## 1 链路聚合

### 1.1 链路聚合概述

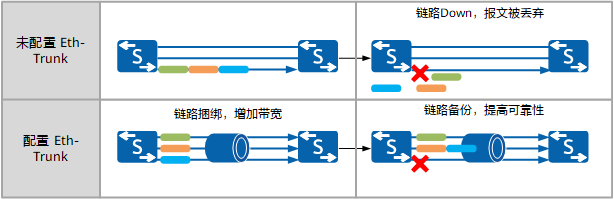
链路聚合通过将多个物理接口捆绑为一个逻辑接口，可以在不进行硬件升级的条件下，达到增加链路带宽的目的。

链路聚合技术主要有以下三个优势：

1. 增加带宽，链路聚合接口的最大带宽可以达到各成员接口带宽之和。

2. 提高可靠性，当某条活动链路出现故障时，流量可以切换到其他可用的成员链路上，从而提高链路聚合接口的可靠性。

3. 负载分担，在一个链路聚合组内，可以实现在各成员活动链路上的负载分担。



**链路聚合组和链路聚合接口**

链路聚合组LAG（Link Aggregation Group）是指将若干条以太链路捆绑在一起所形成的逻辑链路。

每个聚合组唯一对应着一个逻辑接口，这个逻辑接口称之为链路聚合接口。链路聚合接口可以作为普通的以太网接口来使用，与普通以太网接口的差别在于：转发的时候链路聚合组需要从成员接口中选择一个或多个接口来进行数据转发。

**成员接口和成员链路**

组成链路聚合接口的各个物理接口称为成员接口。成员接口对应的链路称为成员链路。

**活动接口和非活动接口、活动链路和非活动链路**

链路聚合组的成员接口存在活动接口和非活动接口两种。转发数据的接口称为活动接口，不转发数据的接口称为非活动接口。

活动接口对应的链路称为活动链路，非活动接口对应的链路称为非活动链路。

**活动接口数上限阈值**

设置活动接口数上限阈值的目的是在保证带宽的情况下提高网络的可靠性。当前活动接口数目达到上限阈值时，再向链路聚合接口中添加成员接口，不会增加链路聚合接口种活动接口的数目，超过上限阈值的链路状态将被置为Down，作为备份链路。

例如，有8条无故障链路在一个链路聚合接口内，每条链路都能提供1G的带宽，现在最多需要5G的带宽，那么上限阈值就可以设为5或者更大的值。其他的链路就自动进入备份状态以提高网络的可靠性。

**活动接口下限阈值**

设置活动接口数下限阈值是为了保证最小带宽，当前活动链路数目小于下限阈值时，链路聚合接口接口的状态转为Down。

例如，每条物理链路能提供1G的带宽，现在最小需要2G的带宽，那么活动接口数下限阈值必须要大于等于2。

**链路聚合模式**

根据是否启用链路聚合控制协议LACP（Link Aggregation Control Protocol），链路聚合分为手工模式和LACP模式。

### 1.2 手工模式链路聚合

手工模式下，链路聚合组建立、成员接口的加入由手工配置，没有链路聚合控制协议LACP的参与。该模式下所有活动链路都参与数据的转发，平均分担流量。如果某条活动链路故障，链路聚合组自动在剩余的活动链路中平均分担流量。

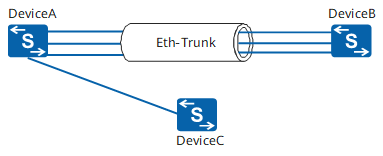
当需要在两个直连设备之间提供一个较大的链路带宽，而其中一端或两端设备都不支持LACP协议时，可以配置手工模式链路聚合。

### 1.3 LACP模式链路聚合

链路聚合控制协议LACP（Link Aggregation Control Protocol），是基于IEEE802.3ad标准的一种实现链路动态聚合与解聚合的协议，以供设备根据自身配置自动形成聚合链路并启动聚合链路收发数据，LACP模式就是采用LACP的一种链路聚合模式。聚合链路形成以后，LACP负责维护链路状态，在聚合条件发生变化时，自动调整链路聚合。

如图所示，DeviceA与DeviceB之间创建链路聚合接口，需要将DeviceA上的四个接口与DeviceB捆绑成一个链路聚合组。由于错将DeviceA上的一个接口与DeviceC相连，这将会导致DeviceA向DeviceB传输数据时可能会将本应该发到DeviceB的数据发送到DeviceC上。而手工模式的链路聚合接口不能及时检测到此故障。

如果在DeviceA和DeviceB上都启用LACP协议，经过协商后，链路聚合接口就会选择正确连接的链路作为活动链路来转发数据，从而DeviceA发送的数据能够正确到达DeviceB。



#### 1.3.1 系统LACP优先级

系统LACP优先级是为了区分两端设备优先级的高低而配置的参数。LACP模式下，两端设备所选择的活动接口必须保持一致，否则链路聚合组就无法建立。此时可以使其中一端具有更高的优先级，另一端根据高优先级的一端来选择活动接口即可。系统LACP优先级值越小优先级越高。

#### 1.3.2 接口LACP优先级

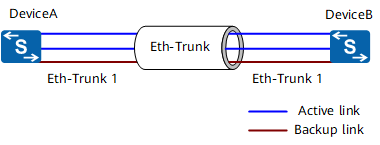
接口LACP优先级是为了区别同一个链路聚合组中的不同接口被选为活动接口的优先程度，优先级高的接口将优先被选为活动接口。接口LACP优先级值越小，优先级越高。

#### 1.3.3 成员接口间M:N备份

LACP模式链路聚合由LACP确定聚合组中的活动和非活动链路，又称为M:N模式，即M条活动链路与N条备份链路的模式。这种模式提供了更高的链路可靠性，并且可以在M条链路中实现不同方式的负载均衡。

如图所示，两台设备间有M+N条链路，在聚合链路上转发流量时在M条链路上分担负载，即活动链路，不在另外的N条链路转发流量，这N条链路提供备份功能，即备份链路。此时链路的实际带宽为M条链路的总和，但是能提供的最大带宽为M+N条链路的总和。

当M条链路中有一条链路故障时，LACP会从N条备份链路中找出一条优先级高的可用链路替换故障链路。此时链路的实际带宽还是M条链路的总和，但是能提供的最大带宽就变为M+N-1条链路的总和。



这种场景主要应用在只向用户提供M条链路的带宽，同时又希望提供一定的故障保护能力时。当有一条链路出现故障，系统能够自动选择一条优先级最高的可用备份链路变为活动链路。

如果在备份链路中无法找到可用链路，并且目前处于活动状态的链路数目低于配置的活动接口数下限阈值，那么系统将会把聚合接口关闭。

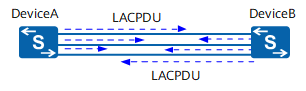
#### 1.3.4 LACP模式实现原理

LACP通过链路聚合控制协议数据单元LACPDU（Link Aggregation Control Protocol Data Unit）与对端交互信息，LACPDU报文中包含设备的系统优先级、MAC地址、接口优先级、接口号和操作Key等信息。

LACP模式链路聚合组建立的过程如下：

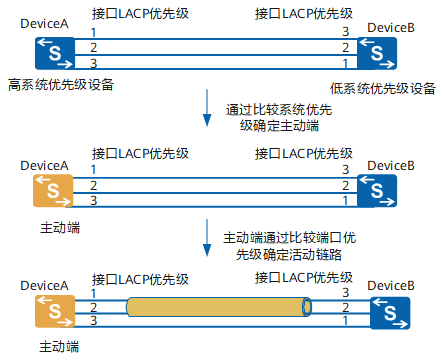
在LACP模式的Eth-Trunk中加入成员接口后，两端互相发送LACPDU报文。

如图所示，在DeviceA和DeviceB上创建链路聚合组并配置为LACP模式，然后向链路聚合组中手工加入成员接口。此时成员接口上便启用了LACP协议，两端互发LACPDU报文。



如图所示，两端设备均会收到对端发来的LACPDU报文。以DeviceB为例，当DeviceB收到DeviceA发送的报文时，DeviceB会查看并记录对端信息，然后比较系统优先级字段，如果DeviceA的系统优先级高于本端的系统优先级，则确定DeviceA为LACP主动端。如果DeviceA和DeviceB的系统优先级相同，比较两端设备的MAC地址，MAC地址小的一端为LACP主动端。

选出主动端后，两端都会以主动端的接口优先级来选择活动接口，如果主动端的接口优先级都相同则选择接口编号比较小的为活动接口。两端设备选择了一致的活动接口，活动链路组便可以建立起来，从这些活动链路中以负载分担的方式转发数据。



#### 1.3.5 手工模式和LACP模式链路聚合的对比

| **维度** | **手工模式** | **LACP模式** |
| --- | --- | --- |
| Eth-Trunk的建立方式 | Eth-Trunk的建立、成员接口的加入由手工配置，没有链路聚合控制协议的参与。 | Eth-Trunk的建立是基于LACP协议的，LACP为交换数据的设备提供一种标准的协商方式，以供系统根据自身配置自动形成聚合链路并启动聚合链路收发数据。聚合链路形成以后，负责维护链路状态。在聚合条件发生变化时，自动调整或解散链路聚合。 |
| 设备是否需要支持LACP协议 | 不需要 | 需要 |
| 数据转发 | 正常情况下，所有链路都是活动链路。所有活动链路均参与数据转发。如果某条活动链路故障，链路聚合组自动在剩余的活动链路中分担流量。 | 正常情况下，部分链路是活动链路。所有活动链路均参与数据转发。如果某条活动链路故障，链路聚合组自动在非活动链路中选择一条链路作为活动链路，参与数据转发的链路数目不变。 |
| 检测故障 | 只能检测到同一聚合组内的成员链路有断路等有限故障，但是无法检测到链路断连、错连等故障。 | 不仅能够检测到同一聚合组内的成员链路有断路等有限故障，还可以检测到链路故障、链路错连等故障。 |

#### 1.3.6 链路聚合负载分担方式

用户可以根据流量模型设置不同的负载分担方式，流量中某个参数变化越频繁，选择对应负载分担方式的流量就越均衡。例如，在网络中，如果报文的IP地址变化较频繁，那么选择基于目的IP地址、源IP地址或源IP和目的IP地址的负载分担模式更有利于流量在各物理链路间合理的负载分担；如果报文的MAC地址变化较频繁，IP地址比较固定，那么选择基于目的MAC地址、源MAC地址或源MAC和目的MAC地址的负载分担模式更有利于流量在各物理链路间合理的负载分担。

交换机可以基于报文的以下参数进行负载分担：

1. 源MAC地址

2. 目的MAC地址

3. 源MAC地址和目的MAC地址

4. 源IP地址

5. 目的IP地址

6. 源IP地址和目的IP地址

7. VLAN、源物理端口等