(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 117957825 A
(43)申请公布日 2024.04.30

(21)申请号 202280061522.5

(22)申请日 2022.09.30

(30)优先权数据
63/250,738 2021.09.30 US

(85)PCT国际申请进入国家阶段日
2024.03.14

(86)PCT国际申请的申请数据
PCT/US2022/045381 2022.09.30

(87)PCT国际申请的公布数据
WO2023/056013 EN 2023.04.06

(71)申请人 华为技术有限公司
地址 518129 广东省深圳市龙岗区坂田华为总部办公楼

(72)发明人 陈怀谟

(51)Int.Cl.
H04L 45/00(2006.01)
H04L 45/02(2006.01)

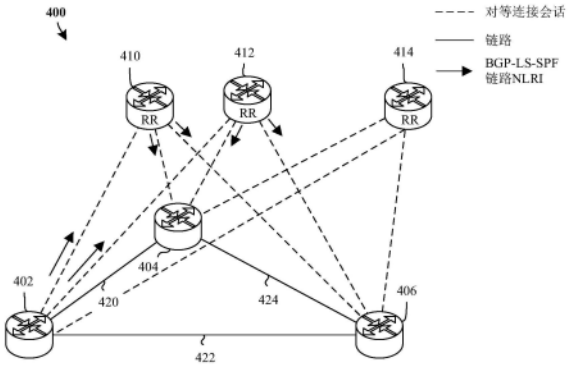
权利要求书6页 说明书20页 附图10页

(54)发明名称

边界网关协议(BGP)-最短路径优先(SPF)泛洪减少

(57)摘要

本发明提供了一种由网络节点实现的用于减少边界网关协议-最短路径优先(Border Gateway Protocol-Shortest Path First,BGP-SPF)域内的泛洪的方法。所述网络节点获取所述BGP-SPF域的泛洪拓扑(flooding topology,FT)。当存在链路变化时,所述网络节点在指示所述链路变化的BGP更新消息中将网络层可达信息(Network Layer Reachability Information,NLRI)发送给所述FT中的与所述网络节点直接连接的网络节点。



CN 117957825 A

1.一种由网络节点实现的用于减少边界网关协议-最短路径优先(Border Gateway Protocol-Shortest Path First,BGP-SPF)域内的泛洪的方法,其特征在于,所述方法包括:

与所述BGP-SPF域内的路由反射器(route-reflector,RR)集合建立外部BGP(external BGP,EBGP)会话以交换路由;

确定与所述网络节点的链路对应的链路变化;

根据确定哪些RR是在所述RR集合中的子集的泛洪行为,通过所述eBGP会话在BGP更新消息中将指示所述链路变化的BGP链路状态SPF(BGP Link-State SPF,BGP-LS-SPF)链路网络层可达信息(Network Layer Reachability Information,NLRI)发送给所述RR集合中的子集。

2.根据权利要求1所述的方法,其特征在于,所述方法还包括:

接收指示所述确定哪些RR是在所述RR集合中的子集的泛洪行为的泛洪行为指令;

在所述网络节点上配置所述泛洪行为。

3.根据权利要求2所述的方法,其特征在于,所述方法还包括:从所述RR集合中的一个RR接收所述泛洪行为指令,其中,所述RR是所述BGP-SPF域内的领导者RR。

4.根据权利要求2或3所述的方法,其特征在于,所述方法还包括:

接收在节点泛洪类型长度值(Type-Length-Value,TLV)中编码的所述泛洪行为指令;

对所述节点泛洪TLV进行解码以确定所述泛洪行为。

5.根据权利要求1至4中任一项所述的方法,其特征在于,所述方法还包括:

根据所述泛洪行为指令,将所述网络节点分配到所述BGP-SPF域内的一组网络节点中;

将指示所述链路变化的所述BGP-LS-SPF链路NLRI发送给所述RR集合中的为所述一组网络节点指定的所述子集。

6.一种由路由反射器(route reflector,RR)实现的用于减少边界网关协议-最短路径优先(Border Gateway Protocol-Shortest Path First,BGP-SPF)域内的泛洪的方法,其特征在于,所述方法包括:

与所述BGP-SPF域内的网络节点建立外部BGP(external BGP,EBGP)会话以交换路由;

为所述网络节点配置泛洪行为;

在BGP更新消息中将指示所述泛洪行为的节点泛洪类型长度值(Type-Length-Value,TLV)发送给所述网络节点。

7.根据权利要求6所述的方法,其特征在于,所述方法还包括:传输所述RR成为所述BGP-SPF域内的领导者的优先级。

8.根据权利要求7所述的方法,其特征在于,所述方法还包括:

在领导者优先级TLV中对所述RR成为所述BGP-SPF域内的领导者的所述优先级进行编码;

将所述领导者优先级TLV发送给所述网络节点和所述BGP-SPF域内的其它RR。

9.根据权利要求6至8中任一项所述的方法,其特征在于,所述方法还包括:

接收所述BGP-SPF域内的所述其它RR成为领导者的优先级;

确定所述RR的所述优先级相对于所述BGP-SPF域内的所述其它RR的所述优先级是成为所述BGP-SPF域内的领导者的最高优先级;

根据所述确定,将所述RR配置为所述BGP-SPF域内的所述领导者。

10.根据权利要求6至9中任一项所述的方法,其特征在于,所述泛洪行为指示所述网络节点将指示链路变化的信息只发送给所述BGP-SPF域内的特定RR。

11.一种由网络节点实现的用于减少边界网关协议-最短路径优先(Border Gateway Protocol-Shortest Path First,BGP-SPF)域内的泛洪的方法,其特征在于,所述方法包括:

获取所述BGP-SPF域的泛洪拓扑(flooding topology,FT),其中,所述FT是连接所述BGP-SPF域的真实网络拓扑(network topology,RT)中的所有节点的子网络拓扑;

确定与所述网络节点的链路对应的链路变化;

在指示所述链路变化的BGP更新消息中将网络层可达信息(Network Layer Reachability Information,NLRI)发送给所述FT中的与所述网络节点直接连接的网络节点。

12.根据权利要求11所述的方法,其特征在于,所述方法还包括:从所述BGP-SPF域内的领导者节点获取所述FT。

13.根据权利要求12所述的方法,其特征在于,所述方法还包括:

从所述领导者节点接收节点索引映射;

使用所述节点索引映射对所述FT的编码进行解码,以得到所述FT。

14.根据权利要求12或13所述的方法,其特征在于,所述方法还包括:

从所述领导者节点接收对所述FT的更新;

根据所述更新修改所述FT。

15.根据权利要求14所述的方法,其特征在于,所述更新包括在路径TLV中编码的新连接,所述路径TLV在BGP更新消息中的多协议可达链路网络层可达信息(MP_REACH_NLRI)路径属性中编码。

16.根据权利要求14或15所述的方法,其特征在于,所述更新包括在路径TLV中编码的移除连接,所述路径TLV在BGP更新消息中的多协议不可达链路网络层可达信息(MP_REACH_NLRI)路径属性中编码。

17.一种由网络节点实现的用于计算泛洪拓扑(flooding topology,FT)的方法,其特征在于,所述方法包括:

从网络中选择节点R0;

使用所述节点R0的节点元素初始化所述FT,其中,所述节点元素包括节点、节点连接数(D)、前跳(previous hops,PHs)列表;

初始化候选队列(candidate queue,Cq),其中,所述Cq包括所述网络中的直接连接到所述节点R0的每个节点的节点元素;

实现所述方法的第一循环,包括:

从所述Cq中移除第一节点的节点元素,并且将所述节点元素附加到所述FT中,其中,所述第一节点的D小于最大连接数(MaxD);

确定所述FT是否包括所述网络中的所有节点;

当所述FT不包括所述网络中的所有节点时,识别所述网络中的连接到所述第一节点且不在所述FT中的一组节点,将所述一组节点中的不在所述Cq中的节点附加到所述Cq中,将

所述第一节点附加到所述一组节点中的在所述Cq中的节点的节点元素的前跳(previous hops,PHs)列表中,并且重复所述第一循环;

当所述FT包括所述网络中的所有节点时,终止所述第一循环;

将链路添加到所述FT中的D等于1的任何节点。

18.根据权利要求17所述的方法,其特征在于,所述方法的所述第一循环在所述Cq不为空时确定所述FT不包括所述网络中的所有节点,在所述Cq为空时确定所述FT包括所述网络中的所有节点。

19.根据权利要求17或18所述的方法,其特征在于,所述将链路添加到所述FT中的D等于1的任何节点包括实现第二循环,所述第二循环包括:

识别所述FT中的单一链路节点,其中,所述单一链路节点在所述FT中的D等于1;

当所述FT中不存在单一链路节点时,终止所述第二循环;

否则,识别所述网络中的连接到所述单一链路节点的一组链路,其中,所述一组链路不包括所述FT中的所述单一链路节点的现有链路;

识别连接到所述一组链路的一组远端节点;

识别所述一组远端节点中的能够支持中转的一组具有中转能力的远端节点;

识别所述一组链路中的连接到所述一组具有中转能力的远端节点中的具有最小D和最小节点标识(identifier,ID)的一个具有中转能力的远端节点的第二链路;

当所述一组具有中转能力的远端节点中不存在具有中转能力的远端节点时,识别附接在所述一组链路中的连接到所述一组远端节点中的具有最小D和最小节点ID的一个远端节点的所述第二链路;

将所述第二链路添加到所述FT中;

将所述FT中的所述单一链路节点的D增加1;

将所述FT中的所述具有中转能力的远端节点或不存在具有中转能力的远端节点时的所述远端节点的D增加1;

重复所述第二循环。

20.根据权利要求17至19中任一项所述的方法,其特征在于,所述节点R0具有所述网络中的最低节点标识(identifier,ID)。

21.根据权利要求17至20中任一项所述的方法,其特征在于,所述Cq使用从最低节点ID到最高节点ID排序的节点进行初始化。

22.根据权利要求17至21中任一项所述的方法,其特征在于,附加到所述Cq中的节点从最低节点ID到最高节点ID进行排序。

23.一种在边界网关协议-最短路径优先(Border Gateway Protocol-Shortest Path First,BGP-SPF)域内的网络节点,其特征在于,所述网络节点至少包括处理器和存储指令的存储器,其中,所述指令在由所述处理器执行时,使得所述网络节点执行以下操作:

与所述BGP-SPF域内的路由反射器(route-reflector,RR)集合建立外部BGP(external BGP,EBGP)会话以交换路由;

确定与所述网络节点的链路对应的链路变化;

根据确定哪些RR是在所述RR集合中的子集的泛洪行为,通过所述eBGP会话在BGP更新消息中将指示所述链路变化的BGP链路状态SPF(BGP Link-State SPF,BGP-LS-SPF)链路网

络层可达信息(Network Layer Reachability Information,NLRI)发送给所述RR集合中的子集。

24.根据权利要求23所述的网络节点,其特征在于,所述指令在由所述处理器执行时,还使得所述网络节点执行以下操作:

接收指示所述确定哪些RR是在所述RR集合中的子集的泛洪行为的泛洪行为指令;

在所述网络节点上配置所述泛洪行为。

25.根据权利要求24所述的网络节点,其特征在于,所述指令在由所述处理器执行时,还使得所述网络节点执行以下操作:从所述RR集合中的一个RR接收所述泛洪行为指令,其中,所述RR是所述BGP-SPF域内的领导者RR。

26.根据权利要求24或25所述的网络节点,其特征在于,所述指令在由所述处理器执行时,还使得所述网络节点执行以下操作:

接收在节点泛洪类型长度值(Type-Length-Value,TLV)中编码的所述泛洪行为指令;

对所述节点泛洪TLV进行解码以确定所述泛洪行为。

27.根据权利要求23至26中任一项所述的网络节点,其特征在于,所述指令在由所述处理器执行时,还使得所述网络节点执行以下操作:

根据所述泛洪行为指令,将所述网络节点分配到所述BGP-SPF域内的一组网络节点中;

将指示所述链路变化的所述BGP-LS-SPF链路NLRI发送给所述RR集合中的为所述一组网络节点指定的所述子集。

28.一种在边界网关协议-最短路径优先(Border Gateway Protocol-Shortest Path First,BGP-SPF)域内的路由反射器(route-reflector,RR),其特征在于,所述RR至少包括处理器和存储指令的存储器,其中,所述指令在由所述处理器执行时,使得所述RR执行以下操作:

与所述BGP-SPF域内的网络节点建立外部BGP(external BGP,EBGP)会话以交换路由;

为所述网络节点配置泛洪行为;

在节点泛洪类型长度值(Type-Length-Value,TLV)中对所述泛洪行为进行编码;

对包括所述节点泛洪TLV的BGP更新消息进行编码;

将所述BGP更新消息发送给所述网络节点。

29.根据权利要求28所述的RR,其特征在于,所述指令在由所述处理器执行时,还使得所述RR执行以下操作:传输所述RR成为所述BGP-SPF域内的领导者的优先级。

30.根据权利要求28或29所述的RR,其特征在于,所述指令在由所述处理器执行时,还使得所述RR执行以下操作:

在领导者优先级TLV中对所述RR成为所述BGP-SPF域内的领导者的所述优先级进行编码;

对包括所述领导者优先级TLV的BGP更新消息进行编码;

将所述BGP更新消息发送给所述网络节点和所述BGP-SPF域内的其它RR。

31.根据权利要求28至30中任一项所述的RR,其特征在于,所述指令在由所述处理器执行时,还使得所述RR执行以下操作:

接收所述BGP-SPF域内的所述其它RR成为领导者的优先级;

根据所述其它RR的所述优先级,确定所述RR的所述优先级是成为所述BGP-SPF域内的

领导者的最高优先级;

将所述RR配置为所述BGP-SPF域内的所述领导者。

32.根据权利要求28至31中任一项所述的RR,其特征在于,所述泛洪行为指示所述网络节点将指示链路变化的信息只发送给所述BGP-SPF域内的特定RR。

33.一种在边界网关协议-最短路径优先(Border Gateway Protocol-Shortest Path First,BGP-SPF)域内的网络节点,其特征在于,所述网络节点至少包括处理器和存储指令的存储器,其中,所述指令在由所述处理器执行时,使得所述网络节点执行以下操作:

获取所述BGP-SPF域的泛洪拓扑(flooding topology,FT),其中,所述FT是连接所述BGP-SPF域的真实网络拓扑(network topology,RT)中的所有节点的子网络拓扑;

确定与所述网络节点的链路对应的链路变化;

在指示所述链路变化的BGP更新消息中将网络层可达信息(Network Layer Reachability Information,NLRI)发送给所述FT中的与所述网络节点直接连接的网络节点。

34.根据权利要求33所述的网络节点,其特征在于,所述指令在由所述处理器执行时,还使得所述网络节点执行以下操作:从所述BGP-SPF域内的领导者节点获取所述FT。

35.根据权利要求34所述的网络节点,其特征在于,所述指令在由所述处理器执行时,还使得所述网络节点执行以下操作:

从所述领导者节点接收节点索引映射;

使用所述节点索引映射对所述FT的编码进行解码,以得到所述FT。

36.根据权利要求34或35所述的网络节点,其特征在于,所述指令在由所述处理器执行时,还使得所述网络节点执行以下操作:

从所述领导者节点接收对所述FT的更新;

根据所述更新修改所述FT。

37.根据权利要求36所述的网络节点,其特征在于,所述更新包括在路径类型长度值(Type-Length-Value,TLV)中编码的新连接,所述路径TLV在BGP更新消息中的多协议可达链路网络层可达信息(MP_REACH_NLRI)路径属性中编码。

38.根据权利要求36或37所述的网络节点,其特征在于,所述更新包括在路径TLV中编码的移除连接,所述路径TLV在BGP更新消息中的多协议不可达链路网络层可达信息(MP_REACH_NLRI)路径属性中编码。

39.一种网络节点,其特征在于,所述网络节点用于计算泛洪拓扑(flooding topology,FT),所述网络节点至少包括处理器和存储指令的存储器,其中,所述指令在由所述处理器执行时,使得所述网络节点执行以下操作:

从网络中选择节点R0;

使用所述节点R0的节点元素初始化所述FT,其中,所述节点元素包括节点、节点连接数(D)、前跳(previous hops,PHs)列表;

初始化候选队列(candidate queue,Cq),其中,所述Cq包括所述网络中直接连接到所述节点R0的每个节点的节点元素;

实现所述网络节点的第一循环,包括:

从所述Cq中移除第一节点的节点元素,并且将所述节点元素附加到所述FT中,其中,所

述第一节点的D小于最大连接数(MaxD)；

确定所述FT是否包括所述网络中的所有节点；

当所述FT不包括所述网络中的所有节点时,识别所述网络中连接到所述第一节点且不在所述FT中的一组节点,将所述一组节点中的不在所述Cq中的节点附加到所述Cq中,将所述第一节点附加到所述一组节点中的在所述Cq中的节点的节点元素的前跳(previous hops,PHs)列表中,并且重复所述第一循环；

当所述FT包括所述网络中的所有节点时,终止所述第一循环；

将链路添加到所述FT中的D等于1的任何节点。

40.根据权利要求39所述的网络节点,其特征在于,所述网络节点的所述第一循环在所述Cq不为空时确定所述FT不包括所述网络中的所有节点;在所述Cq为空时确定所述FT包括所述网络中的所有节点。

41.根据权利要求39或40所述的网络节点,其特征在于,所述将链路添加到所述FT中的D等于1的任何节点的指令包括实现第二循环的指令,所述实现第二循环的指令使得所述网络节点执行以下操作：

识别所述FT中的单一链路节点,其中,所述单一链路节点在所述FT中的D等于1；

当所述FT中不存在单一链路节点时,终止所述第二循环；

否则,识别所述网络中连接到所述单一链路节点的一组链路,其中,所述一组链路不包括所述FT中的所述单一链路节点的现有链路；

识别连接到所述一组链路的一组远端节点；

识别所述一组远端节点中的能够支持中转的一组具有中转能力的远端节点；

识别所述一组链路中的连接到所述一组具有中转能力的远端节点中的具有最小D和最小节点标识(identifier,ID)的一个具有中转能力的远端节点的第二链路；

当所述一组具有中转能力的远端节点中不存在具有中转能力的远端节点时,识别附接在所述一组链路中的连接到所述一组远端节点中的具有最小D和最小节点ID的一个远端节点的所述第二链路；

将所述第二链路添加到所述FT中；

将所述FT中的所述单一链路节点的D增加1；

将所述FT中的所述具有中转能力的远端节点或不存在具有中转能力的远端节点时的所述远端节点的D增加1；

重复所述第二循环。

42.根据权利要求39至41中任一项所述的网络节点,其特征在于,所述节点RO具有所述网络中的最低节点标识(identifier,ID)。

43.根据权利要求39至42中任一项所述的网络节点,其特征在于,所述Cq使用从最低节点ID到最高节点ID排序的节点进行初始化。

44.根据权利要求39至43中任一项所述的网络节点,其特征在于,附加到所述Cq中的节点从最低节点ID到最高节点ID进行排序。

边界网关协议(BGP)-最短路径优先(SPF)泛洪减少

[0001] 相关申请的交叉引用

[0002] 本专利申请要求于2021年9月30日递交的第63/250,738号美国临时专利申请的权益,其指导和公开内容以引用的方式全部并入本文中。

技术领域

[0003] 本发明大体上涉及网络通信,具体涉及用于边界网关协议(Border Gateway Protocol,BGP)-最短路径优先(Shortest Path First,SPF)泛洪减少的各种系统和方法。

背景技术

[0004] 边界网关协议(Border Gateway Protocol,BGP)是一种自治系统(Autonomous System,AS)间路由协议,它有助于自治系统(Autonomous System,AS)之间的路由信息交换,从而实现AS之间的数据包路由。AS是由互联网服务提供商(Internet service provider,ISP)、大型企业技术公司、大学或政府机构等单一管理实体管理的一组路由器或一组连接的互联网协议(Internet Protocol,IP)网络。具体地,AS是一个或多个关联的IP前缀的集合,称为AS的IP地址空间,具有明确定义的路由策略,该策略控制AS如何与其它AS交换路由信息。

[0005] 每个AS使用BGP通告自己负责的IP地址以及与之连接的其它AS。BGP路由器从世界各地的AS获取所有这些信息,并且将这些信息放入到称为路由表的数据库中,以确定AS之间的最快路径。数据包穿越互联网时,通过从一个AS跳转到另一个AS,直到它们到达包括数据包指定的目的地IP地址的AS。例如,当一个数据包到达一个AS时,BGP路由器会参考它们的路由表来确定该数据包下一步应该到达的AS。包括目的地IP地址的AS内的路由器将该数据包发送给与该目的地IP地址对应的网络设备。

[0006] 世界上有如此多的AS,因此BGP路由器需要不断更新它们的路由表。随着网络下线,新网络会上线,AS也会扩大或缩小它们的IP地址空间。所有更新后的信息都必须通过BGP通告,以便BGP路由器能够调整它们的路由表。

发明内容

[0007] 第一方面涉及一种由边界网关协议-最短路径优先(Border Gateway Protocol-Shortest Path First,BGP-SPF)域内的网络节点实现的减少BGP-SPF域内的泛洪的方法。所述方法包括:与所述BGP-SPF域内的路由反射器(route-reflector,RR)集合建立外部BGP(external BGP,EBGP)会话以交换路由;确定与所述网络节点的链路对应的链路变化;根据确定哪些RR是在所述RR集合中的子集的泛洪行为,通过所述eBGP会话在BGP更新消息中将指示所述链路变化的BGP链路状态SPF(BGP Link-State SPF,BGP-LS-SPF)链路网络层可达信息(Network Layer Reachability Information,NLRI)发送给所述RR集合中的子集。

[0008] 可选地,根据所述第一方面中的任一个,在第一种实现方式中,所述方法包括:在所述BGP-LS-SPF链路NLRI中对指示所述链路变化的所述信息进行编码;对包括所述BGP-

LS-SPF链路NLRI的所述BGP更新消息进行编码;将所述BGP更新消息发送给所述RR集合中的子集。

[0009] 可选地,根据所述第一方面或其任一种实现方式中的任一个,在第二种实现方式中,所述方法还包括:接收指示所述确定哪些RR是在所述RR集合中的子集的泛洪行为的泛洪行为指令;在所述网络节点上配置所述泛洪行为。

[0010] 可选地,根据所述第一方面或其任一种实现方式中的任一个,在第三种实现方式中,所述方法还包括:从所述RR集合中的一个RR接收所述泛洪行为指令,其中,所述RR是所述BGP-SPF域内的领导者RR。

[0011] 可选地,根据所述第一方面或其任一种实现方式中的任一个,在第四种实现方式中,所述方法还包括:接收在节点泛洪类型长度值(Type-Length-Value,TLV)中编码的所述泛洪行为指令;对所述节点泛洪TLV进行解码以确定所述泛洪行为。

[0012] 可选地,根据所述第一方面或其任一种实现方式中的任一个,在第五种实现方式中,所述方法还包括:根据所述泛洪行为指令,将所述网络节点分配到所述BGP-SPF域内的一组网络节点中;将指示所述链路变化的所述信息发送给所述RR集合中的为所述一组网络节点指定的所述子集。

[0013] 第二方面涉及一种由BGP-SPF域内的RR实现的减少所述BGP-SPF域内的泛洪的方法。所述方法包括:与所述BGP-SPF域内的网络节点建立外部BGP(external BGP,EBGP)会话以交换路由;为所述网络节点配置泛洪行为;在BGP更新消息中将指示所述泛洪行为的节点泛洪类型长度值(Type-Length-Value,TLV)发送给所述网络节点。

[0014] 可选地,根据所述第二方面中的任一个,在第一种实现方式中,所述方法包括:在所述节点泛洪TLV中对所述泛洪行为进行编码;对包括所述节点泛洪TLV的所述BGP更新消息进行编码;将所述BGP更新消息发送给所述网络节点。

[0015] 可选地,根据所述第二方面或其任一种实现方式中的任一个,在第二种实现方式中,所述方法还包括:传输所述RR成为所述BGP-SPF路由域内的领导者的优先级。

[0016] 可选地,根据所述第二方面或其任一种实现方式中的任一个,在第三种实现方式中,所述方法还包括:在领导者优先级TLV中对所述RR成为所述BGP-SPF路由域内的领导者的所述优先级进行编码;对包括所述领导者优先级TLV的所述BGP更新消息进行编码;将所述BGP更新消息发送给所述网络节点和所述BGP-SPF路由域内的其它RR。

[0017] 可选地,根据所述第二方面或其任一种实现方式中的任一个,在第四种实现方式中,所述方法还包括:接收所述BGP-SPF路由域内的所述其它RR成为领导者的优先级;根据所述其它RR的所述优先级,确定所述RR的所述优先级是成为所述BGP-SPF路由域内的领导者的最高优先级;将所述RR配置为所述BGP-SPF路由域内的所述领导者。

[0018] 可选地,根据所述第二方面或其任一种实现方式中的任一个,在第五种实现方式中,所述泛洪行为指示所述网络节点将指示链路变化的信息只发送给所述BGP-SPF路由域内的特定RR。

[0019] 第三方面涉及一种由BGP-SPF域内的网络节点实现的减少所述BGP-SPF域内的泛洪的方法。所述方法包括:获取所述BGP-SPF域的泛洪拓扑(flooding topology,FT),其中,所述FT是连接所述BGP-SPF域的真实网络拓扑(network topology,RT)中的所有节点的子网络拓扑;确定与所述网络节点的链路对应的链路变化;在指示所述链路变化的BGP更新消

息中将网络层可达信息(Network Layer Reachability Information,NLRI)发送给所述FT中的直接连接到所述网络节点的网络节点。

[0020] 可选地,根据所述第三方面中的任一个,在第一种实现方式中,所述方法还包括:从所述节点连接BGP-SPF域内的领导者节点获取所述FT。

[0021] 可选地,根据所述第三方面或其任一种实现方式中的任一个,在第二种实现方式中,所述方法还包括:从所述领导者节点接收节点索引映射;使用所述节点索引映射对所述FT的编码进行解码,以得到所述FT。

[0022] 可选地,根据所述第三方面或其任一种实现方式中的任一个,在第三种实现方式中,所述方法还包括:从所述领导者节点接收对所述FT的更新;根据所述更新修改所述FT。

[0023] 可选地,根据所述第三方面或其任一种实现方式中的任一个,在第四种实现方式中,所述更新包括在路径TLV中编码的新连接,所述路径TLV在BGP更新消息中的多协议可达链路网络层可达信息(MP_REACH_NLRI)路径属性中编码。

[0024] 可选地,根据所述第三方面或其任一种实现方式中的任一个,在第五种实现方式中,所述更新包括在路径TLV中编码的移除连接,所述路径TLV在BGP更新消息中的多协议不可达链路网络层可达信息(MP_REACH_NLRI)路径属性中编码。

[0025] 第四方面涉及一种由网络设备实现的用于计算泛洪拓扑(flooding topology,FT)的方法。所述方法包括:从网络中选择节点RO;使用所述节点RO的节点元素初始化所述FT,其中,所述节点元素包括节点、节点连接数(D)、前跳(previous hops,PHs)列表;初始化候选队列(Cq),其中,所述Cq包括所述网络中的直接连接到所述节点RO的每个节点的节点元素;实现所述方法的第一循环,包括:从所述Cq中移除第一节点的节点元素,并且将所述节点元素附加到所述FT中,其中,所述第一节点的D小于最大连接数(MaxD);确定所述FT是否包括所述网络中的所有节点;当所述FT不包括所述网络中的所有节点时,识别所述网络中的连接到所述第一节点且不在所述FT中的一组节点,将所述一组节点中的不在所述Cq中的节点附加到所述Cq中,将所述第一节点附加到所述一组节点中的在所述Cq中的节点的节点元素的前跳(previous hops,PHs)列表中,并且重复所述第一循环;当所述FT包括所述网络中的所有节点时,终止所述第一循环;将链路添加到所述FT中的D等于1的任何节点。

[0026] 可选地,根据第四方面中的任一个,在第一种实现方式中,所述方法的所述第一循环在所述Cq不为空时确定所述FT不包括所述网络中的所有节点,在所述Cq为空时确定所述FT包括所述网络中的所有节点。

[0027] 可选地,根据所述第四方面或其任一种实现方式中的任一个,在第二种实现方式中,所述将链路添加到所述FT中的在所述FT中的D等于1的任何节点包括实现第二循环,所述第二循环包括:识别所述FT中的单一链路节点,其中,所述单一链路节点在所述FT中的D等于1;当所述FT中不存在单一链路节点时,终止所述第二循环;否则,识别所述网络中的连接到所述单一链路节点的一组链路,其中,所述一组链路不包括所述FT中的所述单一链路节点的现有链路;识别连接到所述一组链路的一组远端节点;识别所述一组远端节点中的能够支持中转的一组具有中转能力的远端节点;识别所述一组链路中的连接到所述一组具有中转能力的远端节点中的具有最小D和最小节点标识(identifier,ID)的一个具有中转能力的远端节点的第二链路;当所述一组具有中转能力的远端节点中不存在具有中转能力的远端节点时,识别附接在所述一组链路中的连接到所述一组远端节点中的具有最小D和

最小节点ID的一个远端节点的所述第二链路;将所述第二链路添加到所述FT中;将所述FT中的所述单一链路节点的D增加1;将所述FT中的所述具有中转能力的远端节点或不存在所述具有中转能力的远端节点时的所述远端节点的D增加1;重复所述第二循环。

[0028] 可选地,根据所述第四方面或其任一种实现方式中的任一个,在第三种实现方式中,所述节点R0具有所述网络中的最低节点标识(identifier,ID)。

[0029] 可选地,根据所述第四方面或其任一种实现方式中的任一个,在第四种实现方式中,所述Cq使用从最低节点ID到最高节点ID排序的节点进行初始化。

[0030] 可选地,根据所述第四方面或其任一种实现方式中的任一个,在第五种实现方式中,附加到所述Cq中的节点从最低节点ID到最高节点ID进行排序。

[0031] 第五方面涉及一种网络节点。所述网络节点至少包括处理器和存储指令的存储器,其中,所述指令在由所述处理器执行时,使得所述网络节点执行根据上述方面或其实现方式中的任一个所述的方法。

[0032] 第六方面涉及一种计算机程序产品。所述计算机程序产品包括存储在非瞬时性计算机可读介质中的计算机可执行指令,所述指令在由设备中的处理器执行时,使得所述设备执行根据上述方面或其实现方式中的任一个所述的方法。

[0033] 为了清楚起见,任一上述实施例可以与上述其它任何一个或多个实施例组合以创建在本发明范围内的新实施例。

[0034] 根据以下结合附图和权利要求书的具体实施方式,将会更清楚地理解这些和其它特征以及其优点。

附图说明

[0035] 为了更完整地理解本发明,结合附图和具体实施方式,参考以下简要描述,其中,相似的附图标记表示相似的部件。

[0036] 图1是路由反射器(route-reflector,RR)模型下的BGP-SPF泛洪的示意图。

[0037] 图2是节点连接模型下的BGP-SPF泛洪的示意图。

[0038] 图3是本发明一个实施例提供的RR模型下的BGP-SPF泛洪减少的示意图。

[0039] 图4是本发明另一个实施例提供的RR模型下的BGP-SPF泛洪减少的示意图。

[0040] 图5是本发明另一个实施例提供的RR模型下的BGP-SPF泛洪减少的示意图。

[0041] 图6A是本发明一个实施例提供的BGP-SPF泛洪减少节点连接模型的示意图。

[0042] 图6B是用于图6A中的BGP-SPF泛洪减少节点连接模型的FT的示意图。

[0043] 图7是本发明一个实施例提供的FT算法的流程图。

[0044] 图8是本发明一个实施例提供的领导者优先级TLV的示意图。

[0045] 图9是本发明一个实施例提供的节点泛洪TLV的示意图。

[0046] 图10是本发明一个实施例提供的领导者偏好TLV的示意图。

[0047] 图11是本发明一个实施例提供的算法支持TLV的示意图。

[0048] 图12是本发明一个实施例提供的节点ID TLV的示意图。

[0049] 图13是本发明一个实施例提供的路径TLV的示意图。

[0050] 图14是本发明一个实施例提供的用于泛洪的连接(Connection Used for Flooding,CUF)TLV的示意图。

[0051] 图15是本发明一个实施例提供的由BGP-SPF域内的网络节点执行的用于减少所述BGP-SPF域内的泛洪的方法的流程图。

[0052] 图16是本发明一个实施例提供的由BGP-SPF域内的RR执行的用于减少所述BGP-SPF域内的泛洪的方法的流程图。

[0053] 图17是本发明一个实施例的由BGP-SPF域内的网络节点执行的用于减少所述BGP-SPF域内的泛洪的方法的流程图。

[0054] 图18是本发明一个实施例提供的网络节点的示意图。

具体实施方式

[0055] 首先应当理解,尽管下文提供了一个或多个实施例的说明性实现方式,但所公开的系统、计算机程序产品和/或方法可以使用任意数量的技术来实现,无论这些技术是当前已知的还是现有的。本发明绝不限于下文所说明的说明性实现方式、附图和技术,包括本文所说明和描述的示例性设计和实现方式,而是可以在所附权利要求书的范围以及其等效部分的完整范围内修改。

[0056] 本文公开了用于减少BGP-SPF路由域内的泛洪的各种实施例。泛洪是指在一个域内传输(通常也称为发布)链路状态或路由信息的过程,以确保该域内的所有路由器在有限的时间段内最终趋于相同的网络拓扑信息。网络拓扑是网络中的节点和连接或链路的物理和逻辑排列方式。链路是连接两个设备以进行数据传输的通信通道。BGP-SPF路由域或BGP-SPF域是单一管理域下的一组BGP-SPF节点,这些节点使用BGP链路状态SPF(BGP Link-State SPF,BGP-LS-SPF)链路网络层可达信息(Network Layer Reachability Information,NLRI)交换链路状态信息(也称为路由信息)。BGP-LS-SPF链路NLRI是使用分配给BGP-LS-SPF的后续地址族标识(Subsequent Address Family Identifier,SAFI)的链路NLRI。链路NLRI是一种编码格式,用于提供网络层可达信息,例如,描述链路、节点和前缀的信息,包括链路状态信息。BGP-LS-SPF链路NLRI用于确定BGP-SPF路由域内的路由路径。

[0057] BGP-SPF节点是实现BGP-SPF的网络设备(例如,路由器)。BGP-SPF是对BGP的扩展,它利用了BGP链路状态(BGP Link-State,BGP-LS)。BGP-LS是对BGP的扩展,它在2016年3月出版的由H.Gredler等人撰写的标题为“使用BGP北向分发链路状态和流量工程(TE)信息(North-Bound Distribution of Link-State and Traffic Engineering(TE) Information Using BGP)”的互联网工程任务组(Internet Engineering Task Force,IETF)文档请求注释(Request for Comment,RFC)7752中定义。BGP-LS用于使用BGP与外部组件(例如,外部服务器)共享AS的内部网关协议(interior gateway protocol,IGP)链路状态网络拓扑信息。例如,如上所述,BGP-SPF使用BGP-LS-SPF链路NLRI交换路由信息。BGP-LS-SPF链路NLRI重用LS NLRI格式和BGP-LS的地址族标识(Address Family Identifier,AFI)(AFI 16388)。然而,BGP-SPF不使用分配给BGP-LS的SAFI(SAFI 71),而是使用分配给BGP-LS-SPF的SAFI 80来指示链路NLRI是用于BGP-SPF路由计算的BGP-LS-SPF链路NLRI。BGP-LS AFI中使用的所有TLV都适用于BGP-LS-SPF SAFI。另外,与BGP相比,BGP-SPF使用基于SPF算法(例如,计算图形中的某一个节点与图形中的每个其它节点之间的最短路径的Dijkstra算法)的基于SPF的路由选择过程替代现有基于路径向量算法的BGP路由选择决策过程,该路径向量算法在2006年1月出版的由Y.Rekhter等人撰写的标题为“边界网关协议4

(BGP-4)(A Border Gateway Protocol 4(BGP-4))”的IETF文档RFC 4271中有所描述。有关BGP-SPF路由的其它信息,请参见2022年2月15日出版的由K.Patel等人撰写的标题为“BGP链路状态最短路径优先(SPF)路由(BGP Link-State Shortest Path First(SPF) Routing)”的IETF文档(draft-ietf-lsvr-bgp-spf-16)。

[0058] 图1是实现路由反射器(route-reflector,RR)或控制器网络拓扑的BGP-SPF路由域100的现有泛洪过程的示意图。RR或控制器网络拓扑也称为RR对等连接模型、RR模型或稀疏对等连接模型。在RR对等连接模型下,BGP发言者通过外部BGP(external BGP,EBGP)会话仅仅与一个或多个RR或控制器建立对等连接。BGP发言者是使用BGP发布路由信息的路由器。如上所述,在BGP-SPF路由域内,BGP发言者使用BGP-LS-SPF链路NLRI发布路由信息。因此,BGP-SPF路由域内的BGP发言者也可以称为BGP-LS-SPF发言者。例如,在一个实施例中,BGP-SPF路由域内的所有BGP-LS-SPF发言者交换BGP-LS-SPF链路NLRI,运行SPF计算,并且相应地更新它们的路由表。

[0059] RR是一个指定路由器,用于在BGP对等体之间分发/反射路由信息。BGP对等体是用于交换路由信息的BGP发言者。例如,RR可以从第一BGP对等体接收路由信息,并且将这些路由信息反射/分发给其它BGP对等体。通过使用RR,RR对等连接模型可以减少BGP对等连接会话,从而减少从多个BGP对等体接收到相同NLRI的情况。BGP-SPF路由域可以包括任意数量的BGP对等体和RR。例如,BGP-SPF路由域可以包括多个RR,以防止其中一个RR发生故障。

[0060] 作为一个非限制性的示例,在图1中,BGP-SPF路由域100包括3个BGP对等体(节点102、节点104和节点106)和2个RR(RR 110和RR 112)。节点102、节点104和节点106都被配置为BGP-LS-SPF发言者。在所述实施例中,节点102、节点104和节点106都与RR 110和RR 112建立对等连接会话(如图1中的虚线所示)。节点102和节点104之间存在链路120,节点102和节点106之间存在链路122,节点104和节点106之间存在链路124。

[0061] 在所述实施例中,当存在链路变化时,BGP-LS-SPF发言者在BGP更新消息中发送/发布BGP-LS-SPF链路NLRI,以向RR 110和RR 112指示链路变化。BGP更新消息是携带路由信息的消息,用于在BGP邻居之间交换路由信息。链路变化是指影响路由信息的任何变化。例如,当新链路建立/连接时,或者当先前建立的链路撤销/断开时,就会发生链路变化。RR 110和RR 112接收到BGP-LS-SPF链路NLRI之后,在BGP更新消息中将该BGP-LS-SPF链路NLRI发送给与RR 110和RR 112对等连接的其它节点。例如,假设节点102检测到与节点106相连的链路122断开。节点102在BGP更新消息中将指示链路变化的BGP-LS-SPF链路NLRI发送给RR 110和RR 112两者,如箭头所示。RR 110和RR 112从节点102接收到BGP-LS-SPF链路NLRI之后,都将该BGP-LS-SPF链路NLRI发送给节点104和节点106。因此,节点104和节点106都接收相同BGP-LS-SPF链路NLRI的两个副本,一个来自RR 110,另一个来自RR 112。不需要BGP-LS-SPF链路NLRI的第二个副本。因此,在RR对等连接模型下,BGP-SPF路由域的现有泛洪过程不够高效。

[0062] 图2是BGP-SPF路由域200的现有泛洪过程的示意图。BGP-SPF路由域200是节点连接BGP-SPF路由域。本文中提及的节点连接BGP-SPF路由域或节点连接BGP-SPF域是实现本发明描述的节点连接网络拓扑或节点连接模型的BGP-SPF路由域。BGP-SPF路由域200包括4个BGP对等体(节点202、节点204、节点206和节点208),它们都被配置为BGP-LS-SPF发言者。在节点连接模型中,外部BGP(external BGP,EBGP)单跳会话通过将BGP-SPF路由域200内的

节点互连的直接点到点链路进行建立。EBGP用于连接位于不同AS中的节点。一旦建立了单跳BGP会话,并且为该对应的会话交换了BGP-LS-SPF AFI/SAFI能力,则从BGPSPF的角度来看,该链路是连接的。这时,对应的BGP-LS-SPF链路NLRI通过链路上的所有BGP会话发布给BGP-SPF路由域200内的所有节点。如果会话断开,则会撤销对应的BGP-LS-SPF链路NLRI。撤销是通过使用链路上的所有BGP会话将BGP更新消息发布给BGP-SPF路由域200内的所有节点进行的,其中,BGP更新消息包括在多协议不可达NLRI(MP_UNREACH_NLRI)路径属性中编码的BGP-LS-SPF链路NLRI。

[0063] 作为一个非限制性的示例,在图2中,节点202、节点204、节点206和节点208通过六条链路连接。节点202和节点204之间存在两条平行链路,即链路210和链路212,节点202和节点206之间存在链路214,节点202和节点208之间存在链路216,节点204和节点206之间存在链路218,节点206和节点208之间存在链路220。假设除了节点202和节点208之间的链路216上的EBGP会话之外,建立了上述所有链路上的EBGP会话,并且为这些对应的会话交换了BGP-LS-SPF AFI/SAFI能力。当建立了链路216上的EBGP会话,并且为该对应的会话交换了BGP-LS-SPF AFI/SAFI能力时,节点202确定链路216是连接的。然后,节点202通过其EBGP会话(即,节点202和节点204之间的链路210上的EBGP会话、节点202和节点204之间的链路212上的EBGP会话、节点202和节点206之间的链路220上的EBGP会话、节点202和节点208之间的链路216上的EBGP会话)将链路216的BGP-LS-SPF链路NLRI发送给节点204、节点206和节点208(如图2中的箭头所示)。节点204、节点206和节点208从节点202接收到BGP-LS-SPF链路NLRI之后,都将该BGP-LS-SPF链路NLRI发送给与各自建立EBGP会话的其它节点。例如,节点204将BGP-LS-SPF链路NLRI发送给节点206。节点206将BGP-LS-SPF链路NLRI发送给节点204和节点208。节点208将BGP-LS-SPF链路NLRI发送给节点206。

[0064] 类似地,当建立了链路216上的EBGP会话,并且为该对应的会话交换了BGP-LS-SPF AFI/SAFI能力时,节点208认为链路216是连接的,并且通过节点208的EBGP会话(即,节点208和节点206之间的链路220上的会话、节点208和节点202之间的链路216上的会话)将链路216的BGP-LS-SPF链路NLRI发送给节点206和节点202。为简单起见,图2中未示出节点208发起消息的箭头。节点202和节点206从节点208接收到NLRI之后,都将BGP-LS-SPF链路NLRI发送给与各自建立EBGP会话的其它节点(也就是说,节点206将BGP-LS-SPF链路NLRI发送给节点202和节点204,节点202通过与节点204建立的两个并行EBGP会话和与节点206建立的EBGP会话将BGP-LS-SPF链路NLRI发送给节点204和节点206)。

[0065] 如上所示,在BGP-SPF路由域200内,当存在链路变化时,每个节点都会接收到几个冗余BGP-LS-SPF链路NLRI。因此,节点连接BGP-SPF路由域的现有泛洪过程不够高效。

[0066] 类似地,在实现每个节点直接连接到所有其它节点(未示出)的直连节点模型的BGP-SPF路由域(在本文中称为直连节点BGP-SPF路由域)内,会话可以在环回地址之间(即,两跳会话)。因此,即使当节点之间存在多个直接连接时,也会存在单一EBGP会话。只要建立了EBGP会话,并且交换了BGP-LS-SPF AFI/SAFI能力,就会发布BGP-LS-SPF链路NLRI。与节点连接模型相比,由于直连节点BGP-SPF路由域内的每个直连节点之间都会存在EBGP会话,因此只有当节点之间存在平行链路时,EBGP会话才会减少。因此,直连节点BGP-SPF路由域的现有泛洪过程也不够高效。

[0067] 图3是本发明一个实施例提供的实现RR网络拓扑的BGP-SPF路由域300的泛洪减少

过程的示意图。路由域300与图1中的路由域100的相似之处在于,节点302、节点304和节点306都被配置为BGP-LS-SPF发言者,都与RR 310和RR 312建立对等连接会话(如图3中的虚线所示)。节点302和节点304之间存在链路320,节点302和节点306之间存在链路322,节点304和节点306之间存在链路324。

[0068] 根据一个实施例,当一个节点确定存在链路变化时,该节点用于在BGP更新消息中将指示链路变化的BGP-LS-SPF链路NLRI发送给与该节点对等连接的RR中的子集(即一些RR而不是全部RR)。RR或控制器接收到BGP-LS-SPF链路NLRI之后,将该BGP-LS-SPF链路NLRI发送给与RR或控制器对等连接的其它节点/BGP-LS-SPF发言者。

[0069] 在一些实施例中,每个节点的泛洪行为被配置在该节点上。例如,在一些实施例中,每个节点发送BGP-LS-SPF链路NLRI所到的RR的数量可以是用户可配置的(例如,由网络管理员配置)。可替代地,在一些实施例中,每个节点用于将BGP-LS-SPF链路NLRI发送给与该节点对等连接的RR中的一半RR或另外一些比例的RR或与该节点对等连接的仅一个RR。在一些实施例中,每个节点发送BGP-LS-SPF链路NLRI所到的特定RR可以是用户可配置的或者可以是随机选择的。例如,每个节点可以用于随机选择与该节点对等连接的RR中的一半RR,以发送指示链路变化的BGP-LS-SPF链路NLRI。此外,在包括图4和图5中描述的实施例的各种实施例中,在RR模型下,每个节点的泛洪行为被配置在领导者RR等RR或控制器上,领导者RR(例如,使用图9中描述的节点泛洪TLV)将行为发布给网络中或BGP-SPF路由域内的其它RR和每个节点。在一些实施例中,特定RR可以根据该RR成为领导者的优先级选为领导者RR。在一个实施例中,RR成为领导者的优先级由网络管理员分配。在一个实施例中,RR将它们成为领导者的优先级发布给BGP-SPF路由域内的其它RR和网络节点。在一个实施例中,具有成为BGP-SPF路由域内的领导者的最高发布优先级的RR会成为领导者RR。也就是说,如果一个RR将自己发布的优先级与其它RR发布的优先级进行比较,并且确定自己发布的优先级最高,则该RR将自己配置为领导者。类似地,BGP-SPF路由域内的节点在从RR接收泛洪行为指示时,可以根据RR发布的优先级确定一个RR是否是领导者RR。在一些实施例中,一个RR成为BGP-SPF路由域内的领导者的优先级由该RR(例如,使用图8中描述的领导者优先级TLV)发布。在一个实施例中,当具有成为领导者的相同最高优先级的RR不止一个时,将具有最高节点标识(identifier,ID)和最高优先级的RR选为BGP-SPF路由域内的领导者RR。如上所述,领导者RR这时可以指示BGP-SPF路由域内的节点将指示链路变化的BGP-LS-SPF链路NLRI发送到何处(例如,将BGP-LS-SPF链路NLRI发送给BGP-SPF路由域内的哪个或哪些RR)。在一些实施例中,领导者RR可以提供关于将BGP-LS-SPF链路NLRI发送给BGP-SPF路由域内的不同节点的不同指令(例如,可以指示一些节点将这些节点的链路状态泛洪到特定RR,并且可以指示其它节点将这些其它节点的链路状态泛洪到不同的RR或多个RR)。

[0070] 作为一个非限制性的示例,在图3中,假设节点302发现与节点306相连的链路322断开。节点302在BGP更新消息中将指示链路变化的BGP-LS-SPF链路NLRI只发送给RR 310,如箭头所示。RR 310从节点302接收到BGP-LS-SPF链路NLRI之后,将该BGP-LS-SPF链路NLRI发送给节点304和节点306。因此,节点304和节点306都只接收到BGP-LS-SPF链路NLRI的一个副本,而不会接收到相同BGP-LS-SPF链路NLRI的任何冗余副本。因此,与图1中描述的现有泛洪过程相比,图3中的泛洪减少过程将BGP-SPF路由域内的泛洪量减少了一半。

[0071] 图4是本发明一个实施例提供的实现RR网络拓扑的BGP-SPF路由域400的泛洪减少

过程的示意图。路由域400包括节点402、节点404和节点406,它们都被配置为BGP-LS-SPF发言者。节点402、节点404和节点406都与RR 410、RR 412和RR 414建立对等连接会话(如图4中的虚线所示)。节点402和节点404之间存在链路420,节点402和节点406之间存在链路422,节点404和节点406之间存在链路424。

[0072] 在本实施例中,路由域400内的每个节点用于在BGP更新消息中将指示链路变化的BGP-LS-SPF链路NLRI发送给路由域400内的两个相同RR,以在其中一个RR发生故障时提供冗余。例如,节点402、节点404和节点406都用于在BGP更新消息中将指示链路变化的BGP-LS-SPF链路NLRI发送给RR 410和RR 412。例如,假设节点402发现与节点406相连的链路422断开。节点402在BGP更新消息中将指示链路变化的BGP-LS-SPF链路NLRI发送给RR 410和RR 412,如箭头所示。RR 410和RR 412从节点402接收到BGP-LS-SPF链路NLRI之后,将该BGP-LS-SPF链路NLRI发送给节点404和节点406。因此,节点404和节点406都接收到BGP-LS-SPF链路NLRI的两个副本,一个作为冗余备份。即使在本实施例中,节点接收到BGP-LS-SPF链路NLRI的冗余副本,当与图1中描述的现有泛洪过程相比时,图4中的泛洪减少过程将BGP-SPF路由域内的泛洪量减少了三分之一。

[0073] 图5是本发明一个实施例提供的实现RR网络拓扑的BGP-SPF路由域500的泛洪减少过程的示意图。路由域500与图1中的路由域100相似。路由域500包括节点502、节点504和节点506,它们都被配置为BGP-LS-SPF发言者。节点502、节点504和节点506都与RR 510和RR 512建立对等连接会话(如图5中的虚线所示)。节点502和节点504之间存在链路520,节点502和节点506之间存在链路522,节点504和节点506之间存在链路524。

[0074] 在本实施例中,可以将路由域500内的节点平均分成或尽量平均分成多个组。组的数量可以等于路由域内的RR的数量(即,提供了一种最佳负载均衡方法,其中,每组节点具有与每个其它组相同的节点数量,工作负载在路由域内的RR之间均衡)。可替代地,组的数量可以小于路由域内的RR的数量(也就是说,并不是路由域内的每个RR都用于反射BGP-LS-SPF链路NLRI)。可以随机或者根据一个或多个因素(例如,节点和RR之间的距离、节点标识(identifier,ID)或其它因素)选择哪些节点属于哪个组。例如,在一种实现方式中,按照节点的节点ID从小到大对节点进行排序并对节点进行分组,以此将节点(假设共有M个节点)分成N个组。N个组中的每个组都包括M/N个节点。排序后的节点中的前M/N个节点属于第一组,按照节点ID排在第一组之后的M/N个节点属于第二组,按照节点ID排在第二组之后的M/N个节点属于第三组,依次类推。排在倒数第二组之后的节点属于第N组(即最后一组)。这时,每个组中的节点用于在发生链路变化时,在BGP更新消息中将指示链路变化的BGP-LS-SPF链路NLRI发送给分配给这一组中的特定RR。因此,第一组节点将它们的BGP-LS-SPF链路NLRI发送给第一RR,第二组节点将它们的BGP-LS-SPF链路NLRI发送给第二RR,以此类推。一个RR从一个节点接收到BGP-LS-SPF链路NLRI之后,将该BGP-LS-SPF链路NLRI发送给BGP-SPF路由域内的其它节点。

[0075] 例如,假设节点502、节点504和节点506分成两组,其中,节点502属于第一组,而节点504和节点506属于第二组。第一组中的节点502用于在发生链路变化时将BGP-LS-SPF链路NLRI发送给RR 510,第二组中的节点504和节点506用于在发生链路变化时将BGP-LS-SPF链路NLRI发送给RR 512(如图5中的箭头所示)。RR 510从节点502接收到BGP-LS-SPF链路NLRI之后,将该BGP-LS-SPF链路NLRI发送给节点504和节点506。RR 512从节点504接收到

BGP-LS-SPF链路NLRI之后,将该BGP-LS-SPF链路NLRI发送给节点502和节点506。类似地,RR 512从节点506接收到BGP-LS-SPF链路NLRI之后,将该BGP-LS-SPF链路NLRI发送给节点502和节点504。因此,其它节点都只接收来自RR 510或RR 512的相同NLRI的一个副本。相同的NLRI不存在冗余副本。因此,本实施例与图3中的实施例相似,当与图1中描述的现有泛洪过程相比时,将BGP-SPF路由域内的泛洪量减少了一半,同时还均衡了BGP-SPF路由域内的RR之间的工作负载。

[0076] 图6A是本发明一个实施例提供的BGP-SPF路由域600的泛洪减少过程的示意图。BGP-SPF路由域600是与图2中的BGP-SPF路由域200相似的节点连接BGP-SPF路由域。BGP-SPF路由域600包括4个BGP对等体(节点602、节点604、节点606和节点608),它们都被配置为BGP-LS-SPF发言者,而且如图2所述,同样通过六条链路610至620连接。然而,在图2中描述的现有泛洪过程中,一个节点在发生链路变化时将BGP-LS-SPF链路NLRI发送给与该节点建立EBGP会话的所有其它节点,接收节点将该BGP-LS-SPF链路NLRI发送给与该接收节点建立EBGP会话的所有其它节点,这导致在现有泛洪过程中传输相同BGP-LS-SPF链路NLRI的无数个冗余副本,与此不同,BGP-SPF路由域600内的每个节点都会获取BGP-SPF路由域600的泛洪拓扑(Flooding Topology,FT)。FT是真实网络拓扑(network topology,RT)中的子网络拓扑,该子网络拓扑连接RT中的所有节点。例如,图6B示出了图6A中的BGP-SPF路由域600的RT中的FT。如图6B所示,FT是连接RT中的所有节点(节点602、604、606、608)的子网络拓扑(即只包括图6A中的BGP-SPF路由域600的RT中的部分链路(链路610、616、618、620))。满足成为RT中的FT的要求的子网络拓扑可能不止一个。在一个实施例中,FT以分布式模式计算,其中,BGP-SPF路由域600内的每个BGP-LS-SPF发言者使用相同的算法计算BGP-SPF路由域600的FT。图7中描述了计算BGP-SPF路由域600的FT的一种示例性算法。在另一个实施例中,FT以集中式模式计算,其中,BGP-SPF路由域600内的一个节点(即BGP-LS-SPF发言者)被选为领导者节点,并且计算BGP-SPF路由域600的FT。然后,领导者节点将FT发布给BGP-SPF路由域600内的每个节点。每当领导者节点确定BGP-SPF路由域600的RT中存在变化时,领导者节点计算BGP-SPF路由域600的更新后的FT,并且只将上述发布的FT和更新后的FT之间的变化发布给BGP-SPF路由域600内的每个节点。如果所有节点ID与其索引之间的映射中存在变化,则领导者使用节点ID TLV(如图12所述)将所有节点ID与其索引之间的映射中的变化发布给BGP-SPF路由域600内的每个节点。例如,在一个实施例中,对于添加到BGP-SPF路由域600内的新节点,领导者使用在BGP更新消息中的多协议可达NLRI(MP_REACH_NLRI)路径属性中编码的节点ID TLV发布新节点的ID与其索引之间的映射,以添加映射。对于从BGP-SPF路由域600内移除的受损节点,领导者使用BGP更新消息中的MP_UNREACH_NLRI路径属性下的节点ID TLV发布受损节点的ID与其索引之间的映射,以撤销映射。对于添加到当前FT中的新连接/链路,领导者使用BGP更新消息中的MP_REACH_NLRI路径属性下的路径TLV(如图13所述)发布这些新连接/链路,以将这些新连接/链路添加到当前FT中。对于从当前FT中移除的旧连接/链路,领导者使用BGP更新消息中的MP_UNREACH_NLRI路径属性下的路径TLV发布这些旧连接/链路,以将这些旧连接/链路从当前FT中撤销。

[0077] 如果领导者节点发生故障(即不能正常工作),则新领导者根据该节点成为领导者的优先级进行选择,如下文进一步所述。在一个实施例中,在集中式模式下,新领导者计算BGP-SPF路由域600的新FT,并且将该新FT发布给BGP-SPF路由域600内的每个节点。因为这

是一个新FT,而不是一个更新后的FT,所以新领导者还将所有节点与其索引(如图12所述)之间的映射和FT中的每条连接/链路(如图14所述)发布给域内的每个节点。可替代地,在一些实施例中,新领导者将新领导者计算出的第一(新)FT发布给域内的每个节点,作为新领导者先前从旧领导者接收到的当前FT更新后的FT。也就是说,新领导者根据新领导者计算出的新FT,指示BGP-SPF路由域600内的每个节点将新连接/链路添加到当前FT中,并且从当前FT中移除/撤销旧连接/链路。这些新连接/链路在新FT中,但不在当前FT中。类似地,旧连接/链路在当前FT中,但不在新FT中。

[0078] 一旦BGP-SPF路由域600内的每个节点都通过计算FT或通过从另一节点接收FT来获取BGP-SPF路由域600的FT,当一个节点检测到链路变化(例如,链路连接或断开)时,该节点就在指示链路变化的BGP更新消息中将BGP-LS-SPF链路NLRI发送给FT中的对等节点(也就是说,对等节点是通过FT中的链路直接连接到上述节点的一个或多个节点)。接收来自对等节点的BGP更新消息中的BGP-LS-SPF链路NLRI的对等节点用于在BGP更新消息中将BGP-LS-SPF链路NLRI发送给对等节点在FT中的其它对等节点(即,不包括FT中的将BGP-LS-SPF链路NLRI发送给对等节点的一个或多个对等节点)。例如,假设节点602确定链路616是连接的。响应于链路变化,节点602根据图6B中的FT确定节点604和节点608分别通过FT中的链路610和链路616直接连接到节点602。因此,在BGP更新消息中,节点602通过节点602和节点604之间的链路610上的EBGP会话将指示链路616的链路变化的BGP-LS-SPF链路NLRI发送给节点604,并且通过节点602和节点608之间的链路616上的EBGP会话发送给节点608,如图6A中的箭头所示。注意,与图2中描述的现有泛洪过程相比,由于节点606不是直接连接到图6B中的FT中的节点602,即使节点606直接连接到图6A中的RT中的节点602,节点602也不会在BGP更新消息中将指示链路616的链路变化的BGP-LS-SPF链路NLRI发送给节点606。

[0079] 当节点608从节点602接收BGP更新消息中的BGP-LS-SPF链路NLRI时,节点608根据图6B中的FT确定节点606是FT中的对等节点。节点602也是节点608在FT中的对等节点,但由于BGP-LS-SPF链路NLRI是从节点602接收到的,因此被排除在外。节点608通过节点608和节点606之间的链路620上的EBGP会话在BGP更新消息中将指示链路616的链路变化的BGP-LS-SPF链路NLRI发送给节点606,如图6A中的箭头所示。类似地,当节点604从节点602接收BGP更新消息中的BGP-LS-SPF链路NLRI时,节点604根据图6B中的FT确定节点606是FT中的对等节点。

[0080] 节点602也是节点604在FT中的对等节点,但由于BGP-LS-SPF链路NLRI是从节点602接收到的,因此被排除在外。节点604通过节点604和节点606之间的链路618上的EBGP会话在BGP更新消息中将指示链路616的链路变化的BGP-LS-SPF链路NLRI发送给节点606,如图6A中的箭头所示。

[0081] 因此,节点606接收到指示链路616的链路变化的BGP-LS-SPF链路NLRI的两个副本。一个副本是冗余的。此外,当节点606从节点604和608接收指示链路616的链路变化的BGP-LS-SPF链路NLRI时,节点606根据图6B中的FT确定节点604和608是节点606在FT中的对等节点。然而,由于指示链路616的链路变化的BGP-LS-SPF链路NLRI是从节点604和608接收到的,因此节点606不将BGP-LS-SPF链路NLRI发送给任何节点。

[0082] 与图2中描述的现有泛洪过程的情况一样,当节点608确定链路616是连接的(即,通过链路连接的两个节点都检测到链路变化并发出指示链路变化的BGP-LS-SPF链路NLRI)

时,上述泛洪过程也由节点608发起。节点608发起的包括指示链路616的链路变化的BGP-LS-SPF链路NLRI的BGP更新消息类似地由BGP-SPF路由域600内的节点使用如上所述的泛洪减少过程进行传输。为了简单起见,传输由节点608发起的包括指示链路616的链路变化的BGP-LS-SPF链路NLRI的BGP更新消息链路616不再赘述。此外,为了简单起见,图6A中未示出节点608发起消息的箭头。

[0083] 如图6A所示,即使一些节点可以接收冗余BGP-LS-SPF链路NLRI,与图2中描述的节点连接BGP-SPF路由域的现有泛洪过程相比,当实现节点连接BGP-SPF路由域的泛洪减少过程时,在域600内泛洪的BGP更新消息中的BGP-LS-SPF链路NLRI的数量要少得多。因此,使用图6A的泛洪减少过程,提高了节点连接BGP-SPF路由域的性能和效率。

[0084] 图7是本发明一个实施例提供的FT算法700的流程图。可以执行FT算法700以计算节点连接BGP-SPF路由域(例如但不限于图6A中的BGP-SPF路由域600)的RT中的FT。如上所述,在各种实施例中,节点连接BGP-SPF路由域内的每个节点可以执行FT算法700以计算节点连接BGP-SPF路由域的FT,或者,被选的领导者节点或一个或多个特殊配置的节点可以执行FT算法700以计算节点连接BGP-SPF路由域的FT,并且与节点连接BGP-SPF路由域内的其它节点共享计算出的FT。

[0085] 在所述实施例中,在步骤702中,FT算法700包括用于初始化最大度数(MaxD)变量以及FT和候选队列(candidate queue,CQ)的数据结构的指令。度数是与FT中的一个节点相连的连接/链路。因此,MaxD指示FT中的一个节点与FT中的其它节点之间的最大连接数。限制FT中的一个节点与FT中的其它节点之间的最大连接数的原因是FT中的链路数是减少LS泛洪量的关键因素。一般而言,链路数越少,LS泛洪量就越少。在一个实施例中,MaxD设置为初始值3。在一个实施例中,如果在执行的过程中,FT算法700确定FT不能根据MaxD的值使用RT中的所有节点进行构建,则FT算法700包括用于将MaxD的值增加1并重新启动FT算法700以使用增加后的MaxD的值从头开始重建FT的指令。

[0086] 在一个实施例中,FT和CQ数据结构包括形式为(N,D,PHs)的节点元素,其中,N表示节点,D是节点N的度数,PHs包括节点N的前跳。FT数据结构存储FT中的节点的上述信息。CQ数据结构包括可以添加到FT中的潜在节点(即候选节点)。在一个实施例中,FT算法700从初始空FT数据结构FT＝{}和CQ＝{(R0,D＝0,PHs＝{})}开始。在一个实施例中,R0是RT中具有最小节点ID的节点。可替代地,R0可以根据一些其它标准进行选择。在一个替代实施例中,FT算法700可以简单地从R0作为FT中的根节点(即起始节点)(例如,FT＝{(R0,D＝0,PHs＝{})})和初始候选队列CQ包括直接连接到R0的节点的节点元素(例如,CQ＝{(R1,D＝0,PHs＝{R0}),(R2,D＝0,PHs＝{R0}),…,(Rm,D＝0,PHs＝{R0})})开始,其中,节点R1至Rm都具有度数D＝0且直接连接到R0,如前跳PHs＝{R0}所示。在一个实施例中,R1至Rm按照节点ID从小到大进行排列。可替代地,R1至Rm的顺序可以基于不同的因素。

[0087] 在一个实施例中,FT算法700包括用于实现包括步骤704至步骤710的第一循环以将RT中的所有节点添加到FT中的指令。在步骤704中,FT算法700包括用于识别CQ中的节点N不在FT中且PHs中的一个前跳(Previous Hop,PH)的D小于MaxD的第一节点元素并移除该第一节点元素的指令。例如,如果FT＝{(R0,D＝0,PHs＝{})},CQ＝{(R1,D＝0,PHs＝{R0}),(R2,D＝0,PHs＝{R0}),…,(Rm,D＝0,PHs＝{R0})},则(R1,D＝0,PHs＝{R0})从CQ中移除,因为R1不在FT中且R0的D＝0,小于3(MaxD的值)。如上所述,如果CQ中没有节点元素满足步

骤704的上述条件,这表示FT不能使用MaxD的值通过RT中的所有节点进行构建,则FT算法700包括用于将MaxD的值增加1并重新启动FT算法700以使用增加后的MaxD的值从头开始重建FT的指令(例如,如果MaxD最初设置为3,则第一次CQ中没有节点元素满足上述条件时,MaxD增加到4,FT算法700会从头开始构建FT,以判断FT是否可以使用等于4的MaxD进行构建,如果不能,则FT算法700将MaxD增加到5,并且重新开始,以此类推)。

[0088] 在步骤706中,FT算法700包括用于执行以下操作的指令:将节点N的第一节点元素添加到FT中,将节点N的D设置为1,并且将节点N的PH的D增加1。例如,将从CQ中移除的上述节点元素(R1,D=0,PHs={R0})添加到FT={(R0,D=0,PHs={})}中会产生FT={(R0,0,{}),(R1,0,{R0})}。节点N(在本例中为R1的节点元素(或简称为R1节点元素))的D被设置为1,产生了FT={(R0,0,{}),(R1,1,{R0})}。此外,节点N的PH的D增加1。在这个示例中,R1的PH是R0。如上所示,R0的D等于0。因此,R0的D增加到1,产生了FT={(R0,1,{}),(R1,1,{R0})}。因此,FT当前存在两个节点R0和R1,R0和R1之间存在一条链路。

[0089] 在步骤708中,FT算法700包括用于确定RT中的所有节点是否在FT中的指令。在一个实施例中,FT算法700在CQ为空时确定RT中的所有节点都在FT中,在CQ不为空时确定RT中的所有节点都不在FT中。

[0090] 当FT算法700确定RT中的所有节点都不在FT中时,在步骤710中,FT算法700包括用于识别连接到节点N且不在FT中的任何节点(表示为节点Xi,其中,图7中的i=1、2……n)的指令。例如,如上所示,节点N当前是R1。因此,在本步骤中,FT算法700识别RT中的连接到R1但不在FT中的任何节点Xi(i=1、2……n)。对于RT中的连接到R1且不在FT中的任何节点Xi,FT算法700确定节点Xi是否在CQ中,如果不在CQ中,则FT算法700将节点Xi添加到D=0且PHs={N}的CQ的末尾。例如,假设节点X1连接到R1,不在FT中,并且不在CQ中,则FT算法700将(X1,0,{R1})添加到CQ的末尾。如果节点Xi在RT中连接到R1,并且不在FT中,但已经在CQ中,则FT算法700将节点N添加到CQ中的节点Xi的PHs的末尾。然后,FT算法700循环回到步骤704,识别CQ中的节点N不在FT中且PHs中的一个前跳(Previous Hop,PH)的D小于MaxD的第一节点元素,并移除该第一节点元素。使用上述示例,在循环的此次迭代中,FT算法700识别(R2,D=0,PHs={R0})并从CQ中移除,并且将(R2,D=0,PHs={R0})添加到FT中,如上文在步骤706中所述。

[0091] 如果在步骤708中,FT算法700确定RT中的所有节点都在FT中(即,CQ为空),则FT算法700终止第一循环,并且在步骤712中实现第二循环,以将链路添加到FT中的D在FT中等于1的任何节点。例如,在一个实施例中,在步骤712中,FT算法700包括用于对FT中的D等于1的每个节点(在图7中称为节点B)执行for-loop的指令。这表示FT中的每个节点B当前仅链接到FT中的另一个节点(即是单一链路节点)。对于每个节点B,FT算法700包括用于查找附接到节点B的链路(称为链路L)的指令,使得附接到链路L的远端节点(称为节点R)具有最小D、最小节点ID且支持中转(即,基于BGP SPF的非本地流量的转发)。在2022年2月15日发表的由K.Patel等人撰写的IETF文档“BGP链路状态最短路径优先(SPF)路由(BGP Link-State Shortest Path First(SPF)Routing)”中定义的BGP-LS-SPF节点NLRI属性SPF状态TLV具有指示节点是否支持中转的BGP状态。值为2的BGP状态指示节点不支持基于BGPSPF的中转。如果不存在支持中转的节点R,则FT算法700包括用于查找附接到节点B的链路L的指令,使得链路L的节点R具有最小D和最小节点ID。例如,如果RT中存在5条附接到节点B的链路,则FT

算法700识别连接到5条链路(不包括FT中已经存在的链路)的远端节点(即,节点R)中的哪一个具有最小D、最小节点ID并且支持中转。例如,连接到节点B的链路且D＝1的节点R具有最小D,因为FT中的所有节点都存在至少一条链路。如果这是唯一连接到节点B且D＝1的节点R,则选择节点B与节点R之间的链路。如果这不是唯一连接到节点B且D＝1的节点R,则选择节点B与D＝1、具有最小节点ID并且支持中转的任何节点R之间的链路。FT算法700将节点B和节点R之间的所选链路添加到FT中。FT算法700增加FT中的节点B和节点R的D值,以指示为两个节点添加的连接。一旦FT算法700对FT中的每个节点B执行步骤712,则FT完成,并且FT算法700在步骤714中返回FT(即,输出确定的FT)。

[0092] 下面参考图8至图14,本发明描述了可以包括在BGP更新消息中的各种新BGP扩展,以实现上文参考图3至图6B描述的RR模型、节点连接模型和直连节点模型下的BGP-SPF路由域内的泛洪减少。

[0093] 图8是本发明一个实施例提供的领导者优先级TLV 800的示意图。在一些实施例中,在RR模型下,领导者优先级TLV 800可以由RR用于发布RR成为BGP-SPF路由域内的领导者RR的优先级。在一个实施例中,BGP-SPF路由域内的每个RR使用领导者优先级TLV 800发布RR成为领导者RR的优先级。

[0094] 领导者优先级TLV 800包括类型字段802、长度字段804、保留字段806和优先级字段808。类型字段802是指定类型值的2字节字段,该类型值指示TLV是领导者优先级TLV。类型值待定(to be decided,TBD),并且通过互联网号码分配局(Internet Assigned Numbers Authority,IANA)分配。长度字段804是指定等于4的长度值的2字节字段,该长度值指示领导者优先级TLV 800在长度字段804之后使用的字节数(即,领导者优先级TLV 800中的VALUE部分的大小,包括保留字段806和优先级字段808)。保留字段806是目前未使用的3字节字段,在发送时设置为零。在接收时应该忽略保留字段806。优先级字段808是指定优先级值的1字节字段,该优先级值指示RR成为BGP-SPF路由域内的领导者RR的优先级。在一个实施例中,优先级值是一个八位字节中从0到255的无符号整数,指示RR成为领导者RR的优先级。领导者RR是具有成为域内的领导者的最高优先级的RR。在一个实施例中,当具有相同最高优先级的RR不止一个时,具有最高节点ID和最高优先级的RR是域内的领导者RR。

[0095] 图9是本发明一个实施例提供的节点泛洪TLV 900的示意图。在一些实施例中,节点泛洪TLV 900可由RR或控制器用于将泛洪行为指令提供给一个或多个节点,以减少BGP-SPF域内的泛洪。

[0096] 节点泛洪TLV 900包括类型字段902、长度字段904、保留字段906和泛洪行为字段908。类型字段902是指定类型值的2字节字段,该类型值指示TLV是节点泛洪TLV。类型值待定(TBD)。长度字段904是指定等于4的长度值的2字节字段,该长度值指示节点泛洪TLV 900在长度字段904之后的字节数。保留字段906是目前未使用的3字节字段,在发送时设置为零。在接收时应该忽略保留字段906。泛洪行为字段908是指定泛洪行为值的1字节字段,该泛洪行为值指示节点的特定泛洪行为。定义了以下泛洪行为值和对应的泛洪行为。0:保留。

[0097] 1:将链路状态发送给具有最小ID的RR

[0098] 2:将链路状态发送给具有最大ID的RR

[0099] 3:将链路状态发送给具有较小ID的2个RR

[0100] 4:将链路状态发送给具有较大ID的2个RR

[0101] 5:均衡组

[0102] 6:冗余为2的均衡组

[0103] 7至127:为RR模型定义的标准泛洪行为

[0104] 128至254:为RR模型定义的私有泛洪行为。

[0105] 如上所述,领导者RR可以根据该RR成为领导者的优先级(例如,使用图8中的领导者优先级TLV 800发布)为BGP-SPF路由域选择。在一个实施例中,每个节点的泛洪行为在领导者RR上配置,领导者RR使用节点泛洪TLV 900将该行为发布给网络中的其它RR和/或每个节点。例如,在一个实施例中,节点泛洪TLV 900可以用于指示网络中或BGP-SPF路由域内的每个节点通过将节点泛洪TLV 900中的泛洪行为字段908设置为1,将节点的链路状态只发送给一个RR,其中,设置为1指示每个节点将各自的链路状态发送给具有最小ID的RR。在另一个示例中,RR可以用于指示网络中的每个节点通过使用具有设置为3的泛洪行为字段908的节点泛洪TLV 900将行为发布给每个节点,将各自的链路状态发送给两个RR以实现冗余,其中,设置为3指示每个节点将各自的链路状态发送给具有较小ID的两个RR(即具有最小ID的RR和具有第二最小ID的RR)。此外,当希望通过将节点分组并让每个组将各自的链路状态发送给RR来实现RR或控制器之间的流量均衡时,这种泛洪行为配置在RR上,该RR使用具有设置为5的泛洪行为字段908的节点泛洪TLV 900将泛洪行为发布给每个节点,其中,设置为5指示每个节点将网络中的节点分成多个组。一组中的节点将该节点的链路状态发送给指定给这一组的一个或多个RR。

[0106] 图10是本发明一个实施例提供的领导者偏好TLV 1000的示意图。在一个实施例中,领导者偏好TLV 1000可以由BGP-SPF节点用于指示BGP-SPF节点成为实现如上所述的节点连接模型或直连节点模型的BGP-SPF域内的领导者的优先级。在一些实施例中,当FT计算使用如上所述的集中式模式执行时,选为领导者的BGP-SPF节点可以使用领导者偏好TLV 1000将FT由领导者计算的指示发布给BGP-SPF域内的每个BGP-SPF节点,以实现泛洪减少过程,例如但不限于图6A中描述的泛洪减少过程。此外,在一些实施例中,当FT计算使用如上所述的分布式模式执行时,其中,BGP-SPF域内的每个BGP-SPF节点使用相同的算法计算BGP-SPF域的FT,而且FT没有分布,领导者可以使用领导者偏好TLV 1000指示BGP-SPF域内的每个BGP-SPF节点使用特定算法计算BGP-SPF域的FT。在一些实施例中,如果BGP-SPF节点不发布其领导者偏好TLV 1000,则该BGP-SPF节点没有资格成为领导者。

[0107] 在所述实施例中,领导者偏好TLV 1000包括类型字段1002、长度字段1004、保留字段1006、优先级字段1008和算法字段1010。类型字段1002是指定类型值的2字节字段,该类型值指示TLV是领导者偏好TLV。类型值待定(TBD)。长度字段1004是指定等于4的长度值的2字节字段,该长度值指示领导者偏好TLV 1000在长度字段1004之后的字节数。保留字段1006是目前未使用的2字节字段,在发送时设置为零。在接收时应该忽略保留字段1006。优先级字段1008是指定优先级值的1字节字段,该优先级值指示BGP-SPF节点成为BGP-SPF路由域内的领导者的优先级。在一个实施例中,优先级值是一个字节中从0到255的无符号整数,以指示BGP-SPF节点成为领导者BGP-SPF节点的优先级。在一个实施例中,领导者是具有在优先级字段1008中指定的最高优先级的BGP-SPF节点。在一个实施例中,当具有相同最高优先级的BGP-SPF节点不止一个时,具有最高节点ID和最高优先级的BGP-SPF节点是领导者。算法字段1010是1字节字段,可以指定FT是否由领导者计算(即,集中式模式)、由每个节

点计算(即,分布式模式),如果是分布式,使用哪种算法来计算FT。例如,在一个实施例中,算法字段1010可以指定0至254范围内的数字标识,其中,0指示领导者的集中式FT计算,值1至127指示使用特定标准分布式算法的分布式FT计算,值128至254指示使用特定私有分布式算法的分布式FT计算。

[0108] 图11是本发明一个实施例提供的算法支持TLV 1100的示意图。在一个实施例中,节点连接BGP-SPF域内的BGP-SPF节点使用算法支持TLV 1100来指示BGP-SPF节点支持的用于分布式FT计算的算法。

[0109] 在所述实施例中,算法支持TLV 1100包括类型字段1102、长度字段1104、算法字段1106和算法字段1108。类型字段1102是指定类型值的2字节字段,该类型值指示TLV是算法支持TLV。类型值待定(TBD)。长度字段1104是指定长度值的2字节字段,该长度值指示算法支持TLV 1100在长度字段1104之后的字节数。长度值会根据BGP-SPF节点支持的用于计算FT的算法的数量而变化。例如,如果BGP-SPF节点只支持两种计算FT的算法,如图11所示,由于每种算法都在1字节字段中指定,因此长度值为2。假设BGP-SPF节点支持分布式FT计算,算法字段1106指定1至255范围内的第一数字标识,以指示BGP-SPF节点支持的用于计算FT的第一算法。如果BGP-SPF节点支持用于计算FT的第二算法,则算法字段1108指定在范围1至255内的第二数字标识,以指示BGP-SPF节点支持的用于计算FT的第二算法。如果BGP-SPF节点支持用于计算FT的额外算法,则在算法支持TLV 1100中添加额外算法字段,以指示BGP-SPF节点支持的用于计算FT的额外算法。

[0110] 图12是本发明一个实施例提供的节点ID TLV 1200的示意图。在一个实施例中,当启用集中式FT模式时,节点连接BGP-SPF域内的领导者节点使用节点ID TLV 1200来指示节点与其索引之间的映射。在一个实施例中,节点ID TLV 1200在BGP更新消息中的MP_REACH_NLRI路径属性中编码,并且发布给BGP-SPF域内的所有节点。节点ID TLV 1200中的映射信息使得领导者节点能够减小计算出的FT的编码大小,因为FT可以使用分配给节点的索引(即,节点索引)而不是节点ID进行编码。例如,如下所述,节点ID TLV 1200实现2字节节点索引与4字节节点ID之间的映射。因此,通过使用2字节节点索引代替4字节节点ID在FT中表示节点,领导者节点可以大大减少FT的编码大小。领导者节点将节点ID TLV 1200分发给从领导者接收计算出的FT的每个节点,以便节点可以对计算出的FT进行解码并使用FT来减少本文中公开的链路状态信息分发。此外,在一些实施例中,可以修改节点ID TLV 1200以将2字节节点索引映射到6字节节点ID,这进一步增加了使用节点索引对计算出的FT进行编码的好处。

[0111] 在所述实施例中,节点ID TLV 1200包括类型字段1202、长度字段1204、保留字段1206、最后(L)字段1208、起始索引字段1210和节点ID字段1212A至1212N,其中,A是节点ID TLV 1200中的第一个节点,而N是节点ID TLV 1200中的最后一个节点。类型字段1202是指定类型值的2字节字段,该类型值指示TLV是节点ID TLV。类型值待定(TBD)。长度字段1204是指定长度值的2字节字段,该长度值指示节点ID TLV 1200在长度字段1204之后的字节数。长度值根据节点ID TLV中的节点数而变化。保留字段1206是目前未使用的17比特字段,在发送时设置为零。在接收时应该忽略保留字段1206。L字段1208是在保留字段1206之后的1比特字段。在一个实施例中,当在节点ID TLV 1200中的节点ID字段1212N中指定的最后节点ID的索引等于BFP-SPF域的节点ID的完整列表中的最后索引时,设置L字段1208(即,设置

为1)。换句话说,当设置了L字段1208时,接收节点ID TLV 1200的节点可以开始对FT进行解码,因为该节点已经接收到对整个FT进行解码所需的节点映射信息的最后一部分。

[0112] 起始索引字段1210指定在节点ID TLV 1200的节点ID字段1212A中指定的第一节点ID的索引。节点ID字段1212A至1212N指定BGP-SPF域内的一个节点的BGP标识,也称为BGP路由器ID。

[0113] 在节点ID字段1212B(未示出)至节点ID字段1212N中列出的其它节点的索引不在节点ID TLV 1200中编码。相反,为了确定节点ID TLV 1200中的其它节点的索引,一个节点简单地增加在起始索引字段1210中指定的值,以得到在节点ID字段1212B至节点ID字段1212N中列出的节点的索引。例如,如果在起始索引字段1210中指定的值恰好是100,则在节点ID字段1212A中指定的节点的索引是100,在节点ID字段1212B中指定的节点的索引是101,在节点ID字段1212C(未示出)中指定的节点的索引是102,以此类推。

[0114] 图13是本发明一个实施例提供的路径TLV 1300的示意图。在一个实施例中,当启用集中式FT模式时,节点连接BGP-SPF域内的领导者节点使用路径TLV 1300来发布计算出的FT的一部分(即,一条或多条路径)。在路径TLV 1300中,路径被编码为索引序列:(索引1,索引2,索引3……),表示具有索引1的节点与具有索引为2的节点之间的链路/连接,具有索引2的节点与具有索引3的节点之间的链路/连接,以此类推。单一链路/连接至少是只连接两个节点的路径的简单情况。因此,路径TLV 1300至少包括两个节点索引字段,如图13所示。但是,路径的长度各不相同,可能会连接许多节点。因此,路径TLV 1300通常包括额外索引字段。例如,在所述实施例中,路径TLV 1300包括类型字段1302、长度字段1304、索引1字段1306、索引2字段1308,并且可以根据需要包括额外索引字段。类型字段1302是指定类型值的2字节字段,该类型值指示TLV是路径TLV。类型值待定(TBD)。长度字段1304是指定长度值的2字节字段,该长度值指示路径TLV 1300在长度字段1304之后的字节数。长度值根据路径数和在路径TLV 1300中编码的路径长度(即,路径中的节点数)而变化。索引1字段1306指定路径开始处的第一节点的节点索引,索引2字段1308指定沿着路径连接到第一节点的第二节点的节点索引,以此类推。如上所述,一条以上路径可以在一个路径TLV 1300中编码。例如,在一个实施例中,多条路径中的每条路径表示为该路径上的节点的索引序列,不同的路径(例如,两条路径的两个索引序列)由在路径TLV 1300中的两个索引序列之间的索引字段中指定的特殊索引值(例如,0xFFFF)分隔。在这种情况下,长度字段的值为2×(N条路径中的索引数+N-1)。

[0115] 当存在多条单一链路路径(例如,N条单一链路路径)时,使用一个路径TLV来显示这些路径比使用N个路径TLV来显示这些路径高效,其中,每个路径TLV 1300用于对单一链路路径进行编码。使用一个TLV需要4+6×(N-1)+4=6×N+2个字节。使用N个TLV需要N×(4+4)=8×N个字节。当N等于50等大数时,前者所用空间约为后者所用空间的3/4(四分之三)。

[0116] 图14是本发明一个实施例提供的用于泛洪的连接(Connection Used for Flooding,CUF)TLV 1400的示意图。在所述实施例中,CUF TLV 1400包括类型字段1402、长度字段1404、本地节点ID字段1406和远端节点ID 1408。类型字段1402是指定类型值的2字节字段,该类型值指示TLV是CUF TLV。类型值待定(TBD)。长度字段1404是指定等于8的长度值的2字节字段,该长度值指示CUF TLV 1400在长度字段1404之后的字节数。本地节点ID字段1406是4字节字段,指定FT中的用于泛洪链路状态的连接上的会话的本地节点的BGP ID。

远端节点ID 1408是4字节字段,指定FT中的用于泛洪链路状态的连接上的会话的远端节点的BGP ID。

[0117] 在一个实施例中,一个节点使用CUF TLV 1400来指示连接/链路是FT的一部分并用于泛洪,该连接/链路上的会话具有节点的(本地)节点ID和远端节点ID。该节点将CUF TLV 1400发送给BGP-SPF域内的每个节点。接收CUF TLV 1400的节点可以验证在CUF TLV 1400中指定的连接在FT中,以确保接收节点存储和使用的FT是准确的并且没有改变。

[0118] 图15是本发明一个实施例提供的由BGP-SPF域内的网络节点实现的用于减少所述BGP-SPF域内的泛洪的方法1500的流程图。方法1500包括:所述网络节点与所述BGP-SPF域内的RR集合建立eBGP对等连接会话以交换路由(步骤1502)。所述网络节点确定与所述网络节点的链路对应的链路变化(步骤1504)。所述网络节点根据确定哪些RR是在所述RR集合中的子集的泛洪行为,通过所述eBGP会话在BGP更新消息中将指示所述链路变化的BGP链路状态SPF(BGP Link-State SPF,BGP-LS-SPF)链路网络层可达信息(Network Layer Reachability Information,NLRI)发送给所述RR集合中的子集(步骤1506)。

[0119] 此外,如上所述,在一些实施例中,在将指示所述链路变化的所述信息发送给在所述RR集合中的子集时,所述网络节点可以在所述BGP-LS-SPF链路NLRI中对指示所述链路变化的所述信息进行编码;对包括所述BGP-LS-SPF链路NLRI的所述BGP更新消息进行编码;将所述BGP更新消息发送给所述RR集合中的子集。在一些实施例中,方法1500可以包括:接收指示所述确定哪些RR是在所述RR集合中的子集的泛洪行为的泛洪行为指令;在所述网络节点上配置所述泛洪行为。方法1500还可以包括:接收在节点泛洪类型长度值(Node Flood Type-Length-Value,TLV)中编码的所述泛洪行为指令;对所述节点泛洪TLV进行解码以确定所述泛洪行为。方法1500还可以包括:根据所述泛洪行为指令,将所述网络节点分配到所述BGP-SPF域内的一组网络节点中;将指示所述链路变化的所述BGP-LS-SPF链路NLRI发送给所述RR集合中的为所述一组网络节点指定的所述子集。

[0120] 图16是本发明一个实施例提供的由BGP-SPF域内的RR实现的用于减少所述BGP-SPF域内的泛洪的方法1600的流程图。方法1600包括:所述RR与所述BGP-SPF域内的网络节点建立eBGP对等连接会话以交换路由(步骤1602)。所述RR为所述网络节点配置泛洪行为(步骤1604)。所述泛洪行为指示所述网络节点将指示链路变化的信息只发送给所述BGP-SPF路由域内的特定RR。所述RR在BGP更新消息中将指示所述泛洪行为的节点泛洪类型长度值(Node Flood Type-Length-Value,TLV)发送给所述网络节点(步骤1606)。如上所述,所述RR可以在所述节点泛洪TLV中对所述泛洪行为进行编码;对包括所述节点泛洪TLV的所述BGP更新消息进行编码;将所述BGP更新消息发送给所述网络节点。

[0121] 此外,如上所述,在一些实施例中,方法1600可以包括:传输所述RR成为所述BGP-SPF路由域内的领导者的优先级。方法1600还可以包括:在领导者优先级TLV中对所述RR成为所述BGP-SPF路由域内的领导者的所述优先级进行编码;对包括所述领导者优先级TLV的所述BGP更新消息进行编码;将所述BGP更新消息发送给所述网络节点和所述BGP-SPF路由域内的其它RR。方法1600还可以包括:接收所述BGP-SPF路由域内的所述其它RR成为领导者的优先级;根据所述其它RR的所述优先级,确定所述RR的所述优先级是成为所述BGP-SPF路由域内的领导者的最高优先级;将所述RR配置为所述BGP-SPF路由域内的所述领导者。

[0122] 图17是本发明一个实施例提供的由BGP-SPF域中的网络节点实现的用于减少所述

BGP-SPF域内的泛洪的方法1700的流程图。方法1700包括:所述网络节点获取所述BGP-SPF域的FT(步骤1702)。所述FT是连接所述BGP-SPF域的RT中的所有节点的子网拓扑。所述网络节点确定与所述网络节点的链路对应的链路变化(步骤1704)。所述网络节点在指示所述链路变化的BGP更新消息中将NLRI发送给所述FT中的直接连接到所述网络节点的网络节点(步骤1706)。

[0123] 如上所述,在一些实施例中,方法1700可以包括:从所述BGP-SPF域内的领导者节点获取所述FT。方法1700还可以包括:从所述领导者节点接收节点索引映射;使用所述节点索引映射对所述FT的编码进行解码,以得到所述FT。方法1700还可以包括:从所述领导者节点接收对所述FT的更新;根据所述更新修改所述FT。所述更新可以包括在路径TLV中编码的新连接,所述路径TLV在BGP更新消息中的多协议可达链路网络层可达信息(MP_REACH_NLRI)路径属性中编码。所述更新还可以包括在路径TLV中编码的移除连接,所述路径TLV在BGP更新消息中的多协议不可达链路网络层可达信息(MP_REACH_NLRI)路径属性中编码。

[0124] 图18是本发明一个实施例提供的装置1800的示意性架构图。装置1800适合于实现本文中描述的公开实施例。例如,在一个实施例中,装置1800可以用于实现网络节点、路由器或RR。在各种实施例中,装置1800可以部署为网络内的路由器、交换机和/或其它网络节点。

[0125] 装置1800包括:接收单元(Rx)1820或接收构件,用于经由入/输入端口1810接收数据;处理器1830、逻辑单元、中央处理器(central processing unit,CPU)或其它处理构件,用于处理指令;发送单元(TX)1840或发送构件,用于经由数据出/输出端口1850进行发送;存储器1860或数据存储构件,用于存储指令和各种数据。

[0126] 处理器1830可以实现为一个或多个CPU芯片、一个或多个核(例如,作为多核处理器)、一个或多个现场可编程门阵列(field-programmable gate array,FPGA)、一个或多个专用集成电路(application specific integrated circuit,ASIC)和一个或多个数字信号处理器(digital signal processor,DSP)。处理器1830经由系统总线与入端口1810、RX 1820、TX 1840、出端口1850和存储器1860进行通信耦合。处理器1830可以用于执行存储在存储器1860中的指令。因此,处理器1830提供了一种构件。当处理器1830执行合适指令时,这种构件用于确定、创建、指示、执行、提供或与权利要求书对应的任何其它动作。

[0127] 存储器1860可以是任何类型的能够存储数据和/或指令的存储器或组件。例如,存储器1860可以是易失性和/或非易失性存储器,例如,只读存储器(read-only memory,ROM)、随机存取存储器(random access memory,RAM)、三态内容寻址存储器(ternary content-addressable memory,TCAM)和/或静态随机存取存储器(static random-access memory,SRAM)。存储器1860还可以包括一个或多个磁盘、磁带驱动器和固态驱动器,并且可以用作溢出数据存储设备,以在选择程序来执行时存储这些程序以及存储在执行程序的过程中读取的指令和数据。在一些实施例中,存储器1860可以是与处理器1830集成的存储器。

[0128] 在一个实施例中,存储器1860存储BGP-SPF泛洪减少模块1870。BGP-SPF泛洪减少模块1870包括用于实现本发明公开的实施例的数据和可执行指令。例如,BGP-SPF泛洪减少模块1870可以包括用于实现本文中描述的BGP-SPF泛洪减少的指令。通过减少装置1800接收和发送的链路状态消息的数量,将BGP-SPF泛洪减少模块1870包含在内为装置1800的功能提供了实质性的改进,从而提高了装置1800和整个网络的效率。

[0129] 本发明公开的实施例可以是任何可能的集成技术细节层面上的系统、装置、方法和/或计算机程序产品。所述计算机程序产品可以包括一个或多个非瞬时性计算机可读存储介质,所述非瞬时性计算机可读存储介质具有计算机可读程序指令,所述计算机可读程序指令使得处理器执行本发明的各个方面。所述计算机可读存储介质可以是能够保留和存储指令以供指令执行设备使用的有形设备。

[0130] 虽然本发明提供了若干个实施例,但应当理解,在不脱离本发明的精神或范围的情况下,所公开的系统和方法可以通过其它多种具体形式体现。当前的这些示例被认为是说明性的而非限制性的,并且意图不限于本文给出的细节。例如,各种元件或组件可以组合或集成在另一个系统中,或者可以省略或不实现一些特征。

[0131] 另外,在不脱离本发明的范围的情况下,各种实施例中描述和说明为离散或单独的技术、系统、子系统和方法可以与其它系统、模块、技术或方法组合或集成。示出或描述为彼此耦合、或直接耦合、或彼此通信的其它项目可通过某种接口、设备或中间组件以电方式、机械方式或其它方式间接耦合或通信。变化、替换、变更的其它示例可以由本领域技术人员确定,并可以在不脱离本文中公开的精神和范围的情况下举例。

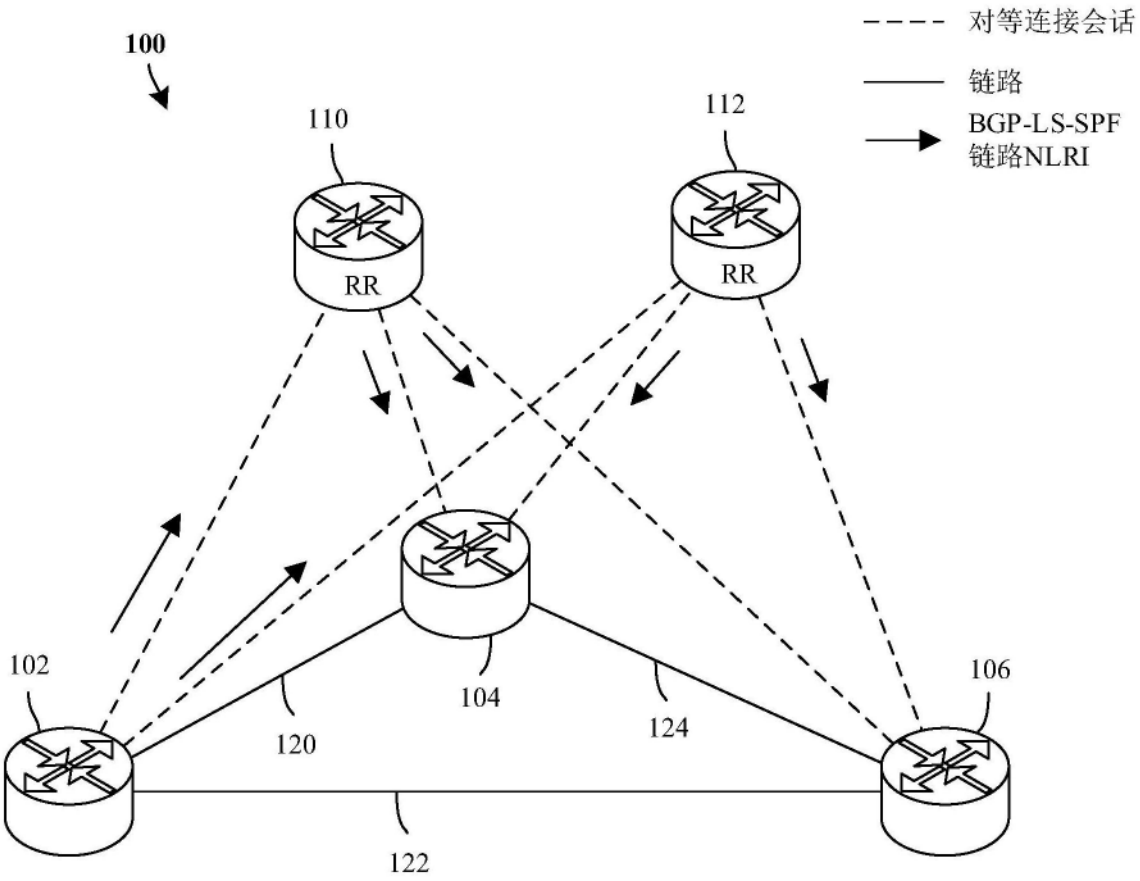

100
对等连接会话
链路
BGP-LS-SPF
链路NLRI
110
112
RR
RR
102
104
106
120
124
122

图1

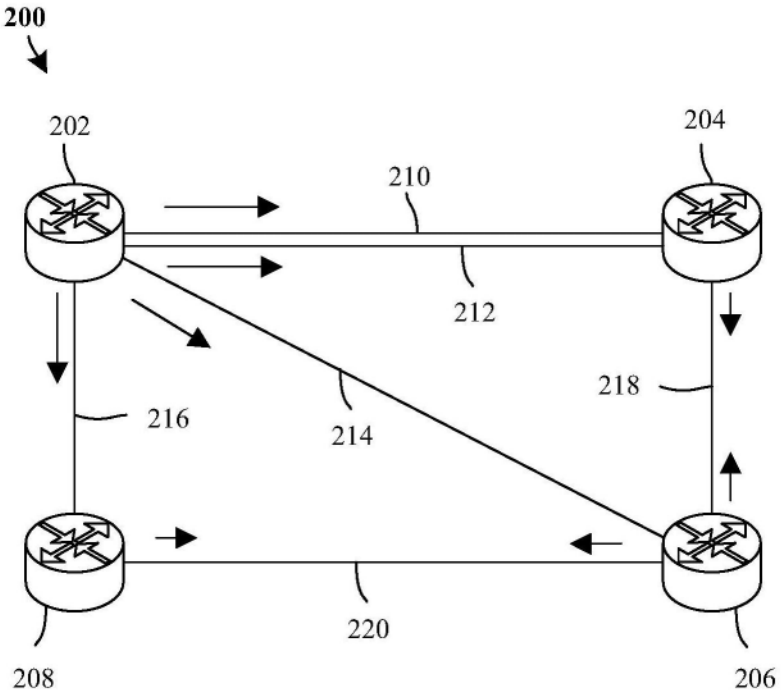

200
202
204
210
212
216
214
218
220
208
206

图2

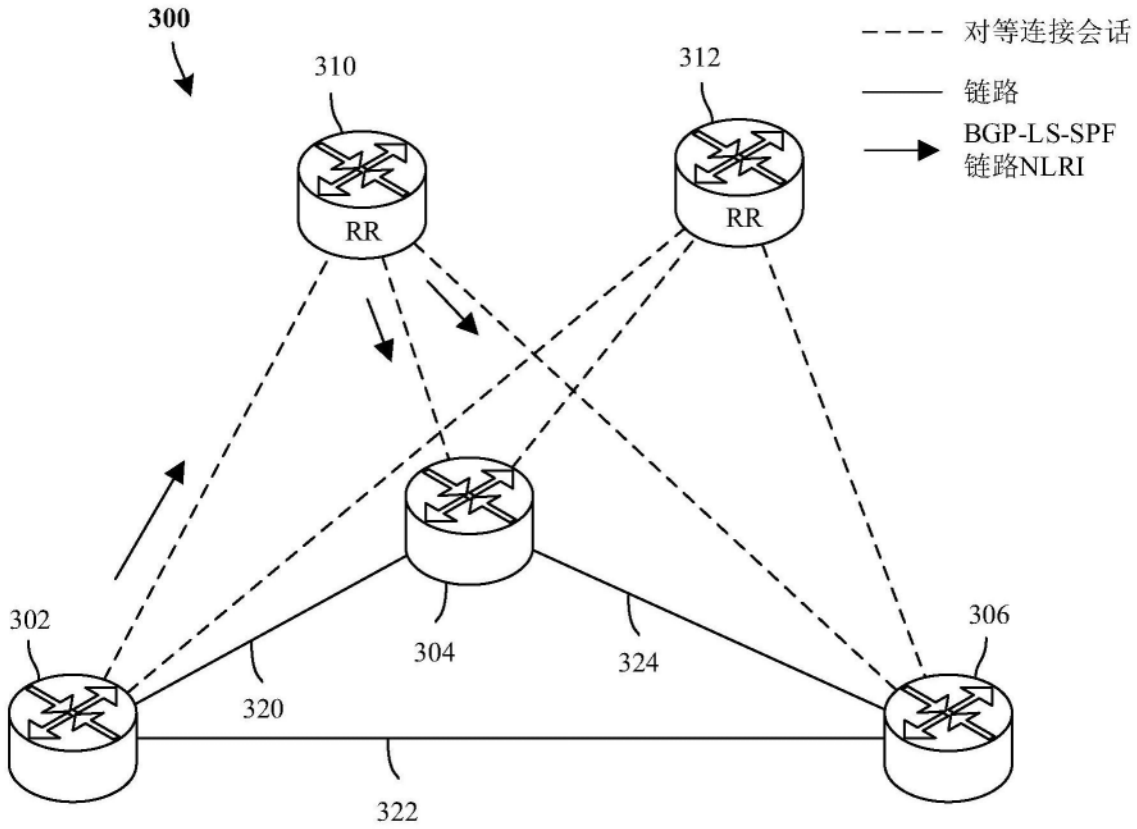

300
310
312
RR
RR
对等连接会话
链路
BGP-LS-SPF
链路NLRI
302
304
306
320
322
324

图3

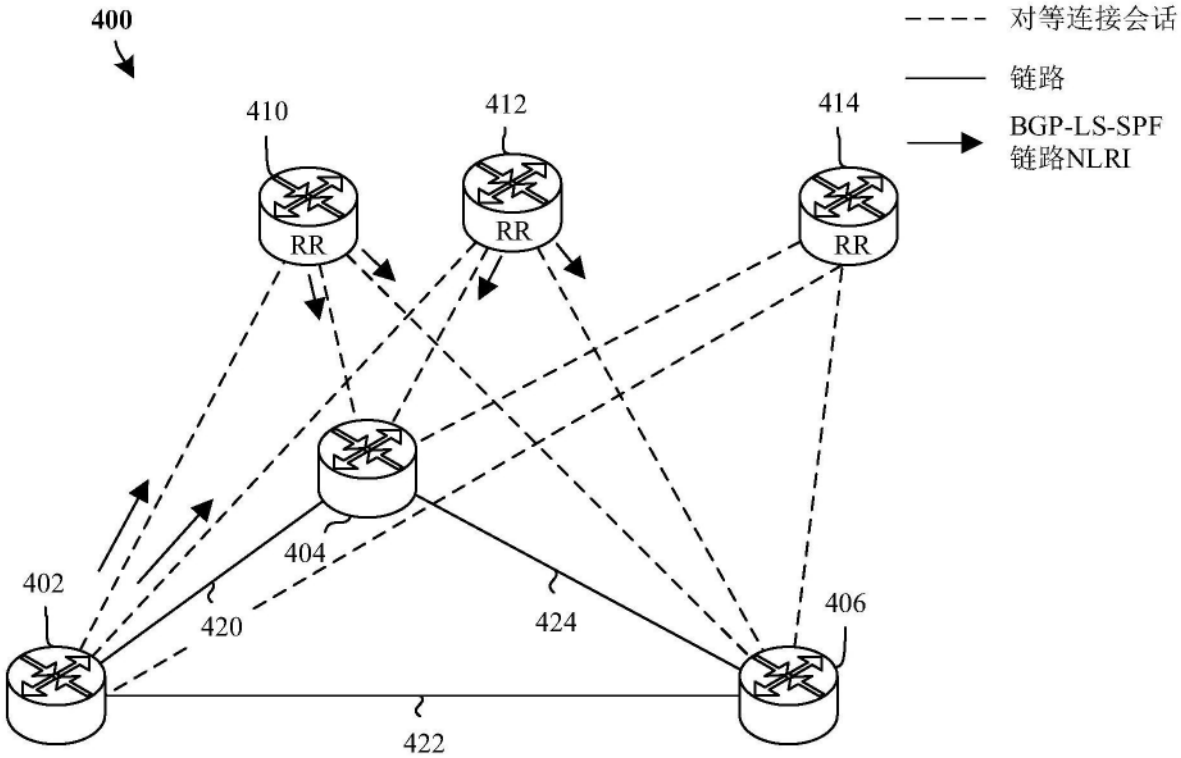

400
410
412
414
RR
RR
RR
对等连接会话
链路
BGP-LS-SPF
链路NLRI
402
404
406
420
422
424

图4

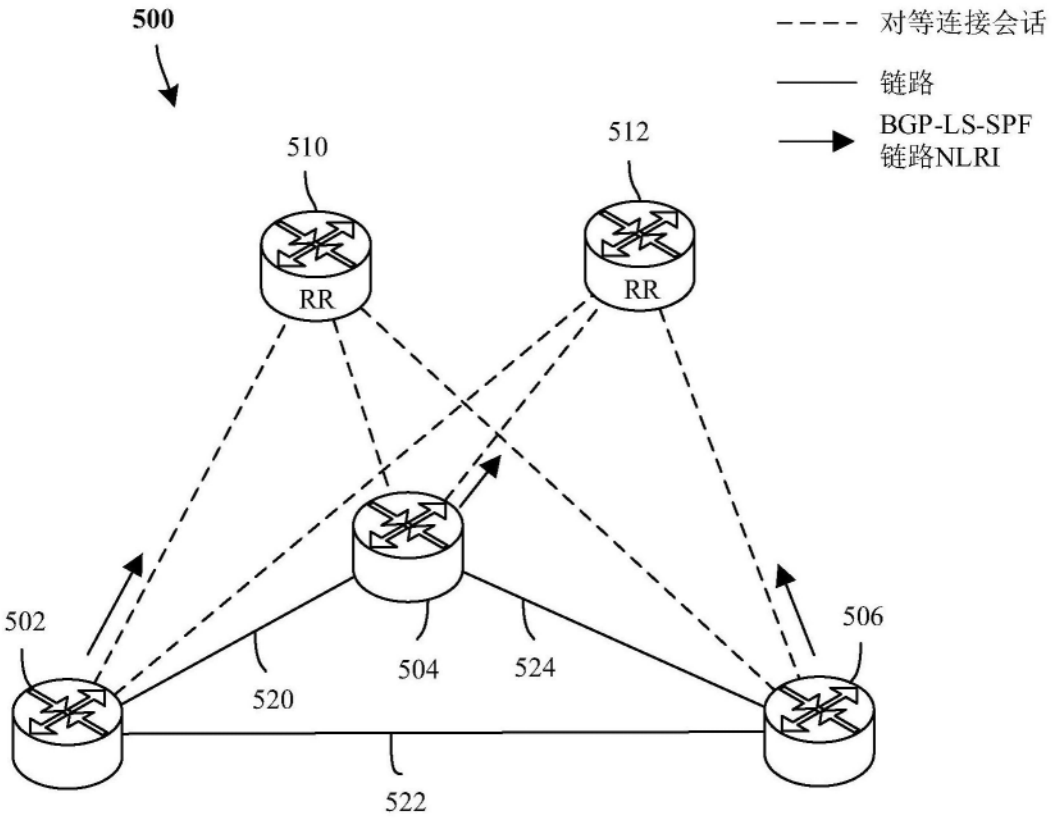

500
对等连接会话
链路
BGP-LS-SPF
链路NLRI
510
512
RR
RR
502
504
506
520
522
524

图5

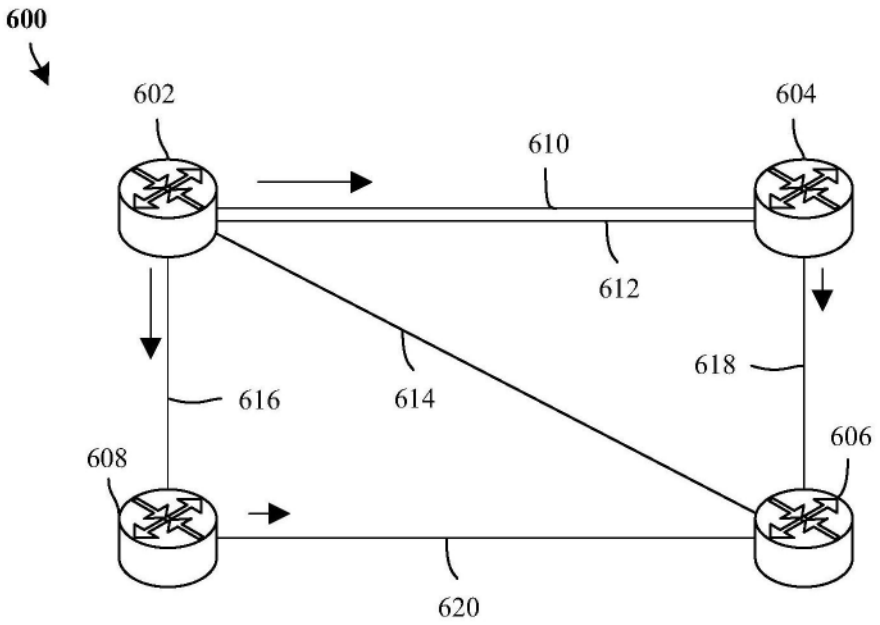

600
602
604
610
612
614
616
618
606
608
620

图6A

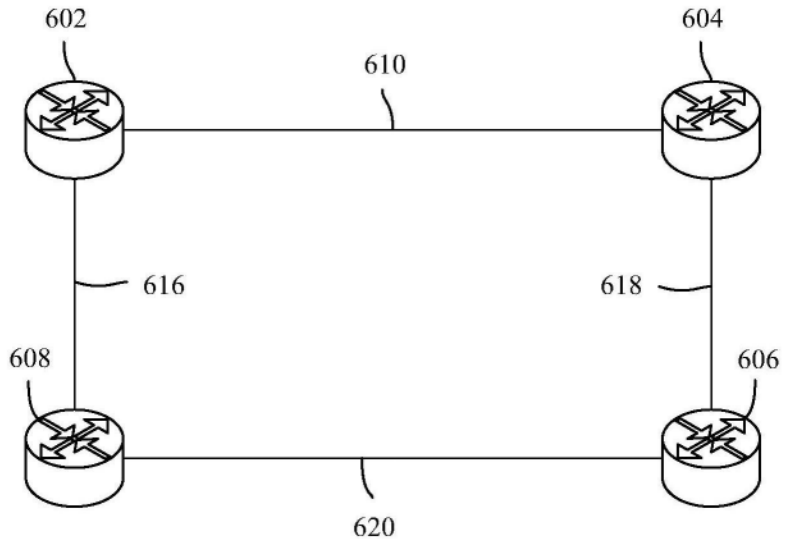


图6B

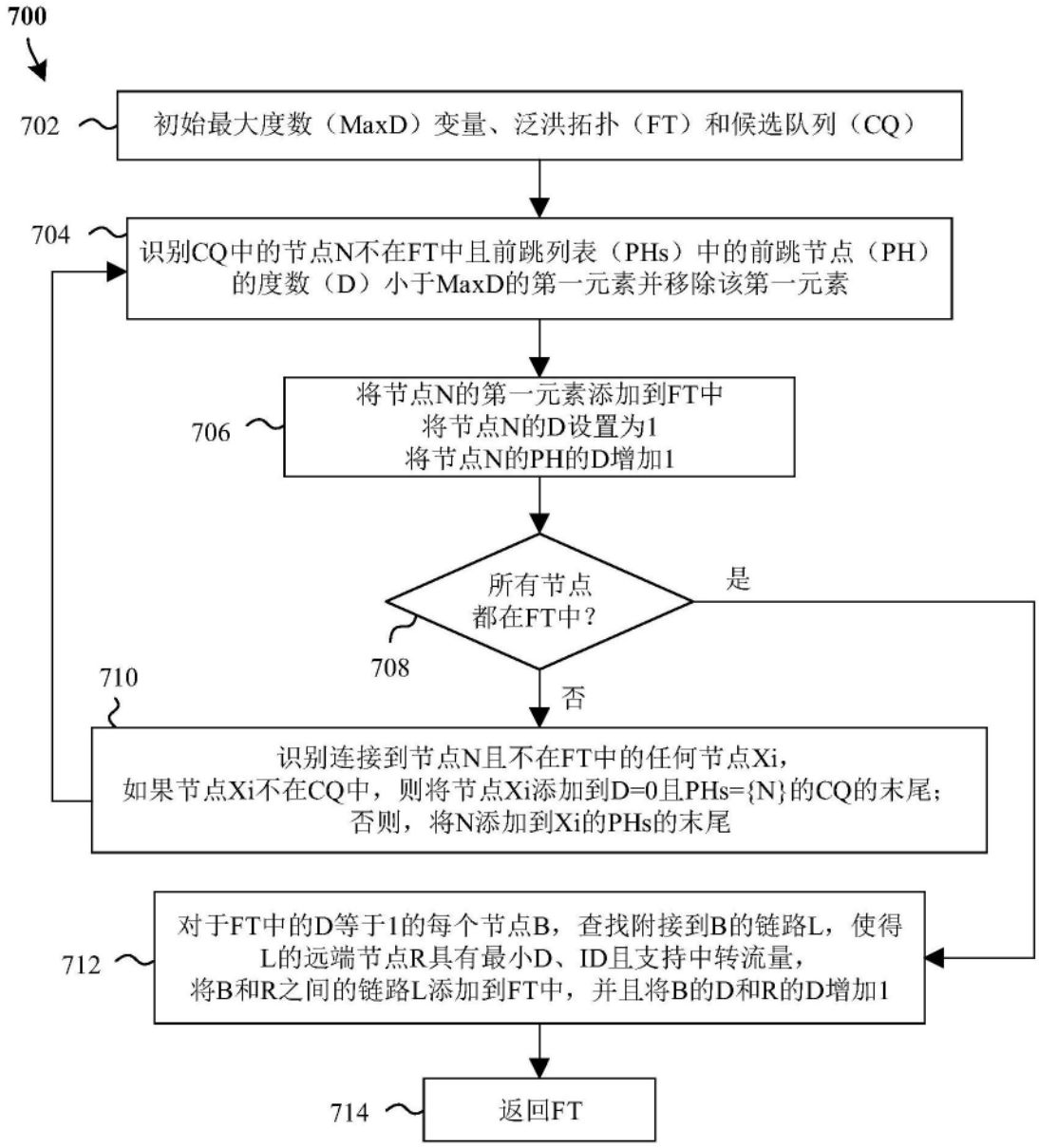

700
702
初始最大度数（MaxD）变量、泛洪拓扑（FT）和候选队列（CQ）
704
识别CQ中的节点N不在FT中且前跳列表（PHs）中的前跳节点（PH）的度数（D）小于MaxD的第一元素并移除该第一元素
706
将节点N的第一元素添加到FT中
将节点N的D设置为1
将节点N的PH的D增加1
所有节点都在FT中？
708
是
否
710
识别连接到节点N且不在FT中的任何节点Xi，
如果节点Xi不在CQ中，则将节点Xi添加到D=0且PHs={N}的CQ的末尾；
否则，将N添加到Xi的PHs的末尾
712
对于FT中的D等于1的每个节点B，查找附接到B的链路L，使得L的远端节点R具有最小D、ID且支持中转流量，
将B和R之间的链路L添加到FT中，并且将B的D和R的D增加1
714
返回FT

图7

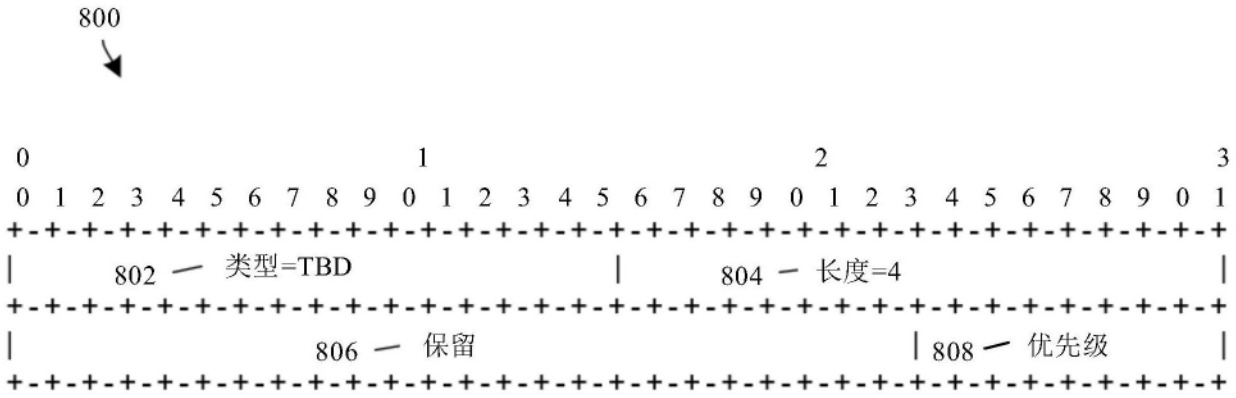

800
0 1 2 3
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
802 类型=TBD
804 长度=4
806 保留
808 优先级

图8

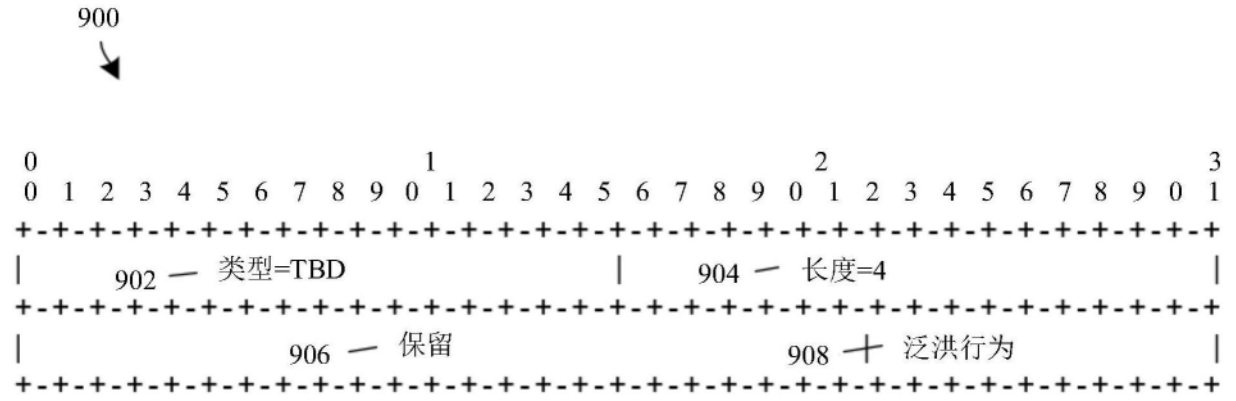

900
0 1 2 3
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
902 类型=TBD
904 长度=4
906 保留
908 泛洪行为

图9

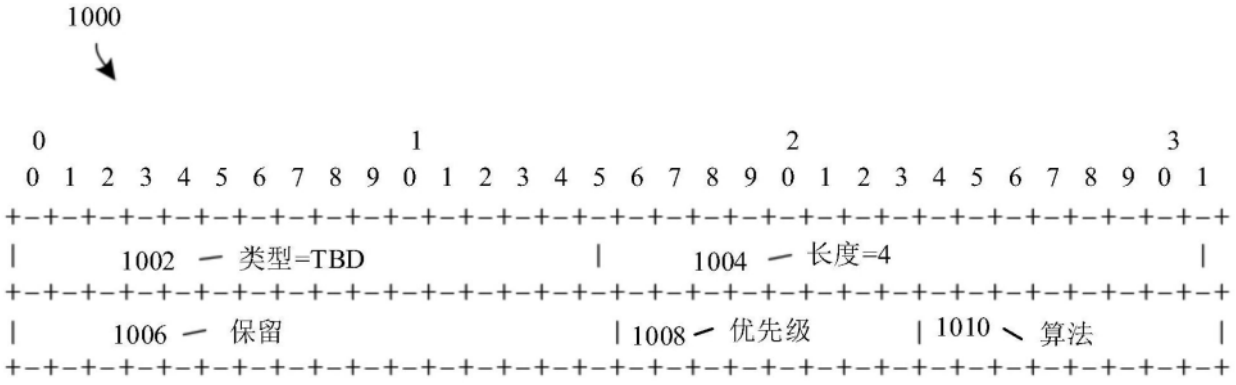

1000
0 1 2 3
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
1002 类型=TBD
1004 长度=4
1006 保留
1008 优先级
1010 算法

图10

图11

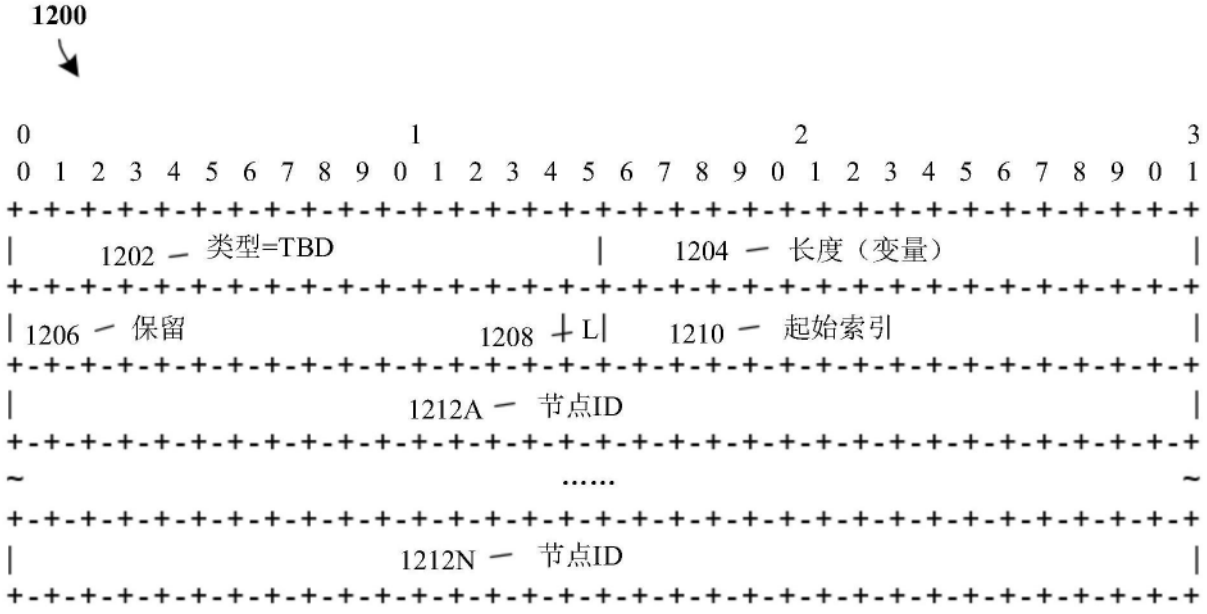


图12

节点 ID TLV

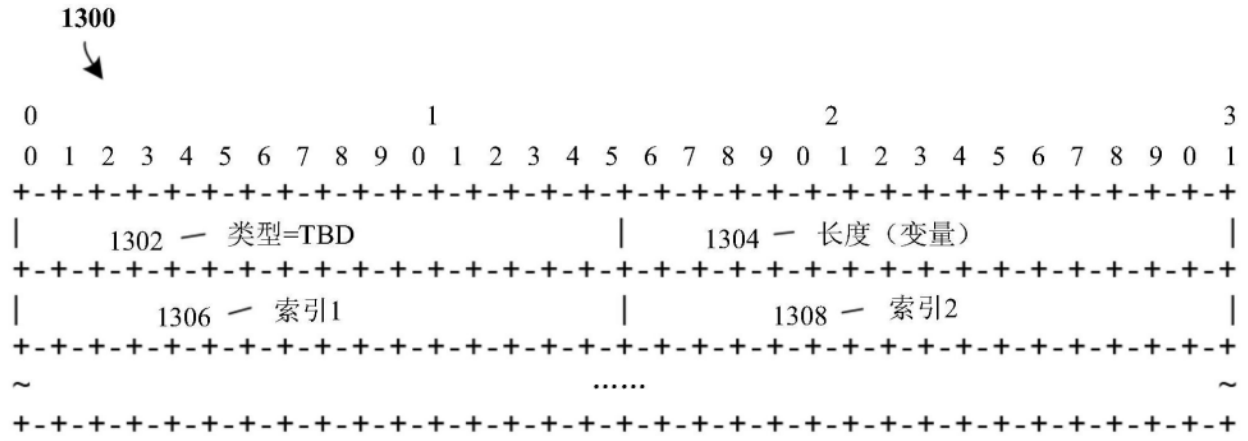

1300
0 1 2 3
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
1302 — 类型=TBD
1304 — 长度（变量）
1306 — 索引1
1308 — 索引2
.....

图13

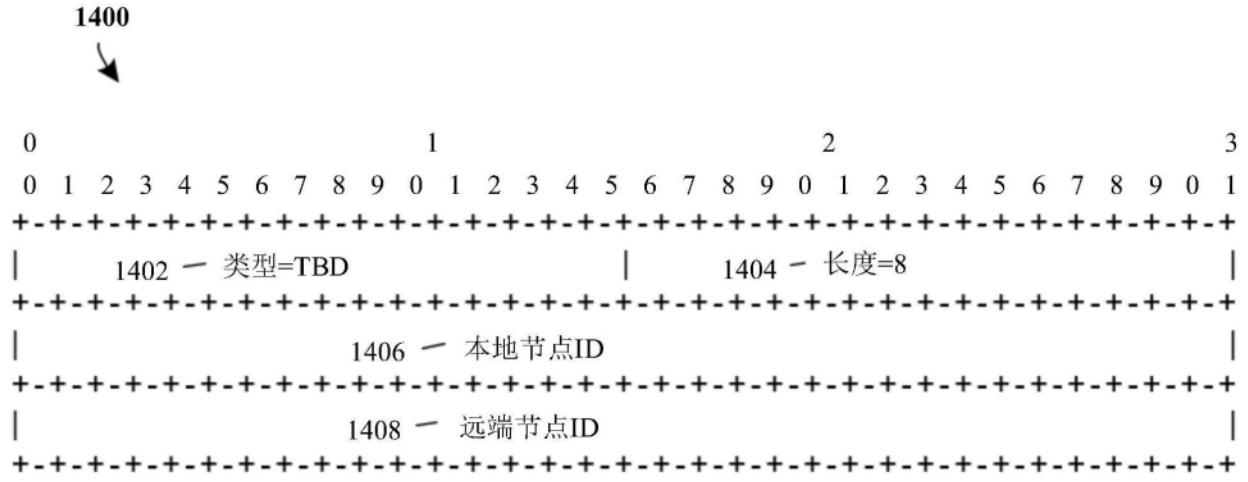

1400
0 1 2 3
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
1402 — 类型=TBD
1404 — 长度=8
1406 — 本地节点ID
1408 — 远端节点ID

图14

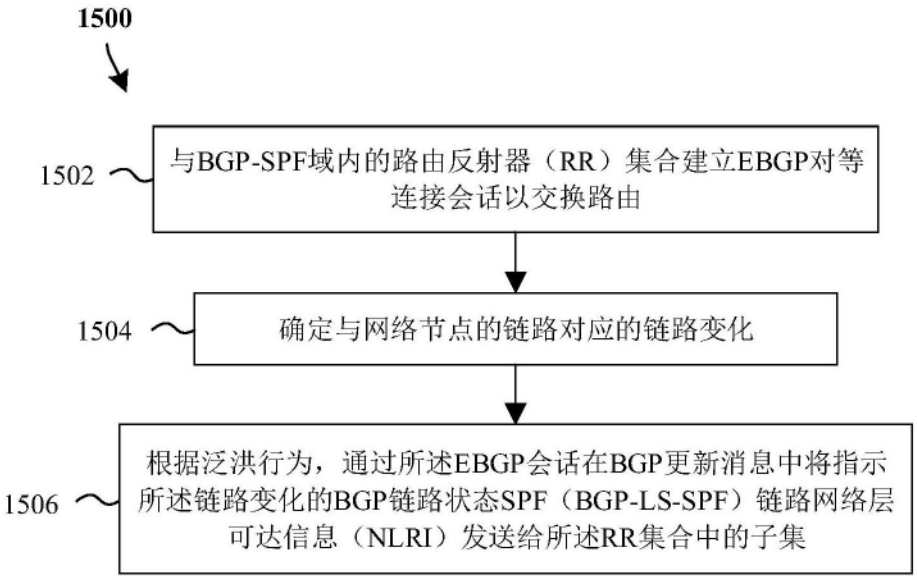

1500
1502
与BGP-SPF域内的路由反射器（RR）集合建立EBGP对等连接会话以交换路由
1504
确定与网络节点的链路对应的链路变化
1506
根据泛洪行为，通过所述EBGP会话在BGP更新消息中将指示所述链路变化的BGP链路状态SPF（BGP-LS-SPF）链路网络层可达信息（NLRI）发送给所述RR集合中的子集

图15

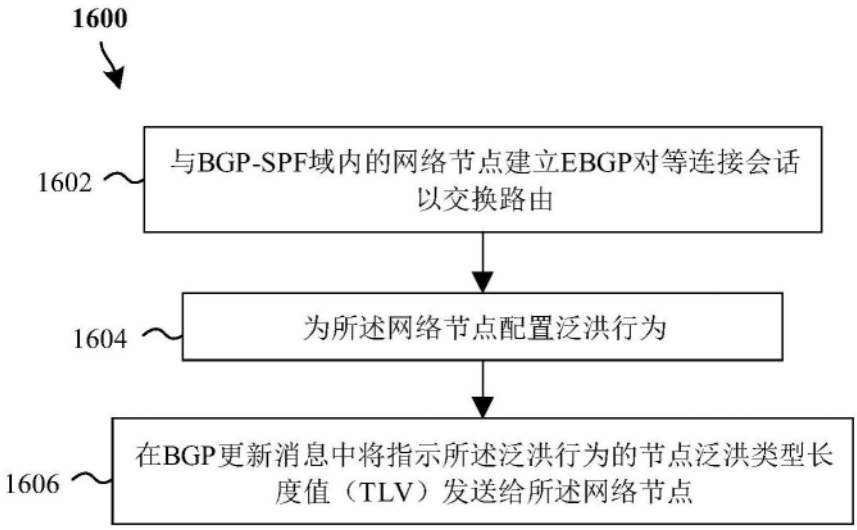

1600
1602
与BGP-SPF域内的网络节点建立EBGP对等连接会话以交换路由
1604
为所述网络节点配置泛洪行为
1606
在BGP更新消息中将指示所述泛洪行为的节点泛洪类型长度值（TLV）发送给所述网络节点

图16

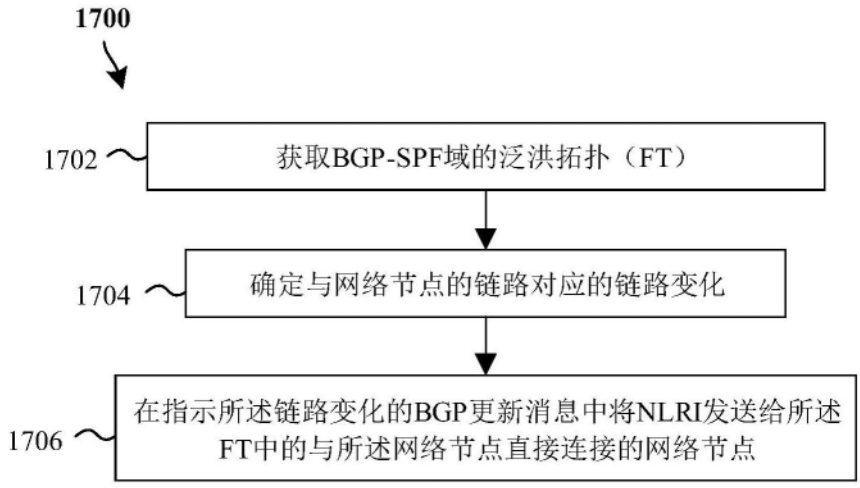

1700
1702
获取BGP-SPF域的泛洪拓扑（FT）
1704
确定与网络节点的链路对应的链路变化
1706
在指示所述链路变化的BGP更新消息中将NLRI发送给所述FT中的与所述网络节点直接连接的网络节点

图17

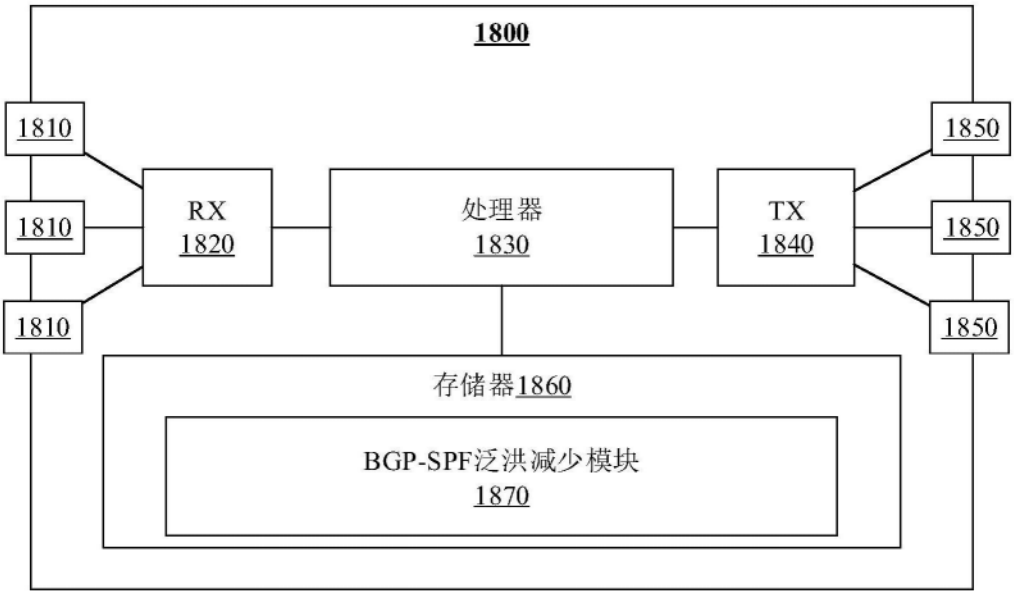

1800
1810
1810
1810
RX
1820
处理器
1830
TX
1840
1850
1850
1850
存储器1860
BGP-SPF泛洪减少模块
1870

图18