某局点S10500设备拥塞丢包处理经验案例

丢包 高佩岩 2017-07-06 发表

某局点反馈S10500设备作为汇聚层交换机，上联S6520-EI设备作为公网出口，下行S7500E设备作为业务流量网关。该局点反馈，在流量高峰期或流量上升期出现业务瞬断现象，严重时在流量高峰阶段半小时内业务闪断数次，非流量高峰期不存在丢包。将业务流量迁移至S10500设备上连的S6520-EI设备则不存在丢包。

无告警信息。

首先从当前故障现象分析，非流量高峰期时不存在丢包，流量高峰期或流量上升期时业务丢包或流量闪断明显，怀疑该问题与流量模型有关。查看设备具体组网方式，S10500设备上联两台S6520-EI，两个上行口分别为6个万兆口链路聚合与8个万兆口链路聚合，下行连接单台S7500E设备，下行端口为4个万兆端口聚合。该组网存在明显的多数端口向少数端口打流的情况，在流量突发时容易形成拥塞丢包，符合该局点反馈的现象。
通过查看设备诊断信息以及对丢包设备的远程登陆查看，在隐藏视图probe模式下通过debug port mapping命令查看业务板面板口与芯片端口的对应关系，通过bcm slot chip show/c命令查看设备有无拥塞丢包（V7设备也可通过在系统视图下使用display packet-drop interface命令查看拥塞丢包）；隐藏视图下的show/c命令对应拥塞丢包关键字为PERQ_DROP_PKT(2)，（2）表示2队列，普通业务报文均走2队列进行转发，packet-drop对应拥塞丢包关键字为Packets dropped due to full GBP or insufficient bandwidth；通过查询该命令，发现设备下联口存在实时拥塞丢包。
我们在处理网上问题中所述的拥塞丢包有两种含义：1、端口带宽打满产生拥塞丢包，这种业务量巨大的情况只能通过扩容解决，通常情况下，带宽占用比持续超过80%则建议现场尽快扩容，防止流量继续增大导致丢包甚至业务中断；2、流量突发产生拥塞丢包，在数据中心应用中，通常情况流量并非是平稳的，在短时间内，由于交换式网络线速转发能力，特定应用与组网条件下，网络流量突然增大，可能会瞬时超过网络设备的实际速率。从微观的角度看，端口实际转发能力都是线速的，端口统计的转发速率快还是慢只是报文转发的间隔时间有所不同。
但是网络突发会在什么情况下拥塞呢？流量突发通常有两种情况会触发：第一种是报文从高速端口转发到低速端口出去，如10GE端口转发到GE端口；第二种情况是端口转发速率相同，但是存在多个端口往其中一个端口转发，如多个GE端口的报文最后汇聚到一个GE端口出去。这两种情况都是在某一瞬间进入设备的报文快于端口出去的报文，这样报文无法及时转发出去，就会在设备内部缓存，而一旦突发的报文超出了端口的缓存能力，这部分报文就会丢弃。
那么如何证明网络流量突发确实存在？不管是通过设备侧执行display interface命令查看端口流量统计的峰值还是通过网管设备监控流量突，想做到将流量突发捕捉到实际非常困难。实际上网管设备获取端口流量图也是通过轮训设备的MIB节点获取值后自行绘制的，设备侧对应的值即为display interface的统计值，所以设备侧捕捉不到的流量突发，网管设备同样捕捉不到。设备侧的端口统计display interface信息实际上并非瞬时值，而是一个平均值；取平均的周期间隔默认为300秒，也就是说当端口统计显示峰值流量为8G时，在过去的5分钟内，端口实际速率既有超过8G的情况，也有小于8G的情况，即平均值为8G；相应的，如果我们将端口的统计周期时长调整为5秒，则能更方便的观察流量突发情况，但是想要抓取到流量的瞬时突发依然不现实。

对于流量突发引起拥塞问题通常有如下两种解决方法:
1、尽量避免网络中出现高速率端口向低速率端口打流、多数端口向少数端口打流的情况。如果设备上下行端口速率相当，则不会出现多数端口/高速率端口出现突发拥塞将少数端口/低速率端口带宽打满的情况，可以从根本上彻底解决拥塞问题。

2、调整端口buffer资源调配、相应增长业务队列长度。由于端口带宽在打满的情况下流量进入buffer缓存转发，缓冲区队列打满后后续后续流量均会丢弃；通常情况下buffer大小为固定值，由芯片类型决定，芯片为每个端口分配相应的缓存区空间，剩余缓存区空间为芯片共用实时调配，对于园区接入设备，可以通过burst-mode enable命令调整缓存区空间分配方式，取消端口分配，整体缓存区空间实时调配，遇到突发流量增强容错能力；而对于园区高端的S10500设备，当前无命令调整缓存区空间分配方式，则可以使用wred方式调整业务队列2的长度（默认情况下业务报文优先级为0，进入端口队列2转发），方法如下：

```
qos wred queue table burst
queue 2 drop-level 0 low-limit 16382 high-limit 16383 discard-probability 1
queue 2 drop-level 1 low-limit 16382 high-limit 16383 discard-probability 1
queue 2 drop-level 2 low-limit 16382 high-limit 16383 discard-probability 1
#
interface Ten-GigabitEthernet0/0/1
 port link-mode bridge
```

```
port access vlan 1000
qos wred apply burst
```

drop-level字段表示丢弃级别，0对应绿色报文、1对应黄色报文、2对应红色报文；low-limit字段表示队列平均长度下限，默认值为100，最大取值16383；high-limit表示队列平均长度上限，默认值为1000，最大取值16383；当队列平均长度小于下限时，不丢弃报文。当队列平均长度在上限和下限之间时，设备随机丢弃报文，队列越长，丢弃概率越高。当队列平均长度超过上限时，丢弃所有到来的报文。通过这样增大2队列长度，也同样可以提高应对突发的容错性。需要注意的是，不管是调整Buffer或者增加2队列长度，在少量突发拥塞的情况下可以缓解拥塞，但是如果网络中出现大量拥塞，只能通过方法1进行端口扩容操作解决。