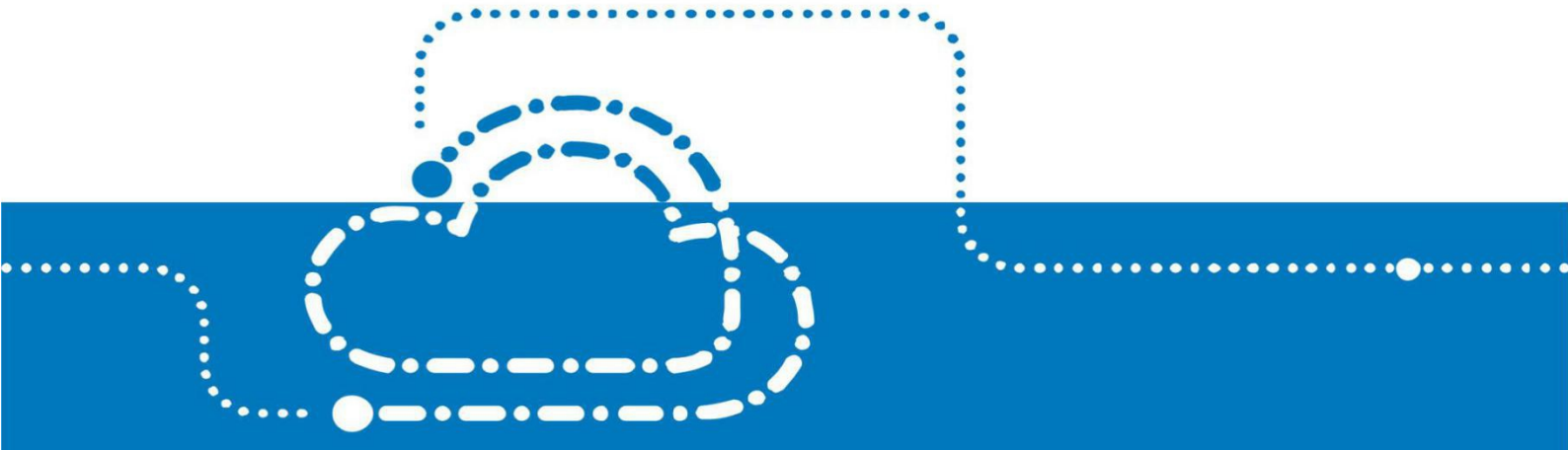
ZTE中兴

BIER 组播技术白皮书



BIER 组播技术白皮书

版本	日期	作者	审核者	备注
V1.0	2020/08/10	ZTE	ZTE	新建
V1.1	2021/03/01	ZTE	ZTE	更新

关键词：无状态、组播、VPN、BIER

摘要：BIER 是新型的无状态组播技术，在组播流量的入节点和出节点保留组播状态信息，中间节点不感知组播流，不建立组播转发树，不维护组播转发状态信息。BIER 组播非常适合大规模组播业务的部署场景，如组播 VPN 业务，IPTV/OTT 等业务。BIER 组播元通过 BFR-prefix 携带 BFR-ID、SD、BSL 以及封装等关键信息。BFR-Prefix 通过传统的 IGP 协议如 ISIS/OSPFv2/OSPFv3 实现全网的洪泛。网络上每台 BIER 路由器通过这些关键信息建立 BIER 转发表，实现对 BIER 封装的报文转发。

目录

图 目录

1. 传统有状态组播的技术局限性

1.1. 传统有状态的组播不适应大规模的组播应用

传统的组播协议如 PIM-SM/PIM-DM 等，为每个组播(Group)建立一个从源到接受者的组播发布树。组播发布树中的每个节点(路由器)维护组播转发状态信息：(Group、Ingress 接口、Egress 接口)。在 IPTV 系统中，一个组播 Group 对应一个 TV 频道，一个大型的 IPTV 系统支持几百个甚至几千订阅频道。传统的组播路由协议为每个组播建立对应的组播发布树。网络中每台路由器对应维护几百到几千的组播转发状态信息,消耗了路由器的宝贵资源如 CPU TCAM 等,现网中的老设备可能面临相当大的压力。当组播订阅者或者网络的拓扑发生变化，导致 IGP 协议重新收敛，IGP 后协议收敛后组播协议才能再次收敛，再重新计算出每个 Group 的组播发布树。组播发布树收敛的时间远大于 IGP 协议的收敛时间。

1.2. 传统有状态组播不适应大型 VPN 组播的部署

大型运营商通过 MPLS L3VPN 给不同的客户提供虚拟专网的业务,每个 VPN 都有独立的地址空间，客户运行独立的单播和组播协议如 ISIS 和 PIM。但运营商不能为所有 VPN 客户都提供大规模的组播服务,主要因为运营商的中间 P 设备无法维护每客户每 VPN 的组播转发状态。如果 N 个 VPN 客户且每个 VPN 客户有 N 个组播业务,运营商的 P 路由器需维护 N^2 个组播流状态信息。运营商网络拓扑变化或者组播源/组播接收者发生变化，每台 P 路由器都要重新计算每个 VPN 的每个组播的组播发布树，这个导致 VPN 组播路由收敛很慢，比全局的组

播路由收敛更慢，严重影响组播 VPN 业务体验。为了减少 P 设备维护的 VPN 组播状态信息的数量，运营商使用各种 VPN 组播技术，都没有彻底解决问题。运营商使用组播头端复制技术，P 节点没有组播状态信息，但要求组播的 PE 节点大带宽和高性能。运营商使用 VPN 组播树的聚合技术，可以减少 P 节点的组播状态的数量，但是会导致 VPN 组播非最优路由，浪费广域网带宽。

1.3. 传统有状态组播不符合网络简化的演进方向

传统有状态的组播使用 mLDP、P2MP、MP2MP 等技术来转发组播流量，这些技术需要部署复杂的 RSVP 协议、LDP 协议等。目前网络向 SRv6 的技术演进趋势，为单播构建一个至简网络，不再部署 LDP、RSVP 等协议或者信令。传统的有状态的组播，无法支持 SRv6 这样的至简网络，不符合网络发展的趋势。

2. 新型无状态 BIER 组播的技术优势

BIER 是一种新型组播转发技术,BIER 为每个组播报文封装了一个 BIER 报文头，BIER 组播接收者信息封装在 BIER 报文头中。 BIER 路由器根据 BIER 报文头中信息转发 BIER 组播报文，不维护每个组播转发状态信息。BIER 封装将具体组播业务与网络层隔离，网络上 P 路由器不再维护每 VPN 每组播的转发状态，BIER 路由器完全不感知上层组播业务，实现 P 路由器对组播的无状态。BIER 组播无状态特性消除大规模部署组播业务对网络的压力。

BIER 组播使用传统的链路状态 IGP 协议的 BIER 扩展，如 ISIS 或者 OSPF 的 BIER 扩展、通过 BIER-prefix 前发布 BIER 的关键信息如 BFR-id、SD、BSL 等洪泛到网络上。每台 BIER 路由器根据这些关键信息生成 BIER 的转发表，

BIER 路由器根据 BIER 转发表转发组播报文，不再查询组播转发表。

3. BIER 的基本原理和架构

3.1. BIER 的基本原理

BIER 是“Bit Index Explicit Replication”的简称，是一种基于位索引显式复制的新型组播技术。BIER 不同于传统的 PIM 组播协议，提供一种无状态的组播转发机制。BIER 在组播首节点（BIER Ingress）确定组播的接收者(BIER Egress)信息，中间节点不需要维护任何组播流转发状态信息（Group、Ingress、Egress），BFIR 是最靠近组播源的 BIER 路由器。BIER 本地转发表根据 IGP 的 BIER 链路状态库计算生成，BIER 链路状态库则由 IGP（ISIS/OSPF）协议的 BIER 扩展洪泛生成。

BIER 基本原理简单高效，每台 BIER 路由器可分配一个不重复的无符号的整数，称之 BFR-id，唯一标识该 BIER 路由器。每台 BIER 路由器都通过特定前缀（BFR-prefix）携带 BFR-id、SD、BSL、封装类型、BIFTBIFT-ID 等重要信息在 IGP 中洪泛。BFR-Prefix 通常为本地的 Loopback 的主机地址。每台 BIER 路由器根据 IGP 算法或者 BIER 算法计算到达其他 BFR-id 的最优路径的 BIER 转发表，类似 IPv4/IPv6 转发表计算生成过程。BIER 组播设计一个特定长度的比特串(BitString)来表示一组 BIER 路由器，BitString 的最第低位开始(right most)，每个比特位对应一个 BFR-id，如使用 BitString 长度（BSL）为 3 的二进制串“101”表示 BFR-id 为 1 和 3 的两台 BIER 路由器，二进制串“011”表示 BFR-id 为 1、2 的两台 BIER 路由器。BSL 长度为 5 的二进制串 “00011”

也可表达 BFR-id 为 1、2 的两台 BIER 路由器。不同的 BSL 影响 BIER 报文头载荷效率，BSL 越大效率越低，但可表达的 BIER 路由器数量越多。目前 RFC8279 标准中要求所有 BIER 路由器都必须支持 BSL 为 256 的值。一个 BIER 路由器可支持多个不同的 BSL，也可在不同网络中使用不同的 BSL。

大型 BIER 网络可以根据网络拓扑或者地理位置设计多个 SD(Sub Domain) 简化管理，如全国性的运营商设立如东部大区 SD 网络、西部大区 SD 网络、南部大区 SD 网络、北部大区 SD 网络，缺省也可以只有一个 SD。每个 SD 内的 BSL 和 BFR-id 是独立的，互不影响。在每个 SD 内，引入 SI（sub identify）概念，用更短长度的 BSL 表达出更多的 BFR-id，如东部大区的 256 台的 BIER 路由器，使用 4 个不同 SI 和 BSL 为 64 的 BitString 来表示，每个 SI 表示东部大区内一个省的 64 台 BIER 路由器，不再需要长度为 256 的 BitString 来标识。SD 内使用 SI 可以减少 BSL 的长度，BIER 报文的负荷效率更高，但每个 SI 都对应节点内一张 BIER 的转发表。

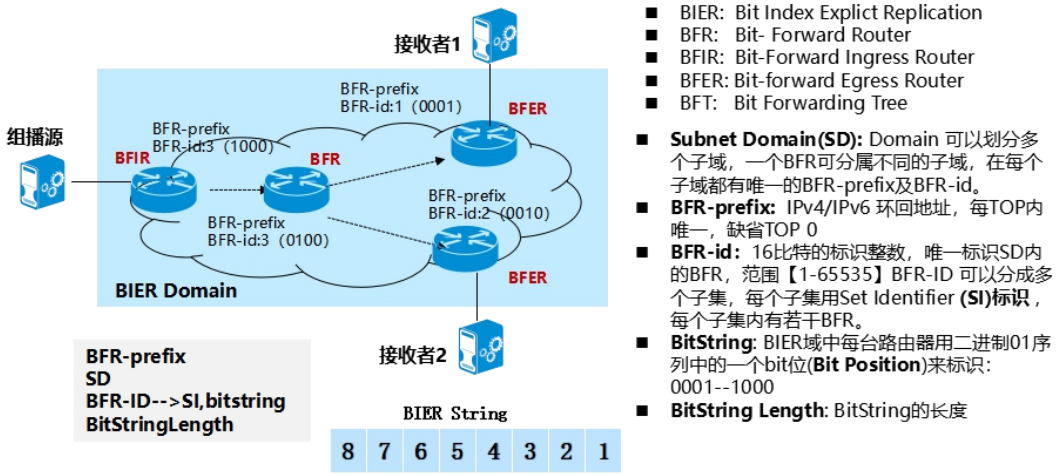


图 1　BIER 的基本概念

3.2. BIER 的三层架构设计

3.2.1. Overlay 层

IETF RFC8279 将 BIER 组播架构分为 Overlay、BIER、Underlay 三层。Overlay 层负责组播业务控制面信息交互，如 BIER Egress 节点和 BIER Ingress 节点之间用户组播的加入和离开、组播流进入和离开 BIER 域的封装和解封装转发等。 Overlay 层可以通过 SDN、MP-BGP（MVPN）、PIM、BMLD（MLD 协议的 BIER 扩展）、静态配置等方式实现，其中 MP-BGP 和 SDN 最为常见。

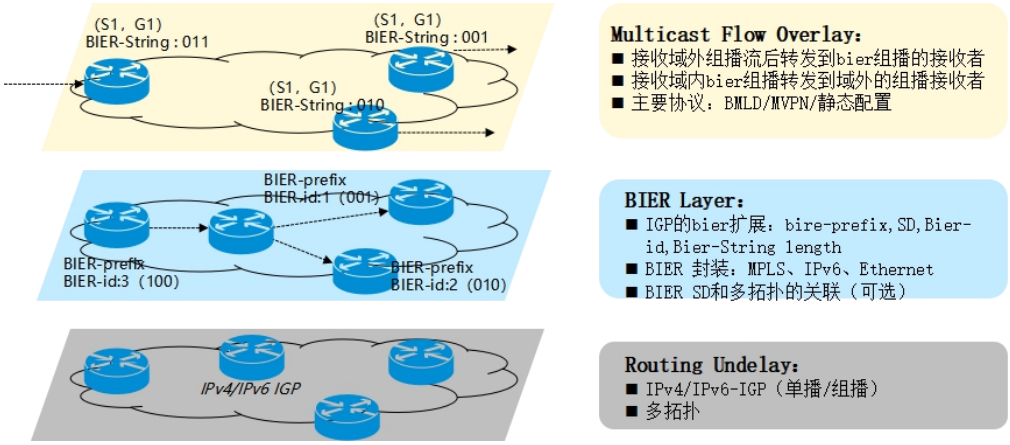


图 2 BIER 三层架构

3.2.2. BIER 层

BIER 层主要完成 BIER 路由信息发布和洪泛以及本地 BIER 转发表计算更新。BIER 层根据 BIER 转发表对 BIER 报文转发，每个转发 BIER 报文的节点对 BIER 报文执行解封装、重新封装的过程。解封装 BIER 报文头，获得 BIER 报文携带关键信息如 BIFT-ID 和 BitString，前者为路由器定位 BIER 转发表的索引，后者是查询 BIER 转发表的键值，类似查询 IPv4 路由表使用 IPv4 目的地址作为键值。BIER 节点根据 BIER 转发表的结果重新封装 BIER 报文头并转发 BIER 报

文。如果该节点为组播的复制点，就有多个不同的查询值，每个值代表节点将复制并重新封装新的 BIER 报文头并转发报文。

一个 BIER 路由器可以有多张 BIER 转发表，每张 BIER 转发表有多个表项内容。每个 BIER 转发表都关联一个 BIFT-ID，BIFT-ID 通过 SD、SI 和 BSL 编码者哈希生成。MPLS 封装的 BIER 报文，BIFT-ID 则是 BIER 的 MPLS 标签值。BIER 转发表项主要内容为一串比特码(RFC8296 标准中被称为 Forwarding bit mask，F-BM) 和一个邻居节点组成，每串 F-BM 表示通过这个邻居以最优路径方式到达其他 BIER 节点的集合。BIER 节点解封装 BIER 报文头中 BitString，用 BitString 与 BIER 转发表中每个表项的 F-BM 进行与运算，根据计算结果决定是否复制 BIER 报文到邻居节点。F-BM 的表示可以参见图-11。

BIER 层屏蔽网络层感知组播业务，中间 BIER 节点不感知组播业务，不建立传统的组播发布树，不维护每个组播的转发状态信息。BIER 路由器仅根据收到 BIER 报文的 BitString 和本地 BIER 转发表进行转发或者复制。

3.2.3. Underlay 层

Underlay 层为传统链路状态路由协议层，通过链路状态协议如 ISIS、OSPF 等扩展 TLV 属性携带本节点的 BIER 信息。BIER 因而继承了 ISIS、OSPF 协议许多特性，如支持 FRR、负荷分担，BIER 转发表收敛与 ISIS 或者 OSPF 协议收敛同步，速度达到毫秒级。 如下图所示，ISIS 的 BIER 扩展新增一个 BIER-SUB-TLV 携带 BFR-id 和 SD 等重要信息。同时也新增了多个 BIER-SUB-SUB-TLV，携带封装类型（如 MPLS 封装，非 MPLS 封装以太，非 MPLS 封装 IPv6 ）、BSL 和最大 SI 值，标签值或者 BIFT-id 等信息。

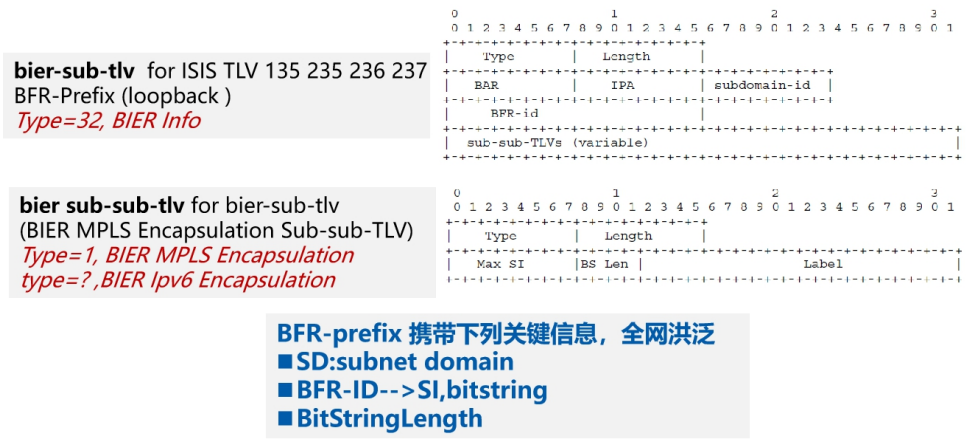


图 3 Underlay 层的 BIER 扩展 TLV

4. BIER 报文格式及封装类型

4.1. BIER 报文格式

IETF 定义 BIER 报文 MPLS 封装、非 MPLS 的以太封装和 IPv6 封装等三种类型，适应不同的组网需求。BIER 不同的封装类型都有一个相同的 BIER 报文头如下图深蓝色部分所示，组播报文进入 BIER 的 Ingress 节点被封装一个 BIER 报文头，组播报文离开 BIER Egress 节点解封装 BIER 报文头还原组播报文。

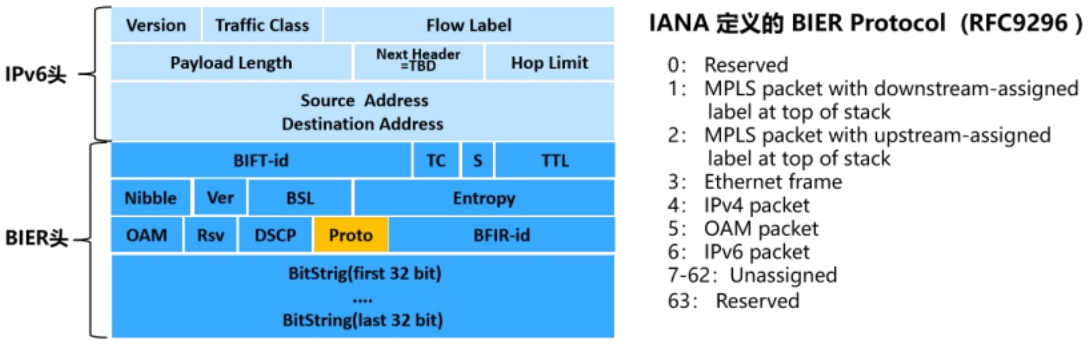


图 4 BIER 的报文头格式

BIFT-id：报文转发使用的 BITF,MPLS 封装对应 Label 值，NO MPLS 封装(以

太或者 IPv6)时使用(SD,SI BSL)来映射或者编码。

TC： 流量类型，同 MPLS 封装的 TC，参考 RFC5462.

S： 标签栈底标识，同 MPLS 封装的 S bit，参考 RFC3032。

TTL： 同 MPLS 封装的 TTL 的使用，参考 RFC3032

Nibble： 固定值 0101,用来区分 BIER 封装和 MPLS 的 ECMP 功能。

Ver： 表示版本号，当前值为 0 表示实验中的版本。

BSL： 表示 BitString 的长度(log2(k)-5)，用于离线分析。

Entropy: 支持 ECMP，相同的 Entropy+BitString，选择相同的路径。

OAM： 缺省为 0，可用 ping/trace，不影响转发和 Qos。

RSV： 保留位，当前不用缺省为 0。

DSCP： MPLS 封装时不使用，no MPLS 封装可使用。

Proto： 表示 Payload 报文的类型，RFC 已经标准化。

BFIR-id：表示组播进入 BIER 域中第一个 BIER 路由器的 BFR-ID 值。

BitStirng：同 SD、SI 表示一组 BFER 路由器。

根据 BIER 中 Protocol 字段定，BIER 的负荷可以是 IPv6 或者 IPv 4 的报文，也可能为 MPLS 或者以太的报文。BIER 既支持 IPv4 组播业务，也支持 IPv6 组播业务。通过上游分配标签的方式，也可以支持支持组播 VPN 业务。

4.2. BIER 封装类型

BIER 支持 MPLS 封装和非 MPLS 的以太封装，其中非 MPLS 封装支持有 o 太封装和 IPv6 的封装。目前，MPLS 封装和非 MPLS 的以太封装已经标准化，而 BIER 的 IPv6 封装还没有完成标准化，存在多个 BIER IPv6 封装版本。

4.2.1. BIER 的以太封装格式

IANA 定义 BIER 的太封装类型 0XAB37，以太头后面直接跟 BIER 报文头。BIER 的以太封装，非常简洁高效，如下图所示。以太类型 0xAB37 标识其 payload 为 BIER 报文，其中 BIER 报文头的格式如 4.1 章节所示。BIER 报文头中的 Protocol 协议字段可以进一步标识上层协议的内容。BIER 报文中的 protocol 字段为 2 标识上游分配的 mpls 标签，通常用来实现组播 VPN 的业务。BIER 的以太封装详细信息可以参考 RFC8296 “Encapsulation for Bit Index Explicit Replication (BIER) in MPLS and Non-MPLS Networks”。

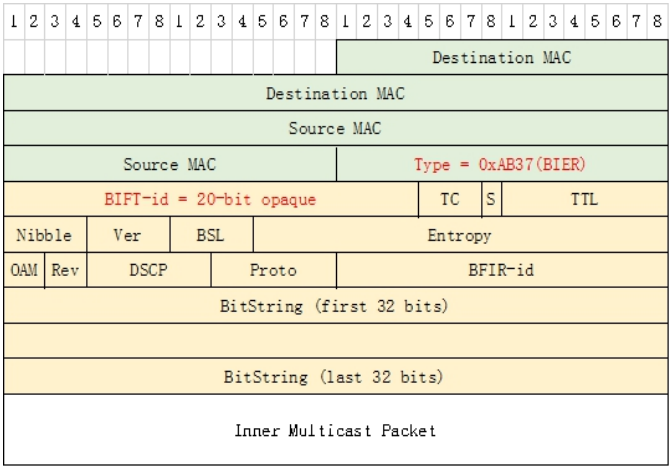


图 5 BIER 的以太封装格式

4.2.2. BIER 的 MPLS 封装格式

BIER 的 MPLS 封装使用 IANA 定义的 MPLS 类型为 0x8847，其中 0x8847 标识以太报文负荷为 MPLS 封装报文，而 BIER 报文类型通过标签值范围来确定。标签管理模块对 BIER 封装如 IPv4 / IPv6 协议一样，分配一段独立的标签范围。MPLS 封装复用了 BIER 报文头的前 4 个字节，其中 BIER 报文头前 20 比特的 BIFT-Id 为标签管理模块分配的 BIER 标签值。BIER 的 MPLS 封装顺序为以太、BIER 报文头和上层协议，没有独立 MPLS 标签，这点需要注意，如下图所示。

MPLS 模块解析 BIER 头前 4 个字节（MPLS-BIER）后，根据标签值进入不同的协议包括 BIER 的处理流程，同时根据标签值确定对应的转发表。BIER 的 MPLS 封装详细信息可以参考 RFC8296 “Encapsulation for Bit Index Explicit Replication (BIER) in MPLS and Non-MPLS Networks”。

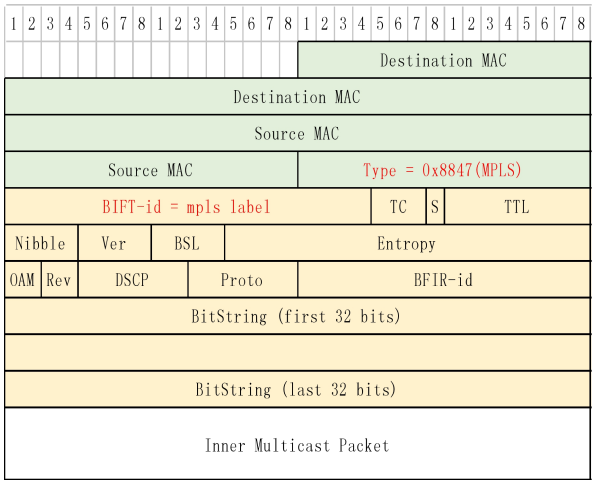


图 6 BIER 的 MPLS 封装格式

4.2.3. BIERin6 封装格式

中兴通讯公司目前支持 BIERin6 封装，支持 draft-zhang-bier-bierin6-04 草案。该草案建议在 IPv6 的 Next Protocol 增加新的 BIER 协议类型，具体类型值等待 IANA 分配，下图中暂用 TBD 表示，BIER 报文头作为 IPv6 负荷。RFC82000 已经定义 IPv6 支持的协议类型有 IGMP、IPv4、TCP、IPv6、UDP、Ethernet、shim6 等。IPv6 支持的上层协议类型和 IPv6 支持的扩展头，使用同一个字段(Next Header)不同值来实现。如 NextHeader 为 0 则标识 IPv6 逐跳的扩展头，NextHeader 为 43 为路由扩展头，SRv6 使用这个扩展头。BIERin6 的封装中的 Protocol 字段遵循 RFC8296 的定义，具体的封装格式如下所示。

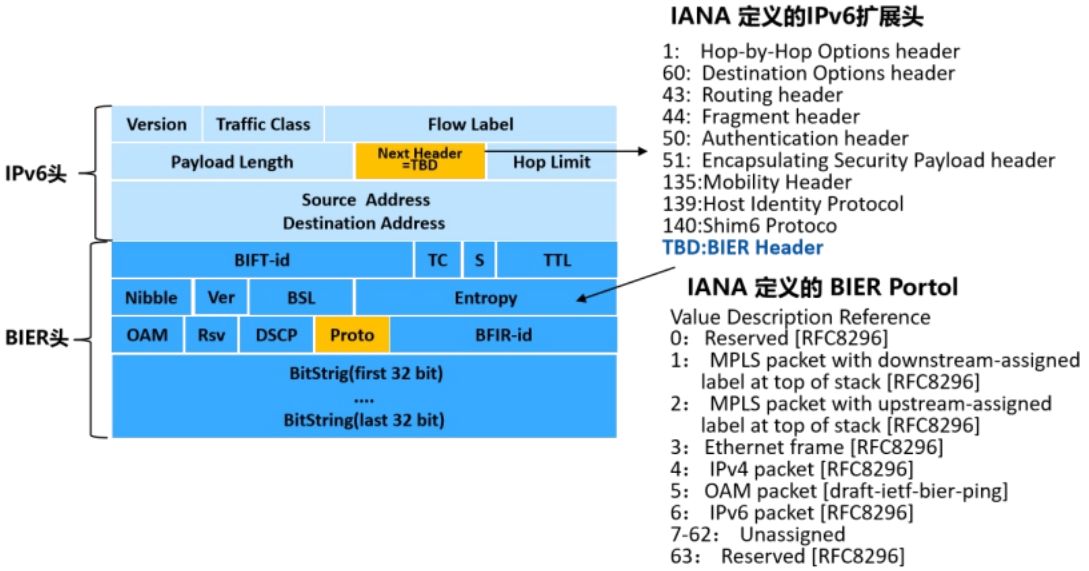


图 7 BIER 的 BIERin6 封装格式

BIERin6 的封装与其他 IPv6 扩展头无缝融合，BIER 报文可以放置在 IPv6 的扩展头如 Hop-by-Hop Options header 和 Destination Options header 的后面，通过扩展头中 NextHead 标识其负荷为 BIER 报文。一般情况下，BIER 报文是逐跳处理的，不建议在 BIER 域内进行报文分片和加密、解密处理。对组播业务的分片、加密和解密报文放在上层业务处理，BFIR 节点仅对组播业务做简单的 BIERin6 封装即可。

5. ISIS 扩展支持 BIER

5.1. ISIS 扩展 sub-tlv 和 sub-sub-tlv 支持 BIER

链路状态协议如 ISIS 和 OSPF 都扩展支持 BIER 且标准化（RFC8401 和 RFC8444）。目前主流厂家如 ZTE/Huawei/Nokiad 都已经实现 ISIS 的 BIER 扩展，OSPF 的 BIER 扩展也都在开发之中。ISIS 扩展的 BIER-SUB-TLV 和 BIER-SUB-SUB-TLV 携带关键的 BIER 信息如 BFR-id、BSL、SD 等。

BIER-SUB-TLV 与 ISIS TLV 135、235、236、237 一起使用，其内容如下图所示。

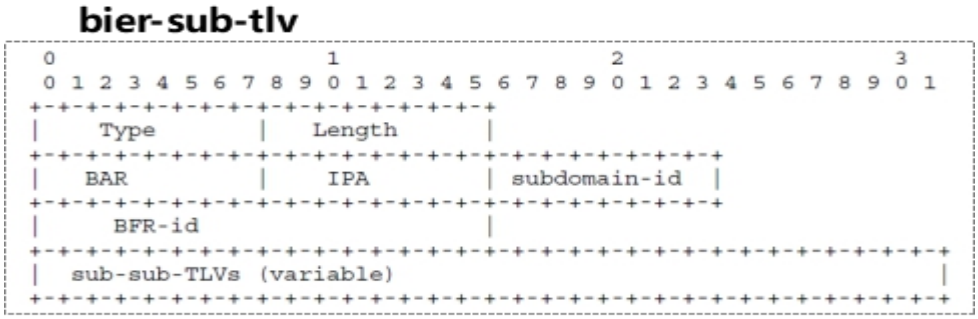


图 8 ISIS -SUB-TLV

TYPE：表示 BIER-SUB-TLV，值为 32

Length: 变化值

BAR: 表示 BIER 算法，用于计算到达 BFER 路径计算

IPA：表示 IGP 算法，表示 IGP 增强或改进算法，可替代 BAR 算法。

Subdomain-id: 表示一个 SD 域

BFR-id：表示该路由分配的 16 比特无符合整数。

Sub-sub-TLV: 表示可选的子子 TLV，具体格式如下所示。

一台路由器可支持多种方式如 MPLS 封装或者以太封装、Bierin6 封装等，不同的封装有对应的参数表达，ISIS 定义 sub-sub-tlv 来表示不同的封装类型，一个 BIER 的 sub-tlv 可以携带多个不同的 sub-sub-tlv。目前已经定义的 MPLS 封装的 sub-sub-tlv 如下：

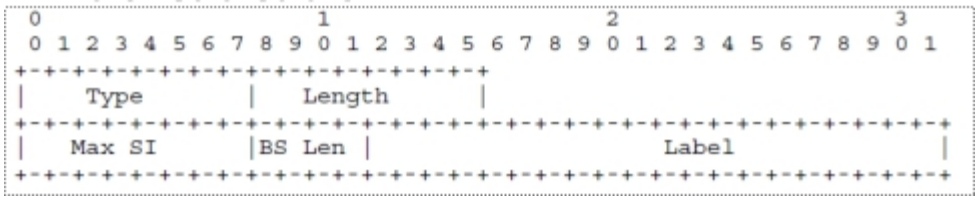


图 9 ISIS -SUB-SUB-TLV（MPLS）

Type：目前值为 1 表示 MPLS 封装

Length：可变化的值

Max SI：表示最达可以支持 SI 的数量。

BS Len：表示 BSL 的长度的编码，4 个比特表示。

Label：表示标签范围内的第一个标签值。

以太封装的 sub-sub-tlv 非常类似 MPLS 的 sub-sub-tlv，主要差别为 type 类型为 2，且 20 比特的 BIFT-id 替代 MPLS 封装的 Label 值，如下图所示。

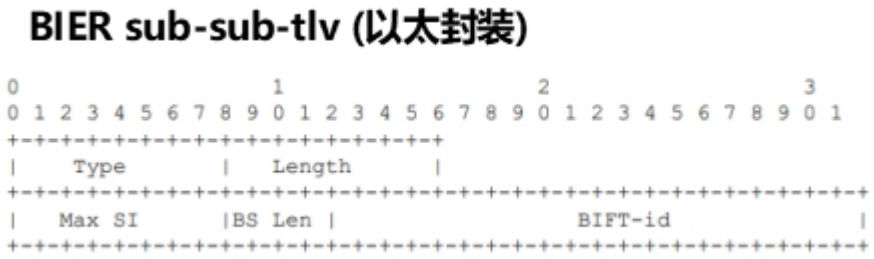


图 10 ISIS -SUB-SUB-TLV（Ethernet）

每个 SI 的 BIFT-id 为 BIFT-id(初始值)+SI(值)。如果 BIFT-id（初始值）+SI 值超出 20 比特值范围则报错。

5.2. ISIS 的 BIER 路由表和转发表

ISIS 扩展的 sub-tlv 和 sub-sub-tlv 携带 BIER 的关键信息:SD、BSL、BFR-ID 通过 BIER-prefix 在 ISIS 网络上洪泛，包括泄露到 ISIS-L1 的路由器。BIER 路由器使用 IGP 的算法或者 BIER 算法生成到 BFR-prefix 前缀的路由，也就是到每个 BFR-id 的路由。每个 BFR-prefix 携带 sub-tlv 对应一个 BFR-id 的信息。如下图所示，R4 上计算除到 R1、R2、R3 的 BFR-prefix 路由表(BIRT)，即对应的 BFR-id 的路由表。R4 通过对图 2 中相同的 R3 出口几个 BFR-id（0:0001/0010/0100）合并成一个 F-BM（Forward Bit-Mask）(0:0111)，即生

成 BFR 的转发表(BIFT)。

BIER 路由器可以根据协议、拓扑、BSL 等多种因素计算或者分配多个 BIFT，每张 BIFT 都有一个独特的索引即 BIFT-id 来标识，其对应 BIER 报文头中的 BIFT-id 字段。通过 BIER 报文中 BIFT-id 索引到正确的 BIFT。BIER 报文中的 BIFT-id 则来源邻居路由器通告的 sub-sub-tlv 中的 BIFT-id 字段，非本地生成生成。MPLS BIER 封装，BIFT-id 则是根据 SI 和起始标签计算出来的 20 位 Label，以太 BIER 封装，BIFT-id 则是根据 SI 和起始 BIFT-id 计算出来的 20 比特 BIFT-id。

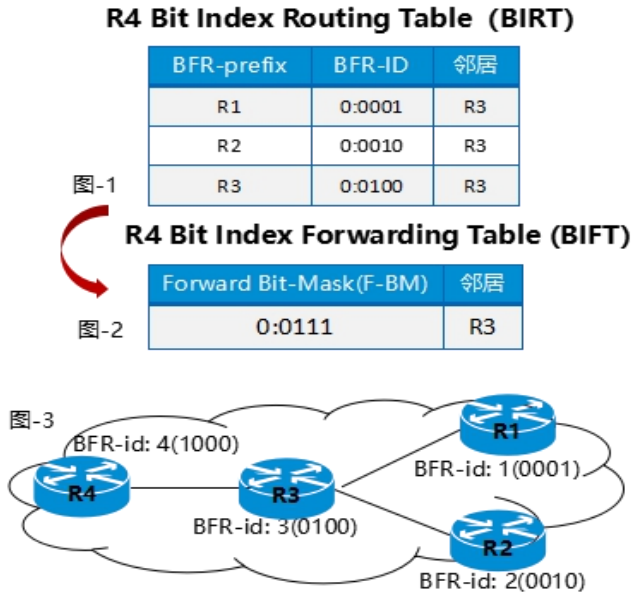


图 11 BIER 的转发表生成

6. BIER 转发过程

组播报文通过 BFIR 路由器封装 BIER 报文头后进入 BIER 域。在 BIER 域内，BIER 路由器根据 BIER 报文头中携带 BIFT-id 定位到 BIFT 表；根据携带的 BitString 查询 BIFT 进行转发或者复制。BIER 报文到达 BFER 节点后解封成组播报文，根据组播地址查找组播路由转发。BFIR 根据 Overlay 层信息确定 BIER

报文内的 BitString，即一条组播全部 BFER 集合。BFIR 根据邻居通告的 BIER 的 BIFT-id 以及本地的 BIER 配置信息，组装 BIER 报文头并使用邻居通告的封装类型发送 BIER 报文，如 MPLS 封装发送，以太封装发送或者 IPv6 封装发送。下面以具体的示例来说明 BIER 转发过程。

BIER 域由六台 BFR 路由器（A、B、C、D、E、F），其中 D、F、E、A 作为 BFIR 或 BFER 其 BFR-id 为 1、2、3、4。BIER 域的 BitString 长度为 4、SI 为 0，四台路由器（D、F、E、A）对应 BFR-id 分别为 0:0001、0:0010、0:0100、0:1000，其他 BFR 路由器不分配 BFR-id，如下图所示。每台路由器根据 IGP 或 BIER 计算生成到其他 BFR-id 的 BIFT。路由器 A 作为 239.1.1.1 组播的 BFIR 节点,D 节点和 E 节点作为 BFER 申请加入组播组 239.1.1.1。BIER 报文在 BIER 域中转发过程如下：

- BFIR（路由器 A）收到 239.1.1.1 的组播报文，查组该播路由的入接口和出接口信息，该组播报文出口为一个 BIER 索引（Index）。该索引指向内容主要包含 SD、SI、BSL、BFER-List 的信息结构。BFIR 根据这些信息生成 BitString 查询本地 BIFT 表获取 BIER 转发的关键信息，如封装类型和邻居通告的 BIFT-id 等。图中的 BFIR 根据 BFER-List（节点 D 和节点 E） 生成 BitString（0:0101)， 通过 Bitstringd 对本地的 BIFT 中表项进行计算获取 BIER 转发的出接口 B 和新的 BitString(0101,没有变化）。BFIR 根据邻居路由器 B 在该接口通告的封装类型，BIER 信息（如 BIFT-id）等完成组播报文的 BIER 封装后发给 BIER 报文给邻居路由器 B。
- 节点 B 收到节点 A 发送 BIER 的报文，解析 BIER 头获取 BIFT-id 和 BitString。根据 BIFT-id 查找指定的 BIFT 转发表，使用 BitString(0101)

与 BIFT 查表获取出每个出接口邻居及其对应 BitString，根据出接口邻居获取 BIER 的封装类型。

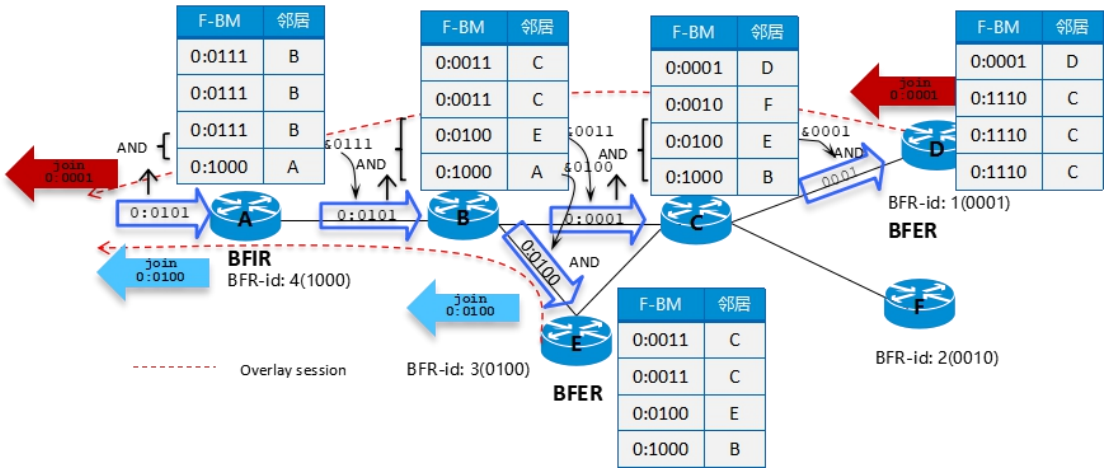


图 12 BIER 的转发示意图

- BFR 节点 C 收到 B 发送 BIER 的报文，处理流程同节点 B，向路由器 D 发送 BIER 封装报文，BIER 报文头携带 BitString 为 0001、节点 D 通告的 BIFT-id 等信息，并使用节点 D 支持的 BIER 封装类型发送。
- BFR 节点 E 收到 B 发送 BIER 的报文，处理流程同节点 B。因为 BitString(0100)与 BIFT 中第三个表项的 F-BM 进行与计算的值为 0100，而该表项对应邻居为 E 为节点本身，确认 E 节点为 BFER。根据 BIFT 的查询结果，节点 E 解封装 BIER 报文，根据组播地址查找组播路由表转发。
- BFR 节点 D 收到 C 发送 BIER 的报文，处理流程同节点 E，确认 D 为 BFER 解封 BIER 报文，根据其组播地址查找组播路由表转发。

7. BIER 组播应用场景

7.1. BIER 在 IPTV 和 OTT 场景中的应用

IPTV 是网络中点到多点的应用服务，IPTV 可以提供视频直播服务和按需

随选的流媒体节目传输服务，比如体育比赛和演唱会的现场直播属于前者，点播重放可以控制播放进度的真人秀类节目属于后者。IPTV 直播服务要求建立从 Egress 节点者到 Ingress 节点视频源的组播发布树，这样可以节省骨干网带宽。2020 年受到疫情的影响，网络上新型直播业务也蓬勃发展，如各种授课平台如腾讯课堂、学而思课堂，各种商业视频会议系统如 Zoom 等。按需随选的流媒体节目则可以通过多级 CDN（Content Delivery Networks）网络，把节目推送离用户最近的边缘 CDN 节点。按需随选的流媒体节目的最终用户通过组播的方式加入离自己最近的边缘 CDN 节点。通过分布式的多级 CDN 网络既可以节省广域网的带宽，又减轻海量并发用户对单个 POP 节点和网络的压力。大型内容提供商或者游戏运营商都可通过多级分布式的 CDN 网络支持海量的并发用户。多级分布式 CDN 网络通过应用层实现可靠的内容传输和管理。OTT（Over The TOP）服务非常类型 IPTV 服务，主要差别 OTT 服务的视频源存放在另外一个网络上，OTT 服务需要跨越两个网络，甚至多个网络。比如通过手机观看腾讯视频，手机终端在运营商的网络而视频源存放在腾讯公司 IDC 机房。腾讯视频流量需要跨域腾讯 IDC 网络和运营商网络有线和无线网络。OTT 服务需要支持跨越的组播服务。

IPTV 服务中应用 BIER 技术，不需要建立从 Egress 节点到 Ingress 节点的组播发布树，中间节点的不运行组播路由协议、不维护组播转发状态。典型的 IPTV 网络中，Egress 节点运行 IGMP/MDP 协议获取终端加入或者离开频道消息后，Egress 节点向 Ingress 节点发送特定频道加入和离开。Egress 节点也可以静态配置加入和离开特定的频道。BIER 处理频道的加入和离开则完全不同，BIER 在 Ingress 节点直接映射每个频道和接收者之间的关系。在 IPTV 的 BIER

方案中增加新的频道非常方便，在 Ingress 节点增加新的组播映射就实现向所有 Egress 节点推送全部频道的应用场景。

OTT 服务中应用 BIER 技术，边缘 CDN 节点利用 BIER 强大伸缩性特点实现组播的跨域发送。BIER 既可以使用 MP-BGP 跨域互通，又可以使用 SDN 方式实现跨域互通，BIER 也支持静态跨越的方式，即一个域的 Egress 节点发出的组播流量，作为另一个域的 Ingess 节点的组播源。CDN 域的 ASBR 路由器终结来自另一个域 OTT 用户的请求后把本地 cache 视频或者距离 ASBR 最近 CDN 视频，发送给 OTT 用户。CDN 作为四层应用，也可以实现跨越任意不同的网络边界的简单应用。位于远端 BIER 域的 CDN 客户，收到组播报文可以作为本域的 Ingress PE 节点的源，实现本域的 BIER 转发。

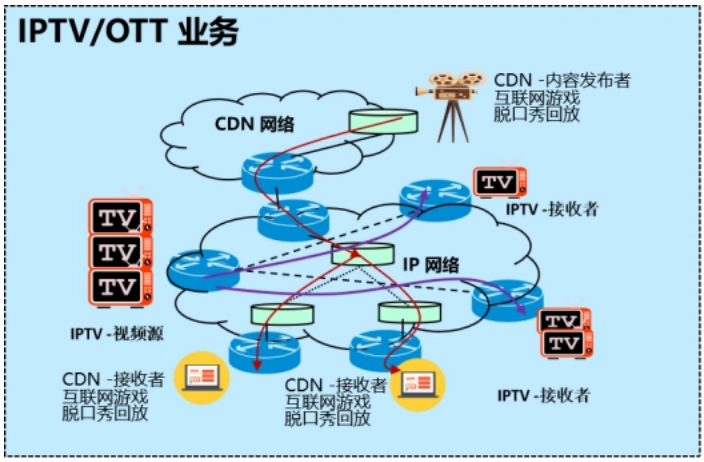


图 13 BIER 在 IPTV/OTT 场景中应用

7.2. BIER 在组播 VPN 场景中应用

传统的组播 VPN 需要公网路由器维护组播 VPN 状态信息。运营商不能为所有 VPN 客户提供大规模的组播服务是因为其网络设备无法维护每 VPN 每组播的转发状态，VPN 客户和组播数量越多运营商设备需要维护的组播状态就越多($O(N^2)$)。为了减低设备维护的组播状态数量，电信运营商使用头端复制或者聚合

组播树等技术，但都没有彻底解决组播状态多而复杂问题。

BIER 不同于传统的组播协议如 PIM 等，提供一种无状态的组播转发机制。在 BIER 组播首节点（BIER Ingress）确定组播流接收者(BIER Egress)信息，中间节点根据 BIER 转发表转发组播流，不需要维护组播状态信息，非常适合运营商的开展 VPN 的组播业务。BIER 的 Overlay 层完成组播 VPN 的控制信息，BIER Ingress 节点根据组播 VPN 接收者的信息，封装 BIER 报文和 VPN 信息，中间节点完成 BIER 转发，不感知组播也不感知 VPN。BIER Egress 节点解封 BIER 信息，根据内层 VPN 信息查询对应 VPN 组播路由表转发组播，如下图所示。

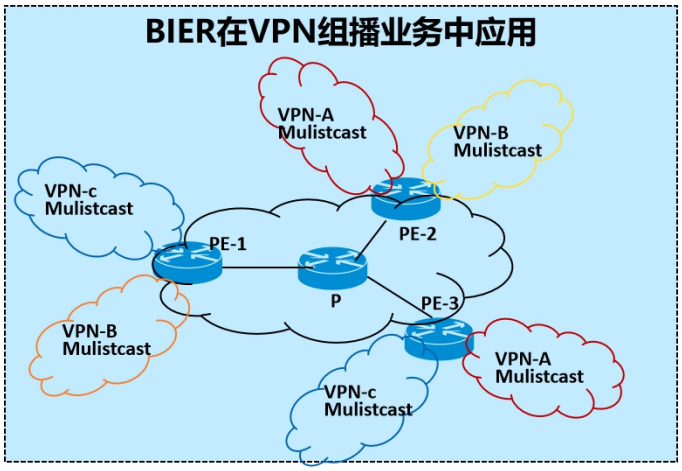


图 14 BIER 在 VPN 组播场景中应用

7.3. BIER 在金融场景中的应用

在金融服务中也大量使用组播业务，最典型的业务是股票业务。股票行情的数据从上海/深圳交易所到达各个地方的证券公司的营业部都是通过组播业务来实现的。这些金融数据通常要求以最优的确定的延时且安全的到达各个节点。没有一家证券公司的营业部能够忍受收到交易所的股票行情数据比其他证券公司晚 100 毫秒的。

传统的组播路由协议如 PIM /mLDP/RSVP-TE 等没有彻底的解决上述问

题。目前的组播树都是组播接收者驱动建立从接收者到组播源的转发树，并不一定是最优路径。当组播的数量增加，网络上组播转发树的增多，将导致组播接收者的延时变化，有的组播接受者收到组播可能晚于其他组播接收者。复杂的组播协议运则可影响了数据的安全性。

BIER 使用已有的单播协议建立从源到接受者的最优路径，即使组播的数量增加更多，从源到目的的最优路径特性都不会改变。无论组播的数量多少，BIER 的收敛时间同单播路由协议一样快，BIER 总是提供的最优，确定性的延时。BIER 保证金融组播服务的公平性比传统的组播路由协议有很大的优势。

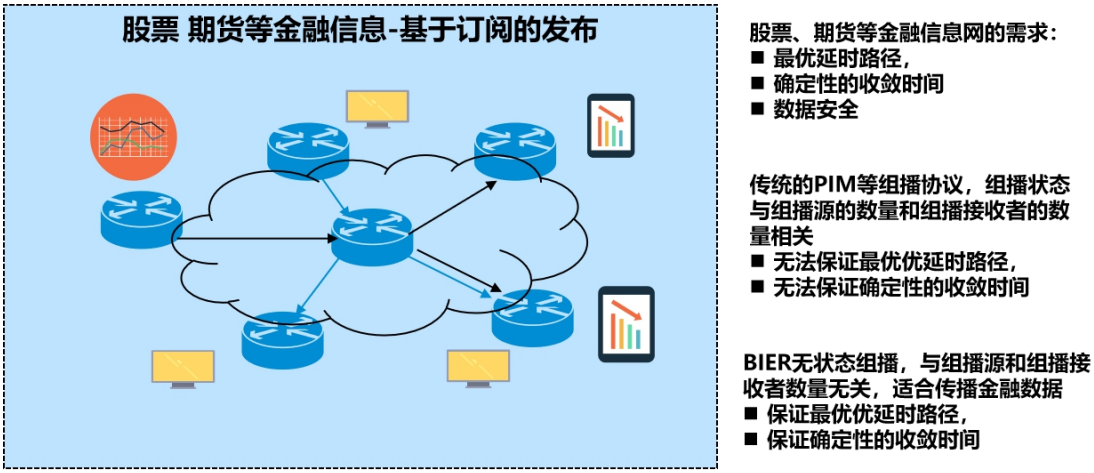


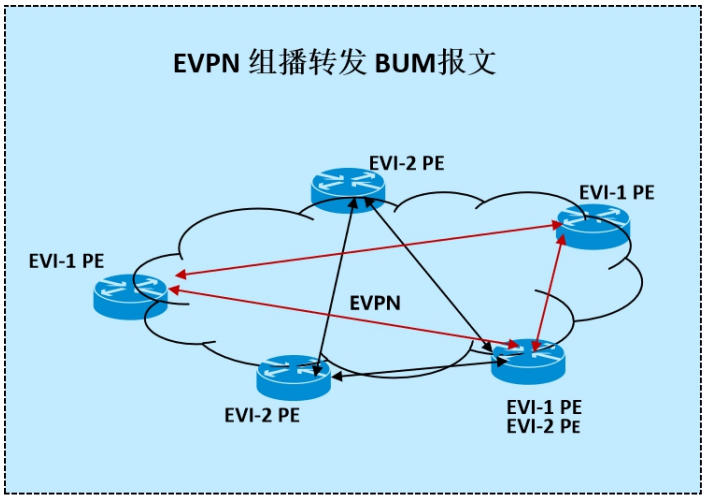
图 15 BIER 在金融场景中应用

7.4. BIER 在 EVPN 场景中的应用

EVPN 的 L2VPN 替代 VPLS 的 L2VPN 目前非常流行，在实际部署中已经被广泛使用。基于 EVPN 的 L2VPN 有支持 AC 的双归接入等优点，但也有处理大量 BUM（ Broadcast, Unknown unicast and Multicast ）的需求。一个 EVI 实例的 BUM 报文需要洪泛到所有支持该 EVI 实例的 PE 路由器上，采用首节点复制可以支持 EVPN 实现在一组 PE 节点之间传输 BUM 报文需求，这也是实际部署中常用的方案。首节点复制在 Ingress PE 节点为同一个 EVI 实例的 Egress

PE 复制 BUM 报文，Ingress PE 节点复制 BUM 报文的数量等于该 EVI 实例 Egress PE 节点数量。因此，要求 Ingress PE 不但组播复制、转发性能高且带宽大。EVPN 中的 PE 节点数量越多，对 PE 节点的性能和带宽的要求就越高。

BIER 通过 EVPN 的 PMSI 属性携带 BIER P-tunnel 功能，实现 BUM 报文的 BIER 封装和转发。BIER 转发不需要 P 节点运行传统 PIM/mLDP/P2MP/MP2MP 等协议维护组播转发状态。P 节点根据 BIER 转发表转发 BIER 封装的 BUM 报文，P 节点不感知组播业务，不运行组播协议，不维护组播转发状态。正常情况下，通过 EVPN 的 I-PMSI 属性，一个 BIER 的子域（sub-domain）可以对应多个 EVI 实例。Ingress PE 为不同的 EVI 实例分配不同的标签并通告给下游的 Egress 节点。Egress 节点解封 BIER 报文识别 EVI 标签转发 BUM 报文进入对应的 EVI 实例。Ingress PE 复用 Inclusive-PMSI 接口发送多个不同 EVI 实例的 BUM 报文，也会导致 BUM 报文到达没有接收者的 Egress PE 节点，浪费带宽。BIER 报文中携带 BitString 可以准确定位 Egress PE 节点，如果 Ingress PE 对接收 BUM 报文的 Egress PE 节点进行准确 BitString 编码和 BIER 封装，则可以避免上述情况的发生。从这个角度看，BIER EVPN 组播同时支持 Inclusive Tree 和 Selective Tree。 虽然一个 BIER 的 sub-domain 可以支持任意多个的 EVI 实例，但如果需要，也可以为每个 EVI 实例建立独立的 BIER sub-domain。



- 部署EVPN在广域网上提供类VPLS的二层的网络服务。
- EVPN 使用mLDP、P2MP/MP2MP等方式为每个zEVI实例建立组播转发树，维护组播转发状态。EVPN可以使用首节点复制的方式实现BUM报文的发送，要求首节点性能好且带宽大。
- EVPN的组播使用组播VPN的PMSI属性来标识各种组播隧道属性。

- BIER沿用组播VPN的PMSI属性，无组播状态维护。
- BIER既可以把所有EVPN 实例映射到同一个BIER的SD，也可以为每个EVPN实力了映射独立的BIER SD

图 16 BIER 在 EVPN 场景中应用

7.5. BIER 在数据中心场景中的应用

VXLAN 作为三层网络上传输二层网络的技术在数据中心内部被大规模的部署。数据中心支持多租户和虚拟化技术能力越大，对 VXLAN 的数量要求就越多，VXLAN 最大支持 16M。每个 VXLAN 都要实现在数据中心内 VTP 节点之间实现二层广播、组播和未知报文（BUM）的转发能力，这要求数据中心的底层网络支持组播功能。在数据中心的底层网络中采用传统的 PIM-SM 等组播路由协议可以支持小规模 VXLAN 的组播需求，VXLAN 多达 16M 时候，底层的网络则无法支持 16M 的组播转发树。即使实现 VXLAN 和组播的 N:1 之间的映射，则可能导致非最优化的 BUM 转发路径，导致不必要的带宽浪费。一般来说，数据中心的管理者不倾向在数据中心内部实现复杂的组播路由协议，小规模的 VXLAN 组播需求管理者通过首节点复制实现的。首节点复制可以避开复杂的组播协议简化网络协议控制，但是对网络带宽和 VTP 节点的性能都有较高的要求。

在数据中心部署 BIER，则可以实现无状态转发 BUM 的报文，不需要建立组播转发树，不需要维护组播转发状态。在首节点上或者 SDN 控制器实现 VXLAN 与组播的映射关系。

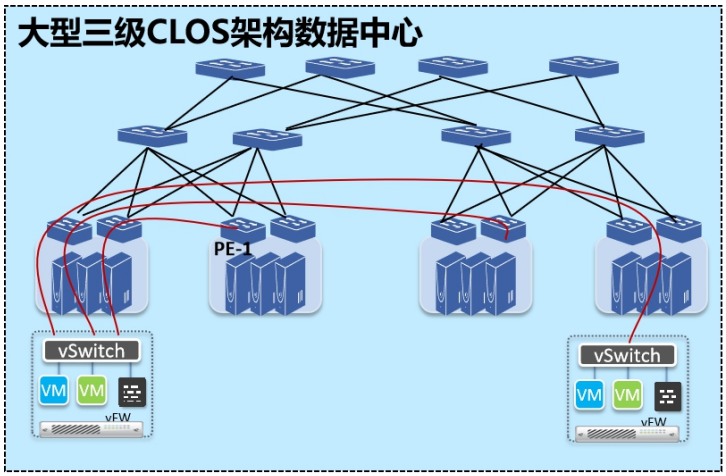


图 17 BIER 在大型数据中心场景中应用

8. BIER 相关的标准

序号	分类	标准	说明
1	架构	RFC8279	BIER 架构，包括概念、术语、工作机制、分层模型等
2	用例	draft-ietf-bier-use-cases-11	BIER 用例描述
3	需求	draft-ietf-bier-oam-requirements-09	BIER OAM 需求描述
4	需求	draft-ietf-bier-ipv6-requirements-04	IPv6 环境下 BIER 需求描述
5	数据面	RFC8296	MPLS 和非 MPLS 网络 BIER 封装，包括 BIER Header 结构以及各字段定义
6	BIER 控制面	RFC8401	ISIS 对 BIER-MPLS 封装的协议扩展,包括相关 sub-TLV 的定义和协议交互处理机制
7		RFC8444	OSPFv2 对 BIER-MPLS 封装的协议扩展，包括相关 sub-TLV 的定义和协议交互处理机制
8		draft-ietf-bier-lsr-ethernet-extensions-01	IGP 对 BIER-Ethernet 封装的协议扩展,包括相关 sub-TLV 的定义和协议交互处理机制
9		draft-ietf-bier-ospfv3-extensions-01	OSPFv3 对 BIER 的协议扩展，包括相关 sub-TLV 的定义和协议交互处理机制
10		draft-ietf-bier-bar-ipa-06	IGP 对 BIER 路径算法和约束条件的协议扩展
11	BIER 控制面	draft-ietf-bier-idr-extensions-07	BGP 对 BIER 的协议扩展
12		draft-zhang-bier-bierin6-04	IGP 对 BIER 通过 IPv6 隧道承载的协议扩展
13		draft-zwzw-bier-prefix-redistribute-05	IGP 对 BIER 前缀域间重分发的协议扩展

14	Overlay 控制面	RFC8556	BGP 对 BIER 承载 L3VPN 和 IP 全局组播的协议扩展
15		draft-ietf-bier-mld-04	MLD 对 BIER Overlay 的协议扩展
16		draft-ietf-bier-pim-signaling-08	PIM 对 BIER Overlay 的协议扩展
17		draft-ietf-bier-evpn-03	BGP 对 BIER 承载 EVPN BUM 组播的协议扩展
18	OAM	draft-ietf-bier-ping-06	BIER ping 和 trace 消息格式和处理机制
19		draft-ietf-bier-path-mtu-discovery-07	BIER path mtu 发现机制
20		draft-ietf-bier-pmmm-oam-07	BIER 染色性能测量方法
21		draft-hu-bier-bfd-05	BIER P2MP BFD 消息格式和处理机制
22	北向接口	draft-ietf-bier-bgp-ls-bier-ext-06	BGP-LS 扩展支持 BIER 拓扑信息上报
23		draft-ietf-bier-bier-yang-06	BIER 相关 YANG 模型

9. 缩略语

术语	全称	说明
BIER	Bit Index Explict Replication	根据 bit 位明确复制组播流量。
BFR	Bit-Forward Router	支持 BIER 转发的路由器。
BFIR	Bit-Forward Ingress Router	Bier 域中的边界路由器，连接组播源的 BFR 路由器。
BFER	Bit-forward Egress Router	Bier 域中的边界路由器，连接组播的接收者的 BFR 路由器。
SD	Sub Domain	一个 BIER 域可设计多个子域 Sub-Domian(SD)对应不同拓扑，一个 BFR 可跨子域。
BFR-prefix	Bit-Forward Router prefix	BFR 的 IPv4/IPv6 环回地址，通过该接口状态通告 BIER 信息。
BFR-id	Bit-Forward Router Identifier	16 比特的标识整数，唯一标识 SD 内的 BFR，范围【1-65535】。
SI	Set Identifier	BFR-id 可分为多个子集合，用 SetIdentifier **(SI)标识** ,子集内有若干 BFR。
BitString	BitString	用来表示 SD 内的 BFR 的一串二进制字符串，每个 Bit 对应 SD 的一个 BFR。

BP	Bit Position	BFR-id 对应 BitString 的特定 bit 的位置(Bit Position)。
BSL	BitString Length	BitString 的长度
BFT	Bit Forwarding Tree	组播流量在 bier 域中的转发路径树
BIRT	Bit Index Routing Table	使用 BRF-prefx 根据 BIER 算法计算出来的路由表
BIFT	Bit Index Routing Forwarding Table	根据 BIRT 按照邻居节点组织合优化的 BFR 转发表。
BIFT-id		用来标识具体的 BIFT 的 ID
BM	Bit-Mask	BIFT 中的 Fowrding Bit Mask，为多个不同 BP 聚合的 Bit string