某局点S12500 LST1CP4RFD2 100G单板接口带宽达不到线速问题经验案例

直通转发 赵跃 2018-03-17 发表

某局点IDC机房两台S12500做堆叠，新扩容的100G LST1CP4RFD2单板分别插在两个框的10槽位和11槽位，每块单板使用1口和3口，两个端口故障现象相同。以H1/11/0/3接口为例，故障发生时，H1/11/0/3接口出方向流量图如图1所示。在每天的流量高峰期19:00开始，流量超过50G后，无法再增加，流量图出现削峰现象，接口带宽达到线速转发。

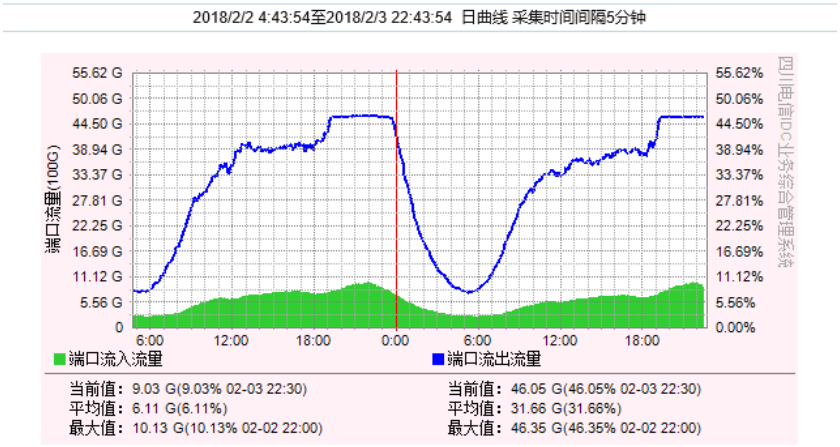

图1 H1/11/0/3出方向流量图

无

1、 首先，出现流量异常的单板LST1CP4RFD2为非线速2:1收敛的单板，1口和2口共提供100G带宽，3口及4口共提供100G带宽。故障发生时，该业务单板只有1口和3口在使用，因而不是因为该单板为非线速单板导致该问题；

2、进一步通过分析客户的现网设备配置和流量模型，确认现网设备配置存在如下特点：

u 堆叠设备的两个框上都存在第一代和第二代业务单板LST1GP48LEC2/ LST1GP48LEC2，且系统工作在standard的标准模式

```
==================================================
 ===============display device===============
Slot No.  Brd Type       Brd Status  Software Version
 1/0      LST1MRPNE1     Standby     S12500-CMW710-R7328P01
 1/1      LST1MRPNE1     Master      S12500-CMW710-R7328P01
 1/2      LST1GP48LEC2   Normal      S12500-CMW710-R7328P01
 1/3      LST1XP16LEC2   Normal      S12500-CMW710-R7328P01
 1/4      LST1XP16LEC2   Normal      S12500-CMW710-R7328P01
 1/5      LST1XP16LEC2   Normal      S12500-CMW710-R7328P01
 1/6      LST1XP16LEC2   Normal      S12500-CMW710-R7328P01
 1/7      LST1XP48LFD1   Normal      S12500-CMW710-R7328P01
 1/8      LST1XP16LEC2   Normal      S12500-CMW710-R7328P01
 1/9      NONE           Absent      NONE
 1/10     LST1CP4RFD2    Normal      S12500-CMW710-R7328P01
 1/11     LST1CP4RFD2    Normal      S12500-CMW710-R7328P01
 1/12     LST1XP16LEC2   Normal      S12500-CMW710-R7328P01
 1/13     LST1XP16LEC2   Normal      S12500-CMW710-R7328P01
 1/14     LST1XP16LEC2   Normal      S12500-CMW710-R7328P01
 1/15     NONE           Absent      NONE
 1/16     NONE           Absent      NONE
 1/17     NONE           Absent      NONE
 1/18     NONE           Absent      NONE
 1/19     NONE           Absent      NONE
 1/20     LST1SF18E1     Normal      S12500-CMW710-R7328P01
 1/21     LST1SF18E1     Normal      S12500-CMW710-R7328P01
 1/22     LST1SF18E1     Normal      S12500-CMW710-R7328P01
 1/23     LST1SF18E1     Normal      S12500-CMW710-R7328P01
 1/24     LST1SF18E1     Normal      S12500-CMW710-R7328P01
 1/25     LST1SF18E1     Normal      S12500-CMW710-R7328P01
```

```
 1/26      LST1SF18E1      Normal       S12500-CMW710-R7328P01
 1/27      LST1SF18E1      Normal       S12500-CMW710-R7328P01
 1/28      NONE            Absent       NONE
 2/0       LST1MRPNE1      Standby      S12500-CMW710-R7328P01
 2/1       LST1MRPNE1      Standby      S12500-CMW710-R7328P01
 2/2       LST1GP48LEC2    Normal       S12500-CMW710-R7328P01
 2/3       LST1XP16LEC2    Normal       S12500-CMW710-R7328P01
 2/4       LST1XP16LEC2    Normal       S12500-CMW710-R7328P01
 2/5       LST1XP16LEC2    Normal       S12500-CMW710-R7328P01
 2/6       LST1XP16LEC2    Normal       S12500-CMW710-R7328P01
 2/7       LST1XP16LEC2    Normal       S12500-CMW710-R7328P01
 2/8       LST1XP16LEC2    Normal       S12500-CMW710-R7328P01
 2/9       LST1XP48LFD1    Normal       S12500-CMW710-R7328P01
 2/10      LST1CP4RFD2     Normal       S12500-CMW710-R7328P01
 2/11      LST1CP4RFD2     Normal       S12500-CMW710-R7328P01
 2/12      LST1XP16LEC2    Normal       S12500-CMW710-R7328P01
 2/13      LST1XP16LEC2    Normal       S12500-CMW710-R7328P01
 2/14      LST1XP16LEC2    Normal       S12500-CMW710-R7328P01
 2/15      NONE            Absent       NONE
 2/16      NONE            Absent       NONE
 2/17      NONE            Absent       NONE
 2/18      NONE            Absent       NONE
 2/19      NONE            Absent       NONE
 2/20      LST1SF18E1      Normal       S12500-CMW710-R7328P01
 2/21      LST1SF18E1      Normal       S12500-CMW710-R7328P01
 2/22      LST1SF18E1      Normal       S12500-CMW710-R7328P01
 2/23      LST1SF18E1      Normal       S12500-CMW710-R7328P01
 2/24      LST1SF18E1      Normal       S12500-CMW710-R7328P01
 2/25      LST1SF18E1      Normal       S12500-CMW710-R7328P01
 2/26      LST1SF18E1      Normal       S12500-CMW710-R7328P01
 2/27      LST1SF18E1      Normal       S12500-CMW710-R7328P01
 2/28      NONE            Absent       NONE
#
 system-working-mode standard
 password-recovery enable
 accelerate chassis 1 slot 7
 accelerate chassis 2 slot 9
#
u  部分万兆口配置了“qos apply policy lp inbound”，将流量映射到本地优先级为5。无跨框流量从100G
端口出去，故现网从100G端口出去的报文会进入2个队列：2队列和5队列。
#
traffic classifier lp operator and
 if-match acl 3000
#
traffic behavior lp
 remark local-precedence 5
#
qos policy lp
 classifier lp behavior lp
#
interface Ten-GigabitEthernet1/3/0/7
 port link-mode bridge
 description To-TengXun-4F-1
 port access vlan 1100
 undo stp enable
 qos apply policy lp inbound
u  100G端口下配置了最小带宽保证功能，即WFQ，保证优先级5的报文最小带宽为2G。
interface HundredGigE1/11/0/3
 port link-mode bridge
 description uT: SC-CD-XH-C-1.NE5000E.IDC:100G0030IDC:100GE/2/1/0/1
 port access vlan 230
 undo stp enable
 packet-filter 3100 inbound
 qos trust dscp
```

```
 qos wfq weight
 qos wfq ef weight 2
 qos bandwidth queue ef min 2000000
 qos apply policy dscp outbound
 port link-aggregation group 40
#
traffic classifier dscp operator and
 if-match acl 3000
#
traffic behavior dscp
 remark dscp cs5
#
qos policy dscp
 classifier dscp behavior dscp
```

基于现网的配置，有如下几个因素，导致100G端口出方向带宽达不到100G：

1） LST1CP4RFD2为100G PUMA3的第三代单板，FAP芯片本身调度粒度较粗，Credit会存在浪费，当出端口为100G时，浪费较严重；

2） 部分端口配置了“qos apply policy lp inbound”，往100G端口出去的流量会进入2个队列，在多队列的情况下Credit浪费也会变大；

3） 100G端口配置了WFQ，当系统工作在Standard模式、配置了WFQ情况下，Credit浪费也会变大

4） 现网工作在Standard模式，该模式下交换网的数据信元工作在FSC模式下(固定大小模式)，交换网带宽也会存在浪费的情况；

因为上述第4）因素导致转发性能会有较大的下降，尤其在上述第2）和3）叠加配置基础上，更恶化了100G板卡的转发性能。

S12500只有一种100G板卡，属于第三代单板。只有配合三代机框+其它三代板卡才能达到最大的性能，在该种硬件配置情况下，可以将设备的系统工作模式修改为grand模式，在该模式下交换网的数据信元会工作在VSC（信元大小可变模式），设备转发性能会提升较大。但现网S12500设备为一/二代板卡+三代板卡混合的硬件配置，因此只能工作在非grand模式（现场是Standard模式），若设备切换为grand模式，则一/二代板卡无法正常启动。

基于现网的设备软硬件配置，做了以下实测，结果如下：

I）虽然有4）因素影响，100G板卡如果没有第2）和3）影响，则报文为单队列，在系统工作模式为standard模式下，可以达到线速转发；

II）在grand系统工作模式下，如果只配置2），100G板卡转发性能也能达到线速转发；

III）在grand系统工作模式下，如果同时配置了2）和3），100G板卡转发性能在91-98G区间内，差异主要与设备流量模型中的2个队列的流量分布大小有关；

IV）在非grand模式（如现网设备运行的standard模式）下，如果只配置2），100G板卡转发性能在现网软件版本（R7328P01）中转发性能为75G左右，如果使用当前最新软件版本（R7377P04）可以达到90G以上；

V）在非grand模式（如现网设备运行的standard模式）下，如果配置了2）和3），100G板卡转发性能只能在50G左右，软件版本上也无法进一步优化。

基于上述分析，可提供如下四个处理方式：

1、建议100G端口去掉WFQ配置。当前配置情况下，100G端口没必要配置WFQ（优先保证优先级队列5出去2G）。因为设备默认就是SP调度，在100G出端口拥塞情况下，设备本身会优先调度转发队列5的流量。

2、在不配置WFQ情况下，当前设备配置和流量模型可以支撑75G左右流量。如果后续流量再增加的话，万兆口去掉“qos apply policy lp inbound”的配置，即去往100G端口的报文只有1个队列，可以支持出去100G流量带宽，接口线速转发

3、通过修改设备模式为grand，保证100G接口达到线速转发。该方式需要更换一二代接口板为可支持grand模式的三代接口板，存在较大成本问题。

4、扩容100G单板，做链路捆绑，同样存在新增单板成本问题。

根据现网设备运行的实际情况，最终采用删除100G接口WFQ配置和万兆口的lp QOS策略后解决。