

(12) 发明专利申请

(10) 申请公布号 CN 101977156 A

(43) 申请公布日 2011. 02. 16

(21) 申请号 201010549943. 1

(22) 申请日 2010. 11. 18

(71) 申请人 北京星网锐捷网络技术有限公司  
地址 100036 北京市海淀区复兴路 29 号中  
意鹏奥大厦东楼 11 层

(72) 发明人 姚辉 吴梦非 林东豪 贾攀

(74) 专利代理机构 北京同达信恒知识产权代理  
有限公司 11291

代理人 郭润湘

(51) Int. Cl.

H04L 12/56(2006. 01)

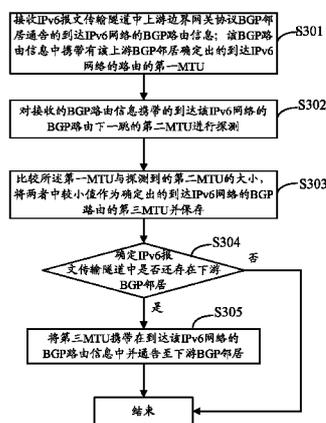
权利要求书 2 页 说明书 8 页 附图 3 页

(54) 发明名称

一种最大传输单元的学习方法、装置及路由设备

(57) 摘要

本发明公开了一种最大传输单元的学习方法、装置及路由设备,该方法包括:接收 IPv6 报文传输隧道中上游 BGP 邻居通告的到达 IPv6 网络的 BGP 路由信息;该 BGP 路由信息中携带有上游 BGP 邻居确定出的到达 IPv6 网络的路由的第一 MTU;对该 BGP 路由信息携带的 BGP 路由下一跳对应的第二 MTU 进行探测;比较第一 MTU 与第二 MTU 的大小,将较小值作为确定出的到达 IPv6 网络的 BGP 路由的第三 MTU 并保存;在确定该隧道中还存在下游 BGP 邻居时,将第三 MTU 携带在到达该 IPv6 网络的 BGP 路由信息中并通告至下游 BGP 邻居。本发明解决了 IPv6 报文在传输过程中传输受阻以及被多次分片导致的报文转发效率低的问题。



1. 一种最大传输单元 MTU 的学习方法,其特征在于,包括:

接收 IPv6 报文传输隧道中上游边界网关协议 BGP 邻居通告的到达 IPv6 网络的 BGP 路由信息;所述 BGP 路由信息中携带有所述上游 BGP 邻居确定出的到达 IPv6 网络的路由的第一 MTU;

对接收的所述 BGP 路由信息携带的到达该 IPv6 网络的 BGP 路由下一跳对应的第二 MTU 进行探测;

比较所述第一 MTU 与探测到的第二 MTU 的大小,将两者中的较小值作为确定出的到达 IPv6 网络的 BGP 路由的第三 MTU 并保存;

确定所述隧道中是否还存在下游 BGP 邻居,若存在,将所述第三 MTU 携带在到达该 IPv6 网络的 BGP 路由信息中并通告至所述下游 BGP 邻居。

2. 如权利要求 1 所述的方法,其特征在于,在到达 IPv6 网络的 BGP 路由信息中携带第一 MTU 或者第三 MTU,通过下述方式实现:

自定义 BGP 路由信息中的扩展团体属性,将自定义的扩展团体属性的类型字段设置为 MTU;将第一 MTU 或第三 MTU 的值封装在所述扩展团体属性的 Value 字段中。

3. 如权利要求 1 所述的方法,其特征在于,所述上游 BGP 邻居为连接该 IPv6 网络的运营商边界设备时,所述第一 MTU 为所述运营商边界设备获取的所述运营商边界设备中与到达 IPv6 网络的路由对应的出接口的 MTU。

4. 如权利要求 1-3 任一项所述的方法,其特征在于,还包括:

保存所述第一 MTU 值与所述探测到的第二 MTU 的值。

5. 如权利要求 4 所述的方法,其特征在于,在当前接收的上游 BGP 邻居通告的 BGP 路由信息中的第一 MTU 与保存的第一 MTU 值不相等时,还包括:

比较当前接收的所述 BGP 路由信息中的第一 MTU 与保存的第二 MTU 的大小确定两者中的较小值,并当所述较小值与保存的第三 MTU 不相等时,使用所述较小值更新保存的第三 MTU,确定所述隧道中是否还存在下游 BGP 邻居,若存在,将更新后的第三 MTU 携带在到达该 IPv6 的 BGP 路由信息中并通告至所述下游 BGP 邻居。

6. 如权利要求 4 所述的方法,其特征在于,将所述第三 MTU 携带在到达该 IPv6 网络的 BGP 路由信息中并通告至所述下游 BGP 邻居的步骤之后,还包括:

周期地对到达该 IPv6 网络的 BGP 路由下一跳的第二 MTU 进行探测,并在当前探测到的第二 MTU 与保存的第二 MTU 不相等时,比较当前探测到的第二 MTU 与保存的第一 MTU 确定两者中的较小值;并当所述较小值与保存的第三 MTU 不相等时,使用所述较小值更新保存的第三 MTU 的值,确定所述隧道中是否还存在下游 BGP 邻居,若存在,将更新后的第三 MTU 携带在到达该 IPv6 的 BGP 路由信息中并通告至隧道中的下游 BGP 邻居。

7. 一种最大传输单元 MTU 的学习装置,其特征在于,包括:

BGP 路由信息接收单元,用于接收 IPv6 报文传输隧道中上游边界网关协议 BGP 邻居通告的到达 IPv6 网络的 BGP 路由信息;所述 BGP 路由信息中携带有所述上游 BGP 邻居确定出的到达 IPv6 网络的路由的第一 MTU;

MTU 探测单元,用于对接收的所述 BGP 路由信息携带的到达该 IPv6 网络的 BGP 路由下一跳对应的第二 MTU 进行探测;

比较单元,用于将比较所述第一 MTU 与探测到的第二 MTU 的大小,将两者中较小值作为

确定出的到达 IPv6 网络的 BGP 路由的第三 MTU 并保存；

路由通告单元,用于确定所述隧道中是否还存在下游 BGP 邻居,若存在,将所述第三 MTU 携带在到达该 IPv6 网络的 BGP 路由信息中并通告至所述下游 BGP 邻居。

8. 如权利要求 7 所述的装置,其特征在于,所述路由通告单元,进一步用于自定义 BGP 路由信息中的扩展团体属性,将自定义的扩展团体属性的类型字段设置为 MTU;将第三 MTU 的值封装在所述扩展团体属性的 Value 字段中。

9. 如权利要求 7 或 8 所述的装置,其特征在于,还包括:MTU 更新单元;

所述比较单元,还用于在 BGP 路由信息接收单元当前接收的上游 BGP 邻居通告的 BGP 路由信息中的第一 MTU 与保存的第一 MTU 值不相等时,比较当前接收的所述 BGP 路由信息中的第一 MTU 与保存的第二 MTU 的大小确定两者中的较小值;

所述 MTU 更新单元,用于当所述较小值与保存的第三 MTU 不相等时,使用所述较小值更新保存的第三 MTU;

所述路由通告单元,还用于确定所述隧道中是否还存在下游 BGP 邻居,若存在,将更新后的第三 MTU 携带在到达该 IPv6 的 BGP 路由信息中并通告至所述下游 BGP 邻居。

10. 如权利要求 7 或 8 所述的装置,其特征在于,还包括:MTU 更新单元;

所述 MTU 探测单元,还用于在路由通告单元将所述第三 MTU 携带在到达该 IPv6 网络的 BGP 路由信息中并通告至所述下游 BGP 邻居的步骤之后,周期地对到达该 IPv6 网络的 BGP 路由由下一跳的第二 MTU 进行探测;

所述比较单元,还用于在所述 MTU 探测单元当前探测到的第二 MTU 与保存的第二 MTU 不相等时,比较当前探测到的第二 MTU 与保存的第一 MTU 确定两者中的较小值;

所述 MTU 更新单元,用于当所述较小值与保存的第三 MTU 不相等时,使用所述较小值更新保存的第三 MTU 的值;

所述路由通告单元,还用于确定所述隧道中是否还存在下游 BGP 邻居,若存在,将更新后的第三 MTU 携带在到达该 IPv6 的 BGP 路由信息中并通告至隧道中的下游 BGP 邻居。

11. 如权利要求 7 所述的装置,其特征在于,当所述装置位于连接该 IPv6 网络的运营商边界设备中时,还包括:

获取单元,用于获取所述运营商边界设备中与到达 IPv6 网络的路由对应的出接口的 MTU;

所述路由通告单元,还用于确定所述隧道中是否还存在下游 BGP 邻居,若存在,将获取到所述出接口的 MTU 携带在到达该 IPv6 网络的 BGP 路由信息中并通告至所述下游 BGP 邻居。

12. 一种路由设备,其特征在于,包括如权利要求 7 ~ 11 任一项所述的最大传输单元 MTU 的学习装置。

## 一种最大传输单元的学习方法、装置及路由设备

### 技术领域

[0001] 本发明涉及计算机网络通信技术领域,尤其涉及一种最大传输单元 MTU 的学习方法、装置及路由设备。

### 背景技术

[0002] 现有互联网协议版本 6(IPv6) 网络作为下一代网络,目前尚未完全普及,仍然作为孤岛网络分散在互联网协议版本 4(IPv4) 网络之中。为了实现这些 IPv6 网络能够通过 IPv4 网络实现互联互通,一般采用隧道技术,如利用互联网协议版本 4-多协议标签交换(IPv4-MPLS) 网络提供 IPv6 报文传输隧道,从而实现 IPv6 孤岛网络的互联互通,互联网工程任务组(The Internet Engineering Task Force, IETF) 为该方案制定了相应的标准,称其为 IPv6 运营商边界(IPv6 Provider Edge, 6PE) 方案。

[0003] 图 1 所示是采用 6PE 方案实现 IPv6 网络互联的一个具体例子,该例子为单个自治域(Autonomous System, AS) 内的 IPv6 网络的互联互通,在图 1 中,路由设备 6PE-1、6PE-2、运营商路由器(Provider, 简称 P) 和 6PE-3 组成了整个 IPv4MPLS 网络,其中 P 设备为网络中 IPv4 单协议栈路由设备,其他各设备均支持 IPv6、IPv4 双协议栈。6PE-1、6PE-2、6PE-3 之间通过内部边界网关协议(Internal Border Gateway Protocol, IBGP) 分发带标签的 IPv6 路由信息,以 6PE-1 通告 IPv6 网络地址 S1 为例,6PE-1 通过多协议内部边界网关协议(Multiprotocol-IBGP, MP-IBGP) 协议向 6PE-3 通告带标签的 IPv6 路由,网络地址为 S1,对应的标签值为 L1,6PE-3 在收到该标签路由之后,将添加 IPv6 路由,出标签为 L1,下一跳为 6PE-1,这样 IPv6 S3 网络中即存在到达 S1 的 IPv6 路由了。同样的原理可以实现 S1、S2、S3 之间 IPv6 路由互联互通。

[0004] 下面描述 IPv6 报文如何在 MPLS 网络中传输,6PE-3 在收到目的地址为 S1 的 IPv6 报文时,将查找本地 IPv6 路由表,确认下一跳为 6PE-1,且出标签为 L1(该报文由 6PE-1 分发),之后 6PE-3 为该 IPv6 报文封装上标签 L1,同时查找下一跳 6PE-1 的出口。根据 MPLS 转发表确定要达到 6PE-1,需要插入标签 T1(为 6PE-3 到 6PE-1 的 MPLS 隧道标签),并转发至 P,此时原来的 IPv6 报文将包含双层标签,内层标签为 IPv6 路由标签(L1),而外层标签为 MPLS 隧道标签(T1)。P 在收到该报文之后,它不关心报文中封装的是 IPv4 数据还是 IPv6 数据,根据外层标签 T1 确定该报文将转发至 6PE-1,并交换标签 T1 为 T2,此时报文的标签栈变为 T2/L1。6PE-1 在收到该报文之后,将弹出标签 T2,同时根据 IPv6 标签 L1 确定转发至正确的网段。

[0005] 上述 IPv6 报文在传输中容易碰到两个问题,一方面,IPv4 网络中的设备 P 在转发报文时将执行最大传输单元(Maximum Transmission Unit, MTU) 检查即检查该报文长度是否超过出接口的 MTU,若当前接收的报文的长度超过其出接口的 MTU 时,P 由于无法识别 IPv6 报文而只能丢弃该报文;另一方面,即使报文的长度未超过 P 出接口的 MTU,由 P 成功转发至 6PE-1,6PE-1 根据该报文的标签 L1 以及 IPv6 路由确认转发至某个出接口,若当前报文的长度超过 6PE-1 出接口的 MTU 时,6PE-1 将为该 IPv6 报文再次执行分片,这样导致

IPv6 报文在经隧道传输过程被多次分片,降低了报文转发效率。

[0006] 图 2 所示是采用 6PE 方案实现 IPv6 网络互联的另外一个具体例子,该例子为跨多个自治域的 IPv6 网络的互联互通,图 2 中 AS1 由 6PE-1、P 和自治系统边界路由器 (Autonomous System Border Router, ASBR) 1 组成, AS2 由 ASBR2 和 6PE-2 组成 (ASBR2 和 6PE-2 之间还可以包括若干 P 设备,在图 2 中未示意出), IPv6 S1 通过 6PE-1 连接自治域 AS1, IPv6 S2 连接自治域 AS2, 6PE-1 和 ASBR1 之间建立 MP-IBGP 连接, 6PE-1 将带标签的 IPv6 路由通告给 ASBR1, ASBR1 与 ASBR2 之间建立多协议 BGP 协议 (Multiprotocol Extensions BGP, MP-EBGP) 连接, ASBR1 将 6PE-1 的带标签 IPv6 路由通过 MP-EBGP 通告至 ASBR-2, ASBR2 与 6PE-2 建立 MP-IBGP 连接, ASBR2 再将 ASBR1 通告的带标签 IPv6 路由通过 MP-IBGP 通告至 6PE-2, 这样 6PE-2 即可学习到 6PE-1 的带标签 Ipv6 路由, S2 和 S1 相互学习到各自的 IPv6 路由。

[0007] IPv6 报文的转发则是通过建立三个 MPLS 隧道来实现的,如图 2 示中的 T3 (6PE-1 和 ASBR1 之间的隧道)、T2 (ASBR1 和 ASBR2 之间的隧道) 和 T1 (ASBR2 和 6PE-2 之间的隧道), 6PE-2 在收到到达 S1 的 IPv6 报文之后,将会先通过隧道 T1 将报文转发至 ASBR2, ASBR2 再通过隧道 T2 转发至 ASBR1, ASBR1 则通过 MPLS 隧道 T3 将报文转发至 6PE-1, 从而实现 IPv6 报文的转发。由于 6PE-2 不需要知道如何到达 ASBR1 和 6PE-1, 它只需要将报文转发至 ASBR2, 剩余的转发操作则有 ASBR2 来执行, 报文转发至 ASBR2 时, ASBR2 会进行 MTU 检查, 如果报文的长度超过其出接口的 MTU, 那么需要执行分片后转发, 类似地, ASBR2 和 6PE-2 在转发过程也同样需要进行 MTU 的检查, 这样可能导致 IPv6 报文在经隧道传输过程被多次分片, 降低了报文转发效率。另外, 报文在经由 AS1 和 AS2 中的 P 设备转发至边界路由设备时, 也有可能因为其长度大于 P 设备出接口的 MTU 而使得 P 设备直接丢弃该报文, 使得 IPv6 报文的传输过程受阻导致 IPv6 网络之间无法实现互通。

## 发明内容

[0008] 本发明实施例提供一种最大传输单元 MTU 的学习方法、装置及路由设备, 用以解决在现有 6PE 技术中 IPv6 报文在传输过程中由于长度大小超过路由设备出接口 MTU 导致传输受阻以及被多次分片导致的报文转发效率低的问题。

[0009] 本发明实施例提供的一种最大传输单元的学习方法, 包括:

[0010] 接收 IPv6 报文传输隧道中上游边界网关协议 BGP 邻居通告的到达 IPv6 网络的 BGP 路由信息; 所述 BGP 路由信息中携带有所述上游 BGP 邻居确定出的到达 IPv6 网络的路由的第一 MTU;

[0011] 对接收的所述 BGP 路由信息携带的到达该 IPv6 网络的 BGP 路由下一跳对应的第二 MTU 进行探测;

[0012] 比较所述第一 MTU 与探测到的第二 MTU 的大小, 将两者中的较小值作为确定出的到达 IPv6 网络的 BGP 路由的第三 MTU 并保存;

[0013] 确定所述隧道中是否还存在下游 BGP 邻居, 若存在, 将所述第三 MTU 携带在到达该 IPv6 网络的 BGP 路由信息中并通告至所述下游 BGP 邻居。

[0014] 本发明实施例提供的一种最大传输单元的学习装置, 包括:

[0015] BGP 路由信息接收单元, 用于接收 IPv6 报文传输隧道中上游边界网关协议 BGP 邻

居通告的到达 IPv6 网络的 BGP 路由信息 ;所述 BGP 路由信息中携带有所述上游 BGP 邻居确定出的到达 IPv6 网络的路由的第一 MTU ;

[0016] MTU 探测单元,用于对接收的所述 BGP 路由信息携带的到达该 IPv6 网络的 BGP 路由由下一跳对应的第二 MTU 进行探测 ;

[0017] 比较单元,用于将比较所述第一 MTU 与探测到的第二 MTU 的大小,将两者中较小值作为确定出的到达 IPv6 网络的 BGP 路由的第三 MTU 并保存 ;

[0018] 路由通告单元,用于确定所述隧道中是否还存在下游 BGP 邻居,若存在,将所述第三 MTU 携带在到达该 IPv6 网络的 BGP 路由信息中并通告至所述下游 BGP 邻居。

[0019] 本发明实施例提供的路由设备,包括本发明实施例提供的上述最大传输单元的学习装置。

[0020] 本发明实施例的有益效果包括 :

[0021] 本发明实施例提供的最大传输单元的学习方法、装置及路由设备,对于 IPv4 网络中形成的 IPv6 报文传输隧道中的每个运行 BGP 的路由设备,执行下述操作即 :接收上游 BGP 邻居通告的到达 IPv6 网络的 BGP 路由信息 ;该 BGP 路由信息携带有该上游 BGP 邻居确定出的到达 IPv6 网络的路由的第一 MTU ;对接收的 BGP 路由信息携带的到达该 IPv6 网络的 BGP 路由由下一跳对应的第二 MTU 进行探测 ;比较第一 MTU 与探测到的第二 MTU 的大小,将两者中较小值作为确定出的到达 IPv6 网络的 BGP 路由的第三 MTU 并保存 ;如果还存在下游 BGP 邻居,则将第三 MTU 携带在到达该 IPv6 网络的 BGP 路由信息中并通告至下游 BGP 邻居,直至该隧道中的每个运行 BGP 的路由设备都学习到了到达该 IPv6 网络的 BGP 路由对应的 MTU。本发明实施例在支持 BGP 协议的运营商边界设备和自治系统边界路由之间通过 BGP 路由传递 MTU 的值,使得传输 IPv6 报文的隧道的终点即连接另一 IPv6 网络的运营商边界设备可以计算出到达目的 IPv6 网络的路由中各跳的 MTU 的最小值,并将其作为发送 IPv6 报文是否需要分片的依据,经过上述过程之后,在该隧道的终点收到发往该 IPv6 网络的 IPv6 报文时,将该报文的大小与自身保存的 MTU 的大小进行比较,决定是否需要分片,如果进行分片,分片后的报文的长度是到达该 IPv6 网络的路由中各跳 MTU 的最小值,因此,对于通过 IPv4 网络实现 IPv6 网络的互联互通的应用场景尤其是跨多个 IPv4 自治域实现 IPv6 网络的互联互通的应用场景来说,不论 IPv4 网络中传输 IPv6 报文的隧道需要跨越多少中间设备 (例如 P 设备) 或边界设备,任何中间设备或者边界设备,都不会因为 IPv6 报文长度的大小超出了该路由设备出接口的 MTU 的大小而丢弃该 IPv6 报文,从而避免了现有技术出现的 IPv6 报文的传输受阻的情况,并且在隧道的终点对收到的 IPv6 报文进行分片操作后,该隧道中其他路由设备也不会再执行报文的分片操作,提高了 IPv6 报文在隧道中转发效率。

#### 附图说明

[0022] 图 1 为现有技术中 6PE 方案实现 IPv6 网络互联的网络连接示意图之一 ;

[0023] 图 2 为现有技术中 6PE 方案实现 IPv6 网络互联的网络连接示意图之二 ;

[0024] 图 3 为本发明实施例提供的最大传输单元的学习方法的流程图 ;

[0025] 图 4 为自定义团体属性的格式示意图 ;

[0026] 图 5 为本发明实施例提供的第一个具体实例的网络连接示意图 ;

[0027] 图 6 为本发明实施例提供的最大传输单元的学习装置的结构示意图。

### 具体实施方式

[0028] 下面结合附图,对本发明实施例提供的一种最大传输单元 (MTU) 的学习方法、装置及路由设备进行详细地说明。

[0029] 本发明实施例提供的 MTU 学习方法,如图 3 所示,包括如下步骤:

[0030] S301、接收 IPv6 报文传输隧道中上游 BGP 邻居通告的到达 IPv6 网络的 BGP 路由信息;该 BGP 路由信息中携带有该上游 BGP 邻居确定出的到达 IPv6 网络的路由的第一 MTU;

[0031] S302、对接收的 BGP 路由信息携带的到达该 IPv6 网络的 BGP 路由下一跳的第二 MTU 进行探测;

[0032] S303、比较所述第一 MTU 与探测到的第二 MTU 的大小,将两者中较小值作为确定出的到达 IPv6 网络的 BGP 路由的第三 MTU 并保存;

[0033] S304、确定 IPv6 报文传输隧道中是否还存在下游 BGP 邻居;若存在,执行下述步骤 S305;若不存在,结束本流程;

[0034] S305、将第三 MTU 携带在到达该 IPv6 网络的 BGP 路由信息中并通告至下游 BGP 邻居。

[0035] 上述步骤 S301 中,IPv6 报文传输隧道指的是连接两个 IPv6 网络的 IPv4-MPLS 网络提供的 IPv6 报文传输通道。

[0036] 上述步骤 S301 和步骤 S305 中,在到达 IPv6 网络的 BGP 路由信息中携带第一 MTU 或携带第三 MTU,可以通过下述方式实现:自定义 BGP 路由信息中的扩展团体属性,将第一 MTU 或第三 MTU 的值封装在所述扩展团体属性的 Value 字段中。

[0037] 扩展团体属性 (Extended Communities Attribute) 是 BGP 路由信息中的路由属性中的一类,其格式如图 4 所示,自定义团体属性,可以自定义其中的 type 字段 (扩展团体属性的类型) 为 MTU,将第一 MTU 或者第三 MTU 的值封装在自定义的扩展团体属性的 Value 字段中。

[0038] 较佳地,在本发明实施例中,还需要同时保存本次 MTU 学习过程中得到的第一 MTU 和第二 MTU 也与到达 IPv6 网络的 BGP 路由信息,以便后续 MTU 发生变化时,可以及时根据保存的第一 MTU 和第二 MTU 对保存的第三 MTU 进行更新。

[0039] 下面结合两个具体的实例,对本发明实施例提供的 MTU 学习方法进行详细地说明。

[0040] 第一个实例:

[0041] 本实例采用图 5 所示的单自治域内 IPv6 网络互联的网络拓扑,运营商边界设备 R1 与 IPv6 S1 网络相连,运营商边界设备 R2 与 IPv6 S2 网络相连,同时 R1 和 R2 同属于 IPv4-MPLS 网络,为了实现将 IPv6 S2 网络的 IPv6 报文发送至 IPv6 S1 网络,需要在报文发送之前,执行 BGP 协议的路由设备 R1 和 R2 按照下面的流程分别执行 MTU 学习的步骤:

[0042] 步骤 1、R1 向 IPv6 报文传输隧道中其上游的 BGP 邻居即 R2 通告到达 IPv6S1 网络的 BGP 路由信息,在通告的 BGP 路由中携带有 R1 确定出的到达 IPv6S1 网络的路由的 MTU1;

[0043] 在本步骤 1 中,由于 R1 是直接与 IPv6 S1 网络相连的运营商边界设备,MTU1 的值, R1 可以通过获取 R1 中与到达 IPv6 S1 网络的路由对应的出接口的 MTU 的值得到 MTU1。

[0044] R1 将 MTU1 的值携带在自定义的扩展团体属性中,具体方法参见前述步骤 S301 的说明。

[0045] 步骤 2、R2 接收到该 BGP 路由信息之后,对该 BGP 路由信息携带的到达 IPv6 S1 网络的路由的 BGP 路由下一跳对应的 MTU2 进行探测;

[0046] R2 从接收的 BGP 路由信息中提取该 BGP 路由的下一跳的 IP 地址, BGP 路由的下一跳是 BGP 路由信息中路由属性的一种,用于指明到达该路由前缀对应的网络 (IPv6 S1 网络) 需要将报文转发至 BGP 下一跳的路由设备 (即 R1)。

[0047] 根据该 BGP 路由的下一跳 (R1) 的 IP 地址, R2 可自动探测该下一跳对应的 MTU 值。具体可采用现有的自动探测方法,例如通过发送 Internet 控制报文协议 (Internet Control Message Protocol, ICMP) 报文来发现 MTU2,如果该传输隧道中 R1 和 R2 之间的路径上还连接有若干 P 设备或其他路由设备,探测 MTU2 的过程,即确定从 R2 经过若干 P 设备或其他路由设备到达 R1 的路径中各跳的 MTU 值的最小值。具体探测方法属于现有技术,在此不再赘述。

[0048] 步骤 3、R2 比较 MTU1 和 MTU2 两者的大小,将两者中的较小值确定为到达 IPv6 S1 网络的 BGP 路由对应的 MTU3 并保存。

[0049] MTU3 可以与到达 IPv6 S1 网络的 BGP 路由一起保存于路由表中,以便在后续在转发该路由的 IPv6 报文时,可以根据 MTU3 的大小判断是否需要分片的操作。

[0050] 步骤 4、R2 判断该隧道中是否存在下游 BGP 邻居,由于 R2 是与 IPv6 S2 网络相连的运营商边界设备,其不存在 BGP 邻居,因此流程到此结束。

[0051] 在经过上述步骤 1~4 之后,R2 中保存的是到达 IPv6 S1 网络的路由各跳 MTU 的最小值,R2 收到 IPv6 S2 网络发送至 IPv6 S1 网络的 IPv6 报文后,将该报文的大小与保存的 MTU3 的大小进行比较,决定是否需要分片,如果进行分片,分片后的报文的长度是到达 IPv6 S1 网络的路由的 MTU 的最小值,因此,在 R2 和 R1 之间的 P 设备不会因为报文长度的大小超出了该路由出接口的 MTU 的大小而丢弃报文,在 R1 中,也不会再执行报文的分片操作,提高了报文的转发效率。

[0052] 第二个实例:

[0053] 本实例采用图 2 所示的跨自治域的 IPv6 网络互联的网络拓扑,其中 IPv6 S1 网络通过 AS1、AS2 与 IPv6 S2 网络相连,AS1 由 6PE-1、P 和 ASBR1 组成,AS2 由 ASBR2 和 6PE-2 组成,AS1 和 AS2 都属于 IPv4-MPLS 网络,6PE-1 和 6PE-2 是运营商边界设备。

[0054] 为了实现将 IPv6 S2 网络的 IPv6 报文发送至 IPv6 S1 网络,需要在报文发送之前,各执行 BGP 协议的路由设备按照下面的流程分别执行 MTU 学习的步骤:

[0055] 步骤 1'、6PE-1 向 ASBR1 通告到达 IPv6 S1 网络的 BGP 路由信息,在通告的 BGP 路由中携带有 6PE-1 确定出的到达该 IPv6 S1 网络的路由的 MTU1;

[0056] 与步骤 S501 相似,6PE-1 通过路由信息中的扩展团体属性携带 MTU1;6PE-1 向 ASBR1 通告的 BGP 路由信息中还包括该 BGP 路由的下一跳的信息,该 BGP 下一跳的信息为 6PE-1 的地址。

[0057] 步骤 2'、ASBR1 根据接收的 BGP 路由信息中的 BGP 下一跳的地址,对该 BGP 路由下一跳对应的 MTU2 进行探测;

[0058] 步骤 3'、ASBR1 比较 MTU1 和 MTU2 的大小,将两者之中较小值作为该 BGP 路由的

MTU3 值并保存。

[0059] 步骤 4'、ASBR1 确定该隧道中还存在下游 BGP 邻居 ASBR2；

[0060] 步骤 5'、ASBR1 向 ASBR2 通告到达 IPv6 S1 网络的 BGP 路由信息，在该 BGP 路由信息中携带 ASBR1 确定出的到达该 IPv6 S1 网络的路由的 MTU3。

[0061] 步骤 6'、ASBR2 接收该 BGP 路由信息之后，与 ASBR1 的处理方式类似，根据该 BGP 路由信息中的 BGP 下一跳 (ASBR1) 的地址，对该 BGP 路由下一跳对应的 MTU4 进行探测，ASBR2 比较 MTU4 和 MTU3 的大小，将两者之中较小值作为该 BGP 路由的 MTU5 并保存。ASBR2 确定该隧道中还存在下游 BGP 邻居 6PE-2，继续向 6PE-2 通告到达 IPv6 S1 网络的 BGP 路由信息，在该 BGP 路由信息中携带 ASBR2 确定出的到达该 IPv6 S1 网络的路由的 MTU5。

[0062] 步骤 7'、6PE-2 接收该 BGP 路由信息之后，与 ASBR2 的处理方式类似，根据该 BGP 路由信息中的 BGP 下一跳 (ASBR2) 的地址，对该 BGP 路由下一跳对应的 MTU6 进行探测；6PE-2 比较 MTU5 和 MTU6 的大小，将两者之中较小值作为该 BGP 路由的 MTU7 并保存。由于 6PE-2 是与 IPv6 S2 网络直接相连的运营商边界设备，因此，MTU 学习的流程到此结束。

[0063] 经过上述流程之后，6PE-2 中保存的是到达 IPv6 S1 网络的路由的各跳 MTU 中的最小值，6PE-2 收到 IPv6 S2 网络发送至 IPv6 S1 网络的 IPv6 报文后，将该报文的大小与保存的 MTU7 的大小进行比较，决定是否需要分片，如果进行分片，分片后的报文的长度是到达 IPv6 S1 网络的路由中各跳的 MTU 的最小值，因此，在 6PE-2 和 6PE-1 之间的 P 设备不会因为报文长度的大小超出了该路由出接口的 MTU 的大小而丢弃该 IPv6 报文，并且在 ASBR2、ASBR1 和 6PE-1 中，也不会再执行报文的分片操作，提高了报文的转发效率。

[0064] 在上述两个实例中，如果连接 IPv6 S1 网络的 R1 或者 6PE-1 发现到达 IPv6 网络的路由对应的出接口的 MTU1 出现变化后，会向其下游邻居 R2 或 ASBR1 通告携带有更新后的 MTU1 的 BGP 路由信息，R2 或 ASBR1 接收到该 BGP 路由信息后，会发现当前接收的 BGP 路由信息中的 MTU1 与保存的 MTU1 值不相等，这时，比较更新后的 MTU1 与本地保存的 MTU2 的大小确定两者之中的较小值，并在确定出的较小值与保存的 MTU3 不相等时，并使用该较小值更新保存的 MTU3；如果是 ASBR1，由于其还存在下游邻居 ASBR2，那么还需要将更新后的 MTU3 携带在到达该 IPv6 的 BGP 路由信息中并通告至 ASBR2，ASBR2 接收到该路由信息后，执行类似的操作，直至 MTU 的计算在整个隧道中得到更新。

[0065] 在上述两个实例中，R2、ASBR1、ASBR2 或 6PE-2 还可以在首次学习 MTU 的过程结束后，周期性对到达该 IPv6 网络的 BGP 路由下一跳对应的 MTU 进行自动探测，并根据自动探测的结果，对 MTU 学习的结果进行更新。以 ASBR1 为例，ASBR1 在当前探测到的 MTU2 与保存的 MTU2 不相等时，则需要将当前探测到的 MTU2 的值与保存的该路由对应的 MTU1 值进行比较确定两者中的较小值，并当两者中的较小值与保存的 MTU3 不相等时，使用该较小值更新保存的 MTU3，将更新后的 MTU3 携带在达到 IPv6 S1 网络的 BGP 路由信息中并通告至 ASBR2，ASBR2 收到后，按照前述更新方法对保存的 MTU 进行更新，直至连接 IPv6 S2 网络的运营商边界设备 6PE-2 也更新了其保存的 MTU7。

[0066] 基于同一发明构思，本发明实施例还提供了一种 MTU 学习装置及路由设备，由于该装置及设备解决问题的原理与前述一种 MTU 学习方法相似，因此该装置和路由设备的实施可以参见方法的实施，重复之处不在赘述。

[0067] 本发明实施例提供的一种最大传输单元 MTU 的学习装置，如图 6 所示，包括：

[0068] BGP 路由信息接收单元 601, 用于接收 IPv6 报文传输隧道中上游边界网关协议 BGP 邻居通告的到达 IPv6 网络的 BGP 路由信息; 所述 BGP 路由信息中携带有所述上游 BGP 邻居确定出的到达 IPv6 网络的路由的第一 MTU;

[0069] MTU 探测单元 602, 用于对接收的所述 BGP 路由信息携带的到达该 IPv6 网络的 BGP 路由下一跳对应的第二 MTU 进行探测;

[0070] 比较单元 603, 用于将比较所述第一 MTU 与探测到的第二 MTU 的大小, 将两者中较小值作为确定出的到达 IPv6 网络的 BGP 路由的第三 MTU 并保存;

[0071] 路由通告单元 604, 用于确定所述隧道中是否还存在下游 BGP 邻居, 若存在, 将所述第三 MTU 携带在到达该 IPv6 的 BGP 路由信息中并通告至所述下游 BGP 邻居。

[0072] 进一步地, 上述路由通告单元 604, 用于自定义 BGP 路由信息中的扩展团体属性, 将自定义的扩展团体属性的类型字段设置为 MTU; 将第三 MTU 的值封装在所述扩展团体属性的 Value 字段中。

[0073] 进一步地, 本发明实施例提供的 MTU 的学习装置, 还包括: MTU 更新单元 605;

[0074] 相应地, 比较单元 603, 还用于在 BGP 路由信息接收单元 601 当前接收的上游 BGP 邻居通告的 BGP 路由信息中的第一 MTU 与保存的第一 MTU 值不相等时, 比较当前接收的所述 BGP 路由信息中的第一 MTU 与保存的第二 MTU 的大小确定两者中的较小值;

[0075] MTU 更新单元 605, 用于当所述较小值与保存的第三 MTU 不相等时, 使用所述较小值更新保存的第三 MTU;

[0076] 路由通告单元 604, 还用于确定所述隧道中是否还存在下游 BGP 邻居, 若存在, 将更新后的第三 MTU 携带在到达该 IPv6 的 BGP 路由信息中并通告至所述下游 BGP 邻居。

[0077] 进一步地, 本发明实施例提供的 MTU 的学习装置中的 MTU 探测单元 602, 还用于在路由通告单元 604 将所述第三 MTU 携带在到达该 IPv6 网络的 BGP 路由信息中并通告至所述下游 BGP 邻居的步骤之后, 周期地对到达该 IPv6 网络的 BGP 路由下一跳的第二 MTU 进行探测;

[0078] 比较单元 603, 还用于在 MTU 探测单元 602 当前探测到的第二 MTU 与保存的第二 MTU 不相等时, 比较当前探测到的第二 MTU 与保存的第一 MTU, 确定两者中的较小值;

[0079] MTU 更新单元 605, 还用于当所述较小值与保存的第三 MTU 不相等时, 使用所述较小值更新保存的第三 MTU 的值;

[0080] 路由通告单元 604, 还用于确定隧道中是否还存在下游 BGP 邻居, 若存在, 将更新后的第三 MTU 携带在到达该 IPv6 的 BGP 路由信息中并通告至隧道中的下游 BGP 邻居。

[0081] 进一步地, 当本发明实施例提供的 MTU 的学习装置位于连接该 IPv6 网络的运营商边界设备中时, 还可以包括:

[0082] 获取单元 606, 用于获取所述运营商边界设备中与到达 IPv6 网络的路由对应的出接口的 MTU;

[0083] 相应地, 路由通告单元 604, 还用于确定所述隧道中是否还存在下游 BGP 邻居, 若存在, 将获取到所述出接口的 MTU 携带在到达该 IPv6 网络的 BGP 路由信息中并通告至所述下游 BGP 邻居。

[0084] 本发明实施例提供的 MTU 的学习装置, 在具体实施时, 可以通过设计下述功能模块实现: BGP 模块、路由表管理模块、MTU 自动探测模块, 其中:

[0085] BGP 模块用于通告 BGP 路由信息, 优选和计算最优路由, 更新路由对应的 MTU 的值以及向路由表安装更新路由表项。

[0086] 路由表管理模块用于管理路由表项, 如确定路由对应的下一跳、出接口, 以及出接口 MTU 的值, 同时也负责向 BGP 模块通告路由更新的信息。

[0087] MTU 自动探测模块用于自动探测到达某个地址的传输 MTU 值, 并向 BGP 模块通告达到某个地址的 MTU 值。

[0088] 为了实现利用上述学习到的 MTU 进行数据转发, 在该装置中, 还可以设置数据转发模块, 用于利用学习到的 MTU 执行数据报文的分片和转发。

[0089] 上述功能模块的划分和实现方式仅是一个具体实例, 本发明实施例提供的 MTU 的学习装置并不限于上述的具体实现方式。

[0090] 本发明实施例还提供了一种路由设备, 该路由设备包含本发明实施例提供的上述 MTU 学习装置。

[0091] 较佳地, 本发明实施例提供的上述路由设备, 在实际组网时, 可以作为 6PE 应用场景中的运营商边界设备或者自治系统边界路由器。

[0092] 本发明实施例提供的最大传输单元的学习方法、装置及路由设备, 对于 IPv4 网络中形成的 IPv6 报文传输隧道中的每个运行 BGP 的路由设备, 执行下述操作即: 接收上游 BGP 邻居通告的到达 IPv6 网络的 BGP 路由信息; 该 BGP 路由信息携带有该上游 BGP 邻居确定出的到达 IPv6 网络的路由的第一 MTU; 对接收的 BGP 路由信息携带的到达该 IPv6 网络的 BGP 路由由下一跳对应的第二 MTU 进行探测; 比较第一 MTU 与探测到的第二 MTU 的大小, 将两者中较小值作为确定出的到达 IPv6 网络的 BGP 路由的第三 MTU 并保存; 如果还存在下游 BGP 邻居, 则将第三 MTU 携带在到达该 IPv6 网络的 BGP 路由信息中并通告至下游 BGP 邻居, 直至该隧道中的每个运行 BGP 的路由设备都学习到了到达该 IPv6 网络的 BGP 路由对应的 MTU。本发明实施例在支持 BGP 协议的运营商边界设备和自治系统边界路由之间通过 BGP 路由传递 MTU 的值, 使得传输 IPv6 报文的隧道的终点即连接另一 IPv6 网络的运营商边界设备可以计算出到达目的 IPv6 网络的路由中各跳的 MTU 的最小值, 并将其作为发送 IPv6 报文是否需要分片的依据, 经过上述过程之后, 在该隧道的终点收到发往该 IPv6 网络的 IPv6 报文时, 将该报文的大小与自身保存的 MTU 的大小进行比较, 决定是否需要分片, 如果进行分片, 分片后的报文的长度是到达该 IPv6 网络的路由中各跳 MTU 的最小值, 因此, 对于通过 IPv4 网络实现 IPv6 网络的互联互通的应用场景尤其是跨多个 IPv4 自治域实现 IPv6 网络的互联互通的应用场景来说, 不论 IPv4 网络中传输 IPv6 报文的隧道需要跨越多少中间设备 (例如 P 设备) 或边界设备, 任何中间设备或者边界设备, 都不会因为 IPv6 报文长度的大小超出了该路由设备出接口的 MTU 的大小而丢弃该 IPv6 报文, 从而避免了现有技术出现的 IPv6 报文的传输受阻的情况, 并且在该隧道的终点对收到的 IPv6 报文进行分片操作后, 该隧道中其他路由设备也不会再执行报文的分片操作, 提高了 IPv6 报文在隧道中转发效率。

[0093] 显然, 本领域的技术人员可以对本发明进行各种改动和变型而不脱离本发明的精神和范围。这样, 倘若本发明的这些修改和变型属于本发明权利要求及其等同技术的范围之内, 则本发明也意图包含这些改动和变型在内。

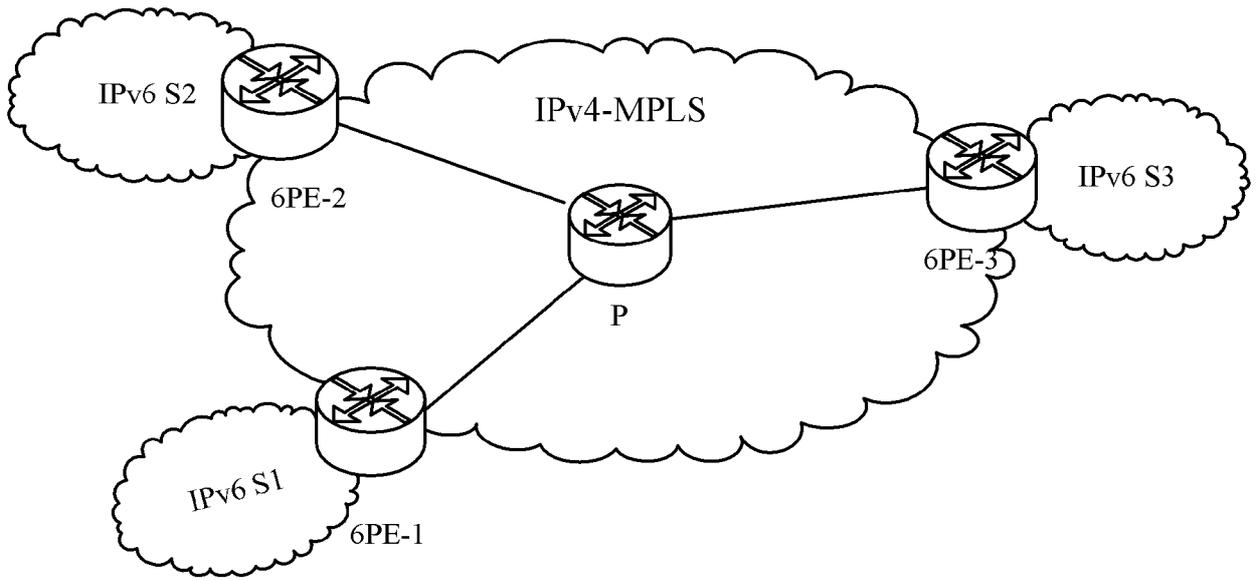


图 1

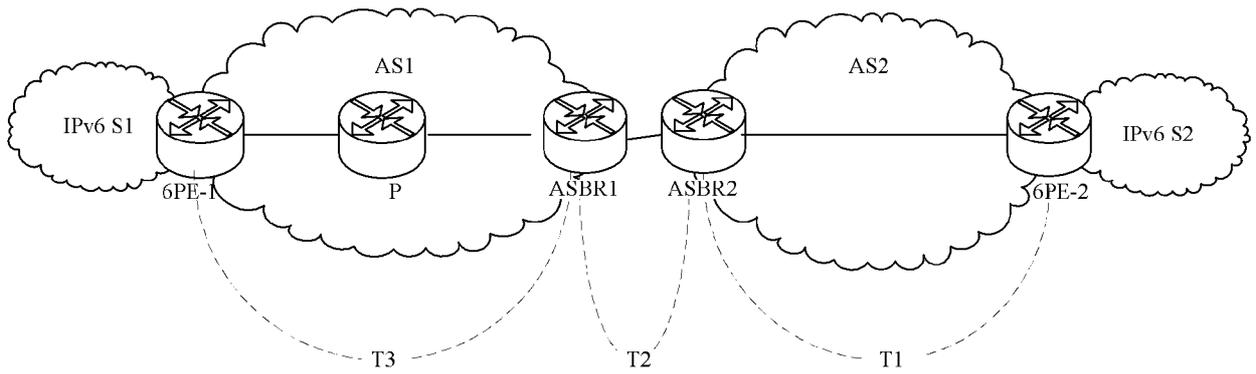


图 2

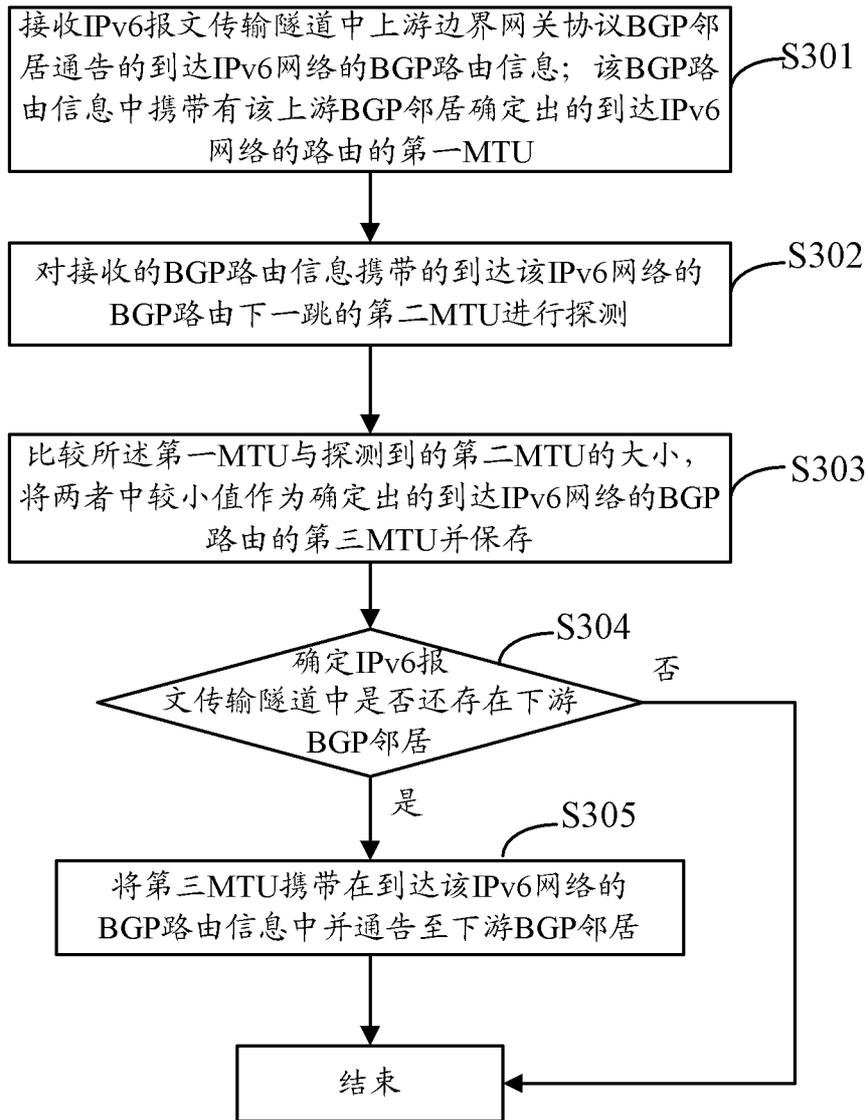


图 3

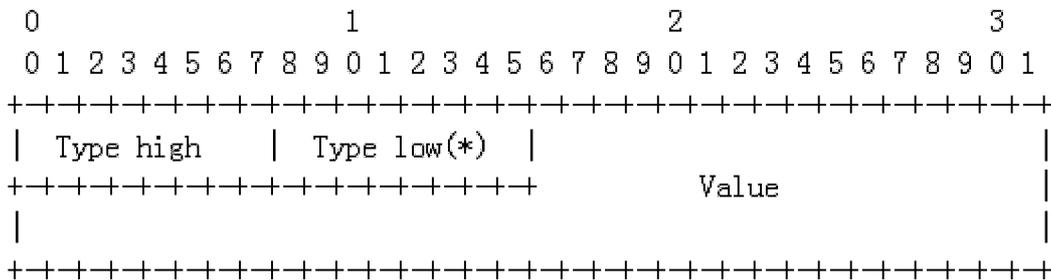


图 4

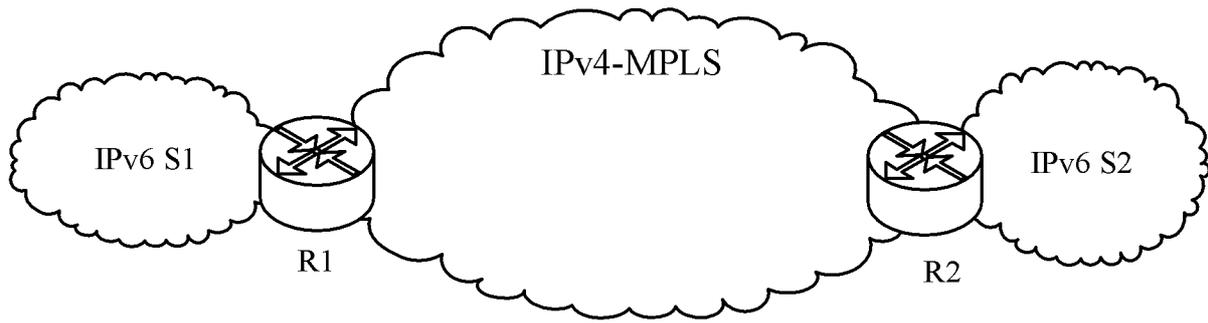


图 5

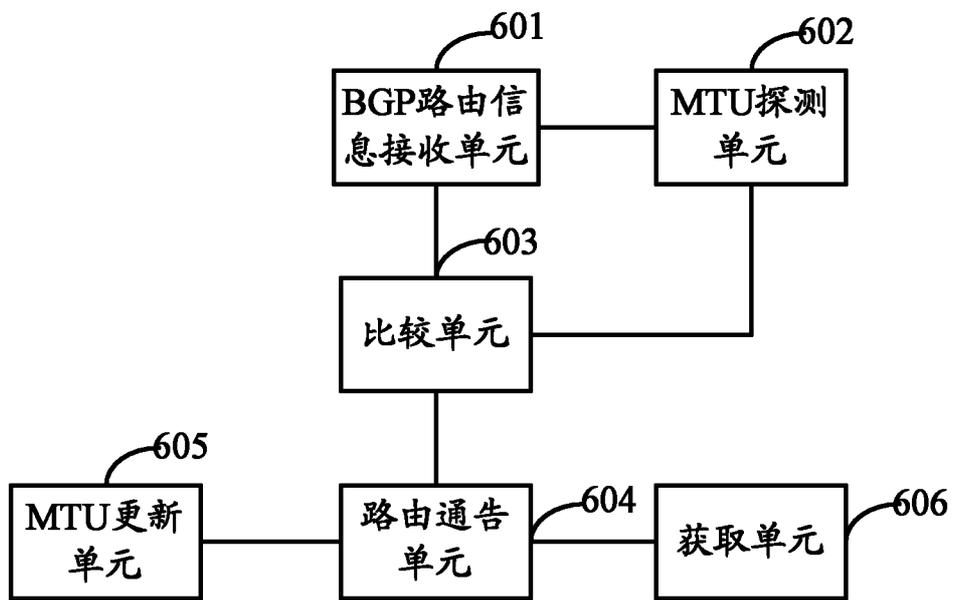


图 6