H3C (/cn/) 数字化解决方案领导者

S-MLAG解决方案介绍

1 摘要

本文介绍了互联网接入层可靠性设计的发展和实现。随着互联网业务的快速的扩张和网络架构的发展，接入层可靠性也经历了从分到合，又从合到分的发展历程。本文讲述这一过程的同时也会详细阐述去堆叠技术的原理和实现方式。

2 服务器接入发展

随着互联网的爆发式增长，数据中心的规模也越来越大，数据中心网络架构有传统的二层架构，过渡成为了OSPF和BGP的全三层架构，理论上BGP三层架构组网中可以承载100000+的服务器。

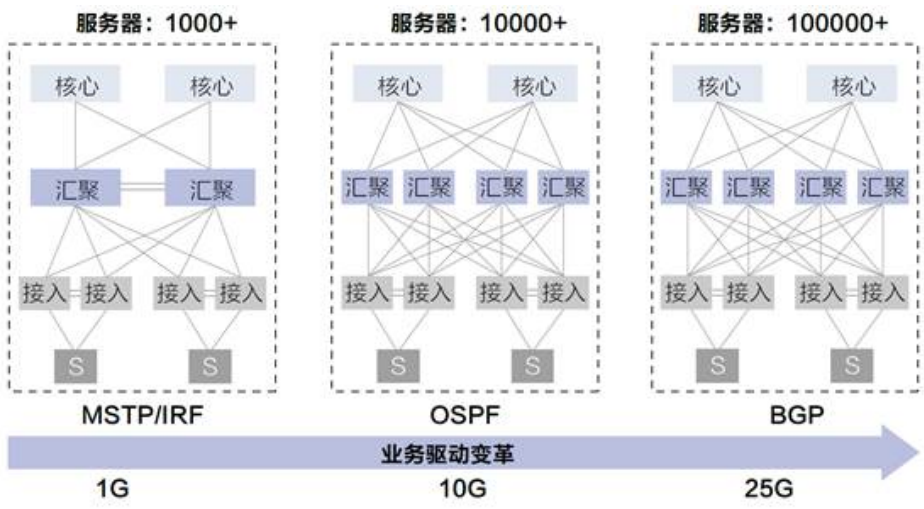


图1 服务器接入发展

随着互联网数据中心架构的发展，服务器接入的发展也经历了三个阶段，这三个阶段分别为：

联系我们

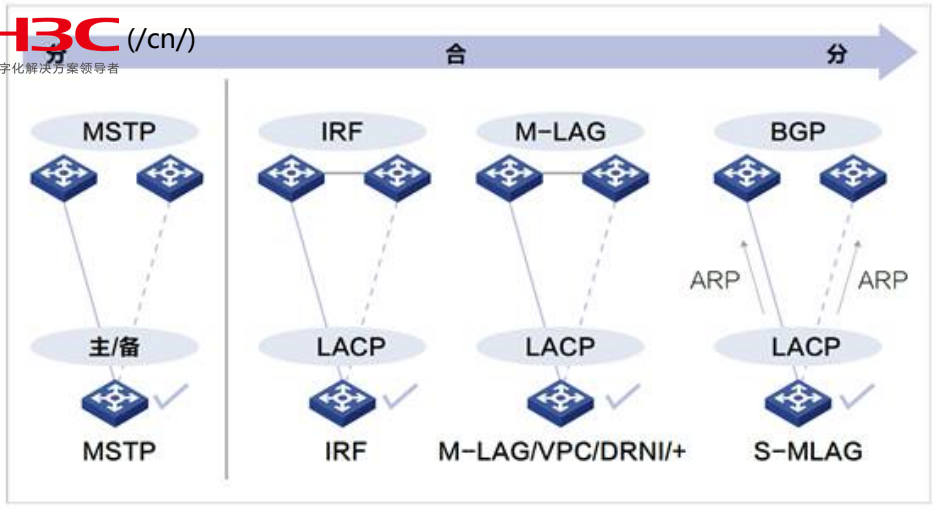


图2 服务器接入的发展阶段

第一阶段，接入层的交换机不支持虚拟化，接入交换机独立运行，服务器网卡工作在主备模式，主用设备故障时网卡会发生切换，交换机表项依靠数据流量进行刷新。

第二阶段，接入层交换机支持IRF等虚拟化技术，支持将多台设备虚拟化为一台，支持与服务器进行链路聚合，从而在提高可用性的同时实现链路双活，提高链路利用率。

第三阶段，由于IRF等虚拟化技术实现门槛较高，同时也存在控制层面唯一，升级困难等问题。在M-LAG和S-MLAG的技术出现后，实现了在接入层交换机控制层面分离的情况下实现了接入层链路双活接入，同时S-MLAG实现相对非常简单，已经被互联网用户所接受。S-MLAG又称之为“去堆叠”，接下来的文章向大家详细阐述S-MLAG技术的原理和实现。

3 去堆叠技术实现原理

链路聚合模式分为静态聚合和动态聚合，与服务器聚合的对应模式如下：

网络聚合	服务器聚合
非聚合模式	mod=1，（active-backup）Active-backup policy
	mod=5，（balance-tlb）Adaptive transmit load balancing
	mod=6，（balance-alb）Adaptive load balancing
静态聚合模式	mod=0，（balance-rr）Round-robin policy
	mod=2，（balance-xor）XOR policy
	mod=3，（broadcas）
动态聚合模式	mod=4，（802.3ad）IEEE 802.3ad Dynamic link aggregation

在静态聚合模式中可以很简单实现跨设备链路聚合，只要接口UP同时关键配置一致时接口就可以处于聚合选中，但是静态聚合缺乏LACP报文对链路的监控和与邻居的协商机制，在聚合模式选择中通常选择动态聚合模式，服务器linux操作系统称为mod5。为了在去堆叠方案中实现跨设备链路聚合，需要解决两个问题：

3.1 在动态链路聚合中，如何让服务器认为连接对端的接入交换机是同一个网络设备？

图3 LACPDU报文

联系我们

图3为LACPDU报文，在动态聚合中当Partner_System_Priority和Partner_System一致时，则认为对端设备为同一个设备。同时本端的不同端口接收LACPDU报文中要求Pantner_Port不一致Partner_key一致时则可以聚合成功。

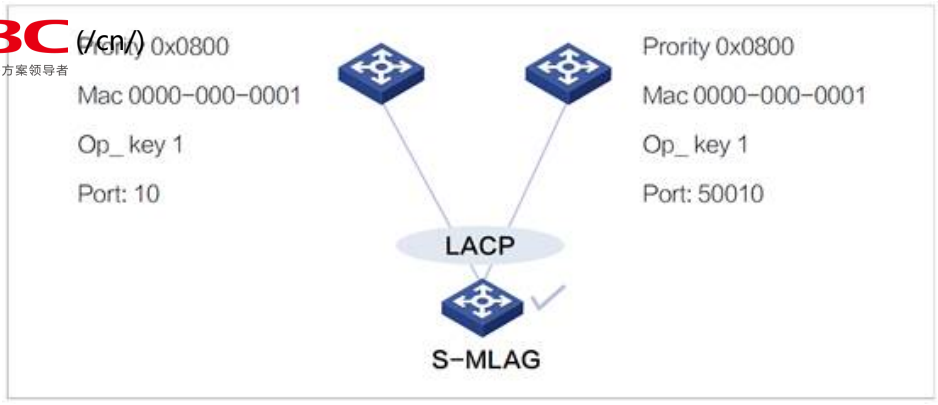


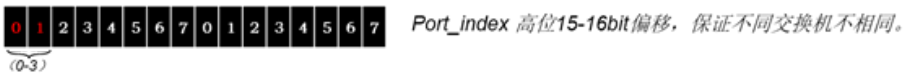
图4 S-MLAG配置实现

S-MLAG配置实现如下:

配置LACP的系统MAC地址：lacp system-mac mac-address（xxxx-xxxx-xxxx）

配置LACP的系统优先级：lacp system-priority priority （0-65535）

配置LACP的系统编号 lacp system-number number（1-3）



配置聚合接口加入S-MLAG组 port s-mlag group group-id（1 ~ 1024）

3. 2 在堆叠方案中两台设备虚拟为一台，控制层面只有一个，两台设备的表项依靠LIPC进行同步，M-LAG方案中依靠M-LAGPDU进行同步，在S-MLAG方案中两台设备控制层面完全独立，路由、ARP、MAC表项是如何同步的呢？

1. 接入层交换机将主机的ARP路由转换成为直连路由，并引入到BGP路由完成路由同步，到达服务器的流量由32位主机路由来引导。

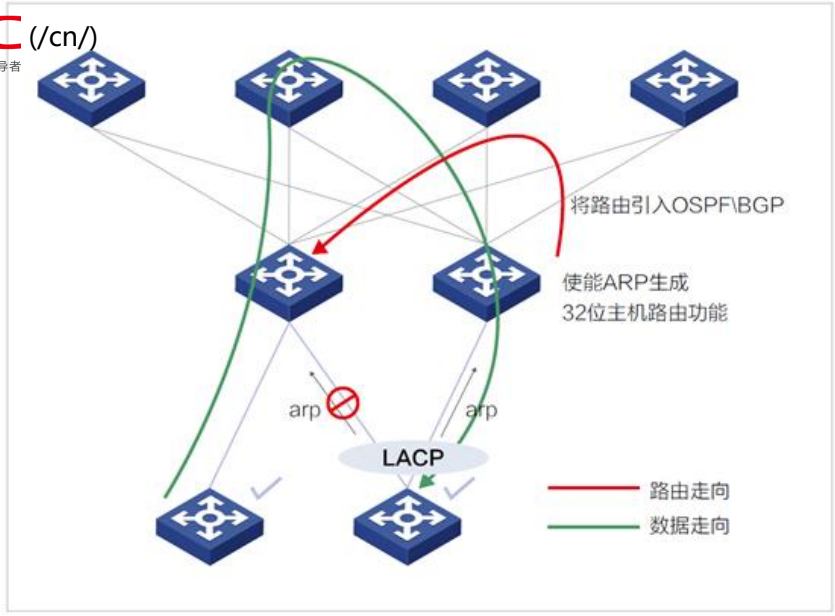


图5 ARP生成直连路由引入OSPF\BGP路由协议中

如图5所示，将ARP生成的直连路由引入到OSPF\BGP路由中，从而达到全网设备主机路由的同步。

2. 服务器在发送ARP请求和应答需要在聚合的所有成员接口网卡进行发送与接收，又叫做“ARP双发”。实现去堆叠设备的ARP和MAC表项同步。按照流量HASH原理，ARP报文会按照算法选择BOND成员网卡中的一个进行发送，这样去堆叠的两台设备ARP表项就不会同步。这时候需要修改服务器操作系统内核，在发送ARP报文时在所有BOND的成员网卡发送。

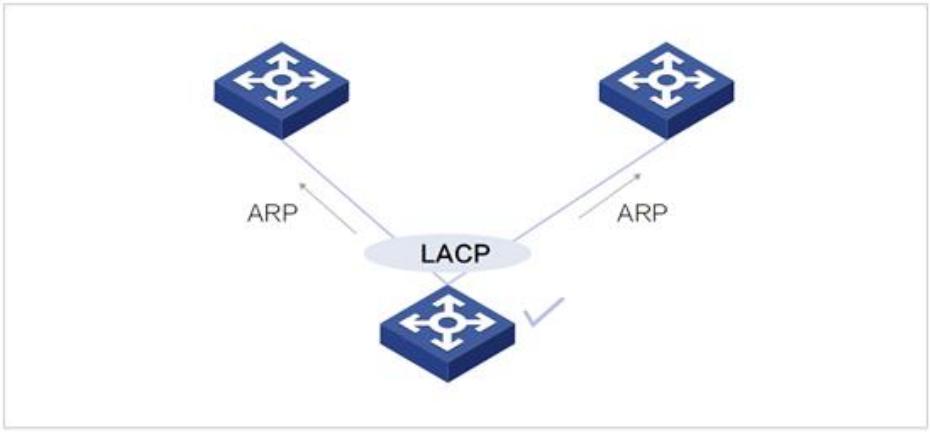


图6 ARP双发

在服务器上看聚合接口状态是两个物理网卡聚合在一起的逻辑接口，单条物理线路的UP/DOWN逻辑的接口状态并没有变化，并不能触发服务器发送免费ARP的更新，如果没有32位主机路由的牵引可能会造成流量负载不均衡。所以需要服务器的OS内核进一步优化，当服务器成员接口发生UP/DOWN时，发送免费ARP进行更新。

3. 两台接入交换机连接服务器的三层网关接口MAC地址更改为相同的MAC地址，避免两台设备发送不同的MAC地址造成服务器侧ARP表项反复切换。

3.3 为了保障故障快速切换，在网络设备的配置上还需要做如下优化配置：

1. 接入设备开启BUM隔离、本地ARP代理，避免在相同TOR接入服务器相互学习到真实MAC的ARP，因为在服务器某一服务器网卡DOWN之后，该TOR的其他服务器访问该服务器还使用真实MAC封装，就会导致访问异常。这时需要TOR交换机开启BUM流量隔离，接入交换机网关开启本地ARP代理功能，在同一TOR下服务器的二层流量也需通过三层转发。

2. 开启TOR上行接口监控。如果TOR的上联接口全部DOWN时，下行接入服务器无法感知到，服务器会继续向故障TOR发送数据。这时需要配置monitor-link监控上行线路，当上行接口全部故障时，关闭下行接口，同时上行接口UP时，下行接口也需要延迟UP。因为上行接口的BGP等路由协议收敛速度远远大于下行接口链路聚合收敛速度，所以需要在上行接口UP时，下行接口延迟UP。

3. 开启ARP主动探测。去堆叠方案中去往服务器的流量都是由32位主机路由进行引导，如果出现静默主机无法生成ARP主机路由的情况，在此种情况下可以通过交换机主动探测ARP功能进行优化。

4 堆叠与去堆叠技术对比

在去堆叠方案中将两个控制层面完全独立的交换机上面实现了服务器的双活接入， 下面表1是S-MALG、M-LAG和堆叠的三种接入方案的详细对比：

	去堆叠（S-MLAG）	M-LAG	堆叠
	两台设备控制面分离，耦合度小 故障只影响一台设备，可靠性较高	两台设备控制面分离，耦合度小 故障只影响一台设备，可靠性较高	两台控制面合一，耦合度高 故障影响整个堆叠体，影响面较大， 可靠性相对较低
配置&维护复杂性	配置有要求两端同步，较为简单	配置要求两边完全一致，否则可能导致功能不可用，较复杂	集中式配置，较简单
收敛性能	<5s	<1s	<1s
升级风险	两台设备完全独立，升级不应影响业务， 升级风险非常低	成员节点单独升级，不影响业务流量的转发，升级风险低	成员节点单独升级会影响控制面， 有黑洞期，影响业务流 ISSU 升级技术复杂度高、 软件工程复杂度高，升级风险较大
三层转发	双活接入，但是需要开启接入设备广播隔离，arp 代理	双活接入，依赖 VRRP 或三层接口 特殊配置	双活接入不依赖 VRRP，无需特殊配置
特性支持能力	较弱，需要分别开发适配	较弱，需要分别开发适配	全面，几乎所有特性
网络方案设计	场景相对单一，适用于全三层场景，服务操作系统需要特殊开发。	相对复杂，在三层网络侧需接双节点设计	简单，接照单节点设计

表1 S-MALG、M-LAG和堆叠的三种接入方案的对比

去堆叠具有良好的兼容性，可以实现不同厂家设备的异构，这是M-LAG和堆叠无法做到的。M-LAG和堆叠在底层实现十分复杂，需要进行大量表项和状态同步工作，去堆叠对交换机LACP协议进行简单的改动就可以实现，三层表项通

过现有路由协议同步，二层表项通过服务器“ARP双发”实现。

去堆叠也有其劣势，为了实现“ARP双发”需要修改操作系统内核代码，对维护和开发人员要求都非常高，互联网使用的操作系统比较单一都是Linux操作系统，操作系统版本统一，容易完成修改和适配。设备发生故障收敛时相对于堆叠收敛还是有一定差距的，同时S-MLAG方案的适配场景相对单一，必须是全三层组网，接入的二层隔离也限制了组播等一些应用。

5 去堆叠技术总结

S-MLAG解决方案在不更改现有服务器接入模式的情况下，经过对交换机LACP协议简单的修改完成跨设备动态链路聚合，但是S-MLAG解决方案也有一定的局限性，需要对操作系统内核ARP部分进行修改，门槛要求较高。在特定组网中经过对协议的简单改造解决复杂的问题，S-MLAG解决方案为我们对网络架构设计提供了另外一种思路。

感谢您对本刊物的关注，如果您在阅读时有何感想，请点击
(/cn/aspx/voteforms/frm50.aspx?doctitle=s-mlag%u89E3%u51B3%u65B9%u6848%u4ECB%u7ECD&magazine=&docurl=https://www.h3c.com/cn/d_20
反馈。