H3C (/cn/)
数字化解决方案领导者

互联网技术详解|数据中心接入双归属方案剖析

【发布时间：2021-06-17】

在网络规划与设计中，为了保证业务的可靠性和连续性，均需要考虑各种冗余设计，如链路冗余、节点冗余等。因此，服务器通常采用双归属接入到网络中，如图1所示，服务器通过两条链路分别连接到两台Leaf交换机上。

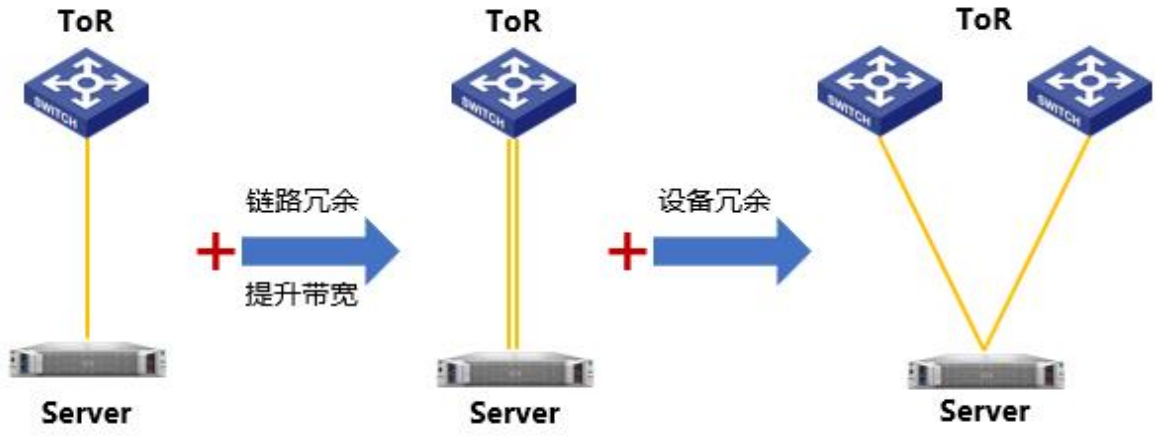

图1：服务器双归接入

【第一阶段：虚拟化堆叠】

此时，就需要解决两个问题，一个是跨设备的冗余链路如何处理，另一个是网关冗余如何解决。网络设备的堆叠技术就是在这种背景下出现的。采用堆叠技术可以很好的解决上述的两个问题，xSTP、VRRP等传统技术也在虚拟化堆叠出现后被逐渐摒弃。

紫光股份旗下新华三集团是最早实现设备堆叠技术的国内厂商，第一代虚拟化堆叠技术IRF最早被应用在2005年推出的基于Comware V3平台的S3900/S5600这两个系列的交换机上。经过持续优化和开发，在2009年新华三集团推出了第二代虚拟化堆叠技术IRF2，IRF2覆盖了低、中、高全系列交换机并沿用至今，在各个行业的数据中心得到了非常广泛的应用，IRF也一度成为了设备虚拟化堆叠技术的代名词。

通过虚拟化堆叠可以将多台设备虚拟化成一台设备，实现多台设备的协同工作、统一管理，如图2所示。

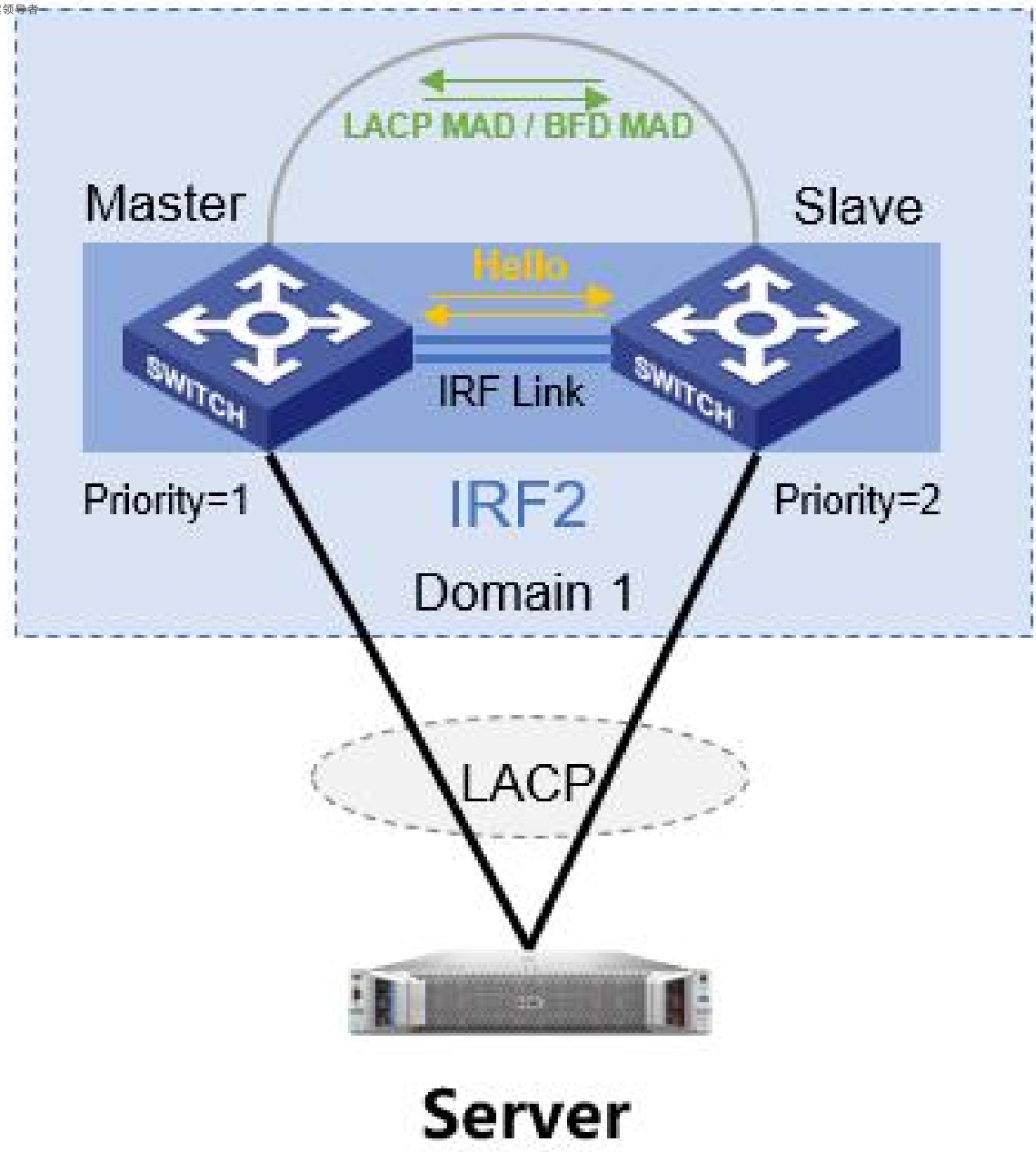

图2：IRF堆叠典型组网

采用虚拟化堆叠有很多的优势，例如：

1.简化了大量接入设备的管理和维护工作量，纳管逻辑网元的数量减少了一半

2.网络方案设计简单，降低了设备冗余部署的难度，堆叠组天然具备和单机一致的功能特性。堆叠组对外呈现与一台设备无异，对下连双归接入的服务器无特殊要求。服务器和交换机之间仅需启用简单的链路聚合即可

3.网络故障能够快速收敛

但是，随着业务需求的变化和快速发展，虚拟化堆叠在应用多年后的弊端也逐渐显现，主要表现在：

1.堆叠系统由于控制平面被整合归一，控制平面一旦出现问题则会影响整个堆叠系统

2.难以实现平滑升级，业务侧对网络的可靠性要求越来越高，同时设备的新特性、新需求在不断迭代，设备的软件版本发布很难严格遵从ISSU，且堆叠系统的在线升级有严格的步骤，操作相对繁琐，一组堆叠体升级往往需要耗费较长的时间

3.无法多厂家设备异构，即便同厂家设备也有同型号或同系列的要求

4.虚拟化虽然部署非常简单，实则原理和设备内部实现复杂，涉及到物理连接、拓扑收集、角色选举、堆叠合并、堆叠分裂、成员加入/退出、冲突检测等多个流程，每个阶段又定义了很多缜密的规则来保证整个堆叠系统的正常运行，开发和测试工作量很大，需要较长的研发和维护经验积累。采用堆叠技术会提高互联网客户自研白盒交换机软件的开发门槛。

5.设备堆叠链路和MAD冲突检测链路需要占用额外的TOR交换机端口资源，数量有限的高速端口无法全部用于上行链路。

【第二阶段：MLAG】

为了解决虚拟化堆叠技术的不足，在组网可靠性要求较高的场景中，轻量级的跨设备链路聚合MLAG技术则是一个更适合的折中方案。

MLAG将两台物理设备在聚合层面虚拟成一台设备来实现跨设备链路聚合，从而提供设备级冗余保护和流量负载分担，如图3所示

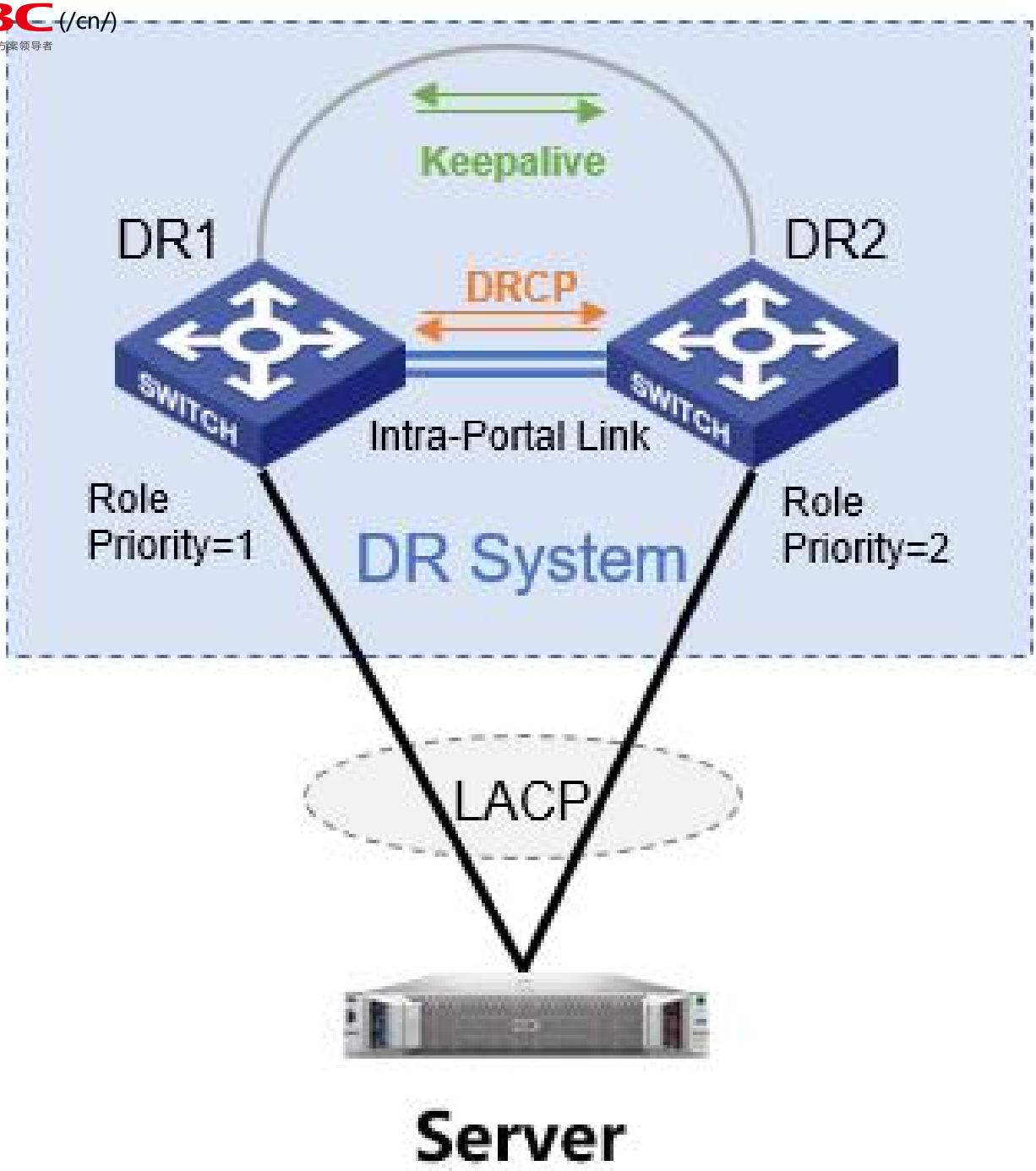

图3：MLAG典型组网

不同于虚拟化堆叠中将控制平面完全耦合，MLAG仅将协议面轻量级耦合，并不需要同步整机所有的设备状态，理论可靠性相对堆叠更高。

MLAG的优势如下：

1.组成MLAG系统的两台成员设备可以独立升级，升级操作相比虚拟化堆叠要简单，风险低、效率高。

2.MLAG和虚拟化堆叠一样，能够实现网络故障的快速收敛。

MLAG技术虽然解决了虚拟化堆叠无法独立升级的最大问题，但依然有如下不足：

不同于虚拟化堆叠中，堆叠组和单机拥有一致的特性能力，MLAG系统的特性支持的能力则需要分别去开发适配，部分特性的使用还会有一些限制

2.MLAG技术同样无法做到多厂家设备异构，均为私有实现（H3C：DRNI、Cisco：vPC、Juniper：MC-LAG、Arista：MLAG、华为：M-LAG），不支持跨厂商组成MLAG系统

3.MLAG系统的两台成员设备配置要时刻保持一致，配置变更等操作后，要确保两台成员设备的配置一致并Save，增加了工作量，部分关键配置不一致则还会影响MLAG系统的转发

4.MLAG只解决了跨设备的冗余链路问题，网关双活则需要用其它方式来实现（例如VLANIF下配置相同的IP和虚拟MAC）

5.Peer-Link和Keepalive-Link，需要占用额外的TOR交换机端口资源。另外， 成员设备之间MAC、ARP、ND等表项同步也会消耗一定的系统性能，尤其是设备存在大量表项的情况

【第三阶段：去堆叠】

去堆叠是互联网公司采用的最新的双归属方案，近两年普及率非常高，已成为当前最主流的方案，如图4所示。

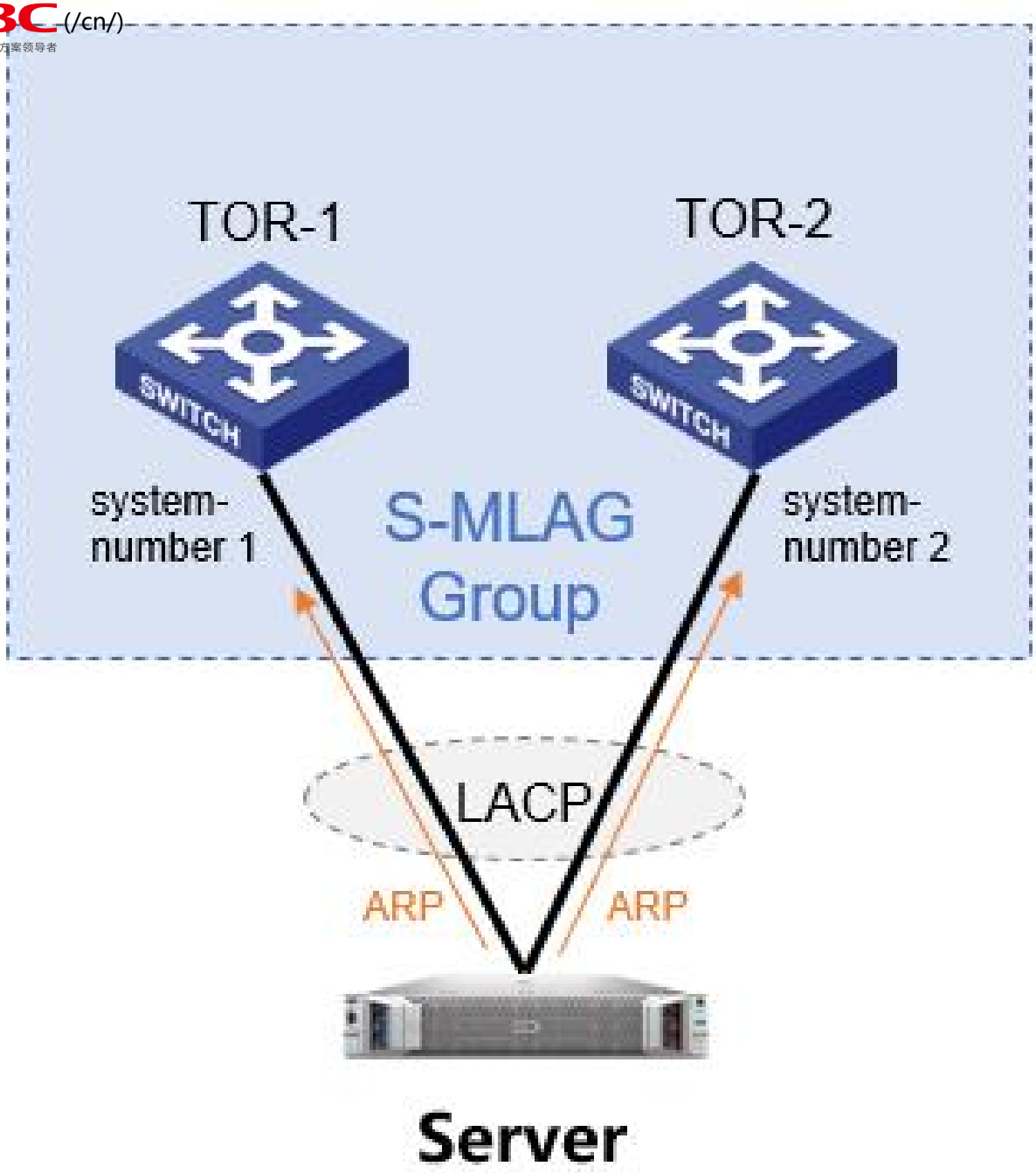

图4：去堆叠典型组网

在设备端，去堆叠由S-MLAG（简单跨设备链路聚合）和ARP直连路由通告两个特性组合实现。

1.通过S-MLAG在两台去堆叠设备上配置一致的LACP系统MAC地址，一致的LACP系统优先级等，形成相同的操作Key，完成与服务器的动态聚合。

2.在两台去堆叠设备的VLANIF下配置相同的IP和虚拟MAC，通过ARP直连路由通告，使设备从ARP表中学习到对应的32位直连路由信息，以便其它路由协议发布该主机路由，形成等价并指导报文转发。

去堆叠方案优势非常明显，主要表现在：

1.不同于虚拟化堆叠和MLAG，去堆叠的两台设备之间无连接，不占用额外的TOR交换机端口资源，所有TOR交换机的高速口均用于连接上行，链路带宽可最大化

2.避免了在TOR交换机之间跨机柜穿线，节约了端口和模块（AOC/DAC）投入，简化了上架布线工作量，降低了成本

3.去堆叠的两台设备之间彻底无表项同步，降低了系统的性能开销

4.由于去堆叠的两台设备之间无耦合，可实现多厂家设备异构

可以看出，去堆叠方案几乎解决了虚拟化堆叠和MLAG中的诸多不足，但是在部署上却引入了一个新的问题，就是对服务器端有特殊要求，一定程度上提高了部署的门槛。

因为去堆叠的两台设备间无耦合，设备之间无表项同步，所以需要借助服务器来解决这个问题，完成虚拟化堆叠和MLAG中表项同步达到的相同效果，实现双活接入。

此时就需要修改服务器的Linux系统内核代码，完成以下工作：

1.让服务器网卡在发送ARP请求和应答时从聚合口的两个成员端口上同时进行，即ARP双发

2.当服务器网卡的聚合成员端口出现Down/UP后，同时发送免费ARP更新

以此来保证连接该服务器的两台去堆叠交换机上ARP和MAC表项同步。

另外，在去堆叠方案中需要配合开启ARP代理、BUM隔离、上下行接口联动、Link-up delay、ARP老化时间调整、ARP探测等相关配置，优化去堆叠组网中的数据转发。

综上，服务器双归接入的发展从时间线上看，由堆叠到MLAG再到去堆叠，是一个逐渐解耦合的过程。三种方案都有各自的特点和侧重，还是需要根据具体的用户场景、需求和实际条件综合评估后进行选择。需求侧的变化也推动着服务器双归接入技术的不断发展和变化，未来也一定会有更优、更完善的技术出现，以满足业务对基础架构的可靠性、连续性和平滑升级等愈来愈高的要求。