

ICS 33.040.40
M 32



中华人民共和国通信行业标准

YD/T 3052-2016

虚拟专用局域网服务（VPLS）中基于边界 网关协议（BGP）的多归连接技术要求

BGP based Multi-homing in VPLS Network

2016-04-05 发布

2016-07-01 实施

中华人民共和国工业和信息化部 发布

目 次

前 言	II
1 范围	1
2 规范性引用文件	1
3 术语、定义和缩略语	1
3.1 术语和定义	1
3.2 缩略语	1
4 VPLS 网络里 BGP-MH 的参考网络模型	2
5 对 BGP 协议的扩展	3
5.1 VPLS-ID	3
5.2 RD	4
5.3 MH-ID	4
5.4 PE-ID	4
5.5 PREF	4
5.6 D-bit 和 F-bit	5
6 BGP-MH 的操作过程	5
6.1 DF 选择操作	5
6.2 DF 切换操作	7
6.3 跨自治系统 VPLS 网络的 BGP-MH 操作	9
附录 A（规范性附录）BGP-MH 消息的编码方式	11
附录 B（规范性附录）基于 BGP 的 VPLS 控制协议里 NLRI 的编码格式	13
参考文献	14

前 言

本部分按照 GB/T 1.1-2009 给出的规则起草。

请注意本文件的某些内容可能涉及专利。本文件的发布机构不承担识别这些专利的责任。

本标准由中国通信标准化协会提出并归口。

本标准起草单位：上海贝尔股份有限公司、中国信息通信研究院、中兴通信股份有限公司、华为技术有限公司、迈普通信技术股份有限公司。

本标准起草人：张立新、陈 端、顾方方、马军锋、陈 然、薛 莉。

虚拟专用局域网服务（VPLS）中 基于边界网关协议（BGP）的多归连接技术要求

1 范围

本标准规定了VPLS网络中基于BGP的多归连接技术要求，主要包括需求和场景描述，多归连接的操作方式，包括冗余的多条接入链路和PE的选择机制，接入链路或PE失效后的切换机制等，以及为支持上述功能而对PE间信令协议BGP进行的扩展。该技术要求的适用范围包括CE通过二条或二条以上的接入链路接入二个或二个以上的PE时，在这些冗余的接入链路和PE之间作1:1或1:N的冗余保护，不包括在这些冗余的接入链路和PE之间作1+1或1+N的负载均衡。

本标准适用于运营商提供的基于LDP的VPLS业务，或基于BGP的VPLS业务，或二者混合的VPLS业务，以及设备制造商提供的用于提供上述业务的网络设备。

2 规范性引用文件

下列文件对于本文件的应用是必不可少的。凡是注日期的引用文件，仅所注日期的版本适用于本文件。凡是不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

- IETF RFC 4761 采用BGP作自动发现和信令的虚拟专用局域网业务
- IETF RFC 4762 采用LDP信令的虚拟专用局域网业务
- IETF RFC 6074 二层虚拟专用网络的配置、自动发现和信令

3 术语、定义和缩略语

3.1 术语和定义

下列术语和定义适用于本文件。

3.1.1

选中的转发设备 Designated Forwarder

从CE多归连接的若干PE里选中的、用于转发该CE业务的特定PE。对应于一个多归连接的CE，在任何时刻有且仅有一个DF；其他所有未被选中的PE都成为非DF。另一种等价的表述是把DF视为PE的一个属性，如果某PE被选中用来转发多归连接CE的业务，就称该PE拥有DF。对应于一个多归连接的CE，在任何时刻有且仅有一个PE拥有DF，其他所有未被选中的PE都不拥有DF。

3.2 缩略语

下列缩略语适用于本文件。

AC	Attachment Circuit	接入电路
AFI	Address Family Identifier	地址族标识
ASBR	Autonomous System Border Router	自治系统边界路由器
BGP	Border Gateway Protocol	边界网关协议
BGP-AD	BGP based Auto-Discovery	基于BGP的VPLS自动发现
BGP-MH	BGP based Multi-Homing	基于BGP的多归连接

YD/T 3052-2016

BGP-VPLS	BGP based VPLS	基于BGP的VPLS
CE	Customer Edge	客户边缘设备
DF	Designatd Forwarder	选中的转发设备
D-bit	Down bit	指示接入电路不能正常工作的标志位
F-bit	Flush bit	指示远端PE刷新MAC地址的标志位
LB	Label Base	标签基
LDP	Label Distribution Protocol	标签分发协议
MAC	Media Access Control	媒体访问控制
MH-ID	Multi-Homed site ID	多归连接的客户站点标识
NLRI	Network Layer Reachability Information	网络层可达性信息
PE	Provider Edge	运营商边缘设备
PREF	Preference	优先级
PW	Pseudowire	伪线
RD	Route Distinguisher	路由区别符
RT	Route Target	路由目标
RR	Route Reflector	路由反射器
SAFI	Subsequent Address Family Identifier	后续地址族标识
VBO	VE Block Offset	VE块偏移量
VBS	VE Block Size	VE块大小
VE	VPLS Edge	VPLS边缘设备
VPLS	Virtual Private LAN Service	虚拟专用局域网业务

4 VPLS 网络里 BGP-MH 的参考网络模型

VPSL网络多归连接的参考网络模型如图1所示，其中CE1双归连接到PE1和PE2，CE2和CE3分别单归连接到PE3和PE4。在PE1、PE2、PE3和PE4上运行VPLS业务。本标准对建立VPLS业务的控制协议（LDP或BGP），对控制客户业务流量在伪线之间互相转发的约束规则都不做限制。

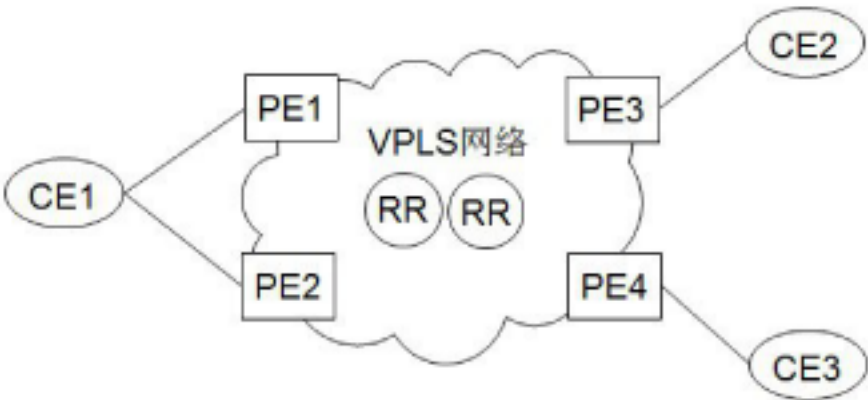


图1 VPLS网络多归连接的参考网络模型

对一个VPLS网络来说，多归连接的冗余控制机制将在PE之间成功建立了伪线连接以后运行，目的是消除多归连接在VPLS实例里引入的环路。以图 1 为例，最终允许的转发拓扑应消除CE1、PE1、PE2之间的环路，即CE1与VPLS网络之间在任何时刻应有且只有一个接入链路和相应的PE处于正常转发状

态，其他的接入链路和相应的PE都应处于备用的阻塞状态。当工作链路或PE出现故障后，业务流量应自动切换到备用链路和PE。

本标准所述的多归连接控制机制基于PE工作，在PE上将运行基于BGP扩展的多归连接控制协议（以下简称BGP-MH协议），以消除CE和VPLS之间的环路。本标准无需在CE上再运行其他的多归连接控制协议，但也不限制CE上运行此类协议。为了本标准所述的机制能够正常工作，要求CE应能支持某种连接检测机制，以便PE能够检测接入链路的可用性；另外还要求CE应能响应PE发出的链路状态通告消息，以便CE能够根据该消息阻塞和/或切换接入链路。例如，CE可支持以太网连接错误管理协议的连接检测和链路状态通知机制，以便与PE协调工作状态。

作为VPLS网络的接入链路和接入节点冗余机制，本标准仅考虑1:1或1:N冗余，即多归连接的多条接入链路和多个接入PE里，仅有一条链路及相应的PE处于正常转发状态，其他的冗余链路和PE都处于阻塞状态。多于一条接入链路和PE处于转发状态的情况，不在本标准的范围内。

5 对 BGP 协议的扩展

为便于与现存的协议和产品相兼容，BGP-MH在BGP-VPLS控制协议上进行扩展，保持BGP消息格式与BGP-VPLS相兼容，对BGP消息里部分原有字段的含义作了新的定义，并定义了若干新的参数。虽然BGP-MH控制协议是在BGP-VPLS控制协议之上进行扩展，但BGP-MH并不要求PE支持BGP-VPLS控制协议。对于仅支持LDP、而不支持BGP作为PE间业务标签分发协议的PE，只要它能够支持本标准所述的BGP-MH控制协议和过程，即可与其他支持BGP-MH的PE互通。更明确地说，BGP-MH与BGP-VPLS是互相独立的协议，虽然二者的协议编码格式互相兼容。

BGP-MH在下列参数上进行操作：VPLS-ID、RD、MH-ID、PE-ID、PREF、D-bit和F-bit等，其中VPLS-ID和RD的含义与BGP-VPLS控制协议保持一致，MH-ID、PE-ID和PREF是在原有字段或参数基础上，根据BGP-MH的需求作了新的或更限定性的定义，D-bit、F-bit是新定义的参数。本章的后续部分将详述这些参数的含义。

上述参数在BGP消息里的编码格式保持与BGP-VPLS控制协议相兼容。详细的BGP-MH消息的编码格式见附录A。

5.1 VPLS-ID

VPLS-ID是6字节的VPLS标识，用于标识一个特定的VPLS实例。在PE属于同一自治系统的情况下，对于自治系统编号为2字节的自治系统，VPLS-ID的前2个字节是自治系统编号，后4个字节由运营商自行分配；对于4字节自治系统编号的自治域，VPLS-ID的前4个字节是自治系统编号，后2个字节由运营商自行分配。

在BGP-MH操作里，VPLS-ID将编码于RT扩展团体属性里。输出RT和输入RT都携带了相同的VPLS-ID，即发出控制消息的PE据此选择属于本VPLS实例的所有PE作为消息目的，接收控制消息的PE也据此选择本VPSL实例的所有PE作为消息源。

对于VPLS跨自治系统的情况，VPLS-ID可能有多种编码方式。见6.3，跨自治系统的VPLS有三种方案。对于方案（a），VPLS-ID（及相应的RT）仍按单自治系统的方式在不同的自治系统内各自独立编码。对于方案（c），VPLS-ID（及相应的RT）应作跨自治系统的全局编码。对于方案（b），VPLS-ID

YD/T 3052-2016

可接单自治系统的方式在不同的自治系统内各自独立编码,然后在ASBR对BGP-MH消息进行跨自治系统的RT属性映射;也可作跨自治系统的全局编码,此时ASBR就无需作跨自治系统的RT属性映射。

跨自治系统的VPLS操作方式(包括VPLS-ID编码方式的约定)由运营商之间协商,本标准不做规定。

5.2 RD

RD是6字节的路由区分符。虽然RD的具体数值不影响DF选择的结果,但如何设置RD值却受DF选择方式的影响。关于RD取值的详细描述,见6.1.1。

5.3 MH-ID

BGP-VPLS控制协议里的VE-ID参数在BGP-MH协议里被重新定义为MH-ID,它是2字节的无符号整数,标识一个多归连接的客户站点。同一CE所连接的各PE所发出的BGP-MH消息里,MH-ID的值应相同。有效的MH-ID值应不为0。

5.4 PE-ID

PE-ID是4字节的发出BGP-MH消息的PE标识,其值是该PE的IPv4系统地址。PE-ID在BGP-MH消息里有三种编码方式:

- 如果发送和接收BGP-MH消息的PE都位于同一自治系统,且BGP-MH消息是不经过RR中继而直达的,那么PE-ID就是发送该BGP-MH消息的PE的BGP标识。
- 如果发送和接收BGP-MH消息的PE都位于同一自治系统,且BGP-MH消息是通过RR中继转发的,那么PE-ID将编码于该BGP-MH消息的ORIGINATOR_ID属性字段里。
- 如果发送和接收BGP-MH消息的PE位于不同的自治系统,那么PE-ID将编码于该BGP-MH消息的Route Origin扩展团体属性字段里。

ORIGINATOR_ID属性和Route Origin扩展团体属性都是BGP-MH消息的可选属性字段。

5.5 PREF

PREF参数是2字节的无符号整数,表示PE被选择为DF的优先级,有效的PREF值应不为0。当同一CE所连接的多个PE作DF选择时,如果不考虑其他可比参数的影响,那么PREF值较大的PE将成为DF。在参与VPLS的PE都位于同一自治系统的情况下,PREF参数编码于BGP-MH消息的LOCAL_PREF属性里。在参与VPLS的PE位于不同自治系统的情况下,由于LOCAL_PREF属性不能穿越自治系统边界,PREF参数将编码于BGP-MH消息的Layer2 Info扩展团体属性里。BGP-MH把BGP-VPLS控制协议所定义的Layer2 Info扩展团体属性里所保留的2字节字段重定义为VPLS Preference字段,用于在跨自治系统的VPLS中进行BGP-MH操作时传送PE优先级信息。

在单自治系统的VPLS中进行BGP-MH操作时,Layer2 Info扩展团体属性的VPLS Preference字段建议仍设置为0,以保持与BGP-VPLS控制协议的一致性,此时PE优先级信息将完全由BGP-MH消息的LOCAL_PREF属性来携带。如果在单自治系统的VPLS中进行BGP-MH操作时,Layer2 Info扩展团体属性的VPLS Preference的值不为0,那么它应与BGP-MH消息的LOCAL_PREF属性的值相同。

通过BGP-MH消息的LOCAL_PREF属性和Layer2 Info扩展团体属性的VPLS Preference字段来唯一提取PREF参数值的计算方法见表1,其中LP是4字节无符号整数,表示BGP-MH消息的LOCAL_PREF属性字段的值;VP是2字节无符号整数,表示BGP-MH消息的Layer2 Info扩展团体属性里VPLS Preference字段的值。如果BGP-MH消息里不含有LOCAL_PREF属性,那么LP=0;如果BGP-MH消息不含有Layer2 Info扩展团体属性,那么VP=0。

表1 计算PREF参数值的方法

VP值	LP值	PREF值	说明
0	0	0	PREF值错误，此PREF值无效
	1~65535	LP	BGP-MH消息的LOCAL_PREF属性字段唯一地编码了PREF参数
	≥65536	65535	4字节LOCAL_PREF属性映射到2字节PREF参数时，所有超出取值范围的值都置为65535
> 0	LP=VP	VP	BGP-MH消息的LOCAL_PREF属性字段和Layer2 Info扩展团体属性的VPLS Preference 字段重复地编码了PREF参数
	LP ≠ VP	0	PREF值错误，此PREF的值无效

5.6 D-bit 和 F-bit

Layer2 Info扩展团体属性的Control Flags字段里设置了二个与BGP-MH操作有关的标志位：D-bit和F-bit。

D-bit指示了在特定VPLS实例里，发送BGP-MH消息的PE与其下连CE之间的接入链路的工作状态。如果此接入链路由于软硬件错误或管理配置等原因而不可转发客户业务帧，那么D-bit=1；如果此接入链路可正常转发客户业务帧，那么D-bit=0。

F-bit指示了在特定的VPLS实例里，发送BGP-MH消息的PE是否拥有DF。如果该PE拥有DF，那么F-bit=1；如果该PE不拥有DF，那么F-bit=0。F-bit从1到0的切换表示发送BGP-MH消息的PE失去了DF，这意味着其他远端PE通过该PE所习得的MAC地址都已失效，应从MAC表里删除这些地址。由此，F-bit的状态变化成为远端PE刷新MAC地址的标志。

6 BGP-MH 的操作过程

VPLS网络的BGP-MH操作包括DF选择操作和DF切换操作二个过程。

同一CE所连接的各PE都发出BGP-MH消息。其他属于同一VPLS实例的所有PE，通过RT扩展属性的过滤，都将收到也只能收到属于同一个VPLS实例的BGP-MH消息，并从中提取出MH-ID、PE-ID、PREF、D-bit和F-bit等参数的值，其中PE-ID、PREF、D-bit的值将用于DF选择，F-bit的值将用于DF切换。

6.1 DF 选择操作

所有VPLS PE之间建立了伪线连接以后，就进入了BGP-MH的DF选择过程。其结果是在CE多归连接的若干PE之间选出一个PE作为DF，负责所有该CE的客户业务帧传送。未被选中为DF的其他PE都不能收发该CE的客户业务帧，应自行阻塞连接该CE的接入链路。被阻塞的接入链路允许在PE-CE间传送运维管理帧。

6.1.1 BGP-MH 中的 DF 选择方式

基于BGP最佳路径选择的DF选择方式

基于BGP最佳路径选择的DF选择方式是利用BGP协议对于冗余路由的选择机制来作DF选择。

连接同一个CE的多个PE所发出的、作VPLS控制的BGP UPDATE消息里，如果NLRI字段里的RD、VE-ID（即MH-ID）、VBO、VBS和LB都相同（其中VBO、VBS、LB全为零），将被接收PE的BGP协议认为是等价路由，并根据LOCAL_PREF属性而选择出最佳路径，即从多个PE中选择一个DF。这种DF选择方式称为基于BGP最佳路径选择的DF选择方式。

在基于BGP最佳路径选择的DF选择方式里，具有相同MH-ID值（即相同VE-ID值）的若干PE所发出的BGP-MH消息里，NLRI字段里的RD应取相同的值，而VBO、VBS、LB应全为零。接收PE基于RT值过

YD/T 3052-2016

滤出属于同一个VPLS实例的所有BGP-MH消息。在同一个VPLS实例里，RD、MH-ID、VBO、VBS、LB全部相同（其中VBO、VBS、LB全为零）的所有BGP-MH消息被视为属于同一个DF选择组。RD的具体数值本身与DF选择结果无关，本标准对RD的具体数值本身不做规定，但要求连接同一个CE的各PE所发出的BGP-MH消息都必须具有相同的RD值。

从BGP协议角度来看，RD、MH-ID、VBO、VBS、LB都相同（其中VBO、VBS、LB全为零）的VPLS控制消息是等价的路由通告，当它们穿越BGP中间节点（例如RR或ASBR）时，可能被选择性过滤或合并，从而可能丢失部分VPLS控制消息，因此基于BGP最佳路径选择的DF选择方式在应用场景上受限。

VPLS DF选择方式

VPLS DF选择方式是BGP-MH所特有的PE选择方式。

连接同一个CE的多个PE所发出的、作VPLS控制的BGP-MH消息里，MH-ID都相同，但RD应取不同的值，这样PE所发出的BGP-MH消息将显式地携带不同的路由信息，从而顺利地穿越所有的中间BGP节点（例如RR或ASBR）。接收PE基于RT值过滤出属于一个VPLS实例的所有BGP-MH消息。在同一个VPLS实例内，MH-ID相同（而不管RD的值不同）的所有BGP-MH消息被视为属于同一个DF选择组。RD的具体数值本身与DF选择结果无关，本标准对RD的具体数值本身不做规定，但要求连接同一个CE的各PE所发出的BGP-MH消息应具有不同的RD值。

从BGP协议角度来看，RD不同、而MH-ID、VBO、VBS、LB都相同（其中VBO、VBS、LB全为零）的VPLS控制消息是不同的路由通告，能够穿越途经的BGP中间节点（例如RR或ASBR），因此VPLS DF选择方式适用于所有场景，包括穿越RR或跨自治系统等。

属于同一VPLS实例的所有PE应采用同一种DF选择模式。如非特殊说明，本标准所述的DF选择算法和DF切换过程适用于上述的二种DF选择方式。

管理员应可配置PE的工作模式。本标准推荐采用VPLS DF选择方式作为PE的默认工作方式。

6.1.2 BGP-MH 中的 DF 选择算法

接收BGP-MH消息的PE以<VPLS-ID, MH-ID>为索引对每个BGP-MH消息所携带的参数（VPLS-ID, MH-ID, PE-ID, PREF, D-bit）进行分组，其中VPLS-ID标识一个VPLS实例，而MH-ID标识该VPLS实例内一个多归连接的客户站点。<VPLS-ID, MH-ID>相同的参数被视为属于同一DF选择组。通过运行下述DF选择算法，将确定一个VPLS实例内每个多归连接客户站点的DF归属。

在每个DF选择组内，PE根据PE-ID、PREF、D-bit的值进行DF选择。选择原则是：D-bit=0者胜过D-bit=1者，PREF值较大者胜过较小者，PE-ID值较小者胜过较大者。

假定D1-bit、PREF1、PE1-ID和D2-bit、PREF2、PE2-ID分别是PE1和PE2所发出的BGP-MH消息里提取的参数。DF选择算法如下：

（1）比较D-bit参数

如果 D1-bit=0 且 D2-bit=1，那么PE1是DF。DF选择过程结束；

如果 D1-bit=1 且 D2-bit=0，那么PE2是DF。DF选择过程结束；

如果比较 D-bit 不能决定DF，则进入（2）比较PREF参数

（2）比较PREF参数

如果 PREF1 > PREF2，那么PE1是DF。DF选择过程结束；

如果 PREF1 < PREF2，那么PE2是DF。DF选择过程结束；

如比较 PREF 不能决定DF, 则进入 (3) 比较PE-ID

(3) 比较PE-ID参数

如果 $PE1-ID < PE2-ID$, 那么PE1是DF。DF选择过程结束;

如果 $PE1-ID > PE2-ID$, 那么PE2是DF。DF选择过程结束;

如果 $PE1-ID=PE2-ID$, 那么这二组参数是来自同一PE的冗余BGP-MH消息。

属于同一个VPLS实例的所有PE都要参与DF选择过程, 包括CE所连接的那些冗余PE。连接同一CE的冗余PE, 如果通过DF选择过程而确定自身拥有DF, 应将PE-CE链路置于正常转发状态, 并通知CE接入链路工作状态正常; 如果通过DF选择过程确定自身不拥有DF, 应将PE-CE链路置于阻塞状态, 并通知CE接入链路已阻塞。PE可采用标准机制来通知CE接入链路的工作状态。例如, 如果CE通过以太网链路接入PE, PE-CE间可采用以太网连接错误管理协议来传递链路工作状态; 如果CE通过伪线接入PE, PE-CE间可采用伪线状态消息机制来传递链路工作状态。本标准不限制PE-CE间采用其他的链路状态消息机制。CE收到来自PE的链路状态消息后, 可选择连接DF的链路作为工作链路。

6.2 DF 切换操作

6.2.1 接入链路故障的 DF 切换操作

6.2.1.1 接入链路故障的 DF 切换场景 1

本场景描述了优先级较大的PE原来是DF, 其接入链路出现故障后的DF切换过程。

参见图1, 设PE1的优先级大于PE2。正常情况下, PE1-CE1链路和PE2-CE1链路都没有故障, 可转发客户业务帧。PE1是DF, PE2不是DF。PE1-CE1链路处于活动状态, 允许转发客户业务帧; PE2-CE1链路处于阻塞状态, 不允许转发客户业务帧, 但可转发运维管理帧。当PE1-CE1链路出现故障后, PE1通过某种机制检测出链路故障, PE1发出BGP-MH消息, 并在Layer2 Info扩展团体属性里设置D-bit=1, F-bit=0, 指示PE1-CE1链路故障, PE1从DF切换为非DF。PE1应立即发出此BGP-MH消息, 而无需等待BGP最小路由通告时间间隔。

如果此VPLS实例是基于LDP建立的, PE1随后将向所有的对等LDP协议体显式地发出MAC地址刷新消息, 撤销所有与PE1-CE1链路有关的MAC地址。如果此VPLS是基于BGP建立的, 那么PE1无需显式地发出MAC地址刷新消息, 上述的BGP-MH消息(F-bit从1变成0, 指示PE1已失去DF)将作为隐含的MAC地址刷新信号。

PE2收到上述BGP-MH消息后, 立即执行DF选择过程, 结果是PE2拥有DF。PE2发出BGP-MH消息, 在Layer2 Info扩展团体属性里设置D-bit=0, F-bit=1, 指示PE2-CE1链路正常, PE2从非DF切换成DF。PE2应立即发出此BGP-MH消息, 无需等待BGP最小路由通告时间间隔。PE2还应通过某种状态通告机制向CE1发出解除PE2-CE1链路阻塞的信号。

远端的PE3和PE4收到来自PE1的、指示PE1失去DF的BGP-MH消息后, 重新进行DF选择过程, 根据D-bit 的值选择PE1为非DF, PE2为DF。如果此VPLS实例是基于LDP建立的, PE3和PE4随后还将收到来自PE1的、显式的MAC地址撤销消息, 并根据其携带的失效MAC地址信息, 从VPLS实例的MAC转发表里删除失效的MAC地址。如果此VPLS实例是基于BGP建立的, PE3和PE4将把来自PE1的、指示PE1失去DF的BGP-MH消息视为隐含的MAC地址刷新信号, 在执行DF选择过程后, 将立即从VPLS实例的MAC转发表里删除所有指向PE1的MAC地址。

PE1、PE3和PE4都会收到PE2发出的、指示PE2从非DF切换为DF的BGP-MH消息，也将重新进行DF选择过程，但不影响DF归属情况，也不引起MAC地址刷新过程。

6.2.1.2 接入链路故障的 DF 切换场景 2

本场景描述了优先级较大的PE原来不是DF，其接入链路从故障恢复后的DF切换过程。

参见图1，初始状态下，PE1的优先级大于PE2，PE1-CE1链路因软硬件故障而不能正常工作，PE2-CE1链路可正常工作。PE1不是DF，PE2是DF。当PE1-CE1链路因故障消除而恢复正常工作后，PE1将立刻通过DF选择过程而从非DF切换为DF，PE1发出BGP-MH消息，并在Layer2 Info扩展团体属性里设置D-bit=0，F-bit=1，指示PE1-CE1链路恢复正常，PE1从非DF切换为DF。PE1应立即发出此BGP-MH消息，无需等待BGP最小路由通告时间间隔。

PE2收到上述来自PE1的、指示PE1从非DF切换为DF的BGP-MH消息后，立即执行DF选择过程，结果是PE2失去DF。PE2发出BGP-MH消息，在Layer2 Info扩展团体属性里设置D-bit=0，F-bit=0，显式地指示PE2-CE1链路仍正常，但PE2失去DF。如果此VPLS实例是基于LDP建立的，PE2随后将向所有的对等LDP协议体显式地发出MAC地址撤销消息，撤销所有与PE2-CE1链路有关的MAC地址。如果此VPLS是基于BGP建立的，那么PE2无需显式地发出MAC地址撤销消息，上述的BGP-MH消息（F-bit从1成0，指示PE2失去DF）将作为隐含的MAC地址刷新信号。PE2应立即发出此BGP-MH消息，无需等待BGP最小路由通告时间间隔。PE2还应通过某种状态通告机制向CE1发出阻塞PE2-CE1链路的信号。

远端的PE3和PE4收到来自PE1的、指示PE1拥有DF的BGP-MH消息后，重新进行DF选择，根据PE的优先级选择PE1为DF，PE2为非DF。PE3和PE4又收到来自PE2的、指示PE2失去DF的BGP-MH消息后，再一次进行DF选择，DF归属不变。如果此VPLS实例是基于LDP建立的，PE3和PE4在收到来自PE2的显式MAC地址撤销消息后，从VPLS实例的MAC转发表里删除失效的MAC地址。如果此VPLS实例是基于BGP建立的，PE3和PE4将把来自PE2的、指示PE2失去DF的BGP-MH消息视为隐含的MAC地址刷新信号，在执行DF选择后，立即从VPLS实例的MAC转发表里删除所有指向PE2的MAC地址。

PE1、PE2分别作DF切换的进度可能不一致，可能出现PE1、PE2同时认为自己是DF的一段时间间隙。例如，PE1已从非DF切换为DF，且发出了BGP-MH消息，PE3和PE4也收到了PE1发出的BGP-MH消息，并重新选择PE1为DF。但此时PE2尚未从DF切换为非DF，仍然维持着与CE的活动连接。为确保在这种情况下，PE3和PE4能够正确地作MAC地址刷新，规定PE3和PE4只有在收到显式的或隐含的MAC地址刷新信号后，才能执行MAC地址刷新过程。具体地，在基于LDP的VPLS里，PE3和PE4只有在收到基于LDP的MAC地址撤销消息后（PE2只有在完成了从DF切换为非DF的过程后才发出MAC地址撤销消息），才能执行相应的MAC地址刷新过程。在基于BGP的VPLS里，PE3和PE4只有在收到PE2失去DF（F-bit从1变为0）的BGP-MH消息后（PE2只有在完成了从DF切换为非DF的过程后才发出此消息），才能执行相应的MAC地址刷新过程。

6.2.2 PE 节点故障的 DF 切换操作

6.2.2.1 PE 节点故障的 DF 切换场景 1

本场景描述了优先级较大的PE原来是DF，该PE出现故障后的DF切换过程。

参见图1，PE1的优先级大于PE2，PE1-CE1链路和PE2-CE1链路都没有故障，可转发客户业务帧。正常情况下PE1是DF，PE2不是DF。当PE1节点出现故障时，PE2、PE3和PE4通过某种快速节点故障检测机制检测到PE1已停止工作，于是PE2、PE3、PE4与PE1之间的伪线全部切换到不可工作状态，引起VPLS

实例的MAC转发表里所有映射到指向PE1伪线的MAC地址都被删除。PE2、PE3、PE4将重新执行DF选择，PE2被选为DF。PE2将发出BGP-MH消息，其中Layer2 Info扩展团体属性的D-bit=0，F-bit=1。

6.2.2.2 PE 节点故障的 DF 切换场景 2

本场景描述了优先级较大的PE原来不是DF，该PE从故障恢复后的DF切换过程。

参见图1，PE1的优先级大于PE2，PE1-CE1链路和PE2-CE1链路都没有故障，可转发客户业务帧。PE1原来处于故障状态，PE2是DF。现PE1从故障中恢复，待PE1恢复（或重建）了与其他PE之间的所有伪线连接后，PE1将发出BGP-MH消息，其中Layer2 Info扩展团体属性的D-bit=0，F-bit=1。PE2接收到来自PE1的BGP-MH消息后，执行DF选择，从DF切换为非DF，然后PE2发出BGP-MH消息，其中Layer2 Info扩展团体属性的D-bit=0，F-bit=0。如果此VPLS实例是基于LDP建立的，PE2随后将向所有的对等LDP协议体显式地发出MAC地址撤销消息，撤销所有与PE2-CE1链路有关的MAC地址。如果此VPLS实例是基于BGP建立的，那么PE2无需显式地发出MAC地址刷新消息，上述的BGP-MH消息（F-bit从1变成0，指示PE2失去DF）将作为隐含的MAC地址刷新信号。PE2可立即发出此DF切换信号，无需等待BGP最小路由通告时间间隔。

远端的PE3和PE4收到来自PE1和PE2的BGP-MH消息后，都将选择PE1为DF，PE2为非DF。如果此VPLS实例是基于LDP建立的，PE3和PE4将等待PE2发出显式的MAC地址撤销消息后，才能在MAC转发表里撤销受影响的MAC地址。如果此VPLS实例是基于BGP建立的，PE3和PE4在收到PE2发出的、显式指示PE2失去DF（F-bit从1变为0）的BGP-MH消息后，即可在MAC表里删除所有指向PE2的MAC地址。

6.3 跨自治系统 VPLS 网络的 BGP-MH 操作

跨自治系统VPLS网络多归连接的参考模型如图2所示。PE1、PE2、ASBR1和ASBR2位于自治系统AS1，PE3、PE4、ASBR3和ASBR4位于自治系统AS2，AS1和AS2通过ASBR1、ASBR2、ASBR3和ASBR4作冗余的E-BGP连接。在跨自治系统的PE1、PE2、PE3和PE4上运行VPLS业务，CE1双归连接至AS1的PE1和PE2上，CE3和CE4分别连接至AS2的PE3和PE4上。

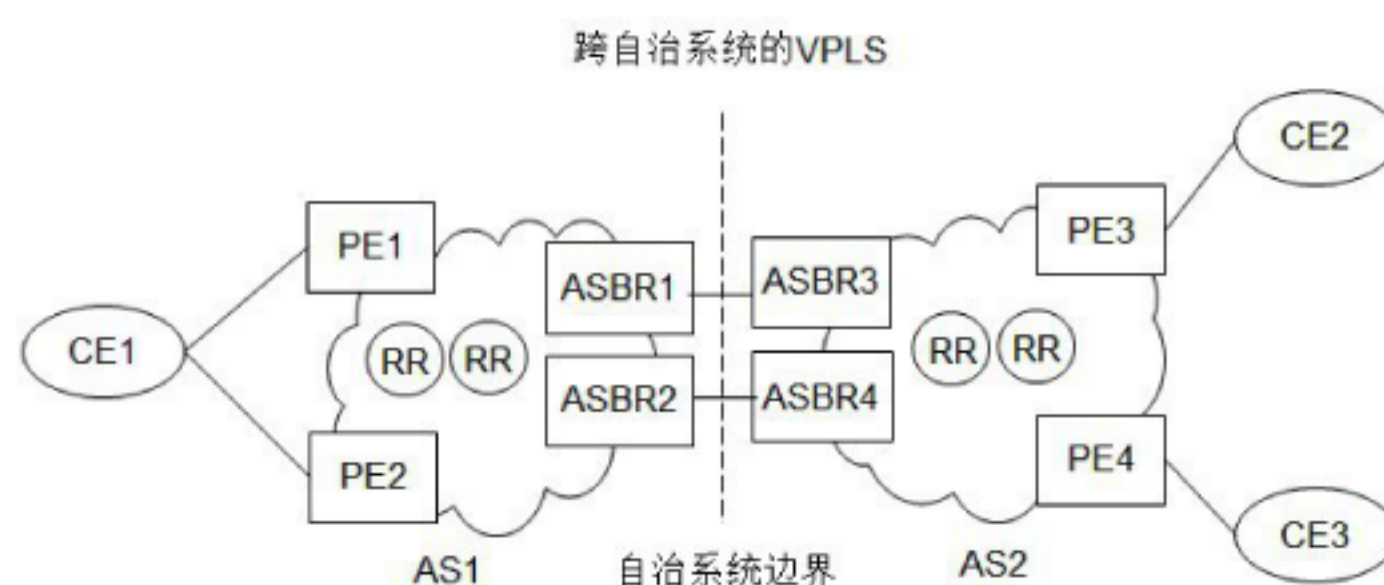


图2 跨自治系统VPLS网络多归连接的参考网络模型

跨自治系统VPLS的BGP-MH有（a）、（b）、（c）等三种方案，分别描述如下：

方案（a）：

方案（a）在不同自治系统的ASBR之间作背靠背的以太网链路连接，VPLS信令无需穿越自治系统边界，各自治系统的BGH-MH消息分别终结于各自的ASBR。本方案的BGP-MH操作与单自治系统的情况相同，ASBR在本自治系统里作PE，而对于对等连接的其它自治系统ASBR，则相当于CE。如果跨自治系统ASBR是冗余配置的，为避免在跨自治系统的多条以太网链路上形成环路，应在这些冗余的跨自治系统以太网链路上运行某种环路检测和阻塞机制。本标准对此机制不作具体规定。

方案（b）：

方案（b）里，ASBR参与VPLS信令过程，在不同自治系统之间中继转发BGP-MH消息。ASBR还需支持基于MPLS来中继连接不同自治系统内的MPLS传送隧道，ASBR之间的连接链路可以是任意的支持MPLS的物理链路。AS1的PE1（或PE2）通过I-BGP发送BGP-MH消息，ASBR1（或ASBR2）通过I-BGP接收到本自治系统的BGP-MH消息以后，通过E-BGP转发到AS2里对等连接的ASBR3（或ASBR4）。ASBR1（或ASBR2）通过E-BGP转发的BGP-MH消息里，NEXT_HOP改为ASBR1（或ASBR2）的IPv4系统地址，NLRI保持不变（RD、MH-ID不变），LOCAL_PREF属性将被删除，而PE优先级信息将重新编码于Layer2 Info扩展团体属性的VPLS Preference字段里，PE1（或PE2）的IPv4系统地址将重新编码于Route Origin扩展团体属性里。ASBR3（或ASBR4）接收到来自AS1的ASBR1（或ASBR2）的BGP-MH消息后，通过I-BGP在AS2转发。ASBR3（或ASBR4）在AS2转发的BGP-MH消息里，NEXT_HOP改为ASBR3（或ASBR4）的IPv4系统地址，NLRI保持不变（RD、MH-ID不变），Route Origin扩展团体属性也保持不变，并增加LOCAL_PREF属性，其中重新编码了来自Layer2 Info扩展团体属性的VPLS Preference字段的PE优先级信息，而Layer2 Info扩展团体属性的VPLS preference字段的PE优先级编码可保留也可删除（即重置为0）。AS2的PE3（或PE4）接收到ASBR3（或ASBR4）转发的、来自AS1的PE1（或PE2）的BGP-MH消息后，即可进行DF选择。以上的I-BGP消息都可直接地或通过RR在一个自治系统内发送。根据VPLS-ID在跨自治系统时是否作全局编码，ASBR可能需要对RT扩展团体属性进行跨自治系统的重映射。

方案（c）：

方案（c）里，VPLS信令透明地穿越自治系统的边界，ASBR不参与VPLS信令过程，但需参与建立跨自治系统的MPLS传送隧道。本方案里，不同自治系统的PE之间作E-BGP多跳连接，通过多跳E-BGP在不同自治系统的PE之间直接发送BGP-MH消息。PE通过多跳E-BGP接收到来自其它自治系统的BGP-MH消息后，即可进行DF选择。本方案要求把PE优先级信息编码于Layer2 Info扩展团体属性的VPLS Preference字段。为增强扩展性，可在各自治系统内分别设置RR，而在不同自治系统的RR之间建立跨自治系统的多跳E-BGP连接。本方案要求VPLS-ID在跨自治系统时作全局编码。

在跨自治系统VPLS的BGP-MH方案（a）、（b）和（c）里，如果在BGP-MH消息的传送路径上有BGP中间节点（如ASBR或RR）作中继转发，就应采用VPLS DF选择方式。

附录 A
(规范性附录)
BGP-MH消息的编码方式

A.1 BGP-MH消息的编码格式

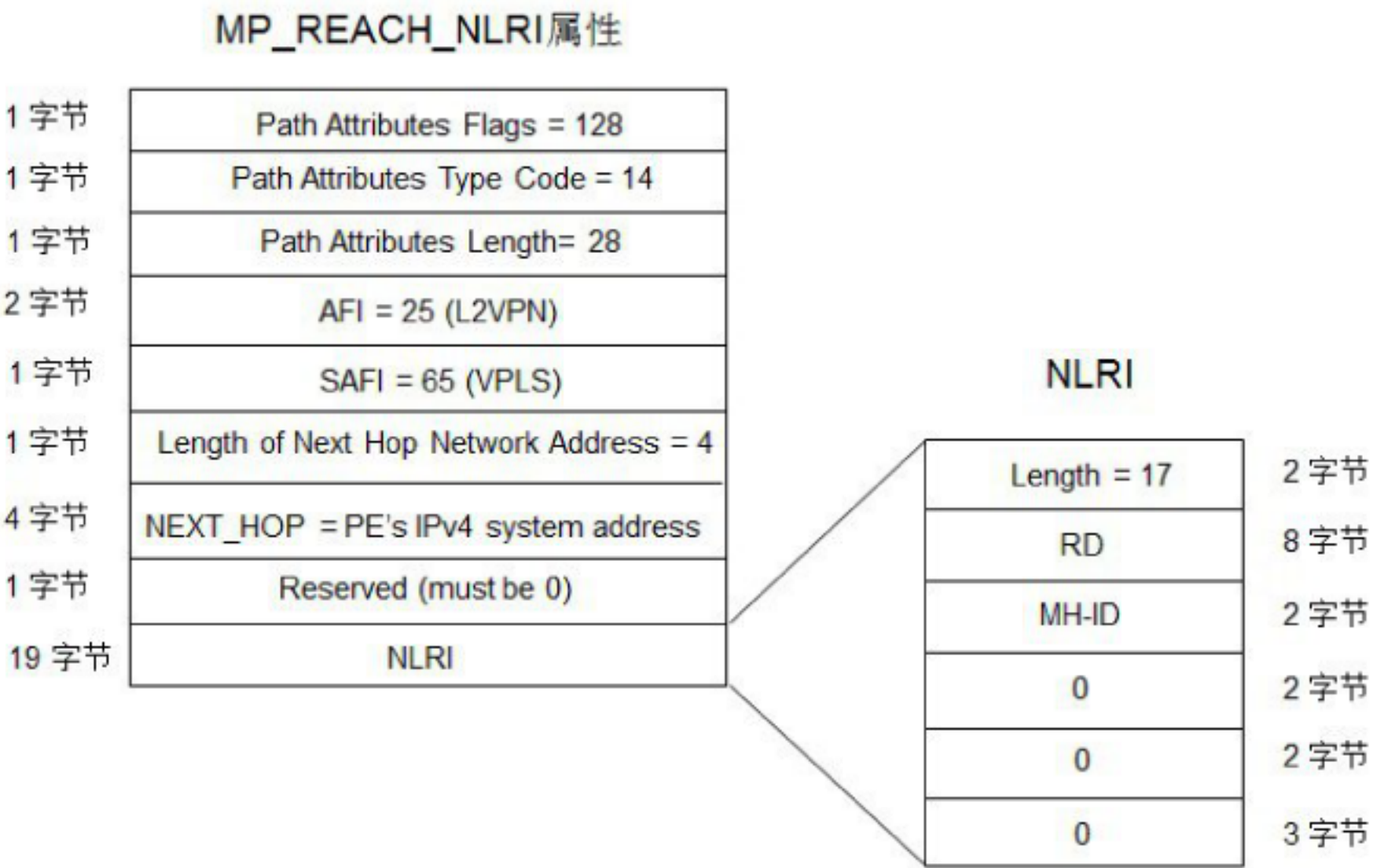
BGP-MH消息是特殊编码的BGP UPDATE消息。PE通过发送一个BGP-MH消息即可完整地传送所有与BGP-MH操作相关的参数。BGP-MH消息的组成如图A.1所示。

Marker (全为0)
BGP Message Length
BGP Message Type = 2
Withdrawn Routes Length = 0
Total Path Attributes Length
LOCAL_PREF (必选)
MP_REACH_NLRI (必选)
Route Target Extended Community (必选)
Layer2 Info Extended Community (必选)
ORIGINATOR_ID (可选)
Route Origin Extended Community (可选)

图A.1 BGP-MH消息的组成

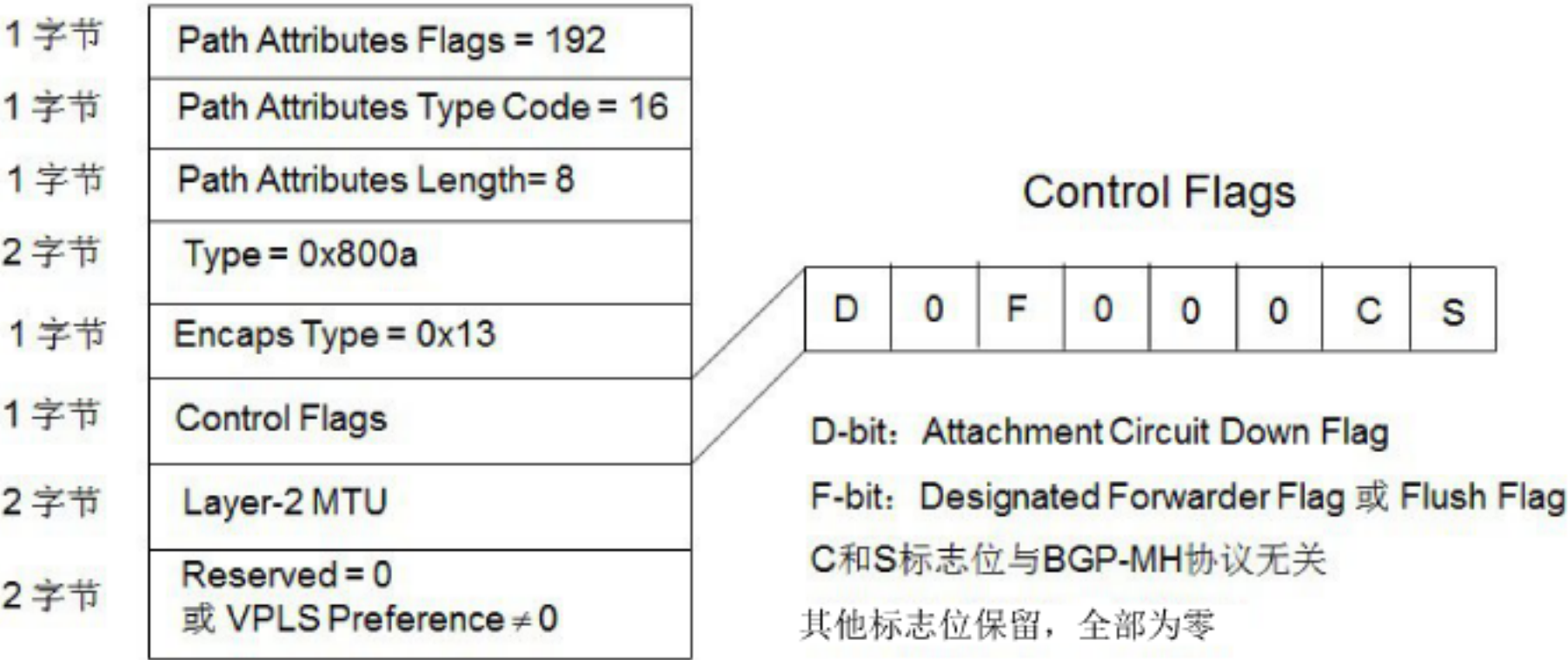
A.2 BGP-MH消息里MP_REACH_NLRI和Layer2 Info扩展团体属性的编码格式

BGP-MH消息里MP_REACH_NLRI属性的编码格式如图A.2所示，Layer2 Info扩展团体属性的编码格式如图A.3所示。



图A.2 BGP-MH消息里MP_REACH_NLRI属性的编码格式

Layer2 Info Extended Community属性



图A.3 BGP-MH消息里Layer2 Info扩展团体属性的编码格式

附录 B
(规范性附录)

基于BGP的VPLS控制协议里NLRI的编码格式

目前基于BGP的VPLS控制协议有BGP-AD、BGP-VPLS和BGP-MH等三种，它们都是通过BGP UPDATE消息来发送相应的控制信息。这三种控制协议是互相独立的，PE可选择支持其中的任意一种、两种或全部。在这三种BGP UPDATE消息里，MP_REACH_NLRI属性在编码格式上的区别完全在于NLRI字段。对于同时支持BGP-AD、BGP-VPLS和BGP-MH三种控制协议的PE，当收到一个BGP UPDATE消息时，应通过识别NLRI字段的编码格式来区分这些不同类型的控制信息。

从VPLS控制的BGP UPDATE消息里（MP_REACH_NLRI属性的AFI=25，SAFI=65），PE可提取出NLRI字段。如果NLRI字段里的Length值为12，那么这个BGP UPDATE消息是BGP-AD消息。如果Length值为17，那么将进一步根据VE Block Offset，VE Block Size和Label Base字段是否为零来判断它是BGP-VPLS消息还是BGP-MH消息。如上述字段不为零，那么它是BGP-VPLS消息；如果上述字段全为零，那么它是BGP-MH消息。

BGP-AD、BGP-VPLS和BGP-MH消息里NLRI字段的编码格式如图B.1所示。

BGP-AD消息里的NLRI	BGP-VPLS消息里的NLRI	BGP-MH消息里的NLRI
2字节: Length = 12	2字节: Length = 17	2字节: Length = 17
8字节: RD	8字节: RD	8字节: RD
4字节: PE系统地址	2字节: VE-ID	2字节: MH-ID
	2字节: VE Block Offset	2字节: 0
	2字节: VE Block Size	2字节: 0
	3字节: Label Base	3字节: 0

图B.1 BGP-AD、BGP-VPLS和BGP-MH里NLRI的编码格式

参 考 文 献

- [1] IETF draft-ietf-l2vpn-vpls-multihoming, BGP based Multi-homing in Virtual Private LAN Service