[2011-06-09]某大型客户S75E+安全多业务插卡部署经验案例

金山 2011-06-09 发表

某大型客户S75E+安全多业务插卡部署经验案例

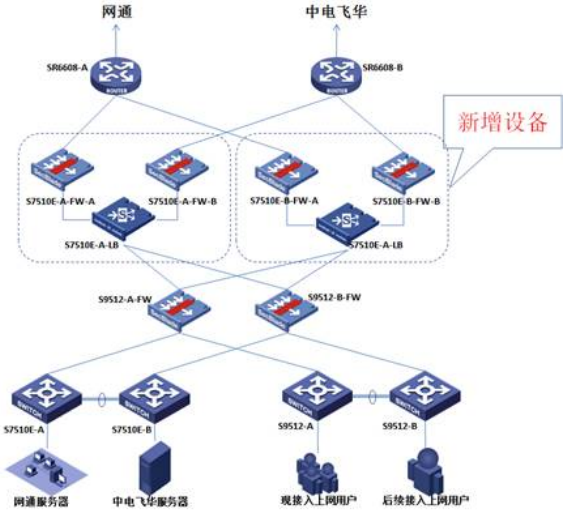
一、 组网概述：

某大型客户采购了SR6608、S7510E、S9500、SecBlade FW、LB、IPS作为信息网出口设备。此前S75E+FW+LB均为单套配置，近期又采购一套，希望实现备份。由于组网复杂，涉及设备非常多，所以编辑本案例分析，供对数据中心安全多插卡方案感兴趣同学参考。

需要首先声明的是，我们并不是推荐所有的客户都采用下面的组网方案。从技术角度组网中有不少设备其实是增加了物理链路和网段、增加了拓扑复杂度、增加了故障点，如果能够精减一部分设备，这个网络可能会简单得多。

客户希望本次割接后的组网如下：

两台SR6608做出口，分别单独连接网通和中华飞电Internet出口。对内网用户做NAT Outbound，对外网用户做NAT Server，服务器位于DMZ区域。LB板卡间接检测两个出口的状态，实现链路负载分担。上层防火墙配置Untrust和Trust区域，下层防火墙配置Untrust、Trust和DMZ区域，进行域单访问控制。由核心交换机做内网的流量汇总，负责用户和服务器的接入。



二、 问题描述：

根据新的组网拓扑，客户要求如下：

1、设备的备份、链路的备份，这是基本需求，要求任意一条链路中断、任意设备故障，都不影响网络使用。

2、现有上网用户通过左边的主设备走，后续部分接入的用户通过右边的设备走。现有的流量在正常情况下全部走左边，右边要求没有流量。后续有用户接入的时候，再走右边的设备。

3、网通和中电飞华都有服务器对外网服务，在SR6608上做了NAT Server。

4、全网静态路由。上图中所有连接均为三层，以30位掩码网段互连。

5、虚线框里为新增设备。

三、 过程分析：

FW、LB与传统三层设备的最大区别——会话表。

正常运行期间，当某个中电地址用户通过网通出口访问内网服务器时，返回的数据包到达LB后，LB检查发现命中某条会话记录，因此不再进行虚服务匹配，而是通过路由转发。这种配置条件上，LB上一般是两条指向不同运营商的默认路由，因此很可能会将数据包从中电下一跳转发出去。在本组网中，S7510E-A-FW-A与S7510E-B-FW-A在网络拓扑中的地位相等，因此它不会同S7510E-A-FW-B进行双机热备。所以S7510E-A-FW-B收到返回报文后，由于没有会话信息，会阻止报文通过。即时我们同意牺牲一些安全性，开启单向流检测将报文放过，公网运营商如果存在防火墙类设备，也很可能将此类报文丢弃。

同理，由于下层防火墙选路时，以S7510E-A-LB为主用，下一跳将选左边，当SR6608与S7510E-A-FW-A之间的链路中断后，基于配置和收敛时间方面的限制，会出现从S7510E-B-LB收到某条数据流首包及后续报文并向内网转发，此时服务器返回的报文，下层防火墙仍按默认选路策略从S7510E-A-LB转发，这样势必又造成来回路径不一致。

“来回路径不一致”问题的简单解决方案：“保存上一跳”。

在我司FW、LB的全局配置中，有个不起眼的功能，“保存上一跳”，我们来说说这个功能。从前文中不难了解到在FW和LB中，与传统三层设备不同，除了路由表和转发表外，FW、LB还会根据每条数据流的五元组+VPN实例等信息，建立会话表。对FW以TCP协议为例，正向SYN报文到达后，查询会话表，没有记录，再查域间策略列表，发现符合permit规则，于是建立会话表项并允许报文通过，查询路由表做转发决策，反向SYN+ACK报文到达防火墙，查询会话表，发现有记录，因此更新会话状态，然后再通过路由表做转发决策，注意，会话查询优先于转发查询，但会话不是转发决策者。对LB以TCP协

议为例，正向SYN报文到达后，查询会话表，没有记录，再查虚服务，发现命中某条虚服务后，根据配置情况，再虚服务所一一对应的某个实服务组中，根据“持续性”、“ACL策略”、“就近性”、“调度算法”等等方法，选择某个实服务（逻辑链路）并进行转发决策（物理链路+下一跳）。反向SYN+ACK报文到达后，查询会话表，有记录，更新会话状态，然后再查询路由表进行转发决策，同样地，会话不是转发决策者。综上所述，当指向会话发起方的路由存在多条时，就容易出现来回路径不一致问题。

要解决这个问题，让会话表项参与转发决策是关键。开启“保存上一跳”功能后，FW和LB在成功建立会话时，会增加记录报文源MAC地址、转发出接口等信息，因此返程报文到达设备后，在查询到命中的会话时，就已经有足够的信息进行转发决策了，设备会将会话所记录的源MAC地址作为转发的目的MAC地址，成帧后直接交给驱动转发。再次提醒，会话查询优先路由查询。按翟运波的名言：“如果是LB，只要开了虚服务和保存上一跳，哪怕是一条路由信息都没有也可以在网络中正常转发流量。”如果这件是交给硬件处理，还能进一步降低设备处理时延，提高效率。

我们再来看刚才的两个来回路径不一致问题，是不是通过在LB和下层FW上开启“保存上一跳”以后，就能解决啦！

棘手的新问题——失去意义的双机热备。

先来说说双机热备，在客户的组网中，S7510E-A-FW-A与S7510E-B-FW-A在拓扑中，处于完全相同的地位，是互相备份的关系，同理我们可以观察出S7510E-A-FW-B与S7510E-B-FW-B、S7510E-A-LB与S7510E-B-LB、S9512-A-FW与S9512-B-FW这三对也是同地位、互备份的关系。如果是传统的路由交换设备，那么一般意义上只要共享了转发决策信息，诸如通过动态路由协议同步全网路由信息，那么两台设备就可以实现两两备份。但对于安全产品FW和LB来说，通过前文兄弟们已经看到了会话对于FW和LB指导报文转发的关键性，当它们两两备份时，除了同步基本的转发决策信息，还必须同步会话信息。否则当报文从左边进来，建立会话后，如果不把会话信息同步给右边的设备，则查询会话表，无记录，查询域间策略或虚服务，发现该报文需丢弃处理，这样就起不到热备效果。那么三层设备有路由协议，四层的会话怎么办？我们有HA接口。通过在两台设备上各取一个物理端口，直接连接并配置后，两台设备在这条线路上交互会话信息。这个接口在网络拓扑中将消失，不转发任何业务报文。损失两个接口换到备份功能，还是值的。这样当报文从右边返回时，因为有了会话信息，防火墙便会更新会话并允许报文通过。就算报文返回的速度快到比左设备将会话备份过来还要快，要右边的设备也不会贸然丢包，而是会先通过HA线将报文送给对端设备再尝试处理一次，不能滥杀无辜啊。最后，提醒大家注意双机热备不会改变它们的邻居向它们转发报文的决策，两台设备运行在主备模式还是负载分担不是它们自己决定的，是其他协议邻居设备来决定的，只不过它们时刻同步会话信息，准备处理发来的报文而已。同时还要注意它与IRF II的区别。

了解了双机热备的原理后，我们再来看组网中的两种可能的情况。

正常情况下报文从S7510E-A进入内网，FW和LB们各自把会话信息备份至S7510E-B对应的兄弟身上。突然，S7510E-A整机Down，网络收敛后上下层设备将报文切换至S7510E-B一侧，由于双机热备，两块FW还保留着先辈们的遗训，继续处理着流量，看似一切正常。再看LB，开着保存上一跳，会话中已记下上一跳的MAC地址，它若无其事地把报文填上二层头转发着，全然不知已经全被对端设备丢弃了。看到这里，可能有兄弟会提出一个方案，把两边的MAC地址修改成一样，这是个不错的方法，我们也用过，但得先在客户主任那里说得过才行。

类似的一个情况，公网用户通过左边这套设备访问DMZ区的服务器，正常情况下都是由S7510E-A-LB进行转发决策，突然S7510E-A-LB“挂”了，流量如果能切换到右边的话，S9512-A-FW还能收到下行报文，但上行报文在“保存上一跳”的指导下，还在向S7510E-A-LB转发。在这个状态下，有会话的数据流全都单通了，反而是那些之前没有建立会话的数据流能正常通信。这也就是我们为什么清除设备上的会话信息后，数据流可以重新建立连接，恢复通信的原因。但这已经和双机热备的预期不符。

最后再看一个情况。内网用户访问公网的流量，正常情况下从S9512-A-FW出，突然，S9512-A-FW挂了，但S7510E-A-LB还会将返回报文转发给S9512-A-FW的接口MAC。如果内网某台PC长ping公网上某个地址，发现拓扑变化后，就算ping一年也始终不通，反而是停下来等一分（ICMP会话默认老化时间60s），再ping就通了。

常见配置错误——最后引入的IPS。

大的框架完成后，准备开始把SecBlade IPS添加至网络中。因为IPS、ACG设备是纯二层设备，报文经过设备时，如果允许通过，那么报文有任何的改动。在与交换机配合时，一般通过OAA或MQC将业务流量重定向给插卡的内联万兆接口，待插卡处理完成后再从这个接口还给交换机，进而由交换机进一步做三层或二层的转发。按照这个思路，代理商将S7510E上连接上行、下行的业务板端口，通过OAA的方式添加为IPS的外、内网安全区域，配置端并激活了策略，一测试发现策略不生效，再看组网，这与S7500E上有两次三层转发有关。如果上下行端口之间只经过了一次三层转发，那么一般是正常的，也是在各个局点常见的配置。既然客户要在S7510E上使用IPS的功能，那么我们就要重新选择它在网络中的位置，既然是二层设备，那么它的位置就可以有三种选择，SR与上层FW之间，上层FW与LB之间、LB与下层FW之间。除非客户坚持，一般会把IPS顶在外面，尽早地实施攻击防护策略，早丢早省心。在这个思路下，外网安全区域仍然是S7510E连接SR6608的端口，而内网安全区域就调整为S7510E连接SecBlade FW的内联接口。配置修改完成后，测试OK。从这个问题中，我们可以看到，二层卡可以在网络三层架构确定后，最后引入，但引入的位置，其上、下行接口安全区域的选择很重要。

四、 解决方法：

简化网络解决问题的终极神器——FW二层转发+端口联动+IRF II

全网有太多的静态路由和NQA，既然网络拓扑是造成路由复杂的原因，保存上一跳是造成双机热备失效的原因，那么我们就来着手简化拓扑、合并MAC地址信息。

首先，第一层防火墙全部改用面板的物理接口，改三层为二层方式运行，虽然对外表现降为二层设备了，但会话和域间策略的实现统统没有变。这样SR6608和LB就变成了直联关系，对LB而言，公网进来的流量，其上一跳是相同的两台SR6608，这样就解决了LB的上一跳不对等的问题。其次，原来的八个点对点网段变成了两个广播网段，这省了多少路由啊。但有样带来一个新问题，如果SR6608-A至S7510E-A-FW-A的线路断了，但下面没断，这样会造成一条三层连接一边UP，另一边Down。没关系，我们有端口联动组，如果上边的接口Down，下边的接口也会跟着Down，这样LB就会感知到了。一不做二不休，我们把4个防火墙的上下行接口和2个LB的上行接口，一共6对口，全部配置联动组，这样就可以实现只要主设备上LB以上任一条链路、任一台设备发生故障，SR6608和LB就能够直接感知到，SR6608从而可以轻松通过浮动静态路由方式切换内网下一跳，LB可以轻松Track到网络变化，切换对内网的VRRP Master状态，此举又省了多少NQA啊。

上层的问题解决了，再来看下层，能不能在LB和下层FW之前再串进来两个三层地址，而且使用一个相同的MAC地址，把它们隔开。如果直的可以，那S9512上的两台FW，“保存上一跳”就不用再开了。没错，就是必杀技IRF II。两台S7510E堆叠以后，对上对下配置两个VLAN虚接口，将原来的4个点对点网段改为两个广播网段，LB通过VRRP维护主备关系，2个S9512 FW也同样实现主备或负载分担。到这里，LB上内网来的流量，和S9512 FW公网方向来的流量，它们的上一跳IP、MAC地址就不会因为网络拓扑的变化而变化了。客户再也不用清会话了，任意的单点故障都不会影响网络连通性，而且TCP连接也不需重新建立，是真的“热备”！

对这个方案客户也很有兴趣，但同时也保留一个意见。那就是这样配置完成后，如果发生网络切换，主切备后，备机侧如果再有一根上联链路 Down，那整网就全断了。而在目前现网运行的方案中，如果主、备两侧各有一条连接运营商的线路发生中断，全网还是保持了一定的与公网的连接性。因此客户表示会再次评估和取舍。如果内网的业务对TCP等连接中断重建立不敏感，可以自动重建恢复，那么现网的方案可以获得更好地链路保证。如果内网业务对TCP等连接中断非常敏感且上层应用不会自动重建连接以恢复，那么新提出的这个方案就非常有优势的。这就应了前辈们的老话，只有了解业务多方比较，才能决策哪种组网更优，在这一点上，多多和客户沟通，开诚布公很重要。