

CloudEngine 12800 系列交换机

TRILL 技术白皮书

文档版本 01

发布日期 2015-05-30

目 录

1 TRILL 简介.....	1
2 原理描述.....	5
2.1 TRILL 基本概念.....	6
2.2 TRILL 报文格式.....	8
2.2.1 TRILL 控制报文.....	8
2.2.2 TRILL 数据报文.....	11
2.3 TRILL 协议机制.....	12
2.4 TRILL 转发流程.....	17
2.5 TRILL Multi-homing 双 active 接入.....	20
2.6 TRILL NSR.....	26
3 应用场景.....	28
3.1 TRILL 在数据中心的应用.....	29
3.2 M-LAG 在 TRILL 双归网络中应用.....	29
4 配置注意事项.....	31
5 配置任务概览.....	34
6 配置 TRILL.....	38
6.1 配置 TRILL 的基本功能.....	39
6.1.1 全局使能 TRILL.....	39
6.1.2 在指定接口上使能 TRILL.....	41
6.1.3 （可选）配置 TRILL 的接口开销.....	42
6.1.4 检查配置结果.....	43
6.2 调整 TRILL 路由选路.....	43
6.2.1 指定 Designated VLAN.....	43
6.2.2 配置 TRILL 负载分担.....	44
6.2.3 配置 RB 进入过载状态.....	45
6.2.4 使能 TRILL 组播组剪枝功能.....	45
6.2.5 检查配置结果.....	46
6.3 控制 TRILL 网络收敛.....	46
6.3.1 调整邻接故障的检测时间.....	46
6.3.2 调整 SNP 报文和 LSP 报文属性.....	47

6.3.3 调整 SPF 的计算时间.....	51
6.3.4 检查配置结果.....	52
6.4 配置 STP/RSTP/MSTP 联动 TRILL 功能.....	52
6.5 配置 M-LAG 双归接入 TRILL.....	58
6.5.1 配置 DFS Group.....	58
6.5.2 配置 peer-link.....	59
6.5.3 配置绑定 DFS Group.....	60
6.5.4 检查配置结果.....	61
6.6 配置 TRILL 网关.....	62
6.7 提高 TRILL 网络的安全性.....	62
6.7.1 配置 TRILL 报文认证.....	62
6.7.2 配置 TRILL 接口认证.....	63
6.7.3 检查配置结果.....	64
7 维护 TRILL.....	65
7.1 复位 TRILL.....	66
7.2 配置 TRILL OAM.....	66
8 配置举例.....	67
8.1 配置 TRILL 基本功能示例.....	68
8.2 配置 TRILL 联动 MSTP 示例.....	74
8.3 配置 M-LAG 双归接入 TRILL 示例.....	80

1 TRILL 简介

介绍TRILL的定义及目的。

定义

TRILL(Transparent Interconnection of Lots of Links)是一种把三层链路状态路由技术应用用于二层网络的协议。TRILL通过扩展IS-IS路由协议实现二层路由，可以很好地满足数据中心大二层组网需求，为数据中心业务提供解决方案。

目的

在云计算时代，数据中心内部一般采用分布式架构处理海量数据的存储、查询、搜索等相关业务。这导致服务器和服务器之间需要进行大量的集群计算，在服务器之间产生了大量的东西向流量。同时集群计算越来越多地利用虚拟机技术来实现，虚拟化的直接后果是使单位计算密度极大提升，物理服务器吞吐量将比虚拟化之前成数倍提升。另外，为了更大幅度地增大数据中心内业务可靠性、降低IT成本、提高业务部署灵活性、降低运维成本，需要虚拟机能够在整个数据中心范围内进行动态迁移。

由于上述云计算下数据中心的新特点，使用传统的二层（xSTP）接入，三层汇聚/核心（路由）的层次化网络划分模型构建数据中心网络时逐渐显得力不从心。目前，业界普遍采用大二层胖树架构。为了部署一个无阻塞的大二层网络并能够实现虚拟机的任意迁移，以及适应目前不断增大的网络规模，TRILL协议应运而生。TRILL相对于传统二层协议xSTP和三层路由协议的优势如下表所示：

表 1-1 TRILL 与 xSTP 对比表

云计算数据中心网络要求	说明	TRILL网络	xSTP网络
虚拟机任意迁移	作为云计算的核心技术之一，服务器虚拟化已经得到越来越广泛的应用。为了更大幅度地增大数据中心内业务可靠性、降低IT成本、提高业务部署灵活性、降低运维成本，需要虚拟机能够在整个数据中心范围内进行动态迁移，而不是局限在一个汇聚或接入交换机范围内进行迁移。	TRILL协议部署在大二层网络，可以支持虚拟机在整个数据中心动态迁移。	对于传统二层xSTP接入+三层IP路由的网络架构，如果虚拟机跨网段迁移，无法实现迁移前后的IP地址保持一致。所以只能支持虚拟机在一个网段内的迁移。
无阻塞、低延迟数据转发	云计算时代下的数据中心流量模型和传统运营商流量模型不同，数据中心中主要是服务器和服务器之间的东西向流量。为保证业务正常开展，需要支持网络数据的无阻塞、低延迟转发。	TRILL网络中每台设备都以自身节点作为源节点，基于最短路径算法计算到达其他所有节点的最短路径，如果存在多条等价链路，在生成单播转发表项时候能够形成负载分担。当存在多路径转发时候，能够充分利用网络带宽，网络中的每个节点都能实现线速转发。	通过阻塞链路实现单路径的转发，极大的浪费了带宽，并且违背了无阻塞网络架构的思想。

云计算数据中心网络要求	说明	TRILL网络	xSTP网络
网络规模大	对于云计算时代下的大型数据中心，支持的服务器数量要能够达到十万甚至百万级别。为了实现无阻塞转发，网络规模要能够达到几百台甚至上千台交换机，在这种大规模组网情况下，组网协议要能够有效避免环路。网络内部的节点和链路故障时，要能够触发整网快速收敛，业务迅速恢复。另外，要求网络维护简单，方便用户业务部署。	● 网络规模：理论上可以支持1000台左右的交换机。 ● 环路避免：引入IS-IS作为控制面协议，天生无环。 ● 收敛速度：采用路由协议生成转发表项，并且TRILL头部有hop-count字段能够允许短暂的临时环路，收敛速度能达到亚秒级。 ● 维护难度：配置比较简单，很多配置参数比如nickname、systemID等都可以自动生成，多数协议参数采用缺省配置即可；其次，单播、组播统一控制协议，用户只需要维护一套协议。	● 网络规模：有网络直径不超过7的限制，支持一百台左右设备。 ● 环路避免：通过阻塞端口破坏。 ● 收敛速度：协议收敛机制设计的比较保守，收敛速度仅能达到秒级。 ● 维护难度：xSTP+三层路由的组网中单播和组播需要维护IGP、PIM等多套路由协议，维护量较大。
多租户	云计算时代下一个物理数据中心不再被一个租户所独享，而是可以同时被多个租户同时使用，每个租户对应一个虚拟数据中心实例，每个租户仿佛享有独有的服务器、存储、网络资源，租户之间数据流量需要进行隔离。	目前TRILL标准采用VLAN ID作为租户标识，通过VLAN对不同租户流量进行隔离，在云计算产业和大二层组网运营处于起步阶段，VLAN ID的4096限制不会形成瓶颈。TRILL后续会演进到通过FineLabel来进行租户标识，FineLabel为24bit，理论上能够支持16M租户规模，足够满足将来租户规模扩展性的需求。	受限于VLAN数量限制，最多只支持4096的规模。

云计算数据中心网络要求	说明	TRILL网络	xSTP网络
可扩展性	为了适应数据中心高速发展的需要，数据中心网络需要具有良好的可扩展性。	采用xSTP协议的传统二层网络，可以无缝接入TRILL大二层网络，同时TRILL支持的网络规模较大，收敛速度较快，可扩展性比较好。	网络规模受限，收敛速度较慢，可扩展性比较差。

受益

- TRILL带来了明显的受益：
- 大二层数据中心支持虚拟机无阻塞迁移，大大降低了网络管理的难度。
 - TRILL设备可以和传统xSTP网络无缝连接，降低了网络升级的成本。

2 原理描述

关于本章

介绍TRILL的基本原理。

[2.1 TRILL基本概念](#)

[2.2 TRILL报文格式](#)

[2.3 TRILL协议机制](#)

[2.4 TRILL转发流程](#)

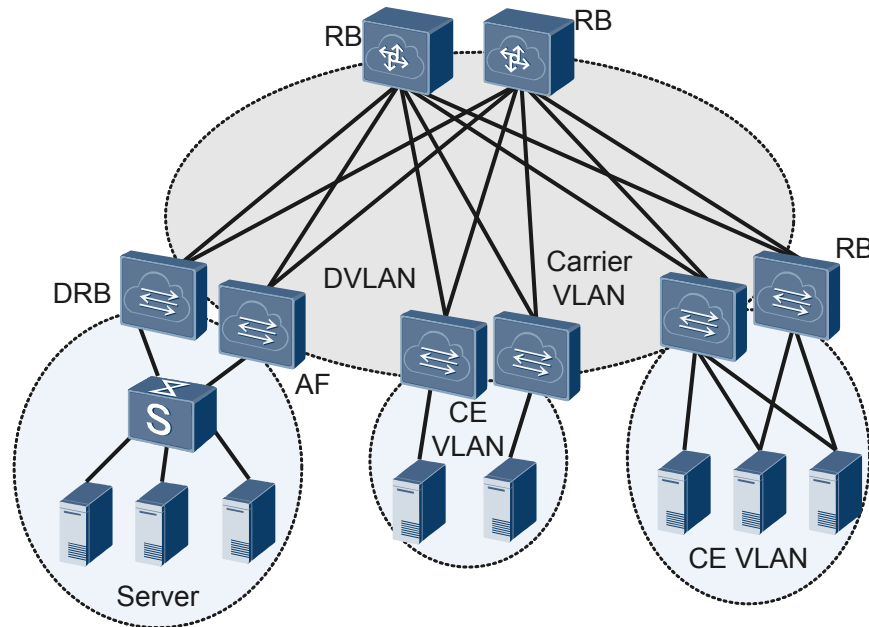
[2.5 TRILL Multi-homing双active接入](#)

[2.6 TRILL NSR](#)

2.1 TRILL 基本概念

本节介绍TRILL协议中的基本概念，如图2-1所示。

图 2-1 大二层 TRILL 组网图



TRILL 网络中的设备名称

RB

RB（Router Bridge）指运行TRILL协议的二层交换机。根据RB在TRILL网络中的位置，又可将其分为Ingress RB、Transit RB和Egress RB三种，分别表示报文进入TRILL网络的入节点、在TRILL网络中经过的中间节点以及离开TRILL网络的出节点。

DRB

DRB（Designated Router Bridge）指在TRILL网络中作为中间设备被指定承担某些特殊任务的RB，它与IS-IS中的指定中间系统DIS（Designated IS）相对应。在TRILL广播网中，两台RB如果处于同一个VLAN（Virtual Local Area Network），在建立邻居关系时需要根据接口的DRB优先级或者MAC地址的大小来选举DRB。DRB负责与网络中每台设备进行通信，最终使整个VLAN的LSDB（Link State DataBase）达到一致状态，减少了多台设备两两通信带来的巨大开销。DRB完成以下工作：

- 当网络中存在多于两台RB时，产生伪节点LSP（Link State PDUs）。
- 发送全序列号报文CSNP（Complete Sequence Number PDUs），同步LSDB。
- 指定Designated VLAN，指定某个Carrier VLAN转发用户报文及TRILL控制报文。
- 指定AF（Appointed Forwarder），每个CE（Custom Edge）VLAN只能由一个RB作为AF。

AF

AF是被指定用来转发流量的RB，由DRB负责选举，只有AF负责转发用户流量，非AF禁止转发用户流量。如图2-1所示，当服务器采用双上行接入TRILL网络时，如果服务器网卡不采用负载分担的方式接入，很容易产生环路。所以需要指定一台RB作为转发用户流量的RB。

TRILL 中的 VLAN

表 2-1 TRILL 中的 VLAN

VLAN名称	说明	承载流量
CE VLAN	用以接入TRILL网络的VLAN，通常配置在TRILL网络边缘设备上。用以生成组播路由。	普通以太报文。
admin VLAN	一种特殊的CE VLAN，用于承载TRILL网络网管流量。	TRILL网管流量。
Carrier VLAN	TRILL网络中用于转发TRILL数据报文和协议报文的VLAN。一台RB最多可配置三个不同的Carrier VLAN。入方向上，普通以太报文在Carrier VLAN中被封装成TRILL报文，出方向上，TRILL报文在Carrier VLAN中被解封装为普通以太报文。	TRILL数据报文和协议报文。
Designated VLAN	在网络规划中，为了实现TRILL网络的合并或分离，可能会在TRILL网络配置多个Carrier VLAN，但是只能由一个Carrier VLAN负责转发TRILL报文。这个被指定转发TRILL数据报文和协议报文的Carrier VLAN被称为Designated VLAN，以下简称DVLAN。	TRILL数据报文和协议报文。

Nickname

Nickname相当于IP地址，用来唯一标识一台交换机。一台RB仅支持配置一个Nickname，且须保证Nickname全网唯一。

Nickname具有优先级和根优先级两种属性，其主要作用分别为：

- 优先级主要用来在TRILL网络中解决Nickname冲突，具体的解决办法是：

1. 如果一个TRILL网络中有两台设备的Nickname相同，则Nickname优先级较大的被发布，较小的被抑制。
 2. 如果优先级相同，则比较两台RB的SystemID，较大的被发布，较小的被抑制。
- 根优先级的主要作用是用来参与选举组播树根，TRILL会选取Nickname根优先级最大的两个RB分别作为两棵组播树的树根。

端口类型

TRILL网络中交换机的端口分为以下几种：

- Trunk端口：交换机之间进行互联的端口，只允许转发TRILL数据报文和协议报文，不允许转发Native Ethernet报文。
- Access端口：接入端口，链路上只收发Native Ethernet报文和协议报文，不收发TRILL数据报文。
- Hybrid端口：混合端口，既可以转发TRILL数据和协议报文，也可以转发Native Ethernet报文。
- P2P端口：通常只有P2P类型的网络中，用于连接两台RB的接口，会设置为p2p类型。P2P Port是一种特殊的Trunk Port，所连接的交换机不进行DRB选举。

缺省情况下，RB的端口为P2P端口。

NET

与IS-IS协议类似，TRILL协议采用网络实体名称NET（Network Entity Title）来标识交换机本身的网络层信息。NET由以下三部分组成：

- Area ID：区域地址用来标识区域。与IS-IS不同的是，TRILL的区域地址规定为“00”。
- System ID：系统ID用来唯一标识一台主机或交换机，在设备的实现中，它的长度固定为48 Bit。

实际应用中，System ID可以自动生成也可以通过配置得到。通过命令**network-entity (TRILL)**配置网络实体名称时，可以指定System ID的值（请保证配置的System ID全网唯一）。若不配置该命令，则设备会自动生成System ID。自动生成的System ID与RB的桥MAC地址相同。

- SEL（Selector）：作用类似IP中的“协议标识符”，不同的传输协议对应不同的SEL。TRILL协议的SEL为“00”。

2.2 TRILL 报文格式

TRILL报文分为TRILL控制报文和TRILL数据报文，以下分别进行介绍。

2.2.1 TRILL 控制报文

TRILL协议通过在交换机之间交换控制消息来实现通信功能，本节介绍TRILL主要的控制报文。TRILL引入了IS-IS作为控制面协议，使用PDU数据包来进行控制信息处理，并对IS-IS PDU进行了扩展。

TRILL PDU的报文格式和IS-IS PDU一致，只是对TLV进行了扩展，这里仅对报文做简单介绍，报文字段详细描述请见《CloudEngine 12800系列 交换机配置指南-IP路由》中的“IS-IS配置”。

TRILL PDU 格式

TRILL协议使用的所有PDU数据包可以归纳为3类：HELLO、LSP、SNP。在所有的TRILL PDU中，起始的8个八位组字节都是公用的，如图2-2所示。

图 2-2 TRILL PDU 结构图

				No. of Octets
Intradomain Routing Protocol Discriminator				1
Length Indicator				1
Version/Protocol ID Extension				1
ID Length				1
R	R	R	PDU Type	1
Version				1
Reserved				1
Maximum Area Address				1
PDU Exclusive				
TLV				

PDU各个主要字段的含义及作用如下：

- **Intradomain Routing Protocol Discriminator:** 域内路由选择协议鉴别符，用来标识网络层协议数据单元。
- **Length Indicator:** 长度标识符，用来标识该固定头部的长度。
- **ID Length:** 用来标识该路由选择域内SystemID的长度。
- **PDU Type:** 用来标识PDU的类型。
- **Maximum Area Address:** 最大区域地址数，表示TRILL所允许的最大区域地址数量。目前，TRILL的区域地址只能为00。
- **PDU Exclusive:** PDU专有字段，根据PDU类型不同该字段也不同。该字段将在下文各类型PDU中介绍。
- **TLV:** 即Type/Length/Value（类型/长度/值），不同PDU类型所包含的TLV是不同的。

TRILL Hello PDU 格式

Hello PDU用于建立和维持邻居关系，广播网中使LAN Hello；非广播网络中则使用P2P Hello。它们的报文格式有所不同。

广播网中的Hello PDU格式如图2-3所示。

图 2-3 LAN Hello 格式

		No. of Octets
Reserved/Circuit Type		1
Source ID		ID Length
Holding Time		2
PDU Length		2
R	Priority	1
LAN ID		ID Length+1
Variable Length Fields		

P2P网络中的Hello PDU格式如图2-4所示。

图 2-4 P2P Hello 格式

		No. of Octets
Reserved/Circuit Type		1
Source ID		ID Length
Holding Time		2
PDU Length		2
Local Circuit ID		1
Variable Length Fields		

从图中可以看出，P2P Hello中的多数字段与LAN Hello相同。不同的是没有Priority和LAN ID字段，而多了一个Local Circuit ID字段，表示本地链路ID。

TRILL LSP PDU 格式

链路状态报文LSP（Link State PDUs）用于交换链路状态信息。LSP的报文格式，如图2-5所示。

图 2-5 TRILL LSP PDU 格式

				No. of Octets
PDU Length				2
Remaining Lifetime				2
LSP ID				ID Length+2
Sequency Number				4
Checksum				2
R	ATT	OL	IS Type	1
Variable Length Fields				

TRILL SNP PDU 格式

SNP（Sequence Number PDUs）通过描述全部或部分数据库中的LSP来同步各LSDB（Link-State DataBase），从而维护LSDB的完整与同步。SNP又可以分为如下两种：

- 全序列号报文CSNP（Complete SNP）：包括LSDB中所有LSP的摘要信息，从而可以在相邻交换机间保持LSDB的同步。在广播网络上，CSNP由DRB定期发送（缺省的发送周期为10秒）；在点到点链路上，CSNP只在第一次建立邻接关系时发送。

CSNP的报文格式如[图2-6](#)所示。

图 2-6 TRILL CSNP 格式

	No. of Octets
PDU Length	2
Source ID	ID Length+1
Start LSP ID	ID Length+2
End LSP ID	ID Length+2
Variable Length Fields	

- 部分序列号报文PSNP（Partial SNP）：只列举最近收到的一个或多个LSP的序号，它能够一次对多个LSP进行确认，当发现LSDB不同步时，也用PSNP来请求邻居发送新的LSP。

PSNP的报文格式如[图2-7](#)所示。

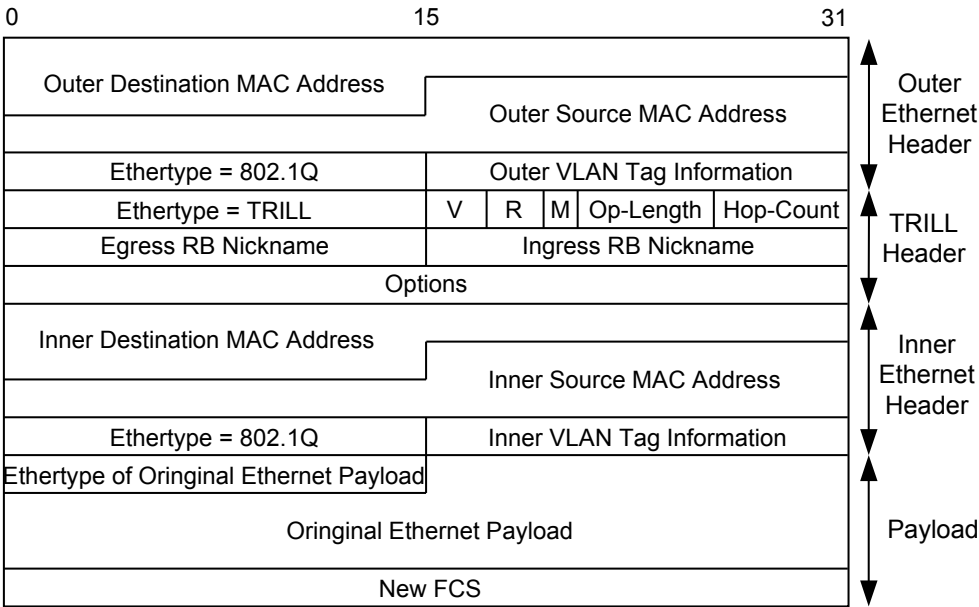
图 2-7 TRILL PSNP 格式

	No. of Octets
PDU Length	2
Source ID	ID Length+1
Variable Length Fields	

2.2.2 TRILL 数据报文

TRILL数据报文格式如[图2-8](#)所示。

图 2-8 TRILL 数据报文格式



TRILL数据报文在原始以太网报文之前添加了TRILL头和外层以太网头。其中，TRILL头中各字段的含义如下：

- **Ethertype**：类型固定为TRILL。
- **V**：版本号，当前为0。每个RB收到TRILL报文时都必须检查此位，若不正确则将其丢弃。
- **R**：为扩展而保留。Ingress RB打TRILL头时将其设为0，Transit RB和Egress RB均不关注此位。
- **M**：多目的属性。0表示已知单播报文，1表示组播、广播或未知单播报文。
- **Op-Length**：Options字段的长度，0表示没有Options字段。
- **Hop-Count**：跳数，用于防止环路。当TRILL报文的跳数为0时，RB将丢弃该报文。
- **Egress RB Nickname**：单播报文中，该字段为Egress RB的Nickname。组播报文中，该字段为用于转发的组播树树根的Nickname。
- **Ingress RB Nickname**：Ingress RB的Nickname。
- **Options**：可选字段，当Op-Length的取值为非0时才会有本字段。

关于TRILL数据报文在TRILL网络中如何转发，请参见[TRILL转发流程](#)。

2.3 TRILL 协议机制

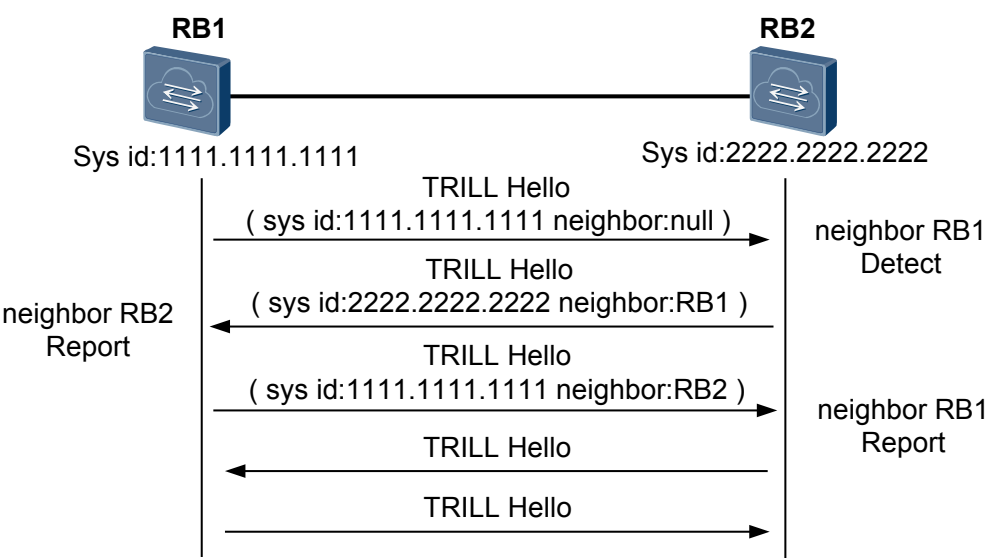
利用TRILL协议实现二层网络中RB之间的通信，需要完成以下几个步骤：

1. [TRILL邻居关系的建立](#)
2. [LSDB同步](#)
3. [路由计算](#)

TRILL 邻居关系的建立

TRILL协议通过发送Hello报文（即TRILL Hello PDU）建立邻居。由于端口模式不同，广播链路和P2P链路发送的Hello报文也有所不同，但是建立邻居的过程是相似的，图2-9所示为两台RB建立邻居的过程。

图 2-9 TRILL 邻居建立过程图



由图2-9可以看出，TRILL网络中两台RB建立邻居关系的过程可以概括为：

- 1. RB1发送TRILL Hello报文，RB2收到此报文后，检测到neighbor字段不包含本地MAC地址，则将自己和RB1的邻居状态标识为Detect；
- 2. RB2回复TRILL Hello报文给RB1，RB1收到该报文，检测到neighbor字段中包含了本地的MAC地址，于是将自己和RB2的邻居状态标识为Report；
- 3. RB1再次发送TRILL Hello报文给RB2，RB2收到这个neighbor字段包含本地MAC地址的Hello报文后，将RB1邻居状态标识为Report。此时，邻居关系正式建立；
- 4. 邻居关系建立之后，两台RB会定期发送Hello报文来维持邻居关系，如果一端连续发送3个Hello报文而没有收到对方的应答报文，则视为邻居失效。

在广播网中，为了提高收敛速度及通信效率，TRILL协议提供了以下机制：

● 选举DRB

在广播网中，任意两台RB之间都要传递信息。如果网络中有n台RB，则需要建立n*(n-1)/2个邻接关系。这使得任何一台RB的状态变化都会导致多次传递，浪费了带宽资源。为解决这一问题，TRILL协议定义了DRB，规定所有RB都只将信息发送给DRB，由DRB将网络链路状态广播出去。DRB的选举过程发生在TRILL邻居达到Detect状态之后。

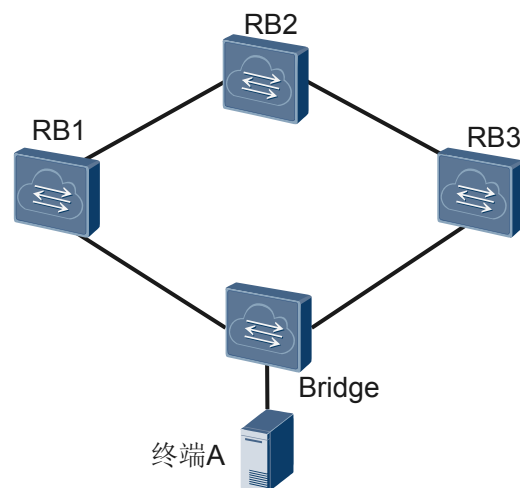
DRB的选举通过以下原则进行：

- 1. 首先比较邻居相连接口的DRB优先级，优先级较高的被优选为DRB；
- 2. 如果两端接口的DRB优先级相同，则比较邻居的MAC地址的大小，MAC地址较大的被优选为DRB。

- 指定AF

当未知单播流量或组播流量经过TRILL网络进行转发时，由于同一个VLAN内的流量可以广播，很容易形成环路。如图2-10所示，当用户A的用户组播流量通过普通二层交换机转发到TRILL网络中时，如果RB1与RB3同属一个VLAN，那么该组播流量就会同时分发到这两个RB，从而造成了环路。指定AF可以解决这个问题。AF由DRB针对CE VLAN选举而出，只有AF可以作为Ingress和Egress，非AF只能作为Transit RB。如果图2-10中RB1被指定为AF，则用户流量由RB1负责转发，不会经过RB3，从而避免了环路。

图 2-10 指定 AF 组网图



AF由DRB负责指定，其选取过程为首先检查TRILL网络入口的两台RB所使能的CE VLAN，具有与用户流量相同VLAN的RB被指定为AF；如果符合条件的RB不止一台，则依据如下原则进行选取：

1. 首先比较DRB优先级，DRB优先级较高的RB被选举为AF；
2. 如果DRB优先级相同，则选取系统MAC地址较大的RB为AF；
3. 如果系统MAC地址相同，则选取接口PortID较大的RB为AF；
4. 如果PortID相同，则选取SystemID较大的RB为AF。

 说明

- 只有接入RB的TRILL端口类型为Access或Hybrid时此RB才可以被指定为AF。
 - 在广播网中，如果有两台RB的Nickname发生冲突，则不会选取任何一台作为AF。
 - 在广播网运行过程中，如果DRB发生变化，则删除所有AF信息，并重新选举AF。
- 指定DVLAN

TRILL网络中会存在多个Carrier-VLAN的情况，此时网络接口需要指定一个DVLAN用来转发流量。TRILL在接口上发送协议报文、转发TRILL数据流量时，外层以太头的VLAN填充为链路上的DVLAN。DVLAN可以手动配置，也可以由DRB指定。

LSDB 同步

DRB选举完成之后，网络中每台RB上维护的LSDB达到统一的过程，称为LSDB同步。LSDB是生成转发表的根据，因此，要实现网络中的数据流量能够正确转发，LSDB同步是必不可少的过程。不同网络类型同步LSDB的过程也不相同。

以上内容仅为本文档的试下载部分，为可阅读页数的一半内容。如要下载或阅读全文，请访问：<https://d.book118.com/148054143032006046>