



基于 LACP 多激活检测方法和处理机制的研究

王培英^{1,2}, 肖志辉², 李本源²

(1.西南交通大学信息科学与技术学院 成都 610031; 2.迈普通信技术股份有限公司 成都 610041)

摘要: 虚拟交换机技术可将汇聚层的两台物理设备虚拟为一台虚拟设备, 该虚拟设备通过跨设备链路聚合与接入层设备相连。这种简化后的组网因无需配置 MSTP、VRRP 而简化了网络配置。当成员设备出现故障时, 接入层设备依靠跨设备链路聚合, 提高了网络连接可靠性。但虚拟设备内, 成员设备间链路出现故障时, 虚拟设备发生分裂, 形成两个虚拟设备, 出现多激活, 导致网络瘫痪。因此如何快速检测到多激活及检测后的处理, 对虚拟交换技术的发展是极其重要的, 介绍了一种多激活检测方法, 可较快速地检测到多激活, 是一种高效的多激活处理方法; 还提出了一种多激活处理方法, 可有效保证网络业务正常运行。

关键词: 链路聚合控制协议; 虚拟交换技术; 虚拟交换技术分裂; 多激活检测; 多激活处理

doi: 10.3969/j.issn.1000-0801.2013.05.017

Research on LACP Multi-Active Detection and Processing Method

Wang Peiying^{1,2}, Xiao Zhihui², Li Benyuan²

(1.School of Information Science and Technology, Southwest Jiaotong University, Chengdu 610031, China;

2.Maipu Communication Technology Co., Ltd., Chengdu 610041, China)

Abstract: Virtual switching technology allows two physical devices of convergence layer to be together into a single virtual device, and the device in the access layer connects virtual device by link aggregation across devices. This simplified network is able to simplify network configuration without using MSTP and VRRP. When the member device goes wrong, the devices of the access layer is able to improve the network connection reliability relying on cross-device link aggregation. But when the link between members of the virtual device goes wrong, the virtual device will split and form two virtual devices, resulting in multi-active and causing network paralysis. Therefore, how to quickly detect multi-active and handle multi-active is extremely important for virtual switching technology development. Based on this, a fast multi-active detection method was described, which could detect multi-active quickly, and an efficient multi-active processing method was also proposed, which could effectively guarantee the normal operation of network business.

Key words: link aggregation control protocol, virtual switching technology, virtual switching technology splitting, multi-active detection, multi-active processing

1 引言

随着互联网业务类型的不断增多和用户数的不断增长, 网络承载的数据业务流量越来越大。实时通信对服务质量的要求也不断提高, 导致网络的高可用性日益成为网络

使用者关注的焦点, 因此如何提高网络可用性是目前最热门的话题之一。网络的可用性可通过以下 3 种方法提高:

- 增加网络拓扑中的冗余节点和备份设计;
- 增加网络设备自身冗余部件和热备份机制设计;
- 增加网络中传输线路备份设计。



目前,传输线路备份设计是提高网络可用性的必备方法,其中应用最广泛的技术为生成树协议(STP)和链路聚合(link aggregation)。生成树协议通过阻塞某些链路达到传输线路备份的目的,属于链路间冗余设计;链路聚合通过聚合技术提供逻辑链路内的冗余方案,不但没有阻塞某些链路,而且扩展了逻辑链路带宽,已成为网络部署者必采用的组网方案。

链路聚合技术的实质是将两台设备间的多条物理链路组合成一条逻辑链路,称为链路聚合组(link aggregation group, LAG),链路聚合组中的端口称为成员端口。

链路聚合组分为静态聚合组和动态聚合组。静态聚合组内的成员端口禁止启动 LACP(link aggregation control protocol, 链路聚合控制协议),因此不与对端交互 LACPDU(link aggregation control data unit, 链路聚合控制协议数据单元);动态聚合组内的成员端口启动 LACP,与对方交互 LACPDU。

目前,层次化的网络结构一般分为核心层、汇聚层和接入层。网络中主要存在两种形态的交换机通信设备:盒式设备和框式分布式设备。由于框式分布式设备具有高性能、高密度端口的组件及高可用性,因此经常被应用于重要场合,如核心层、汇聚层。汇聚层与接入层之间的常见组网方式如图 1 所示。

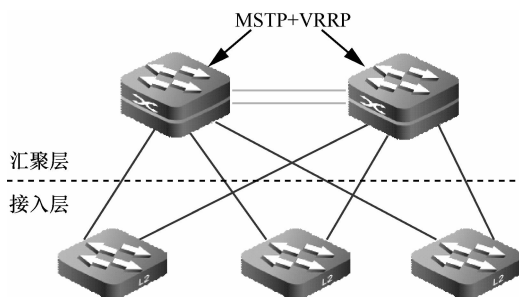


图 1 汇聚层与接入层之间的简化组网示意

因为增加网络拓扑中的冗余节点和备份设计可以提高网络的可用性,所以在部署网络环境时,通常在汇聚层部署两台交换机,两台交换机之间使用两条物理链路相连,然后所有接入层交换机都通过两条链路分别与汇聚层交换机相连。

当采用这种网络结构进行 2 层技术组网时,设计了一些冗余链路,虽然增强了网络的可靠性,但导致网络出现了环路,因此必须配置 MSTP 阻塞一些链路,达到消除环路的目的,这样就导致这些阻塞链路的带宽没有被利用,从而造成带宽资源的浪费。在实际应用中,MSTP 往往由于设备故障或者链路中断等原因发生拓扑震荡,且 MSTP 的收敛时间较长,可能在收敛时间内导致网络不能传输业务

流量,从而影响网络的服务质量。

当采用这种网络结构进行 3 层技术组网时,通常采用 VRRP 实现冗余备份。当处于主控(master)状态的交换机发生故障时,处于备份(slave)状态的交换机将切换成主控状态,切换时间一般为几秒钟,在故障恢复时间内仍会导致网络不能传输业务流量。同时 VRRP 会导致网络拓扑复杂,给网络维护者带来管理困难等问题。

为了克服上述问题,近年来业界一些设备厂商提出了针对框式分布式设备应用的虚拟交换技术(virtual switching technology, VST)。

2 虚拟交换技术

虚拟交换技术是一种基于内部分布式交换网络,把多台物理设备虚拟为一台虚拟设备的技术。如图 2 所示,汇聚层的两台设备通过虚拟交换链路(virtual switching link, VSL)连接,形成一个虚拟交换域(virtual switching domain, VSD),接入层的设备通过跨设备链路聚合,上联到虚拟设备。

2.1 VST 概述

虚拟设备内的成员设备有两种角色:主控设备负责整个虚拟设备的管理和控制,备份设备负责本地设备的基本硬件管理,并接受主控设备的管理和控制。虚拟设备内每一个成员设备都需要一个成员编号(member ID)进行唯一标识。虚拟设备内每个成员设备都有一个成员优先级,优先级越高(数值越大),当选为主控设备的可能性越大。

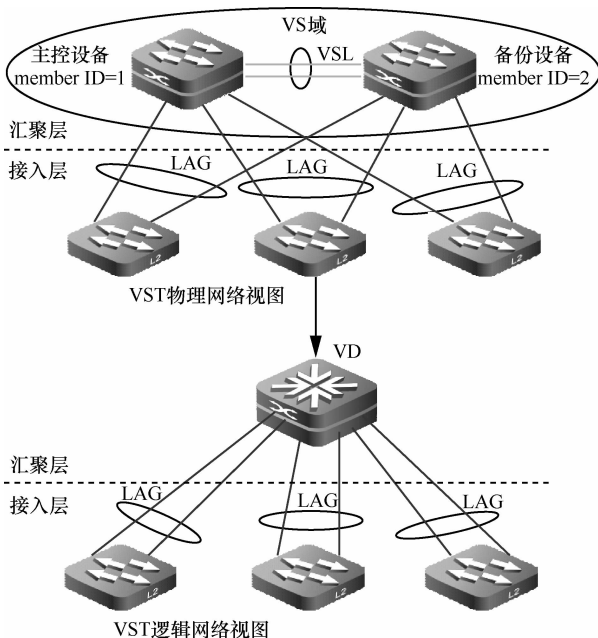


图 2 虚拟化视图

VSL 是两台成员交换机之间,由一条或者多条物理链路聚合而成的逻辑链路。当多台设备通过 VST 形成一台虚拟设备时,该虚拟设备根据选举规则,先选举出主控设备。主控设备使用自己的启动配置文件启动,启动完成后,将配置批量同步给所有备份设备,使主控设备完成初始化,至此,虚拟设备形成。

虚拟设备运行过程中有新的成员设备加入时,虚拟设备也会将配置批量同步给新设备,新设备以同步过来的配置完成初始化,而不再读取本地的启动配置文件^[4]。

所有设备初始化完成后,虚拟设备作为一台逻辑设备在网络中运行。主控设备作为虚拟设备系统的管理中枢,负责响应用户的登录需求,即用户无论使用何种方式登录虚拟设备中的成员设备,其实质都是登录到主控设备;对成员设备的配置,其实质都是对主控设备进行配置。主控设备负责将用户的配置实时同步给虚拟设备内的各个备份设备,使虚拟设备内各成员设备的配置随时保持高度一致性^[4]。

VST 因具有以下优势而逐渐被广泛使用。

(1) 高可靠性

虚拟设备的成员设备间采用 VSL 聚合相连,接入层设备通过跨设备链路聚合与虚拟设备相连,既提供了冗余链路,又实现了动态负载均衡,充分利用所有带宽。

(2) 简化网络拓扑

通过 VST 形成的虚拟设备在网络中相当于一台设备,该虚拟设备通过跨设备链路聚合与下联设备相连,由于不存在环路,所以无需配置 MSTP 和 VRRP。各种控制层协议运行在一台虚拟设备上,减少了设备间大量协议报文的交互,缩短了路由收敛时间。

(3) 简化管理

两台或者多台设备形成一台虚拟设备后,管理员可以对多台设备进行统一管理,即管理员登录虚拟设备内的任意一台成员设备,即可对虚拟设备中的多台设备进行配置和管理。

2.2 VST 分裂和多激活

当虚拟设备内的 VSL 链路出现故障,导致两个相邻成员设备在物理上不连通时,一个虚拟设备变为两个虚拟设备,这个过程称为 VST 分裂。

由于虚拟设备中各成员设备的配置信息实时高度一致,因此网络中将会出现配置完全相同的两台设备,这将导致业务流量无法正确转发和选路,如图 3 所示,两台设备配置了完全相同的 IP 地址,出现 IP 地址冲突,导致 3 层网络不可使用,这种现象称为多激活(multi-active)。

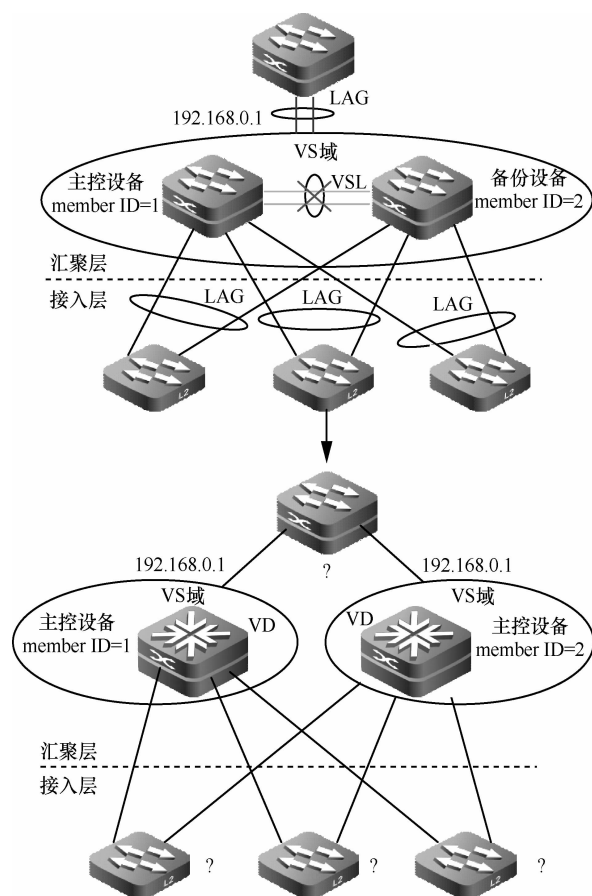


图3 虚拟设备分裂和多激活产生示意

因此,当虚拟设备发生分裂时,保证网络正常运行对 VST 技术的快速发展是非常重要的。

如何快速检测多激活的产生以及检测到多激活后应如何解决其给网络带来的影响问题是本文研究的重点。本文介绍了一种可较快速检测多激活的快速多激活检测(multi-active detection, MAD)方法和可有效保证网络业务正常运行的高效多激活处理方法。

3 基于 LACP 的多激活检测方法和处理机制

3.1 多激活检测方法

多激活检测方法就是通过某些手段检测到多激活。业界一些厂商提出的比较常见的多激活检测方法有以下两种。

(1) 基于 BFD 的多激活检测(BFD MAD)方法

该方法实现的前提条件是,在一个 VLAN 虚拟接口配置 BFD 和主从 MAD IP 地址。虚拟设备正常运行时,只有主控设备上配置的 MAD IP 地址生效,BFD 会话处于 down 状态;当 VST 分裂后会形成两个虚拟设备,不同虚拟设备中主控设备上配置的 MAD IP 地址均会生效,BFD 会话被激



活,此时检测到多激活。基于 BFD 的多激活检测方法需要占用端口资源,使用专用的检测端口和 3 层 VLAN 接口,从而使这些端口和接口不能再用于传输普通业务流量。

(2) 基于 LACP 的多激活检测(LACP MAD)方法

该方法通过扩展 LACP 报文内容实现。实现的前提条件是,虚拟设备与周边设备采用动态汇聚方式相连,借助业务动态汇聚组成员端口发送 LACPDU 报文实现多激活检测,不占用额外的设备端口资源。

网络管理者需要根据实际网络环境部署多激活检测方法。在设备端口资源紧张的情况下,LACP MAD 检测方法彰显出优势。

3.2 基于 LACP 的多激活检测原理

第 2 节提及,虚拟设备采用跨设备链路聚合与接入层设备相连。如果采用动态聚合方式,那么接入层设备会与虚拟设备交互 LACPDU 报文。

LACP MAD 通过扩展 LACPDU 报文实现,即在 LACPDU 的扩展字段内定义一个新的 TLV(type length value)数据域,用于交互虚拟设备的 MAD ActiveID (以下简称 ActiveID,表示虚拟设备中主控设备的成员编号)。

(1) LACP MAD 报文格式

LACP MAD 报文格式如图 4 所示。

Field	Octet
IEEE 802.3 标准 LACPDU 结构	74
TLV_type=MAD information	1
MAD_information_length = 10	1
MAD DomainID	1
MAD ActiveID	1
MAD Bridge MAC	6
Reserved	40
FCS	4

图 4 LACP MAD 报文格式

其中,MAD DomainID 为虚拟交换域,MAD ActiveID 为当前虚拟设备内主控设备的成员编号,MAD Bridge MAC 为虚拟设备的桥 MAC 地址。

(2) LACP MAD 检测机制原理

虚拟设备中各成员设备开启 LACP MAD 功能后,在发送 LACPDU 时会携带 MAD TLV 扩展域,其中最重要的字段是 ActiveID,此值在一个虚拟设备内具有唯一性,非虚拟设备内的设备在收到带有 MAD TLV 的 LACPDU 后,会将此报文向聚合组中的其他成员端口进行透传。虚拟设备正常工作时,通过这种方式,虚拟交换系统中的所有成员设备

都能收到其他成员设备发送的带有 MAD TLV 的 LACPDU,其中备份设备将收到的带有 MAD TLV 的 LACPDU 通过 VSL 发送到主控设备,由主控设备统一处理。主控设备在对接收到的 LACPDU 进行处理的过程中,发现各成员设备发送的 LACPDU 具有统一的 ActiveID,则认为虚拟设备正常。虚拟设备正常运行时 ActiveID 的交互如图 5 所示。

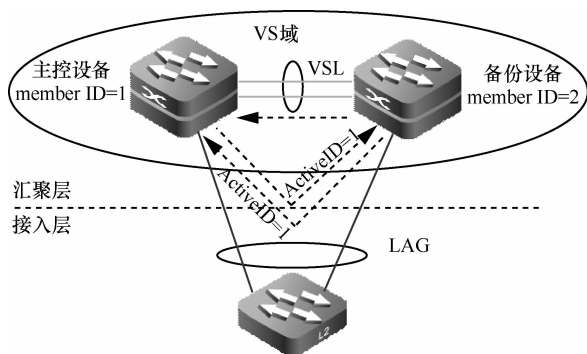


图 5 虚拟设备正常运行时 ActiveID 的交互

当虚拟设备发生分裂时,一个虚拟设备变为两个虚拟设备,不同虚拟设备内的成员设备发送的 LACPDU 携带的 ActiveID 不同。当这些 LACPDU 到达其他虚拟设备的主控设备时,主控设备通过比较发现 ActiveID 与自身的 ActiveID 不同,即检测到多激活,具体如图 6 所示。同时动态汇聚组的各成员端口会迅速发送 LACPDU 报文,便于其他虚拟设备在最短的时间内快速检测到多激活。

3.3 多激活处理

检测到多激活后,主控设备将从接收的 MAD TLV 中解析出 ActiveID,并按照下述算法步骤判断本虚拟设备是否进入禁用状态。

(1) 比较虚拟设备运行时间:主控状态运行时间小的虚拟设备进入禁用状态,执行步骤(4)。如果各虚拟设备的主控运行时间相等,则执行步骤(2)。

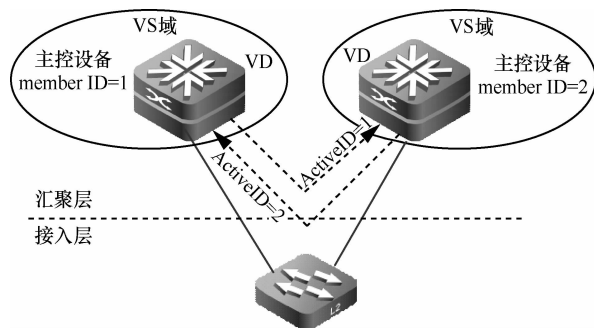


图 6 虚拟设备分裂后 ActiveID 的交互

(2)比较成员优先级:主控成员优先级低的虚拟设备进入禁用状态,执行步骤(4)。如果各虚拟设备的主控成员优先级相同,则执行步骤(3)。

(3)比较主控的成员编号:主控成员编号大的虚拟设备进入禁用状态,执行步骤(4)。

(4)进入禁用状态的虚拟设备需要完全从网络中移除,相关动作如下:

- 关闭该虚拟设备上所有业务的以太网端口(包括检测端口);
- 关闭该虚拟设备上的所有 VLAN 接口。

经过步骤(4)的处理之后,网络环境中不会出现两台配置相同的虚拟设备同时运行的情况,确保了网络业务的正常运行。

3.4 实验过程

检验上述 LACP MAD 方法及处理机制的实验环境,验证过程具体介绍如下。

实验环境:两台支持 VST 的交换机 A 和交换机 B,一台支持 MAD LACP 报文透传的交换机 C。物理连接如图 7 所示。交换机 A 和交换机 B 之间通过支持 VSL 的端口相连;交换机 C 通过普通业务端口分别与交换机 A 和交换机 B 相连。

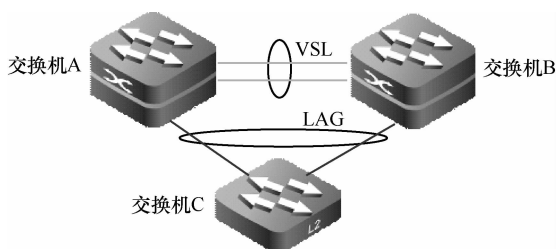


图 7 实验环境

配置及操作步骤具体介绍如下。

(1)在交换机 A 和交换机 B 上,完成 VST 的基本配置,使两台设备形成一台虚拟设备。

(2)在主控设备上配置动态聚合组,并将与交换机 B 和交换机 C 相连的端口加入聚合组;在交换机 C 上配置动态汇聚组,并将与交换机 A 和交换机 B 相连的端口加入聚合组。

(3)在主控设备上开启该聚合组的 LACP MAD 功能。

(4)将 VSL 断开,模拟 VST 分裂。

观察交换机 C 上与交换机 B 相连接的端口的物理链路状态是否为 down,即端口的 up 指示灯是否由亮变灭。若为 down,表示多激活检测成功,处于禁用状态的交换机 B 将业务端口关闭。

4 结束语

按照上述实验过程,经过测试发现,本文提出的多激活检测方法可以检测到多激活发生;在检测到多激活后,通过多激活处理方法计算选择处于禁用状态的虚拟设备,并将该虚拟设备的所有业务端口关闭,可解决网络中出现两台配置相同的虚拟设备后,引起的网络瘫痪问题。由此保证了网络的正常运行,对于推进 VST 的快速发展和应用具有重要意义。

另外,通过实际应用环境发现,在 VST 分裂到多激活被检测到的时间段内,网络仍处于瘫痪状态,仍会导致网络短时间内业务中断,这是网络使用者不能接受的。因此缩短业务中断时间、提高 LACP MAD 多激活检测的速度仍是后续工作和学习中需要继续探讨和研究的内容。

参考文献

- 1 陆敏飞. 链路聚合技术在分布式以太网交换机上的研究与实现. 浙江工业大学硕士学位论文, 2009
- 2 Mark A D, McDonald R, Antoon W R. Network Fundamentals CCNA Exploration Companion Guide. Cisco Press, 2009
- 3 Wayne Lewis, Ph D. LAN Switching and Wireless CCNA Exploration Companion Guide. Cisco Press, 2009
- 4 杭州华三通信技术有限公司. IRF2.0 技术白皮书, 2009
- 5 杭州华三通信技术有限公司. H3C S12500 IRF2 技术白皮书, 2012
- 6 迈普通信科技股份有限公司. VST 技术白皮书(V0.2), 2011
- 7 Cisco. Cisco Catalyst 6500 Series Virtual Switching System (VSS) 1440 White Paper, 2010
- 8 锐捷网络. 锐捷 VSU 技术白皮书, 2011
- 9 IEEE 802.1AX. IEEE Standard for Local and Metropolitan Area Networks-Link Aggregation, 2008

[作者简介]



王培英,女,西南交通大学硕士研究生,主要研究方向为通信与信息系统。

肖志辉,男,迈普通信科技股份有限公司首席执行官、研究院院长,中国通信标准化协会(CCSA)理事。

李本源,男,迈普通信科技股份有限公司系统级工程师。

(收稿日期:2013-04-12)