Nexus 3500输出丢弃和缓冲区QoS

目录

简介

本文档介绍用于排除Nexus 3500平台上丢弃的流量类型和丢弃此流量的输出缓冲区(OB)故障的命令。

方法

1. 检查输出丢弃
2. 确定丢弃是单播还是组播
3. 确定使用哪个输出缓冲区
4. 检查活动缓冲区监控

检查输出丢弃

检查物理接口统计信息以确定流量是否在出口方向丢弃。确定TX方向上的“输出丢弃”计数器是否递增和/或为非零。

```
Nexus3548# show interfce Eth1/7
Ethernet1/7 is up
 Dedicated Interface
 Hardware: 100/1000/10000 Ethernet, address: a44c.116a.913c (bia a44c.116a.91ee)
```

```
 Description: Unicast Only
 Internet Address is 1.2.1.13/30
 MTU 1500 bytes, BW 1000000 Kbit, DLY 10 usec
 reliability 255/255, txload 35/255, rxload 1/255
 Encapsulation ARPA
 full-duplex, 1000 Mb/s, media type is 1G
 Beacon is turned off
 Input flow-control is off, output flow-control is off
 Rate mode is dedicated
 Switchport monitor is off
 EtherType is 0x8100
 Last link flapped 00:03:48
 Last clearing of "show interface" counters 00:03:55
 1 interface resets
 30 seconds input rate 200 bits/sec, 0 packets/sec
 30 seconds output rate 0 bits/sec, 0 packets/sec
 Load-Interval #2: 5 minute (300 seconds)
   input rate 40 bps, 0 pps; output rate 139.46 Mbps, 136.16 Kpps
 RX
   1 unicast packets  118 multicast packets  0 broadcast packets
   119 input packets  9830 bytes
   0 jumbo packets  0 storm suppression bytes
   0 runts  0 giants  0 CRC  0 no buffer
   0 input error  0 short frame  0 overrun   0 underrun  0 ignored
   0 watchdog  0 bad etype drop  0 bad proto drop  0 if down drop
   0 input with dribble  0 input discard
   0 Rx pause
 TX
   23605277 unicast packets  0 multicast packets  0 broadcast packets
   23605277 output packets  3038908385 bytes
   0 jumbo packets
   0 output errors  0 collision  0 deferred  0 late collision
   0 lost carrier  0 no carrier  0 babble 11712542 output discard
   0 Tx pause
```

确定丢弃是单播还是组播

确定接口丢弃流量后，输入**show queuing interface <x/y>**命令以确定丢弃的流量是组播还是单播。在6.0(2)A3(1)之前的版本中，输出如下所示：

```
Nexus3548# show queuing interface Eth1/7
Ethernet1/7 queuing information:
 TX Queuing
   qos-group  sched-type  oper-bandwidth
       0       WRR            100

 RX Queuing
   Multicast statistics:
       Mcast pkts dropped                    : 0
   Unicast statistics:
   qos-group 0
   HW MTU: 1500 (1500 configured)
   drop-type: drop, xon: 0, xoff: 0
   Statistics:
        Ucast pkts dropped                     : 11712542
```

在版本6.0(2)A3(1)及更高版本中，输出如下所示：

```
Nexus3548# show queuing interface Eth1/7
Ethernet1/7 queuing information:
   qos-group  sched-type  oper-bandwidth
```

```
      0       WRR            100
   Multicast statistics:
       Mcast pkts dropped                    : 0
   Unicast statistics:
   qos-group 0
   HW MTU: 1500 (1500 configured)
   drop-type: drop, xon: 0, xoff: 0
   Statistics:
Ucast pkts dropped                      : 11712542
```

注意：如果为端口配置了组播慢速接收器，请参阅以了解功能信息，则由于硬件限制，不会使用**show queuing interface Eth<x/y>命**令跟踪丢包。请参阅Cisco Bug ID CSCuj21006。

确定使用哪个输出缓冲区

在Nexus 3500中，出口方向使用三个缓冲池。show hardware internal mtc-usd **info port-mapping命令的输出提**供映射信息。

```
Nexus3548# show hardware internal mtc-usd info port-mapping
OB Ports to Front Ports:
========= OB0 =========    ========= OB1 =========    ========= OB2 =========
45 47 21 23 09 11 33 35    17 19 05 07 41 43 29 31    13 15 37 39 25 27 01 03
46 48 22 24 10 12 34 36    18 20 06 08 42 44 30 32    14 16 38 40 26 28 02 04

Front Ports to OB Ports:
=OB2= =OB1= =OB0= =OB2=    =OB1= =OB0= =OB2= =OB1=    =OB0= =OB2= =OB1= =OB0=
12 14 04 06 08 10 00 02    00 02 04 06 08 10 12 14    12 14 04 06 08 10 00 02
13 15 05 07 09 11 01 03    01 03 05 07 09 11 13 15    13 15 05 07 09 11 01 03

Front port numbering (i.e. "01" here is e1/1):
=OB2= =OB1= =OB0= =OB2=    =OB1= =OB0= =OB2= =OB1=    =OB0= =OB2= =OB1= =OB0= 01 03 05 07 09 11
13 15    17 19 21 23 25 27 29 31    33 35 37 39 41 43 45 47 02 04 06 08 10 12 14 16    18 20 22
24 26 28 30 32    34 36 38 40 42 44 46 48

Note: Text in Red font is _not_ CLI output, it's purely to help those reading
the document faster match the actual front port instead of having to manually
count up.
```

结果的第一部分表明，OB池0被前端端口（如45、46、47、48等）使用，OB1被前端端口（如17、18等）使用。

结果的第二部分显示Eth1/1映射到OB2端口12,Eth1/2映射到OB2端口13，依此类推。

讨论中的端口Eth1/7映射到OB1。

有关详细信息，请参阅本文档的缓冲区管理部分。

检查活动缓冲区监控

有关此功能的详细信息，请参阅Cisco Nexus 3548活动缓冲区监控白皮书和本文档的部分。

计数器主动增量

如果输出丢弃主动增量，请使用此命令启用主动缓冲区监控(ABM)。请注意，该命令允许您监控单播或组播，但不能同时监控两者。此外，它还允许您配置采样间隔和阈值。

```
hardware profile buffer monitor [unicast|multicast] {[sampling
```

简要输出

启用ABM后，您可以使用此命令查看结果。

```
Nexus3500# show hardware profile buffer monitor interface e1/7 brief
Brief CLI issued at: 09/30/2013 19:43:50

                  Maximum buffer utilization detected
                 1sec     5sec     60sec     5min      1hr
                ------   ------   ------   ------   ------
Ethernet1/7      5376KB   5376KB   5376KB      N/A      N/A
```

这些结果表明，过去60秒内离开Eth1/7的单播流量已使用5.376 MB的OB1缓冲区（共6 MB）。

详细输出

```
Nexus3500# show hardware profile buffer monitor interface Eth1/7 detail
Detail CLI issued at: 09/30/2013 19:47:01

Legend -
384KB  - between   1 and 384KB of shared buffer consumed by port
768KB  - between 385 and 768KB of shared buffer consumed by port
307us  - estimated max time to drain the buffer at 10Gbps

Active Buffer Monitoring for Ethernet1/7 is: Active
KBytes                 384  768 1152 1536 1920 2304 2688 3072 3456 3840 4224 4608 4992 5376 5760
6144
us @ 10Gbps            307  614  921 1228 1535 1842 2149 2456 2763 3070 3377 3684 3991 4298 4605
4912
                    ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ----
----
09/30/2013 19:47:01      0    0    0    0    0    0    0    0    0    0    0    0    0  250
0    0
09/30/2013 19:47:00      0    0    0    0    0    0    0    0    0    0    0    0    0  252
0    0
09/30/2013 19:46:59      0    0    0    0    0    0    0    0    0    0    0    0    0  253
0    0
09/30/2013 19:46:58      0    0    0    0    0    0    0    0    0    0    0    0    0  250
0    0
09/30/2013 19:46:57      0    0    0    0    0    0    0    0    0    0    0    0    0  250
0    0
09/30/2013 19:46:56      0    0    0    0    0    0    0    0    0    0    0    0    0  250
0    0
09/30/2013 19:46:55      0    0    0    0    0    0    0    0    0    0    0    0    0  251
0    0
09/30/2013 19:46:54      0    0    0    0    0    0    0    0    0    0    0    0    0  251
0    0
09/30/2013 19:46:53      0    0    0    0    0    0    0    0    0    0    0    0    0  250
0    0
09/30/2013 19:46:52      0    0    0    0    0    0    0    0    0    0    0    0    0  253
0    0
```

```
09/30/2013 19:46:51      0    0    0    0    0    0    0    0    0    0    0    0    0 249
0    0
...
```

每行中的信息以第二间隔记录。每列表示缓冲区使用情况。如命令结果中所述，如果列“384”报告了非零值，则表示当ABM轮询OB使用时，缓冲区使用介于0-384 KB之间。非零数字是报告使用次数。

这些结果表明，Eth1/7在过去10秒内，OB1平均每秒使用5.376 MB的次数为249 - 253次。要清除此流量的缓冲区，需要4298微秒(us)。

超过阈值时生成日志

如果丢弃计数器和缓冲区使用率定期增加，则可以设置阈值并在阈值超过时生成日志消息。

```
logging level mtc-usd 5
hardware profile buffer monitor unicast sampling 10 threshold 4608
```

该命令设置为以10纳秒间隔监控单播流量，当它超过缓冲区的75%时，会生成日志。

您还可以创建调度程序，以便每小时收集ABM统计信息和接口计数器输出，并将其附加到bootflash文件。此示例用于组播流量：

```
hardware profile buffer monitor multicast

feature scheduler
scheduler job name ABM
show hardware profile buffer monitor detail >> ABMDetail.txt
show clock >> ABMBrief.txt
show hardware profile buffer monitor brief >> ABMBrief.txt
show clock >> InterfaceCounters.txt
show interface counters errors >> InterfaceCounters.txt
scheduler schedule name ABM
time start now repeat 1:0
job name ABM
```

值得注意的思科漏洞ID

- 思科漏洞ID CSCum21350:快速端口摆动会导致同一QoS缓冲区中的所有端口丢弃所有TX组播/广播流量。这在版本6.0(2)A1(1d)及更高版本中已修复。
- 思科漏洞ID CSCuq96923:组播缓冲区块被卡住，导致出口组播/广播丢弃。此问题仍在调查中。
- 思科漏洞ID CSCva20344:Nexus 3500缓冲区块/锁定 — 无TX组播或广播。不可修复的问题，可能在版本6.0(2)U6(7)、6.0(2)A6(8)和6.0(2)A8(3)中得到修复。
- 思科漏洞ID CSCvi93997: Cisco Nexus 3500交换机输出缓冲区块卡住。这在版本7.0(3)I7(8)和9.3(3)中已修复。

常见问题

ABM是否影响性能或延迟？

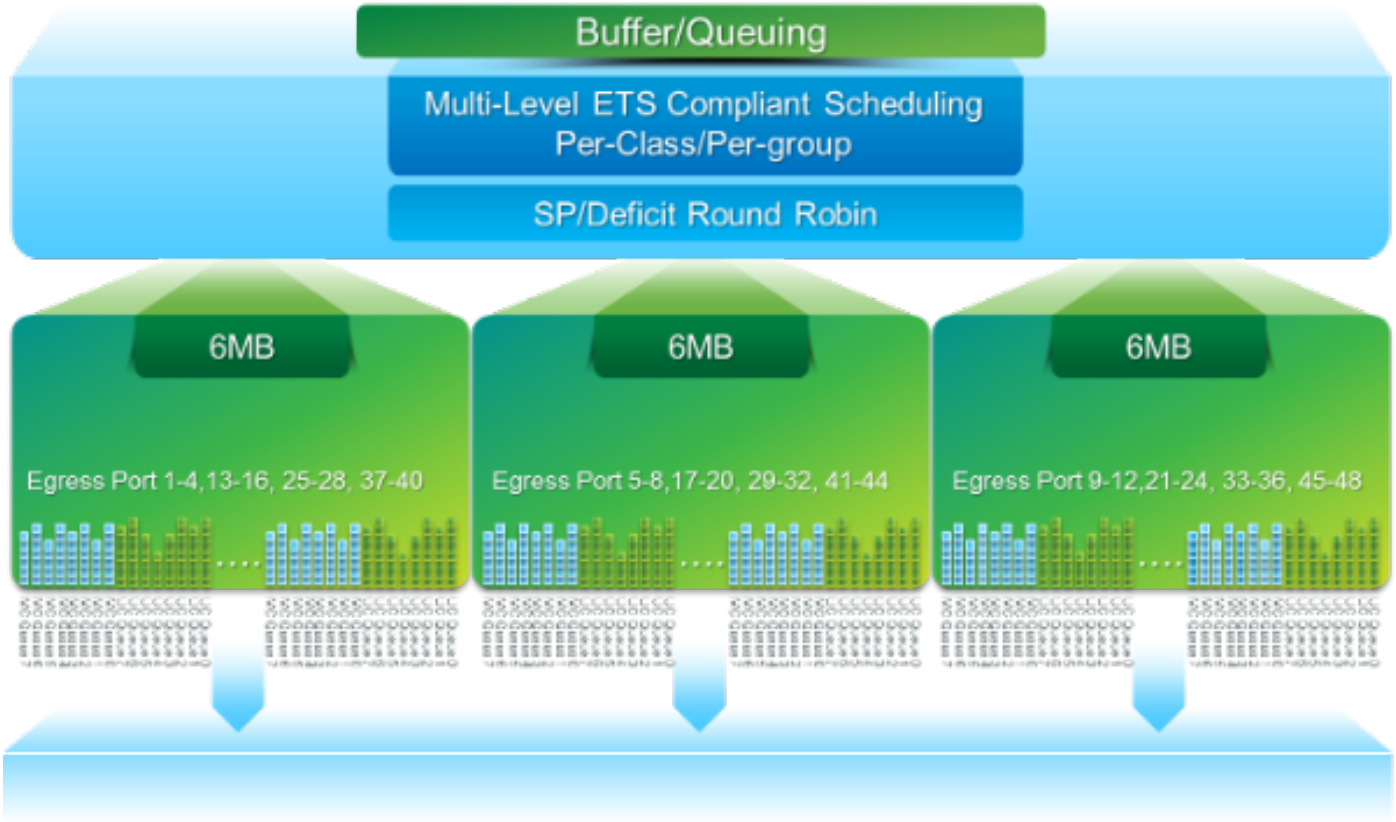
否，此功能不会影响设备的延迟或性能。

较低的ABM硬件轮询间隔有何影响？

默认情况下，硬件轮询间隔为4毫秒。您可以将此值配置为低至10纳秒。由于硬件轮询间隔较短，因此不会影响性能或延迟。选择4毫秒的默认硬件轮询以确保在软件每一秒轮询一次之前不会使直方图计数器溢出。如果降低硬件轮询间隔，则可能会使硬件计数器饱和255个样本。由于CPU和内存限制，设备无法处理低于一秒的软件轮询，以便匹配较低的硬件轮询。本白皮书举例说明了较低的硬件轮询间隔及其使用案例。

附录 — 功能信息

缓冲区管理

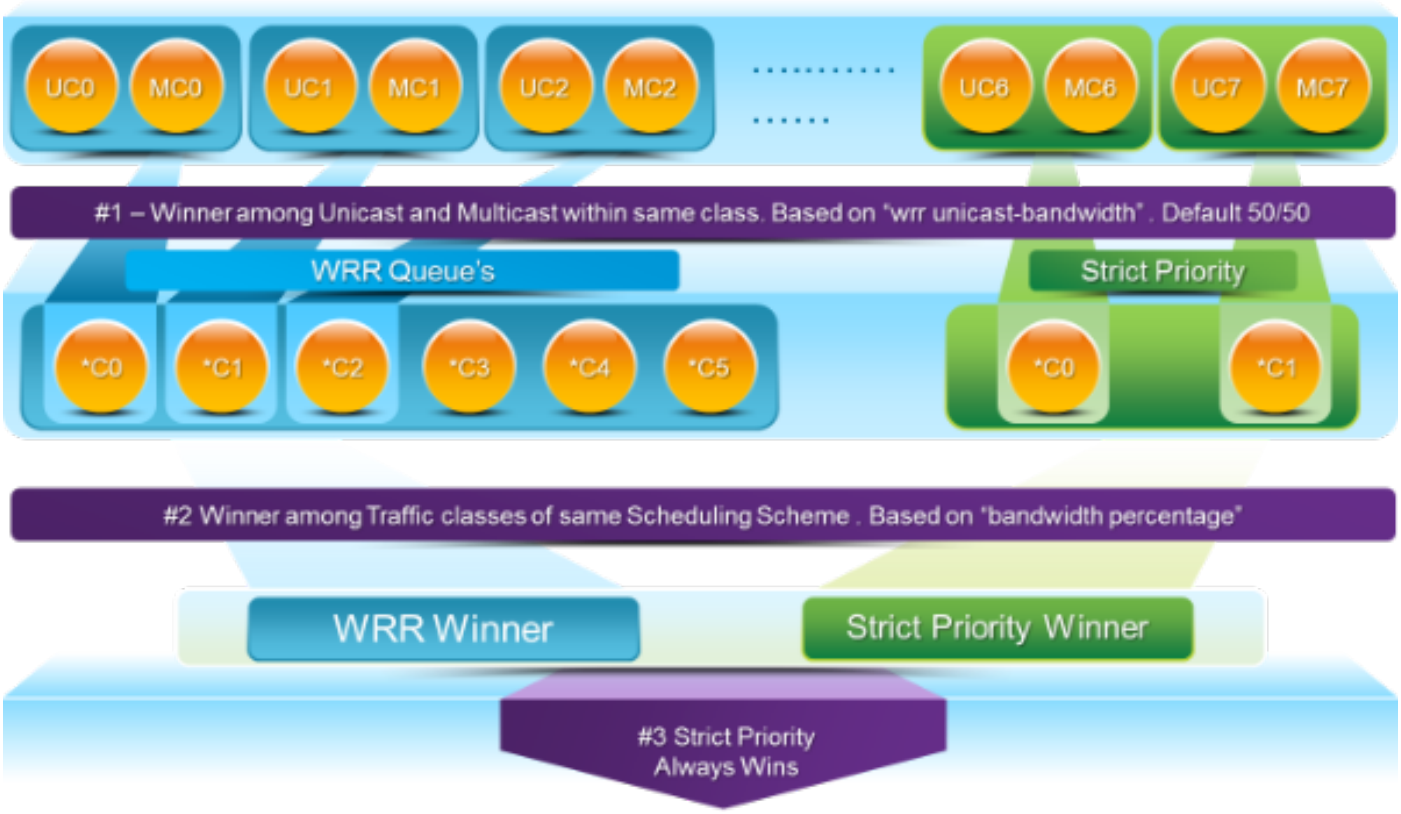


- 由三个OB块共享的18 MB数据包缓冲区： 约4 MB保留：根据已配置的最大传输单位(MTU)大小（每端口总和2 x MTU大小x启用的QoS组数）约14 MB共享：总缓冲区的余数约767 KB的OB:0用于CPU发往的数据包
- 每个OB的6 MB由一组16个端口共享(**show hardware internal mtc-usd info port-mapping命令**)

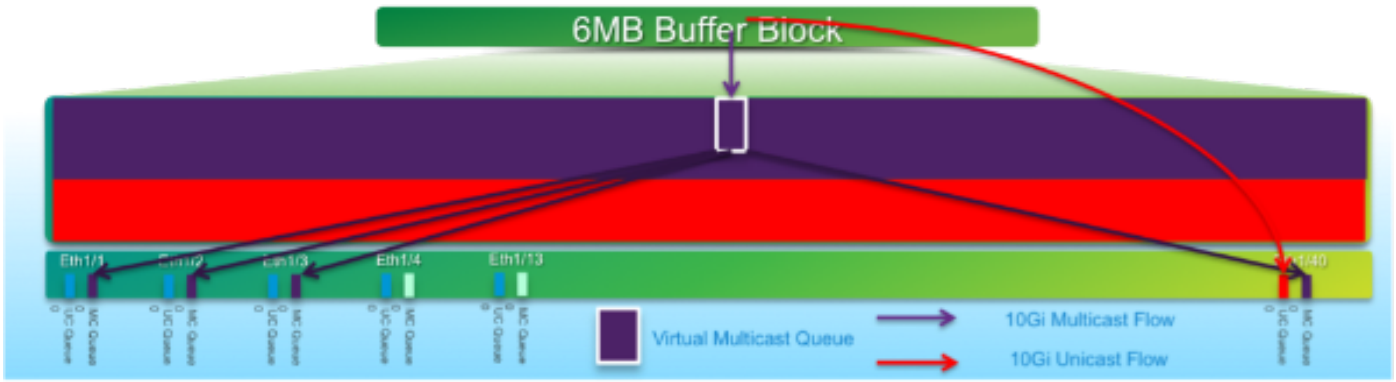
计划

三层调度：

- 单播和组播 当与LTL和CBL相连接的EARL将某个端口或端口组识别为目的地的时候，它指示目的端口上的SAINT ASICs通过控制总线继续接收帧，随后从它的端口发送出去。
- 同一调度方案的流量类
- 方案中的流量类



组播慢速接收器



在本图中：

- 1 G Eth1/40上引入持续拥塞。
- 缓冲区块上的其他组播接收器(Eth1/1-3)受到组播调度行为的影响。 *其他缓冲区块上的接收器不受影响。*
- “组播慢速接收器”可应用于e1/40，以避免非拥塞端口上的流量丢失。
- “组播慢接收器”以10 G速率在Eth1/40上排除组播。*拥塞端口仍预计会发生丢包。*
- 使用硬件配**置文件组播慢速接收器端口<x>命**令配置。

活动缓冲区监控

有关此功能的概述，请参阅Cisco Nexus 3548活动缓冲区监控白皮书。

硬件实施

- ASIC有18个桶，每个桶对应一系列缓冲区利用率（例如0-384KB、385-768KB等）。
- ASIC每4毫秒轮询一次所有端口的缓冲区利用率（默认）。 此ASIC轮询间隔可配置为低至10纳秒。
- 根据每个硬件轮询间隔的缓冲区利用率，相应范围的桶计数器将递增。即，如果端口25消耗500 KB的缓冲区，桶#2(385-768KB)计数器将递增。
- 此缓冲区利用率计数器以直方图格式维护每个接口。
- 每个存储桶都用8位表示，因此计数器的最大值为255，一旦软件读取数据，就会重置。

软件实施

- 每隔一秒，软件会轮询ASIC以下载和清除所有直方图计数器。
- 这些直方图计数器在内存中以一秒的粒度维护60分钟。
- 软件还确保每小时将缓冲区直方图复制到bootflash，然后再复制到分析器进行进一步分析。
- 实际上，这可以为所有端口保留2小时的缓冲区直方图数据，在内存中保留最近1小时，在bootflash中保留第二小时。