见，终端站 150A-N 的 FEC 将被描述为 IP 地址前缀，但应该理解，其中一个或多个 FEC 可以是不同的(例如主机地址、PWID)。可以是订户终端站或服务器终端站的终端站 160A-L 在图 1 和随后附图中被描述为要发送到终端站 150A-N 的分组的源，然而应该理解，它们也可以是分组的目的地。

[0026] 在一个实施例中，LSR 110、120、130 和 140 各具有类似于网络单元 200 的架构，而在其它实施例中，中间 LSR 120 和 130 不包含 H 标记模块。LSR 110、120、130 和 140 中的每个都运行确定从源到目的地的最佳路径的 IGP 实现(例如链路状态协议，诸如 OSPF 或 IS-IS、RIP、静态)。LSR 110、120、130 和 140 中的每个还运行遵循由 IGP 协议所确定的最佳路径的 LDP 实现。参考图 1，从 LSR 140 到达 LSR 110 的最佳路径通过 LSR 120(由虚线最佳路径线 196 表示)。从 LSR 140 到达 LSR 110 的辅助路径通过 LSR 130(由虚线辅助路径线 198 表示)。因而，从由终端站 160A-L 所发送的业务到终端站 150A-N 的最佳路径通过 LSR 120。

[0027] LSR 110 包含分层标记(H 标记)模块 112，并且 LSR 140 包含 H 标记模块 142(在一个实施例中分层标记(H 标记)模块 112 和 H 标记模块 142 是 LSR 110 和 LSR 140 的相应控制平面的一部分)。H 标记模块 112 生成标记映射消息并使该标记映射消息包含它是其出口的 FEC 的分层标记映射。H 标记模块 142 处理接收的标记映射消息(其包含分层标记映射)，所述处理包含在一个或多个结构中安装分层标记条目，并安装 FEC 标记与到出口 LSR 110 的递归下一跳。因而，H 标记模块 112 将现有 LDP 标记映射始发的功能性扩展成支持分层标记映射，并且 H 标记模块 142 将现有 LDP 标记管理的功能性扩展成支持分层标记映射。应该理解，尽管 H 标记模块 112 被描述为始发分层标记映射(实质上充当出口 H 标记模块)并且 H 标记模块 142 被描述为处理接收的分层标记映射(实质上充当入口 H 标记模块)，但其中一个或多个模块可包含这两个功能性。根据一个实施例，H 标记模块 112 和 142 是运行 LSR 110 和 140 的 LDP 的 LDP 模块的一部分。根据一个实施例，中间 LSR 120 和 130 不包含特定 H 标记模块，并且相反运行标准 LDP 机构。LSR 110 还包含 FTN 结构 162 和 NHLFE 结构 180，它们各由 H 标记模块 142 管理。在一个实施例中，FTN 结构 162 和 NHLFE 结构 180 是 LSR 140 的数据平面的一部分，并且可至少部分存储在 LSR 140 的一个或多个线卡上。

[0028] 当生成标记映射消息以向它充当其出口的每一个不同 FEC 的对等体广告标记绑定时,出口 LSR 110 的 H 标记模块 112 包含分层标记映射,该分层标记映射将那个 FEC 映射到非空标记并指示该标记由出口 LSR 110 始发。此外,出口 LSR 110 广告标记映射消息,该标记映射消息包含出口 LSR 标记映射消息,该出口 LSR 标记映射消息用标记(例如空标记)映射出口 LSR 110 的 IP 地址。这些标记映射被传送到 LSR 140 的每一个邻居(例如 LSR 120 和 130)。

[0029] 通过 LDP 传播的分层标记映射和出口 LSR 标记映射的组合提供出口 LSR 与 FEC 之间的关联性。因而,甚至在 IGP 实现是非链路状态协议、诸如 RIP 或静态路由的情况下,仍可获得出口 LSR 与 FEC 之间的关联性。这消除了对于 IGP 使用链路状态协议、诸如 OSPF 或 IS-IS 来承载信息的必要性;消除了由于承载大量 FEC/ 路由而引起 IGP 收敛变慢的问题(例如在本发明的实施例中,IGP 仅能承载最少的信息,诸如环回地址和标记交换路由器之间的链路,并且没有 FEC),由此导致更快速的 IGP 收敛;并且还避免了运行其它协议、诸如 BGP 或具有 LDP 对等体的全网格以在整个网络上传递 FEC 信息的复杂性。因而,标记交换基于出口 LSR 的可到达性,而不是 FEC 装置的可到达性。

[0030] 图 3 例证了根据一个实施例在标记映射消息中使用的分层标记映射的示范消息格式。分层标记 TLV(类型长度值)310 包含通用标记子 TLV 320 和 FEC 子 TLV 330 作为值 340。分层标记 TLV 310 的类型 335 指示该消息包含分层标记映射。通用标记子 TLV 320 类似于在 RFC 5036 中描述的通用标记 TLV。通用标记子 TLV 320 用于对 FEC 的标记进行编码。通用标记子 TLV 320 包含指示它是通用标记的类型和作为值 350 的标记(该标记通常是非空标记)。FEC 子 TLV 330 类似于在 RFC 5036 中描述的 FEC TLV。FEC 子 TLV 330 用于对出口 LSR 的 FEC 进行编码(例如出口 LSR 的环回地址)。FEC 子 TLV 330 包含指示它是 FEC TLV 的类型 355 和作为值 360 的出口 LSR 地址前缀 FEC 单元(例如出口 LSR 的环回地址)。在一个实施例中,仅存在单个地址作为值 360。

[0031] 在一些实施例中,度量 TLV 也可包含在分层标记映射消息中。度量 TLV 用于传播与 FEC 相关联的度量。图 4 例证了根据一个实施例的度量 TLV 的示范格式。度量 TLV 410 包含指示该消息是度量 TLV 的类型 420、长度 430、度量类型 440 和度量值 450。度量类型 440 指示度量的类型(例如,0 指示区域内,1 指示区域间,0xf 指示高于任何内部度量的度量的外部路由)。度量类型 440 的具体值可取决于正在使用的 IGP 实现的类型。度量值 450 指示度量的值。在一些实施例中,如果度量 TLV 未包含在包含分层标记映射的标记映射消息中,则该标记映射消息被看作具有为 0 的度量类型以及为 0 的度量值。

[0032] 这些 LSR 使用度量 TLV 来确定当为单个 FEC 接收到已经从不同出口 LSR 始发的多个分层标记映射时安装哪个分层标记映射。例如,如果入口 LSR 为单个 FEC 从不同邻居接收到始发自不同出口 LSR 的多个分层标记映射,则该 LSR 将选择那些分层标记映射之一来安装。在一个实施例中,具有度量类型 N 的度量 TLV 低于具有度量类型 N+1 的度量 TLV,而不管度量值如何。如果度量 TLV 具有相同度量类型(除了类型 0x0f 之外),则对通过将来自该路由的 RIB 的度量添加到(在 H 标记 TLV 中标识的)出口地址前缀 FEC 单元获得的值与度量 TLV 的度量值进行比较。如果度量 TLV 具有相同度量类型 0x0f,则仅使用这些度量 TLV 的度量值进行该比较。如果这些值在比较之后是相同的,则它们被视为等价的(然后可随机选择它们中的一个)。

[0033] 参考回图 1, 考虑 LSR 110 生成并传送与终端站 150A 相关联的 FEC1 的标记映射消息。在操作 1.1, H 标记模块 112 生成标记映射消息 114, 标记映射消息 114 包含出口 LSR 标记映射消息 116, 并且还包含分层标记映射消息 118, 出口 LSR 标记映射消息 116 包括 LSR 110 的 IP 地址 (例如 LSR 110 的环回地址) 映射到空标记, 分层标记映射消息 118 映射被映射到由 LSR 110 始发的非空标记 (标记 1) 的 FEC1 并且包含关于 LSR 110 始发那个非空标记的指示。在生成和 / 或传送分层标记映射消息 118 之前, 可生成并传送出口 LSR 标记映射消息 116。出口 LSR 标记映射消息 116 和分层标记映射消息 118 一起允许建立分层 LSP, 其包含出口 LSR LSP 和 FEC LSP。当标记交换指向该 FEC 的分组时, 将使用出口 LSR LSP, 并且 FEC LSP 将由该出口 LSR 用于标识分组并将分组转发到那个 FEC。在出口 LSR LSP 内遂穿该 FEC LSP。出口 LSR LSP 对分别与终端站 150A-N 相关联的 FEC 1-N 中的每个而言将是公用的。

[0034] 在一个实施例中, 代替用于分层标记映射的标记 TLV (如在 RFC 5036 中所定义的), 使用在图 3 中描述的分层标记 TLV, 或除了用于分层标记映射的标记 TLV (如在 RFC 5036 中所定义的) 之外, 还使用在图 3 中描述的分层标记 TLV。图 5 例证了分层标记映射消息 500 的示范格式, 其包含堆叠在 FEC TLV 510 顶上的分层标记 TLV 310。当用在图 1 中描绘的示例中时, 分层标记 TLV 310 指示非空标记 (标记 1) 与出口 LSR 110 的 IP 地址之间的映射, 并且 FEC TLV 510 指示 FEC1 (例如终端站 150A 的 IP 地址前缀)。

[0035] 在一些实施例中, LSR 110 响应于接收到标记请求消息而生成标记映射消息 114 (例如出口 LSR 标记映射消息 116 和 / 或分层标记映射消息 118)。该标记请求消息用于请求一个或多个 FEC 的标记绑定 (映射) (例如它可含有通配 FEC TLV 单元以请求多个标记绑定)。图 6 例证了根据一个实施例的标记请求消息 600 的示范格式。标记请求消息 600 包含 FEC TLV 610, 其标识为其请求标记的 FEC (FEC TLV 610 可包含通配符), 并且包含可选参数 620。可选参数 620 可包含分层标记 TLV。如果 FEC TLV 610 是通配符, 则在响应中返回所有标记映射 (这导致多个标记映射消息被生成和传送)。响应于标记请求消息, LSR 110 可在标记映射消息中包含“更多标记 TLV”。“更多标记 TLV”的存在指示响应于标记请求消息将对于该 FEC 发送更多标记映射。

[0036] 在生成标记映射消息 114 之后某一时间, 向 LSR 110 的对等体传送消息。因而, 在操作 1.2A, 所生成标记映射消息 114 (其包含出口 LSR 标记映射消息 116 和分层标记映射消息 118) 通过链路 115 传送到 LSR 120, 并在操作 1.2B 通过链路 125 传送到 LSR 130。

[0037] 在接收到标记映射消息之后, 在操作 1.3 和 1.4, 中间 LSR 120 和 130 分别至少安装出口 LSR 标记映射。例如, 它们在它们的 NHLFE 结构中安装标记和出口 LSR 110 的 IP 地址, 并在它们的 ILM 结构中创建条目, 使得它们能执行标记对换 (或倒数第二弹出)。例如, LSR 120 和 130 中的每个都在它们的 NHLFE 中安装条目, 其包含出局标记 (属于分层 LSP 的外部标记) (例如空标记)、对换动作 (或在执行倒数第二弹出的情况下是弹出动作)、到达 LSR 110 的下一跳 (在此情况下其是 LSR 110)、出局接口, 并且可包含其它数据处理信息。LSR 120 和 130 中的每个都从它们的相应标记空间为出口 LSR 标记映射分配标记。例如, LSR 120 为出口 LSR 标记映射分配标记 2, 并且 LSR 130 为出口 LSR 标记映射分配标记 3。LSR 120 和 130 在它们的相应 ILM 结构中创建条目, 以便将为出口 LSR 标记映射分配的标记 (其在分组转发期间将被作为入局标记接收) 映射到 LSR 110 的 IP 地址的 NHLFE。因

而,当 LSR 120 接收到将标记 2 作为入局标记的所标记分组时(将相对于图 8 对其进行更详细描述),访问标记 2 的 ILM 条目,其包含指向 NHLFE 结构中 LSR 110 的 IP 地址的条目的指针。类似地,当 LSR 130 接收到将标记 3 作为入局标记的所标记分组时(将相对于图 11 对其进行更详细描述),访问标记 3 的 ILM 条目,其包含指向 NHLFE 结构中 LSR 110 的 IP 地址的条目的指针。在一些实施例中,中间 LSR 120 和 130 并未在超出向它们的相对对等体(其在图 1 中描绘的示例中是 LSR 140) 转发消息之外的范围处理分层标记映射消息 118。

[0038] 虽然图 1 例证 LSR 120 和 LSR 130 每 FEC 各接收单个分层标记映射消息,但应该理解,存在它们为单个 FEC 从多个邻居接收多个分层标记映射消息的情形。在一个实施例中,当中转 LSR 为单个 FEC 从多个邻居接收多个分层标记映射消息时,它选择这些标记映射中的一些用于广告和 / 或安装(中转 LSR 将在它具有没有能力处理本文中描述的分层标记映射的邻居的情况下安装这些标记映射)。在一个实施例中,选择从针对分层标记中所指示的 IP 地址作为下一跳的邻居接收的分层标记映射,并将所述分层标记映射广告给所有 LDP 邻居(假设那些 LDP 邻居能够处理分层标记映射)。如果存在不能够处理分层标记映射的邻居,则中间 LSR 应该以与入口 LSR 将安装那个分层标记映射类似的方式来安装该分层标记映射(如果适用时包括比较度量 TLV 中指示的度量类型和度量值)。

[0039] 在安装出口 LSR 标记映射之后某一时间,分别在操作 1.5 和 1.6,中间 LSR 120 和 130 分别生成标记映射消息 122 和 132 并将标记映射消息 122 和 132 传送到 LSR 140。如图 1 中所例证的,标记映射消息 122 包含出口 LSR 标记映射消息 124,并且包含分层标记映射消息 118,其中出口 LSR 标记映射消息 124 映射出口 LSR 110 的 IP 地址和标记 2。标记映射消息 132 包含出口 LSR 标记映射消息 134,并且包含分层标记映射消息 118,其中出口 LSR 标记映射消息 134 映射出口 LSR 110 的 IP 地址与标记 3。因而,虽然 LSR 120 和 130 通过将出口 LSR 标记改变成它们标记空间中的标记已经各修改了出口 LSR 标记映射,但分层标记映射 118 保持不变。

[0040] LSR 140 可将包含在标记映射消息 122 和标记映射消息 132 中的信息存储在 IGP 数据结构中。例如,在一个实施例中,该信息被存储在链路状态数据库(LSDB)中。接收到这些标记映射消息之后,在操作 1.7,LSR 140 的 H 标记模块 142 确定选择在其数据平面中在从 LSR 130 接收的标记映射消息 132 之上安装从 LSR 120 接收的标记映射消息 122。在一个实施例中,这个判定基于由 IGP 所确定并由度量值进一步精炼(如果有必要的话)的最佳路由(例如,如果度量 TLV 包含在这些标记映射消息中)。

[0041] 在操作 1.8,H 标记模块 142 安装标记映射消息 122,使得 LSP 分层将用在传输属于 FEC1 的分组时。因而,当接收到 FEC1 的分组时,入口 LSR 140 会将由 LSR 110 为 FEC1 始发的标记推送到标记栈上(内部标记),并且会将到达出口 LSR 110 的标记(在所描绘的示例中是标记 2)(外部标记)推送到标记栈上。例如,H 标记模块 142 使条目安装在 NHLFE 结构 180 中,用于包含在出口 LSR 标记映射消息 124 中的出口 LSR 标记映射和包含在分层标记映射消息 118 中的分层标记映射。出口 LSR 标记映射的 NHLFE 包含标记 2、推送动作和下一跳(到 LSR 120 的 IP 地址)。这个 NHLFE 将用在放上分层 LSP 的外部标记时。属于 FEC1 的标记的 NHLFE 包含标记 1、推送动作和到出口 LSR 110 的 IP 地址的递归下一跳。H 标记模块 142 还使条目安装在 FTN 结构 162 中,用于与终端站 150A 相关联的 FEC1。

[0042] 如图 1 中所例证的,FTN 结构 162 包含 FEC 索引 164 和 NHLFE 索引 166,NHLFE 索

引 166 充当到 NHLFE 结构 180 中的指针。NHLFE 结构 180 包含对应于 NHLFE 索引 166 的索引 182、出局标记字段 184、动作字段 186 和下一跳字段 188。NHLFE 结构 180 还可包含附加信息（例如出局接口、其它数据处理信息）。如图 1 中所例证的，FTN 结构 162 包含与终端站 150A-N 相关联的每一个 FEC 的条目。此外，NHLFE 结构 180 包含出口 LSR 标记映射的出口 LSR 标记条目 190（例如根据出口 LSR 标记映射消息 124 生成的）和分层标记映射的分层标记条目 192（例如在图 1 所描绘的示例中，基于分层标记映射消息 118 生成 FEC1 的条目）。虽然例证了 FTN 和 NHLFE 的分开结构，但应该理解，在一些实施例中，存在表示 FTN 和 NHLFE 的单个结构。虽然图 1 例证了安装标记映射消息 122 使得 LSP 分层将用在传输属于 FEC1 的分组时的具体方式，但应理解，它是示范性的，并且标记映射消息 122 可以不同方式安装，而当传输属于 FEC1 的分组时仍创建 LSP 分层。此外，虽然在图 1 中描绘的示例是特定于 FEC1 的，但应理解，对于每一个 FEC 1-N 都执行类似操作。

[0043] 图 7 是例证根据一个实施例入口 LSR 安装标记映射消息使得当传输属于具体 FEC 的分组时将使用 LSP 分层的示范操作的流程图。现在将参考图 1 描述图 7 的操作。然而，应该理解，图 7 的操作可由本发明的不同于相对于图 1 所讨论实施例的实施例来执行，并且相对于图 1 讨论的实施例可执行与相对于图 7 所讨论的那些操作不同的操作。

[0044] 在块 710，LSR 140 从 LSR 120 接收出口 LSR 标记映射消息和分层标记映射消息。可在不同时间接收这些标记映射消息。出口 LSR 标记映射消息指示出口 LSR 110 的 FEC（例如 IP 地址前缀）与标记的映射。分层标记映射消息指示与终端站（例如终端站 150A-N 之一）相关联的 FEC 与由出口 LSR 始发的标记的映射，并且还标识始发了那个标记的出口 LSR。

[0045] 流程然后移动到块 720，并且 H 标记模块 142 为出口 LSR 110 与所标识标记的映射创建 NHLFE，其包含所标识标记的推送操作和到 LSR 120 的下一跳。流程然后移动到块 730，并且 H 标记模块 142 为在分层标记映射消息中所指示的 FEC 创建 NHLFE，其包含分层标记映射消息中所指示的标记的推送操作和到出口 LSR 110 的 NHLFE 的递归下一跳。到出口 LSR 110 的 NHLFE 的递归下一跳将引起执行另一查找（访问出口 LSR 110 的 NHLFE）。

[0046] 流程然后移动到块 740，并且 H 标记模块 142 为分层标记映射创建与分层标记映射消息中指示的 FEC 对应的 FTN 条目，并且可选地为出口 LSR 标记映射创建与出口 LSR 140 的 IP 地址前缀对应的 FTN 条目。流程然后移动到块 750，并且所创建的条目被下载到 LSR 140 的一个或多个线卡。例如，NHLFE 条目被下载到 NHLFE 结构 180（其可存储在一个或多个线卡上），并且 FTN 条目被下载到 FTN 结构 162（其可存储在一个或多个线卡上）。

[0047] 如图 1 中所例证的，出口 LSR 条目 190 包含标记 2、推送动作和 LSR 120 的下一跳，标记 2 在出口 LSR 标记映射消息 124 中被映射到出口 LSR 110 的 IP 地址。为与终端站 150A 相关联的 FEC 安装的 FEC 标记条目包含对应的标记 1、推送动作和到出口 LSR 110 的 NHLFE 的递归下一跳，标记 1 包含在分层标记映射消息 118 中。递归下一跳指示：当转发分组时，将执行基于出口 LSR 110 的 IP 地址的附加查找。

[0048] 图 8 例证了根据一个实施例使用 LSP 的分层从终端站 160A 到终端站 150A 的示范分组流。将参考图 1 和图 9 来描述图 8，图 9 例证了当接收到送往远程 FEC 的分组时由入口 LSR 执行的示范操作。然而，应该理解，图 9 的操作可由本发明的不同于参考图 1 和 9 所讨论的那些实施例的实施例来执行，并且参考图 1 和 9 讨论的实施例可执行与参考图 9 讨论

的那些操作不同的操作。

[0049] 终端站 160A 传送包含 IP 标题（具有终端站 150A 的 IP 目的地地址）和数据有效载荷的分组 810。参考图 9，在块 910，入口 LSR 140 接收分组 810。流程然后移动到块 920，并且入口 LSR 140 确定分组 810 的目的地 IP 地址的 FEC，例如，入口 LSR 140 包含将目的地 IP 地址映射到 FEC 的 IP 到 FEC 规则的集合。所得到的 FEC 对应于终端站 150A 的 IP 地址前缀。流程然后移动到块 930。

[0050] 在块 930，入口 LSR 140 访问对应于该 FEC 的 NHLFE。例如，首先基于该 FEC 来访问 FTN 结构 162，以确定到对应于该 FEC 的 NHLFE 的 NHLFE 索引（指针）。在图 1 的示例中，对应于 FEC IP 前缀 1 的 NHLFE 索引是 2。访问对应于所确定索引 2 的 NHLFE。该 NHLFE 包含标记、推送动作和到出口 LSR 110 的 NHLFE 的递归下一跳。流程然后移动到块 940，并且标记 1 被推送到标记栈上。流程然后移动到块 950。

[0051] 在块 950，入口 LSR 140 因为该 FEC 的 NHLFE 的递归下一跳而访问出口 LSR 110 的 NHLFE（出口 LSR 标记条目 190）。在图 1 的示例中，出口 LSR 标记条目 190 包含标记 2、推送动作和 LSR 120 的下一跳。流程然后移动到块 960，并且标记 2 被推送到标记栈上。流程然后移动到块 970，并且分组 815 被传送到下一跳 LSR 120。如图 8 中所例证的，分组 815 在标记 1 顶上包含标记 2。

[0052] LSR 120 接收分组 815。基于入局标记 2，LSR 120 确定如何转发该分组。例如，LSR 120 访问它的用于入局标记 2 的 ILM 条目，并确定对应 NHLFE（其将指示如何处理该分组）。在一些实施例中，中间 LSR 120 执行倒数第二跳弹出，使得它在向 LSR 110 传送该消息之前移除最外部标记（到达出口 LSR 110 的标记）。在其它实施例中，中间 LSR 120 执行标记对换。例如，它将入局标记与出口 LSR 110 所广告的标记（例如空标记）对换。如图 8 中所例证的，LSR 120 在向出口 LSR 110 传送分组 820 之前将标记 2 弹出该分组的标记栈。入口 LSR 110 接收分组 820，并基于标记 1 来确定该分组的目的地。例如，入口 LSR 110 基于入局标记 1 来检查其 ILM 条目，并确定目的地是终端站 150A。LSR 110 弹出标记 1，并向终端站 150A 传送分组 810。

[0053] 因而，使用 LSP 的分层来传送标记的分组，其中外部标记的 LSP 属于出口 LSR，并且内部标记的 LSP 标识在外部 LSP 的出口处的 FEC。

[0054] 响应于影响出口 LSR 110 的可到达性（因此还有每一个终端站 150A-N 的可到达性）的拓扑改变，仅需要修改出口 LSR 标记条目。图 10 是例证当拓扑改变影响出口 LSR 110 的可到达性时执行的示范操作的数据流程图。在操作 10.1，已经发生了拓扑改变，其已经将最佳路径从路径 196 改变到路径 198。发生拓扑改变可能由于数个原因，其包括：链路（例如链路 145 和 / 或链路 115）已经被破坏、度量的改变使路径 198 更优化、静态路由已经改变、节点已经被破坏（例如 LSR 120 已经被破坏）或者已经建立了更优化的新路径。在一个实施例中，IGP 模块通知 LDP 模块关于该拓扑改变。

[0055] 在图 10 中例证的示例中，该拓扑改变已将从终端站 160A-L 发送的送往终端站 150A-N 的分组的下一跳从 LSR 120 改变到 LSR 130。由于分组是基于它们的出口 LSR 的可到达性而不是基于目的地 FEC 的可到达性被标记交换的，因此仅需要更新对应于出口 LSR 的条目。因而，响应于该拓扑改变，在操作 10.2，H 标记模块 142 仅改变出口 LSR 标记条目 190，以使这些分组被标记交换到 LSR 130，来代替 LSR 120。具体地说，H 标记模块 142 将

该标记改变成标记 3(其之前由 LSR 130 在出口 LSR 标记映射消息 134 中广告了) 并且将下一跳改变到 LSR 130。应该理解,分层标记条目 192 都不需要改变。尽管图 10 中未例证,但在一些实施例中, H 标记模块 222 基于其控制卡来改变其 MPLS 信息中的条目,并仅将改变的条目下载到存储在 LSR 140 的线卡上的 NHLFE 结构 180。

[0056] 因而,代替改变和下载受(影响出口 ISR 的可到达性的)拓扑改变影响的 FEC IP 前缀的每一个条目(例如 NHLFE 结构中的条目,可能还有 FEC IP 前缀的 ILM 结构中的条目),仅需要改变出口 LSR 的条目,由此减少收敛时间。因此,当到出口 LSR 的路径改变了并导致下一跳改变时,在数据平面仅需要重新编程对应于到达出口 LSR 的路径的下一跳。使用分层标记还在改变到达出口 LSR 的拓扑的链路故障或节点故障状况期间减少了业务损耗的持续时间。例如,考虑当链路 145 和 / 或链路 115 或 LSR 120 被破坏时的情况。在不使用分层标记映射的现有技术解决方案中,通过 IGP 收敛(至少在控制平面中)之后是 LSR 计算并下载更新的 NHLFE 条目(可能还有 ILM 条目)到数据平面,来选通(gate)业务恢复。随着 IP 前缀数量的增加,IGP 收敛(至少在控制平面中)的持续时间以及计算并下载所有更新条目到数据平面的持续时间增加了。因而,在不使用分层标记映射的现有技术解决方案中,链路故障或节点故障导致随着网络中的前缀 / 标记的数量增加而增加的业务损耗持续时间。相比之下,使用分层标记允许仅更新出口 LSR 的条目。因而,使用本发明的实施例,仅通过 IGP 收敛(至少在控制平面)的时间量以及只改变和下载更新条目到数据平面的时间来选通业务恢复。

[0057] 图 11 例证了根据一个实施例在图 10 中所例证的拓扑改变之后从终端站 160A 到终端站 150A 使用 LSP 分层的示范分组流。终端站 160A 传送分组 1110,分组 1110 包含 IP 标题(具有终端站 150A 的 IP 目的地)和数据有效载荷。LSR 140 接收该分组,并确定包含在 IP 标题中的目的地 IP 地址的 FEC。基于该 FEC,访问对应的 NHLFE,其包含标记 1、推送动作和到出口 LSR 110 的 NHLFE 的递归下一跳。LSR 140 将标记 1 推送到标记栈上,并访问对应于该递归下一跳的 NHLFE。那个 NHLFE 包含标记 3、推送操作和到 LSR 130 的下一跳。LSR 140 将标记 3 推送到标记栈上,并向 LSR 130 传送分组 1115。

[0058] LSR 130 接收分组 1115。基于入局标记 3, LSR 130 确定如何转发该分组。例如,LSR 130 访问它的用于入局标记 3 的 ILM 条目,并确定对应的 NHLFE。在一些实施例中,中间 LSR 130 执行倒数第二跳弹出,使得它在向 LSR 110 传送消息之前移除最外部标记(到达出口 LSR 110 的标记)。在其它实施例中,中间 LSR 130 执行标记对换。例如,它将入局标记与出口 LSR 110 广告的标记(例如空标记)对换。如图 11 中所例证的,LSR 130 在向出口 LSR 110 传送分组 1120 之前将标记 3 弹出分组的标记栈。入口 LSR 110 接收分组 1120,并基于标记 1 来确定该分组的目的地。例如,入口 LSR 110 基于入局标记 1 来检查其 ILM 条目,并确定目的地是终端站 150A。LSR 110 弹出标记 1,并向终端站 150A 传送分组 1110。

[0059] 在一些实施例中,网络中的 LSR 交换信息以确定是否支持分层标记能力。例如,可通过 LDP 在这些 LSR 之间交换分层标记能力 TLV(其指示是否支持分层标记能力)。在一个实施例中,如果沿被路由路径的 LSR 不支持处理本文描述的分层标记映射,则外部 LSP 继续,直到沿到出口 LSR 的被路由路径在出现不能够进行分层标记处理的 LSR 之前支持分层标记处理的最远下游 LSR。在一个实施例中,外部 LSP 是 TE(业务工程)LSP。在一个实施例中,如果由没有分层标记映射的 LSR 广告存在分层标记映射的 FEC(例如,如果邻居不能

够处理分层标记映射),则应该利用到分层 LSP 的对换操作来安装那个 LSR 上的 ILM 条目,并且分层标记映射应该安装在那个 LSR 上,其方式与入口 LSR 安装那个分层标记映射的方式类似。

[0060] 在一些实施例中,当广告分层标记映射时,可使用独立的 LSP 控制或有序的 LSP 控制。例如,有 H 标记能力的 LSR 可使用独立的标记分发控制以在它期望的任何时间向其对等体广告分层标记映射。因而,如果该 LSR 未接收到分层标记映射的话(假设该 LSR 未从 FEC 的其中一个下一跳接收到分层映射),该 LSR 可广告具有其地址的 FEC 的分层标记映射。

[0061] 有分层标记能力的 LSR 还可使用有序标记分发控制进行操作。在有序的标记分发控制模式中,LSR(其从其邻居接收到到 FEC 的多个分层标记映射)选择具有到该 FEC 的最低成本路径的 H 标记映射。为该 FEC 选择的分层标记映射被广告给其邻居(至少是有分层标记能力的邻居)。在向相邻 LSR 广告分层标记映射之前,对应于这些 FEC 的路由不需要经由 IGP 出现在 RIB(或 FIB) 中。

[0062] 在一些实施例中,LSR 可使用标记保留模式(诸如保守标记保留模式)来为 FEC 保持从邻居(其不是针对该 FEC 的其下一跳)处获知的标记绑定。在保守标记保留模式,如果用于该 FEC 的所有路径 / 下一跳都具有公用共享风险链路组(SRLG),则该 LSR 可作为备份具有不共享 SRLG 的备选下一跳,其可能需要向另一标记请求标记。公用 SRLG 是两个链路共享的风险。作为示例,如果通过公用管道承载多个光纤,则它们共享 SRLG,这是因为如果该管道被切断,则两个光纤也可能被切断。

[0063] 在一些网络拓扑中,这些终端站可被多次返回(multi-homed)到多个出口 LSR。在这种情况下,每一个出口 LSR 都广告与那些终端站相关联的 FEC 的分层标记映射,以及出口 LSR 标记映射。当主要出口 LSR 变得不可到达(例如那个 LSR 被破坏)时,入口 LSR 改变到达辅助出口 LSR 的外部 LSP 以及内部 LSP。然而,由于在本发明实施例中属于这些终端站的 FEC 未在 IGP 中广告(例如仅环回地址以及在标记交换路由器之间的链路在 IGP 中广告),因此入口 LSR 上的 IGP 数据库(例如链路状态数据库(LSDB))将比较小,这将减少 IGP 收敛所需的时间量且允许入口 LSR 更快速切换到辅助 LSR。

[0064] 可使用在一个或多个电子装置(例如终端站、网络单元等)上存储和执行的代码和数据来实现图中所示的技术。这种电子装置使用机器可读介质、诸如机器可读存储介质(例如磁盘、光盘、随机存取存储器、只读存储器、闪存装置、相变存储器)和机器可读通信介质(例如电、光、声或其它形式的传播信号-诸如载波、红外信号、数字信号等)存储代码和数据以及(在内部和 / 或通过网络与其它电子装置)传递代码和数据。此外,这种电子装置通常包含耦合到一个或多个其它组件(诸如一个或多个存储装置、用户输入 / 输出装置(例如键盘、触摸屏和 / 或显示器)以及网络连接)的一个或多个处理器的集合。例如,在网络单元包含控制卡和线卡的情况下,这些卡中的每个卡都包含一个或多个处理器的集合(例如,线卡包含一个或多个分组处理实体(例如分组处理 ASIC)的集合)。处理器集合与其它组件的耦合通常通过一个或多个总线和桥(也称为总线控制器)。承载网络业务的存储装置和信号分别表示一个或多个机器可读存储介质和机器可读通信介质。因而,给定电子装置的存储装置通常存储代码和 / 或数据以便在那个电子装置的一个或多个处理器的集合上执行。当然,可使用软件、固件和 / 硬件的不同组合来实现本发明实施例的一个或多个部分。

[0065] 虽然附图中的流程图示出了通过本发明某些实施例执行的具体操作顺序,但应该理解,这种顺序是示范性的(例如备选实施例可按不同顺序执行这些操作、组合某些操作、交叠某些操作等)。

[0066] 虽然已经根据多个实施例描述了本发明,但本领域技术人员将认识到,本发明不限于描述的实施例,可以在所附权利要求书的精神和范围内用修改和改变来实施。描述由此被视为例证性的,代替限制性的。

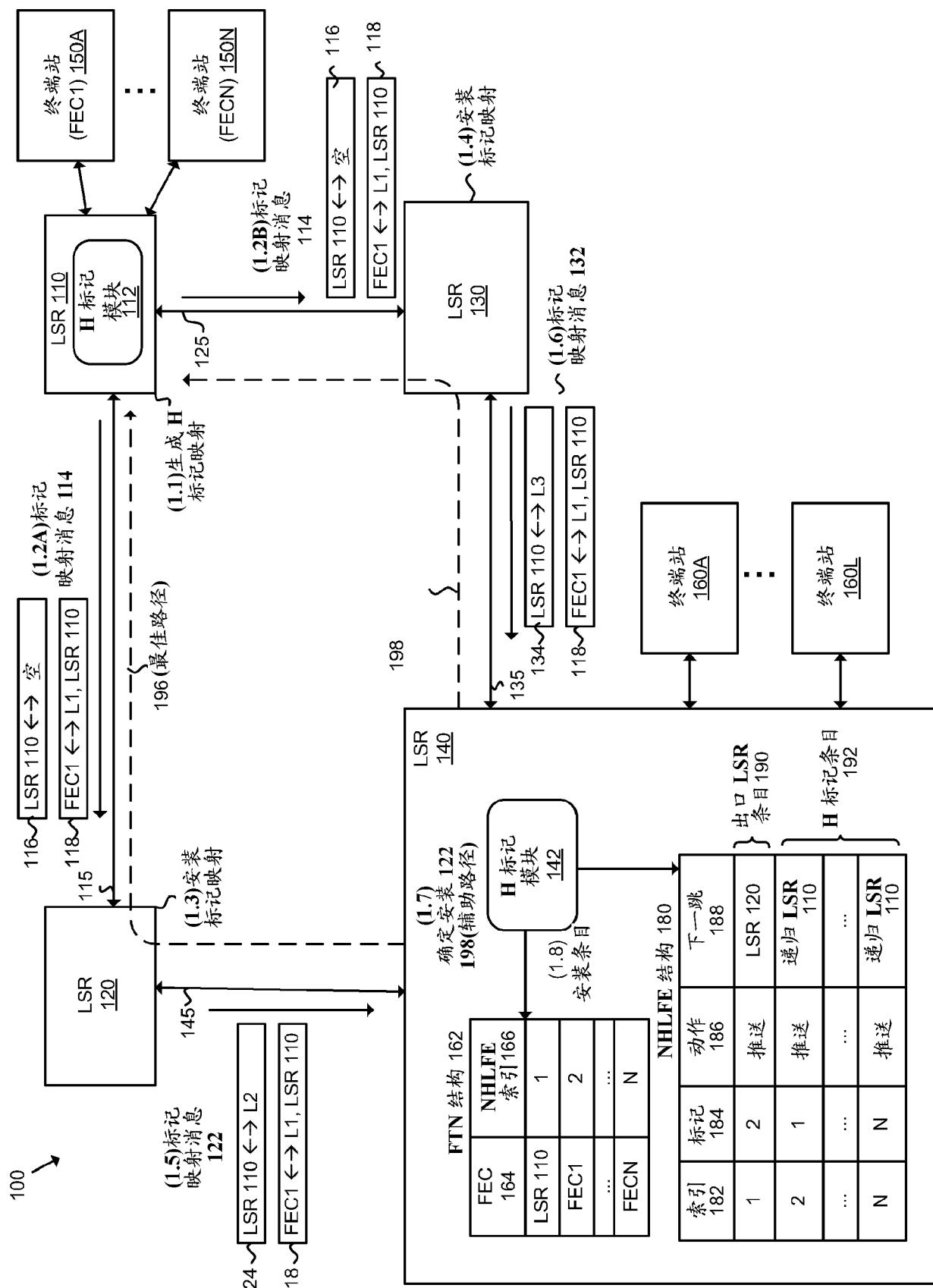


图 1

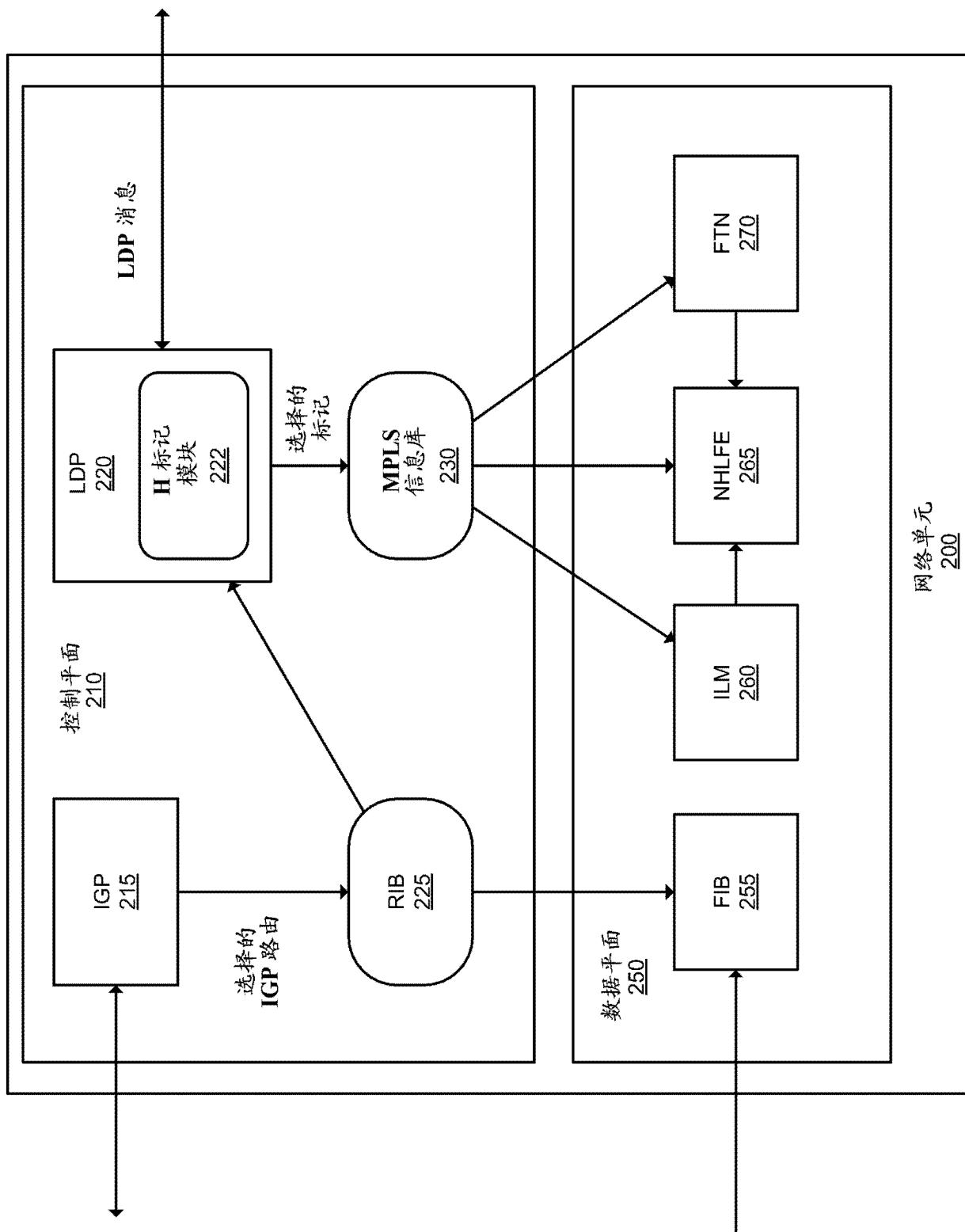


图 2

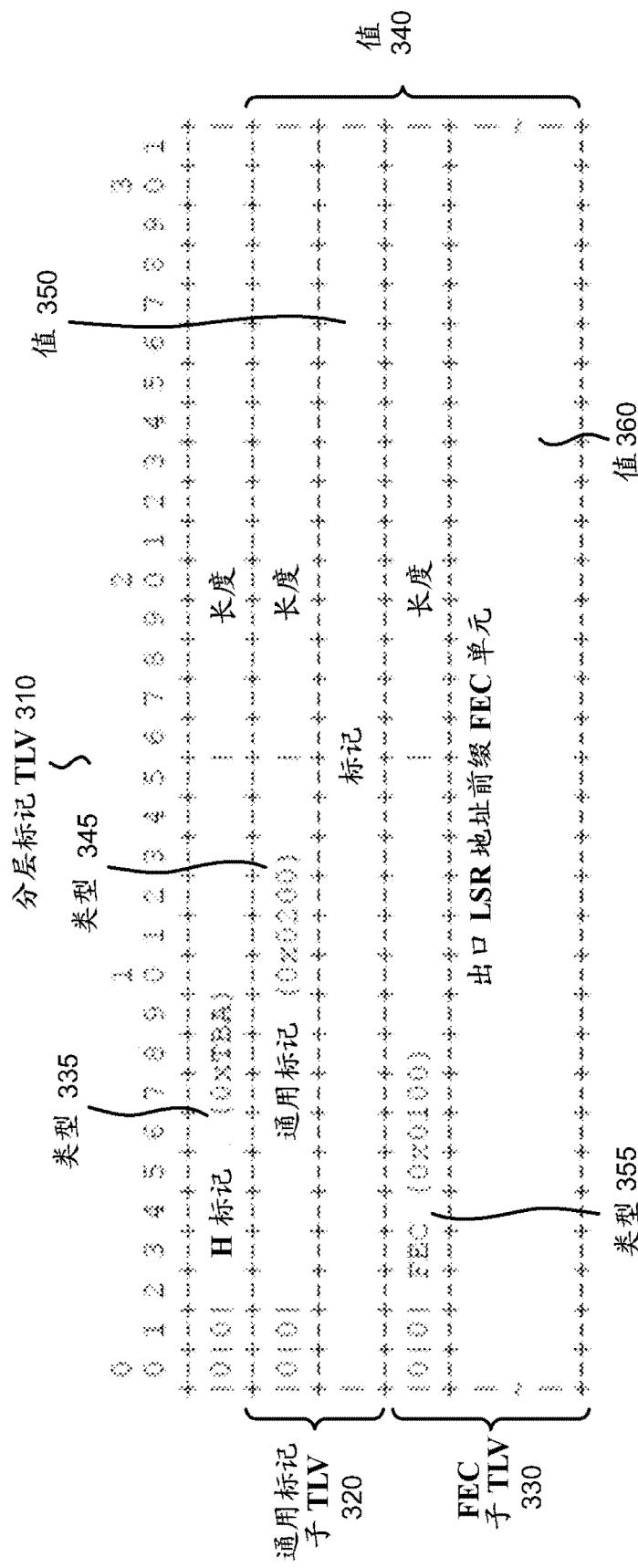


图 3

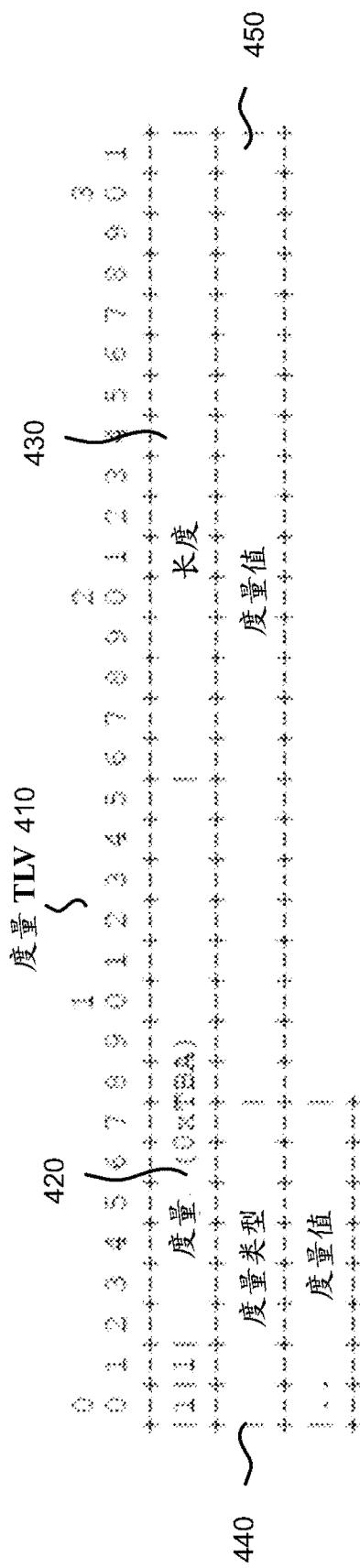


图 4

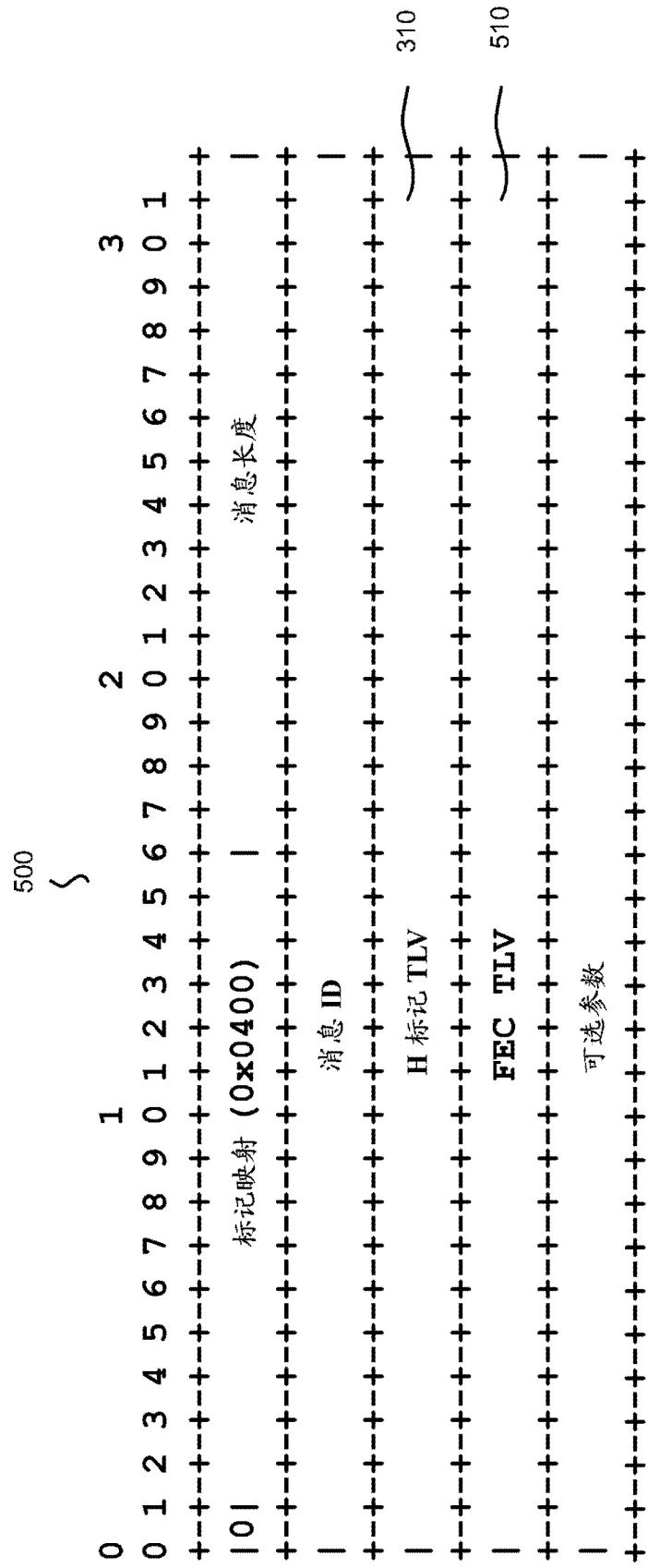


图 5

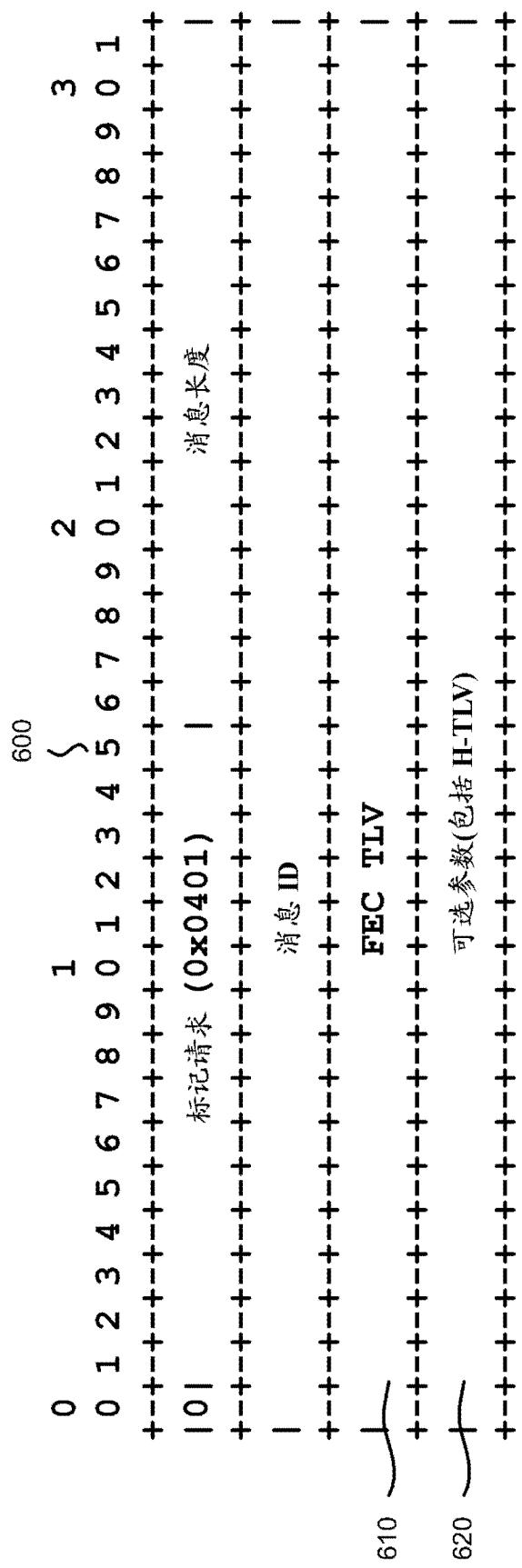


图 6

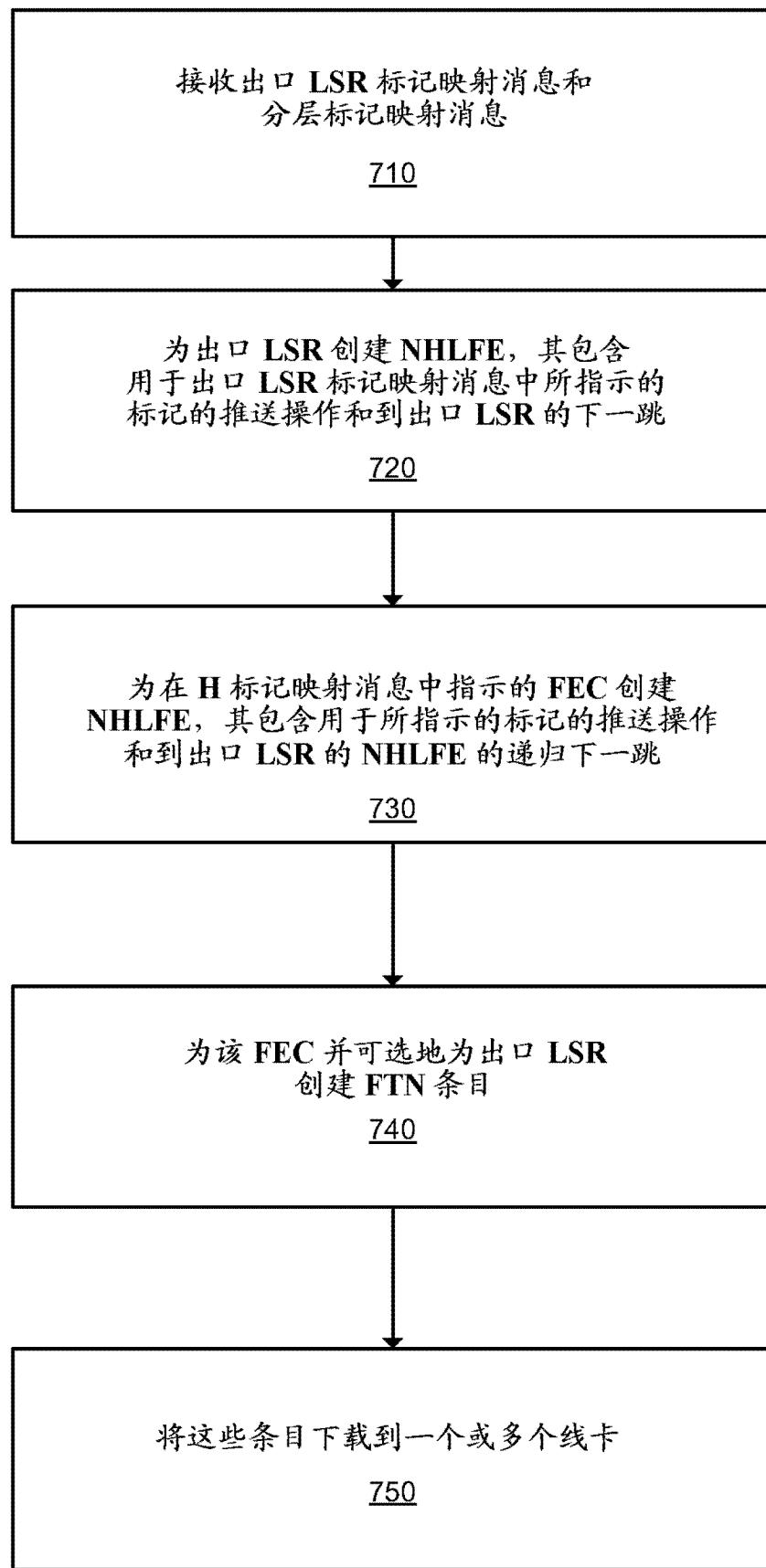


图 7

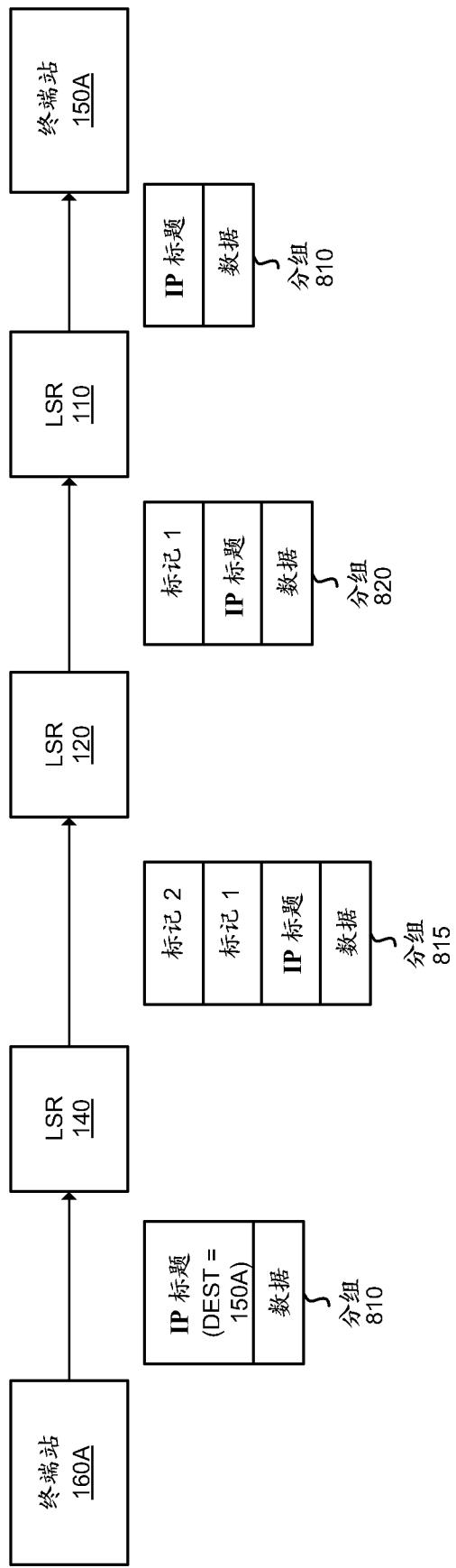


图 8

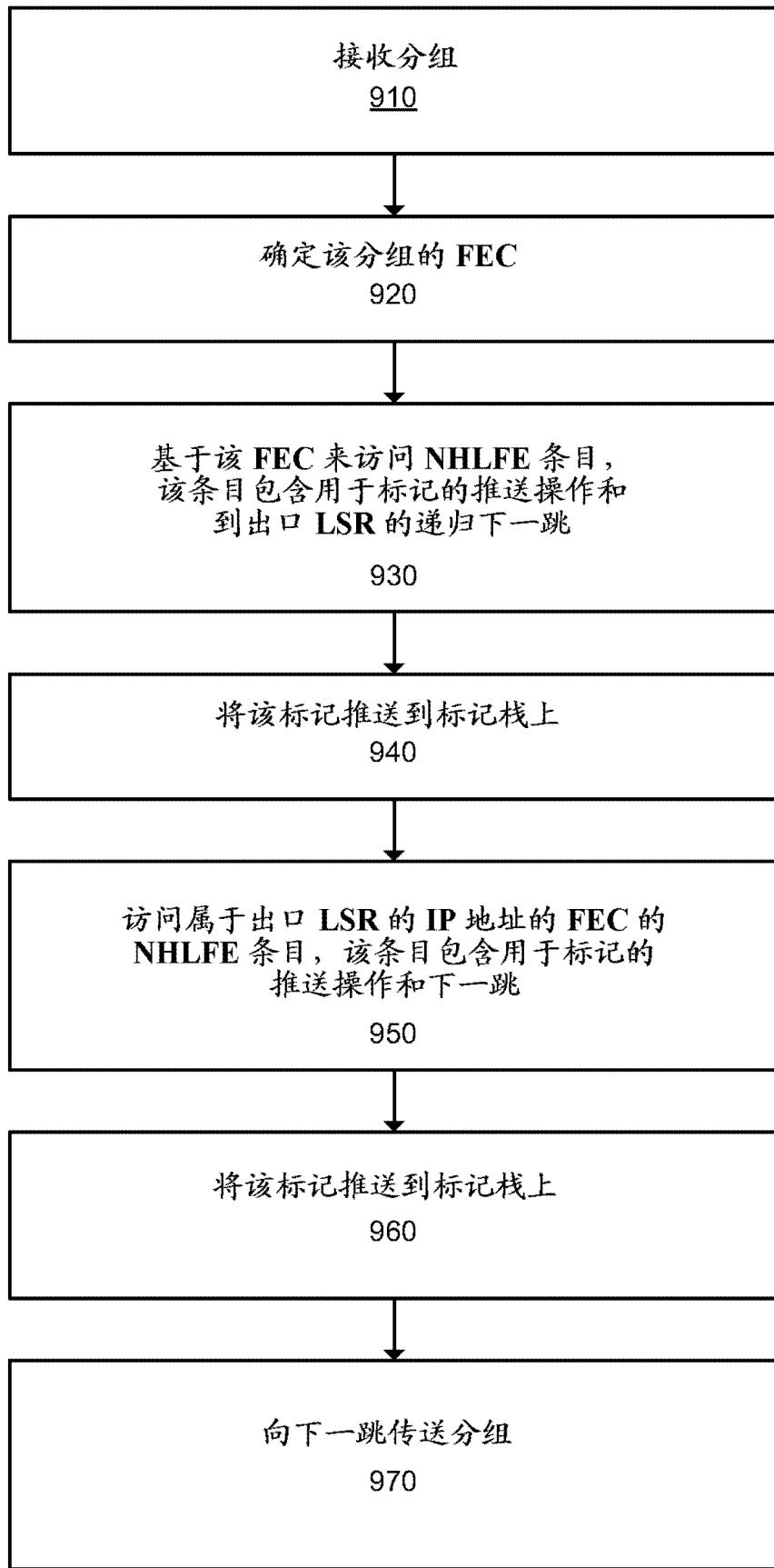
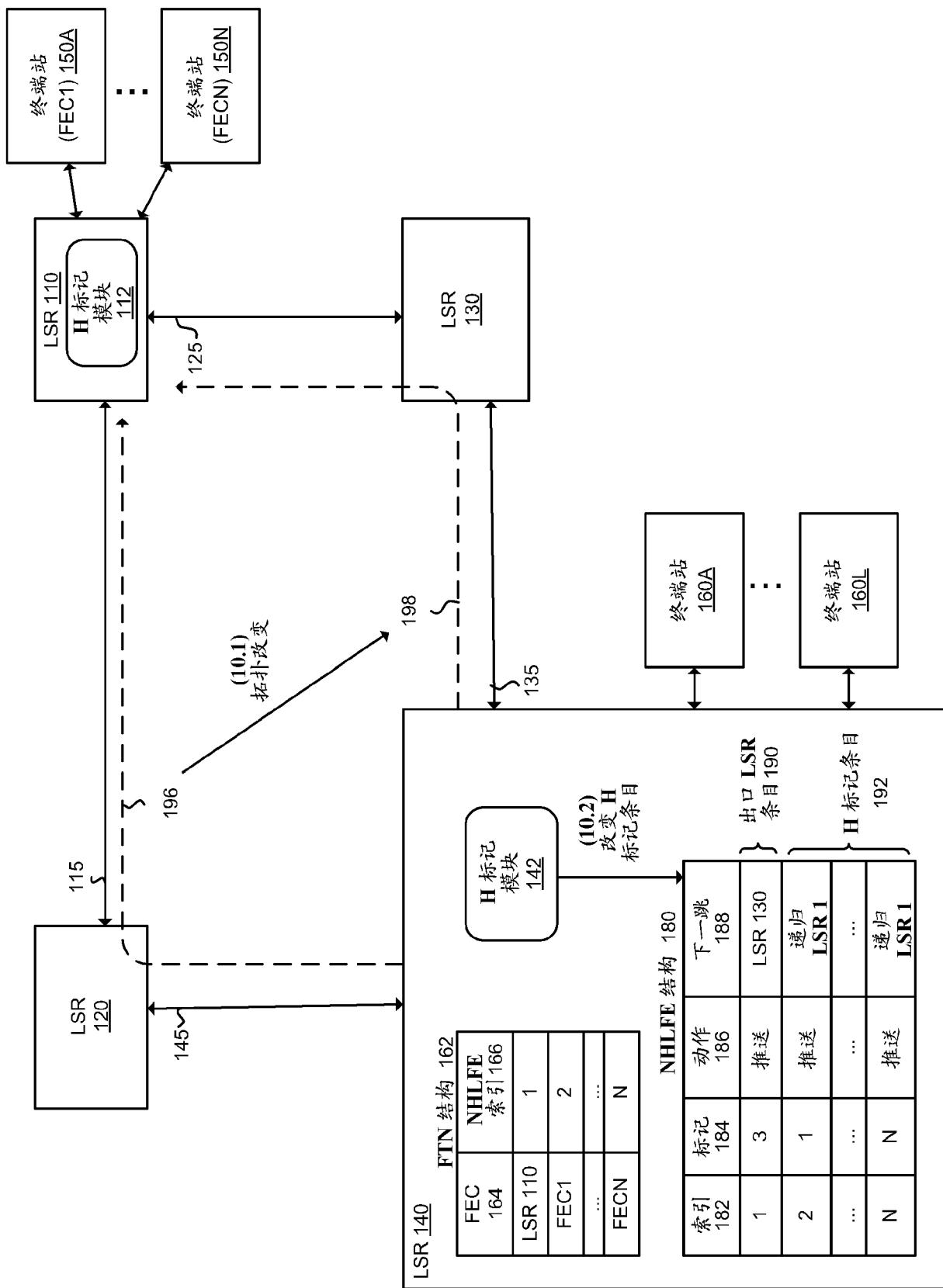


图 9



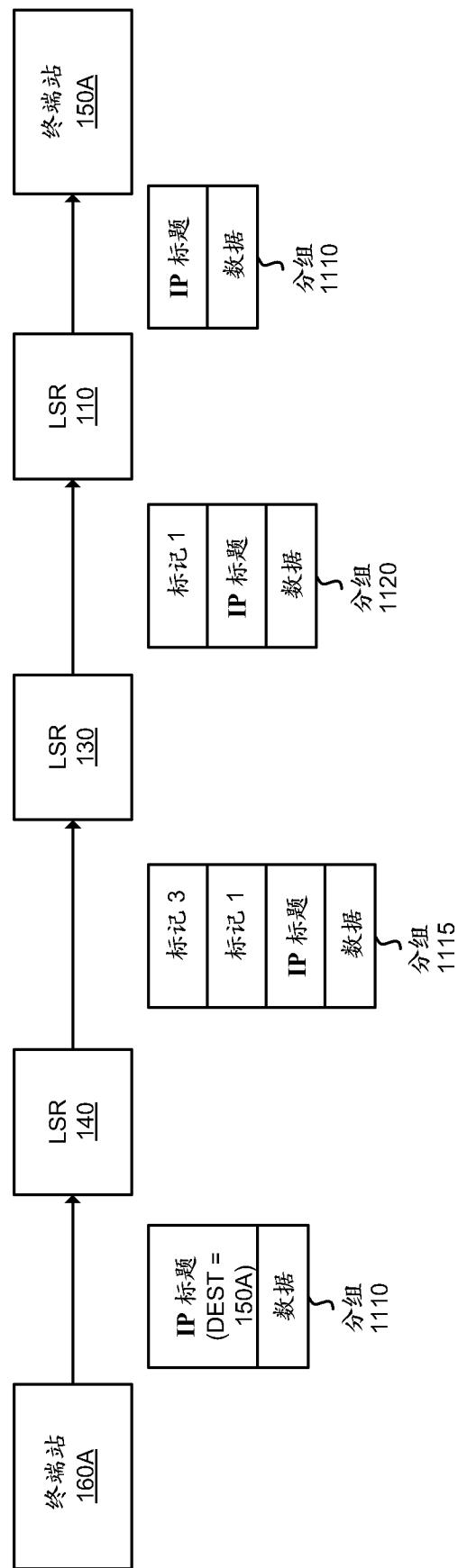


图 11