排除ACI交换矩阵内转发故障 — 间歇丢弃

目录

简介

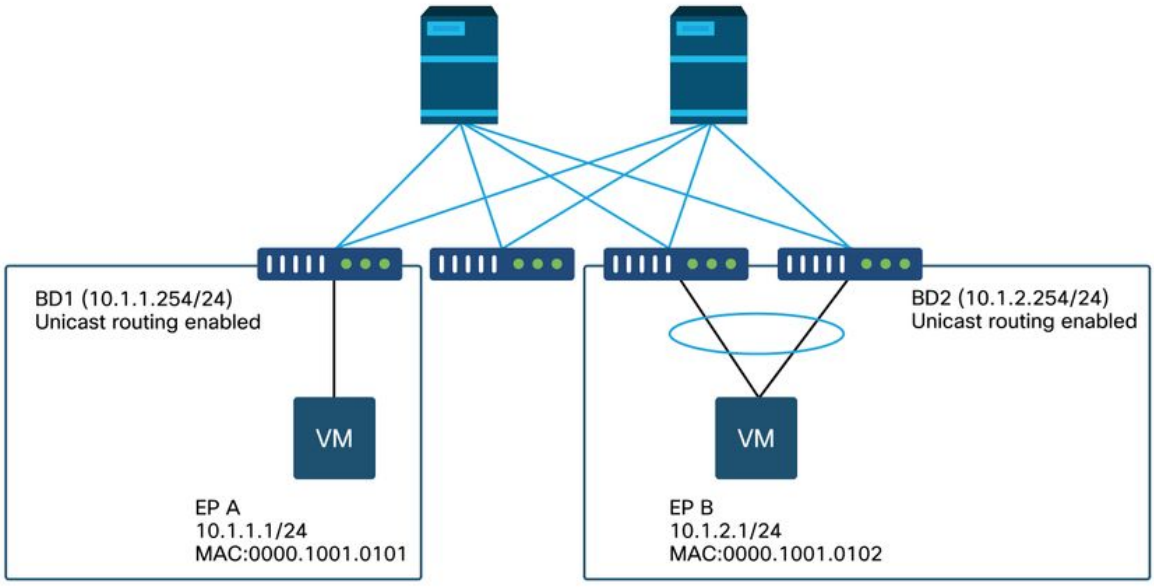
本文档介绍对ACI中的间歇性丢弃进行故障排除的步骤。

背景信息

本文档中的资料摘自Troubleshooting Cisco Application Centric Infrastructure， Second Edition一书，特别是Intra-Fabric forwarding - Interrupt drops一章。

排除ACI交换矩阵内转发故障 — 间歇丢弃

拓扑示例



在本例中，从EP A(10.1.1.1)到EP B(10.1.2.1)的ping操作遇到间歇性丢弃。

```
[EP-A ~]$ ping 10.1.2.1 -c 10
PING 10.1.2.1 (10.1.2.1) 56(84) bytes of data.
64 bytes from 10.1.2.1: icmp_seq=1 ttl=231 time=142 ms
64 bytes from 10.1.2.1: icmp_seq=2 ttl=231 time=141 ms
                                <-- missing icmp_seq=3

64 bytes from 10.1.2.1: icmp_seq=4 ttl=231 time=141 ms
64 bytes from 10.1.2.1: icmp_seq=5 ttl=231 time=141 ms
64 bytes from 10.1.2.1: icmp_seq=6 ttl=231 time=141 ms
                                 <-- missing icmp_seq=7

64 bytes from 10.1.2.1: icmp_seq=8 ttl=231 time=176 ms
64 bytes from 10.1.2.1: icmp_seq=9 ttl=231 time=141 ms
64 bytes from 10.1.2.1: icmp_seq=10 ttl=231 time=141 ms

--- 10.1.2.1 ping statistics ---
10 packets transmitted, 8 received, 20% packet loss, time 9012ms
```

故障排除工作流程

1.确定哪个方向导致间歇性跌落

在目标主机(EP B)上执行数据包捕获（tcpdump、Wireshark等）。 对于ICMP，请关注序列号，以查看EP B上观察到的间歇性丢弃的数据包。

```
[admin@EP-B ~]$ tcpdump -ni eth0 icmp
11:32:26.540957 IP 10.1.1.1 > 10.1.2.1: ICMP echo request, id 3569, seq 1, length 64
11:32:26.681981 IP 10.1.2.1 > 10.1.1.1: ICMP echo reply, id 3569, seq 1, length 64
11:32:27.542175 IP 10.1.1.1 > 10.1.2.1: ICMP echo request, id 3569, seq 2, length 64
11:32:27.683078 IP 10.1.2.1 > 10.1.1.1: ICMP echo reply, id 3569, seq 2, length 64
11:32:28.543173 IP 10.1.1.1 > 10.1.2.1: ICMP echo request, id 3569, seq 3, length 64   <---
11:32:28.683851 IP 10.1.2.1 > 10.1.1.1: ICMP echo reply, id 3569, seq 3, length 64     <---
11:32:29.544931 IP 10.1.1.1 > 10.1.2.1: ICMP echo request, id 3569, seq 4, length 64
11:32:29.685783 IP 10.1.2.1 > 10.1.1.1: ICMP echo reply, id 3569, seq 4, length 64
11:32:30.546860 IP 10.1.1.1 > 10.1.2.1: ICMP echo request, id 3569, seq 5, length 64
...
```

- 模式1 — 在EP B数据包捕获中观察到所有数据包。

丢弃应位于ICMP回应应答中（EP B到EP A）。

- 模式2 — 在EP B数据包捕获中观察到间歇性丢弃。

丢弃应位于ICMP回应中（EP A到EP B）。

2.检查源/目标IP相同的其它协议是否有相同的问题

如果可能，请尝试使用两个终端之间合同允许的不同协议（例如ssh、telnet、http、..）测试两个终端之间的连接

- 模式1 — 其他协议具有相同的间歇性丢弃。

问题可能出在终端抖动或队列/缓冲中，如下所示。

- 模式2 — 只有ICMP出现间歇性丢弃。

转发表（如终端表）应该没有问题，因为转发基于MAC和IP。排队/缓冲也不应该是原因，因为这会影响其他协议。ACI根据协议做出不同转发决策的唯一原因就是PBR使用案例。

一种可能是其中一个主干节点出现问题。当协议不同时，具有相同源和目标的数据包可以通过入口枝叶被负载均衡到另一个上行链路/交换矩阵端口（即另一个主干）。

原子计数器可用于确保数据包不会在脊柱节点上丢弃并到达出口枝叶。如果数据包未到达出口枝叶，请检查入口枝叶上的ELAM以查看数据包从哪个交换矩阵端口发出。要将问题隔离到特定主干，可以关闭枝叶上行链路，强制流量流向另一个主干。

3.检查它是否与终端学习问题相关

ACI使用终端表将数据包从一个终端转发到另一个终端。间歇性可达性问题可能由终端抖动引起，因为不适当的终端信息会导致数据包被发送到错误的目的地，或由于被分类为错误的EPG而被合同丢弃。即使目标应该是L3Out而不是终端组，请确保不会将IP作为终端跨任何枝叶交换机学习为同一VRF中的终端。

请参阅本部分的“终端抖动”子部分，了解有关如何对终端抖动进行故障排除的更多详细信息。

4.通过更改流量频率检查是否与缓冲问题相关

增大或减小ping间隔，查看丢弃率是否发生变化。间隔差应该足够大。

在Linux中，“ — i”选项可用于更改间隔（秒）：

```
[EP-A ~]$ ping 10.1.2.1 -c 10 -i 5      -- Increase it to 5 sec
[EP-A ~]$ ping 10.1.2.1 -c 10 -i 0.2    -- Decrease it to 0.2 msec
```

如果丢弃率在间隔缩短时增加，则可能与终端或交换机上的排队或缓冲有关。

要考虑的丢弃比是（丢弃数/发送的数据包总数）而不是（丢弃数/时间）。

在这种情况下，请检查以下项。

1. 检查交换机接口上的任何丢弃计数器是否随着ping而增加。有关详细信息，请参阅“交换矩阵内转发”一章中的“接口丢弃”部分。
2. 检查Rx计数器是否随目标终端上的数据包一起增加。如果Rx计数器的计数与传输的数据包相同，则数据包很可能在终端本身被丢弃。这可能是由于TCP/IP堆栈上的终端缓冲。

例如，如果100000以尽可能短的时间间隔发送ping，则可以观察到终端上的Rx计数器随着时间的100000加。

```
[EP-B ~]$ ifconfig eth0
eth0: flags=4163<UP,BROADCAST,RUNNING,MULTICAST>  mtu 1500
      inet 10.1.2.1  netmask 255.255.255.0  broadcast 10.1.2.255
      ether 00:00:10:01:01:02  txqueuelen 1000  (Ethernet)
      RX packets 101105  bytes 1829041
      RX errors 0  dropped 18926930  overruns 0  frame 0
      TX packets 2057  bytes 926192
      TX errors 0  dropped 0 overruns 0  carrier 0  collisions 0
```

5.检查ACI是否正在向外发送数据包，或者目的设备是否正在接收数据包

在枝叶交换机的出口端口上捕获SPAN，以从故障排除路径中消除ACI交换矩阵。

目标上的Rx计数器也可用于从故障排除路径中排除整个网络交换机，如前面的缓冲步骤所示。

终端抖动

本节介绍如何检查终端抖动。其他详细信息可在以下文档中找到：

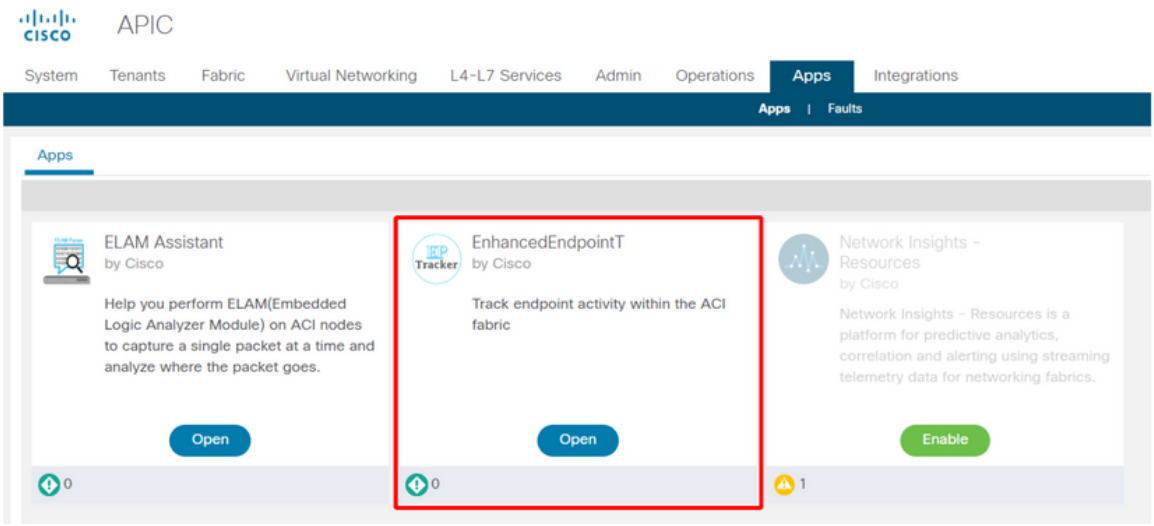
- www.cisco.com上的“ACI交换矩阵终端学习白皮书”
- “Cisco Live BRKACI-2641 ACI故障排除：终端”(www.ciscolive.com)

当ACI在多个位置获取同一MAC或IP地址时，终端看起来已移动。这也可能由欺骗设备或配置错误导致。此类行为称为终端抖动。在这种情况下，流向移动/摆动端点的流量（桥接流量的MAC地址，路由流量的IP地址）将间歇性地发生故障。

检测终端抖动的最有效方法是使用增强型终端跟踪器。此应用可以作为ACI AppCenter应用运行，也可以在外部服务器上作为独立应用运行，以便管理更大的交换矩阵。

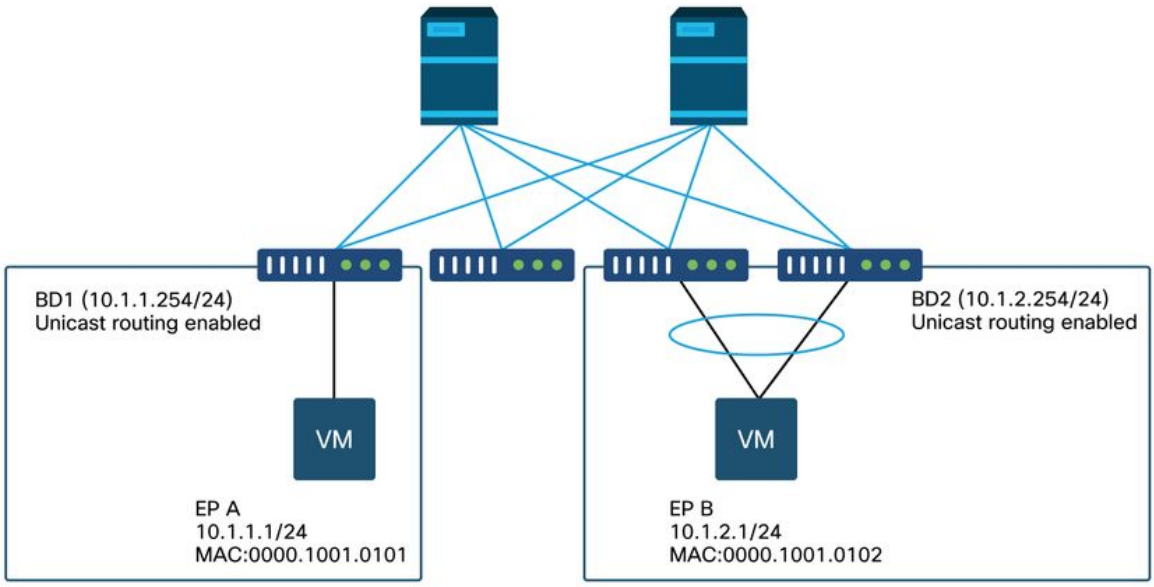
增强的终端跟踪器

弃用警告！本指南写于4.2;此后，Enhanced Endpoint Tracker应用已弃用，转而支持Nexus Dashboard Insights的功能。有关详细信息，请参阅Cisco Bug ID CSCvz59365 .



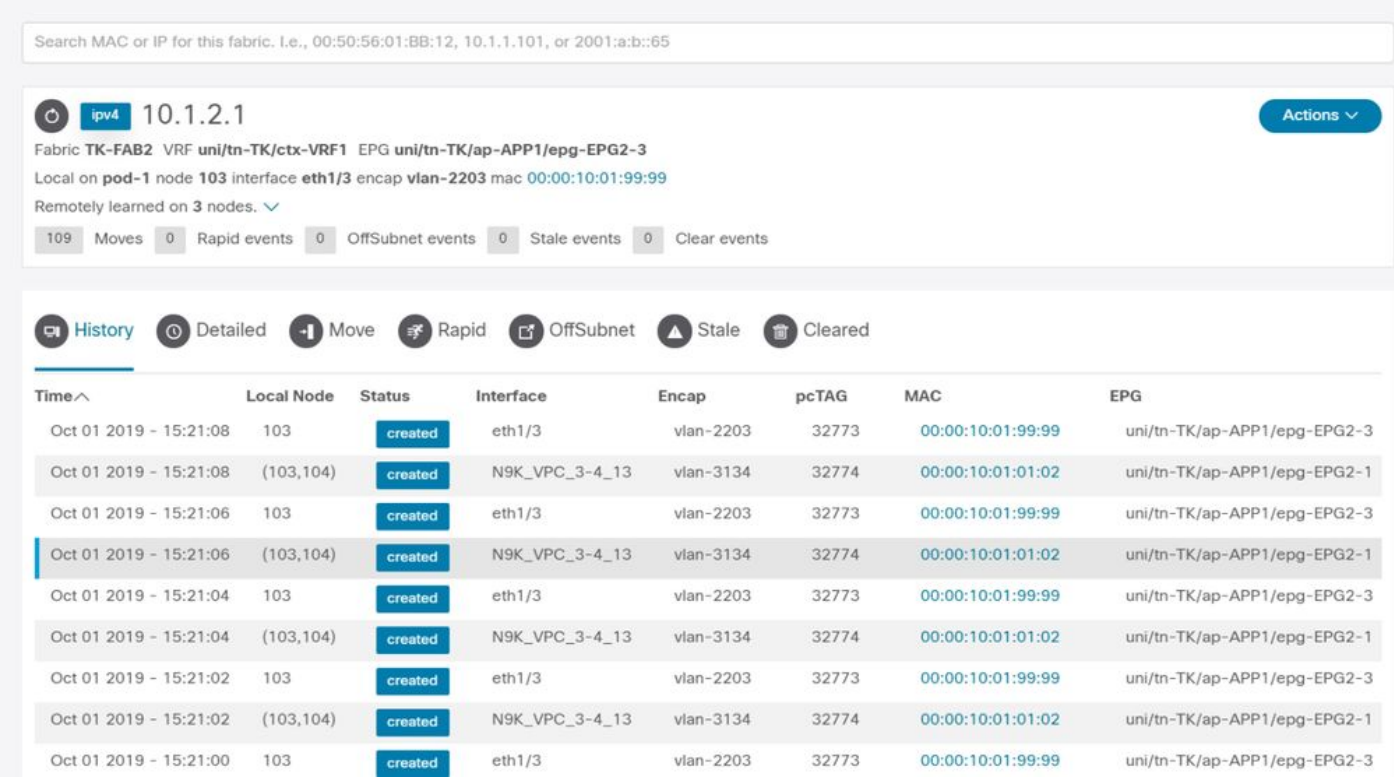
上图显示了AppCenter中的增强型终端跟踪器。以下显示如何使用增强型终端跟踪器查找摆动终端的示例。

终端抖动示例



在本例中，IP 10.1.2.1应属于具有MAC 0000.1001.0102的EP B。但是，具有MAC 0000.1001.9999的EP X由于配置错误或可能存在IP欺骗，也使用IP 10.1.2.1来采购流量。

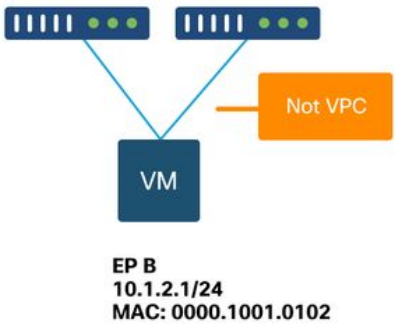
增强的终端跟踪器输出 — 移动



Enhanced Endpoint Tracker显示何时何地获知IP 10.1.2.1。如上面的屏幕截图所示，10.1.2.1在带有MAC 0000.1001.0102（预期）和0000.1001.9999（非预期）的两个终端之间摆动。 这将导致发往IP 10.1.2.1的可达性问题，因为当在错误的MAC地址上获知数据包时，数据包将通过错误的接口发送到错误的设备。要解决此问题，请采取措施防止意外的VM使用不合适的IP地址发起流量。

以下显示由于配置不当导致终端摆动的典型示例。

可能导致终端抖动的拓扑示例



当服务器或VM通过没有VPC的两个接口连接到ACI枝叶节点时，服务器需要使用主用/备用NIC组合。否则，数据包将负载均衡到两个上行链路，并且从ACI枝叶交换机的角度来看，终端看起来好像在两个接口之间摆动。在这种情况下，需要主用/备用或等效的NIC组合模式，或者仅使用ACI端的VPC。

接口丢弃

本章介绍如何检查与入口接口丢弃相关的主计数器。

硬件丢弃计数器类型

在ACI模式下运行的Nexus 9000交换机上，ACI上有三个主要硬件计数器用于入口接口丢弃。

转发

丢弃的主要原因包括：

- SECURITY_GROUP_DENY:由于缺少允许通信的合同而丢弃。
- VLAN_XLATE_MISS:由于VLAN不当而丢弃。例如，帧进入带有802.1Q VLAN 10的交换矩阵。如果交换机端口上有VLAN 10，它会检查内容并根据目的MAC做出转发决策。但是，如果端口上不允许VLAN 10，它会丢弃该端口并将其标记为VLAN_XLATE_MISS。
- ACL_DROP:由于SUP-TCAM而下降。ACI交换机中的SUP-TCAM包含要在正常L2/L3转发决策之上应用的特殊规则。SUP-TCAM中的规则是内置的，用户不可配置。SUP-TCAM规则的主要目的是处理某些异常或某些控制平面流量，而不是由用户检查或监控。当数据包达到SUP-TCAM规则且规则为丢弃数据包时，丢弃的数据包将计为ACL_DROP，它将增加转发丢弃计数器。

转发丢弃实质上是指由于有效已知原因而丢弃的数据包。它们通常可以忽略，不会导致性能下降，这与实际数据流量丢弃不同。

Error

当交换机收到无效帧时，该帧将作为错误被丢弃。示例包括带有FCS或CRC错误的帧。有关详细信息，请参阅后面的“CRC — FCS — 直通交换”部分。

缓冲区

当交换机收到帧时，如果没有缓冲区可用于入口或出口，该帧将带有“Buffer”标记。 这通常提示网络中的某处存在拥塞。显示故障的链路可能已满，或者包含目标的链路已拥塞。

使用API收集计数器

值得注意的是，通过利用API和对象模型，用户可以快速查询这些丢包的所有实例（从apic运行这些实例）。

```
# FCS Errors (non-stomped CRC errors)
moquery -c rmonDot3Stats -f 'rmon.Dot3Stats.fCSErrors>="1"' | egrep "dn|fCSErrors"

# FCS + Stomped CRC Errors
moquery -c rmonEtherStats -f 'rmon.EtherStats.cRCAlignErrors>="1"' | egrep "dn|cRCAlignErrors"

# Output Buffer Drops
moquery -c rmonEgrCounters -f 'rmon.EgrCounters.bufferdroppkts>="1"' | egrep "dn|bufferdroppkts"

# Output Errors
moquery -c rmonIfOut -f 'rmon.IfOut.errors>="1"' | egrep "dn|errors"
```

在CLI中查看丢弃统计信息

如果发现故障，或者需要使用CLI检查接口上的丢包，最好通过查看硬件中的平台计数器来检查丢包

。并非所有计数器都使用“show interface”显示。 三个主要丢弃原因只能使用平台计数器查看。要查看这些信息，请执行以下步骤：

枝叶

通过SSH连接到枝叶并运行这些命令。本示例适用于ethernet 1/31。

```
ACI-LEAF# vsh_lc
module-1# show platform internal counters port 31
Stats for port 31
(note: forward drops includes sup redirected packets too)
IF          LPort                    Input                        Output
                                Packets       Bytes          Packets    Bytes
eth-1/31    31  Total          400719   286628225      2302918    463380330
                Unicast     306610   269471065       453831     40294786
                Multicast        0           0      1849091    423087288
                Flood        56783     8427482            0            0
              Total Drops     37327                        0
                Buffer            0                        0
                Error             0                        0
                Forward      37327
                LB                0
                AFD RED                                    0
...
```

主干

可以使用与枝叶交换机相同的方法检查固定主干（N9K-C9332C和N9K-C9364C）。

对于模块化主干（N9K-C9504等），必须先将线卡连接到才能查看平台计数器。通过SSH连接到主干，然后运行这些命令。本示例适用于ethernet 2/1。

```
ACI-SPINE# vsh
ACI-SPINE# attach module 2
module-2# show platform internal counters port 1
Stats for port 1
(note: forward drops include sup redirected packets too)
IF          LPort                    Input                        Output
                                Packets       Bytes          Packets    Bytes
eth-2/1      1  Total        85632884  32811563575     126611414   25868913406
                Unicast   81449096  32273734109     104024872   23037696345
                Multicast  3759719   487617769     22586542   2831217061
                Flood            0           0            0            0
              Total Drops         0                        0
                Buffer            0                        0
                Error             0                        0
                Forward           0
                LB                0
                AFD RED                                    0
...
```

使用“show queuing interface”显示排队统计信息计数器。 本示例适用于ethernet 1/5。

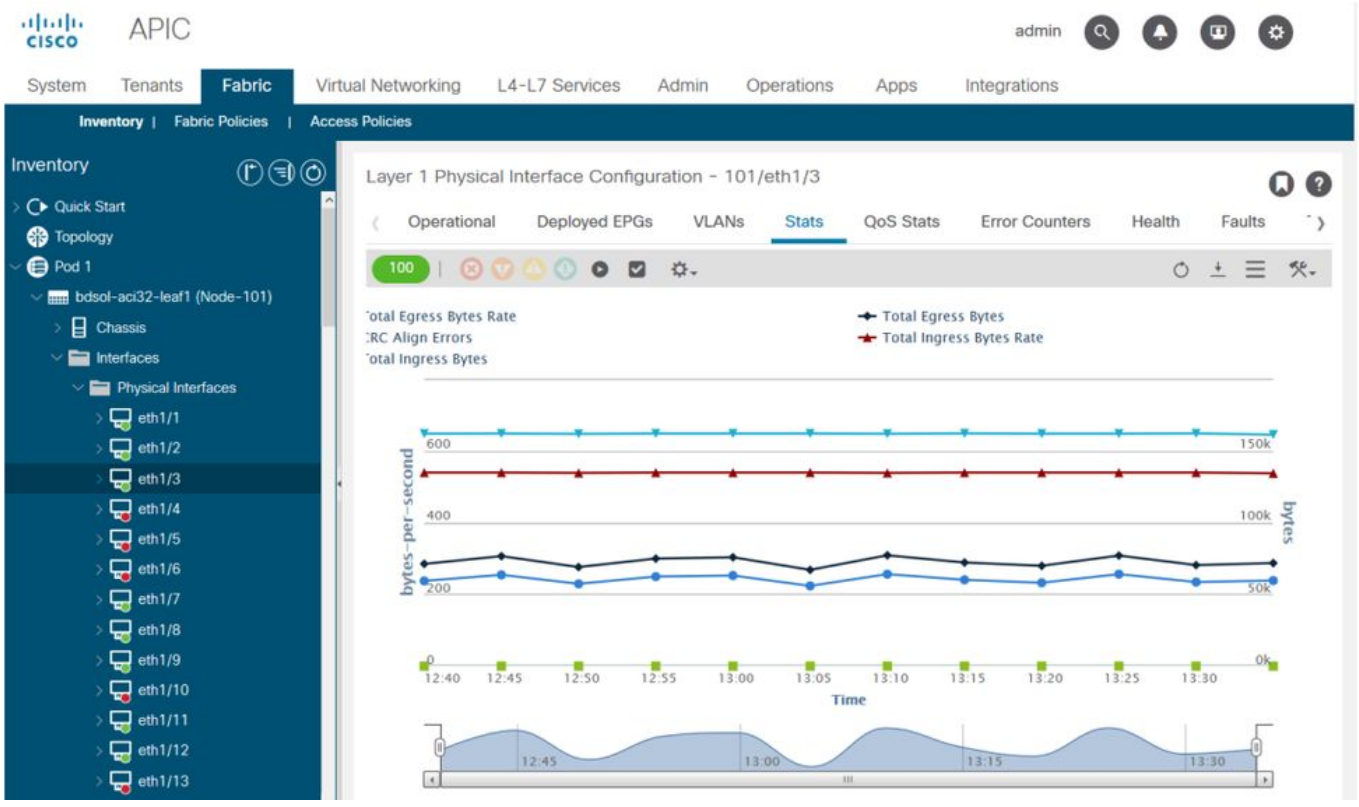
```
ACI-LEAF# show queuing interface ethernet 1/5
================================================================================
               Queuing stats for ethernet 1/5
================================================================================
```

```
================================================================================
                Qos Class level1
================================================================================
Rx Admit Pkts :  0                      Tx Admit Pkts :  0
Rx Admit Bytes:  0                      Tx Admit Bytes:  0
Rx Drop Pkts  :  0                      Tx Drop Pkts  :  0
Rx Drop Bytes :  0                      Tx Drop Bytes :  0

================================================================================
                Qos Class level2
================================================================================
Rx Admit Pkts :  0                      Tx Admit Pkts :  0
Rx Admit Bytes:  0                      Tx Admit Bytes:  0
Rx Drop Pkts  :  0                      Tx Drop Pkts  :  0
Rx Drop Bytes :  0                      Tx Drop Bytes :  0

================================================================================
                Qos Class level3
================================================================================
Rx Admit Pkts :  1756121                Tx Admit Pkts :  904909
Rx Admit Bytes:  186146554              Tx Admit Bytes:  80417455
Rx Drop Pkts  :  0                      Tx Drop Pkts  :  22
Rx Drop Bytes :  0                      Tx Drop Bytes :  3776

...
```

在GUI中查看统计信息

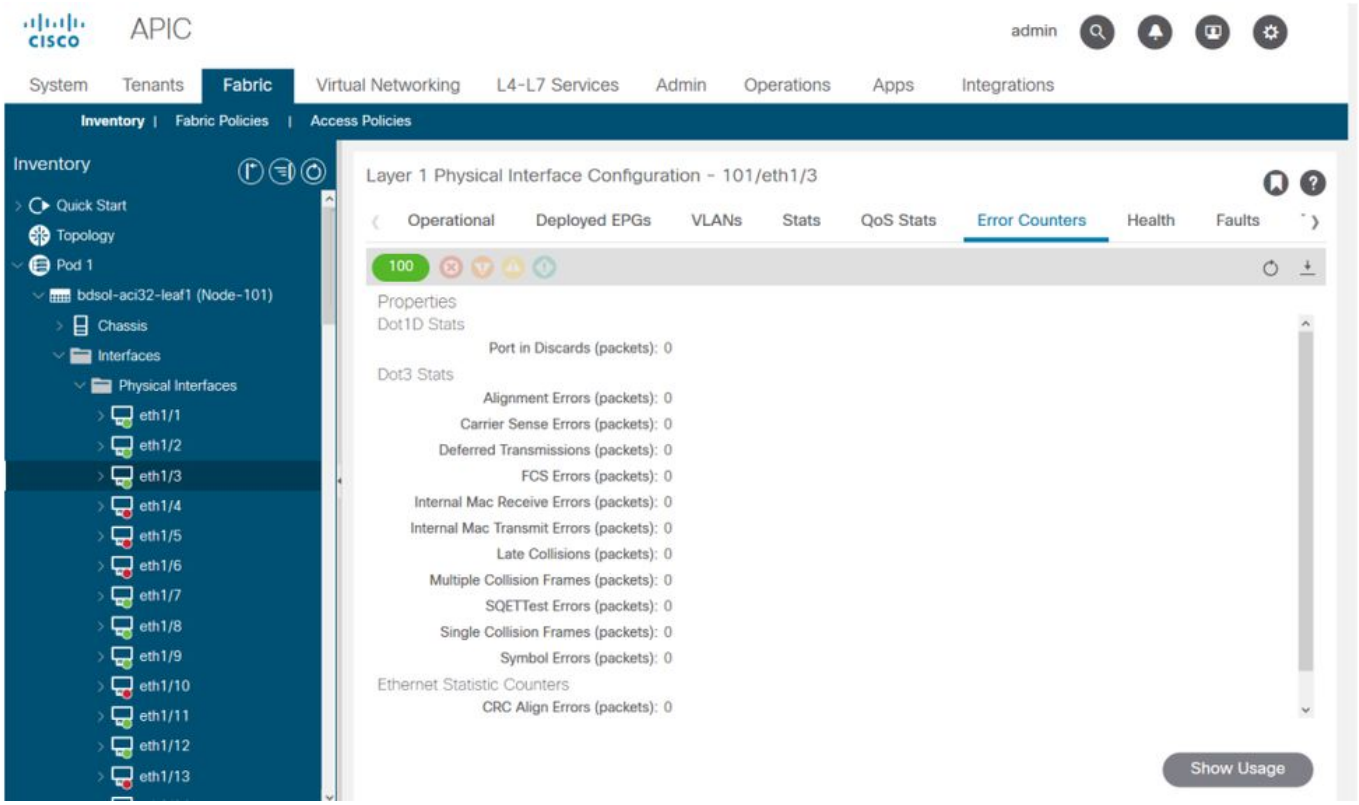
位置为“交换矩阵>资产>枝叶/主干>物理接口>统计信息”。

GUI界面统计信息



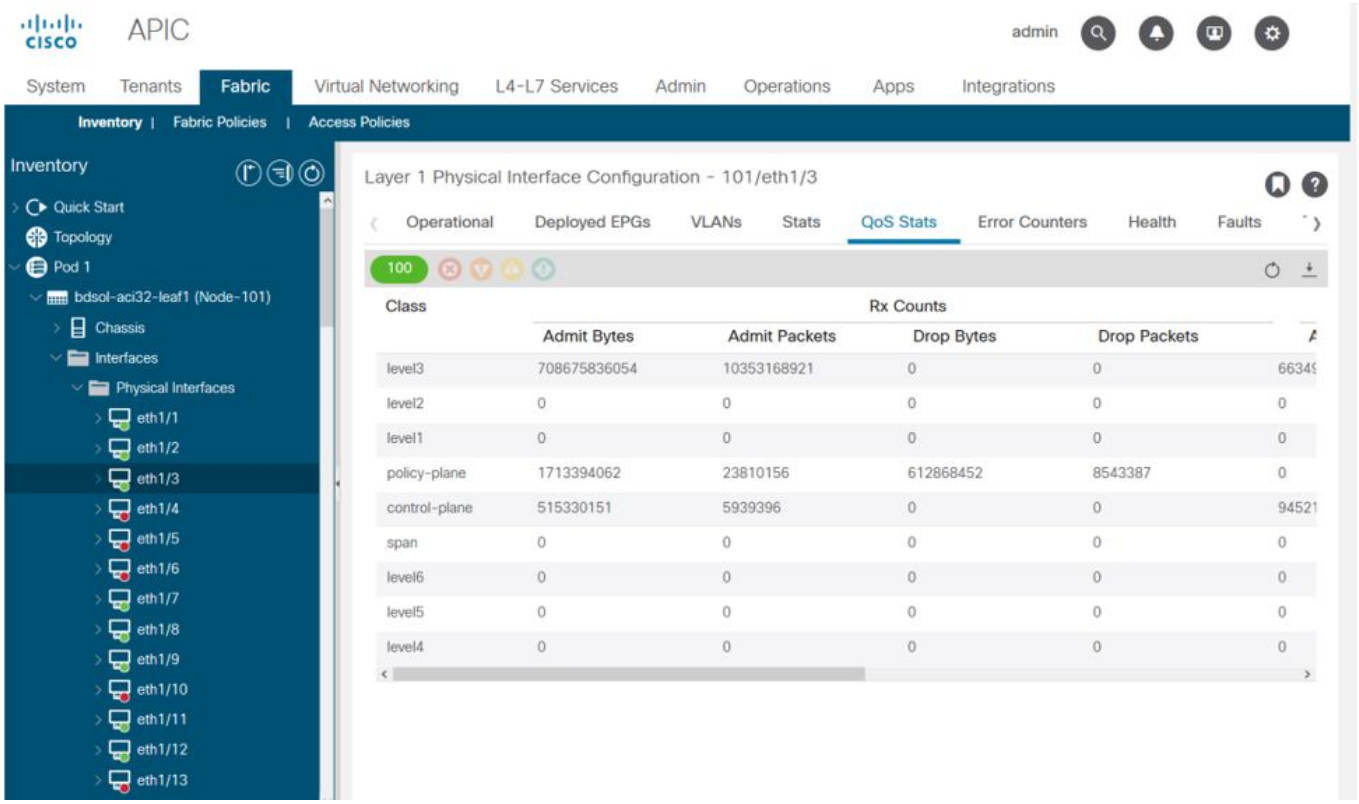
错误统计信息可在同一位置看到：

GUI接口错误



最后，GUI可以显示每个接口的QoS统计信息：

GUI接口QoS计数器



CRC - FCS — 直通交换

什么是循环冗余校验(CRC)?

CRC是帧上的多项式函数，在以太网中返回4B数字。它将捕获所有单比特错误和相当比例的双比特错误。因此，其目的是确保帧在传输过程中未损坏。如果CRC错误计数器增加，这意味着当硬件在帧上运行多项式函数时，结果为4B编号，不同于在帧本身上找到的4B编号。帧可能由于多种原因而损坏，例如双工不匹配、布线故障和硬件损坏。但是，应该会遇到一定级别的CRC错误，该标准允许以太网上的最高10-12位错误率（1012位中的1位可以翻转）。

存储转发交换与直通交换

存储转发和直通第2层交换机都根据数据包的目的MAC地址做出转发决策。当站点与网络上的其他节点通信时，它们还会检查数据包的源MAC(SMAC)字段来学习MAC地址。

存储转发交换机在收到整个帧并检查其完整性之后，会对数据包做出转发决策。直通交换机在检查传入帧的目标MAC(DMAC)地址后不久就开始进行转发过程。但是，直通交换机必须等到查看完整个数据包后才能执行CRC检查。这意味着当CRC验证时，数据包已经转发，如果检查失败，则无法丢弃。

传统上，大多数网络设备都基于存储转发运行。直通交换技术往往用于需要低延迟转发的高速网络。

具体而言，对于第2代及更高版本的ACI硬件，如果入口接口速度较高，而出口接口速度相同或较低，则执行直通交换。如果入口接口速度低于出口接口，则执行存储转发交换。

踩踏

具有CRC错误的数据包需要进行丢弃。如果在直通路径中交换帧，则在转发数据包后进行CRC验证。因此，唯一的选择是停止以太网帧校验序列(FCS)。停止帧涉及将FCS设置为一个已知值，该值不会通过CRC校验。因此，一个未通过CRC的坏帧在它经过的每个接口上都会显示为CRC，直到它到达将丢弃它的存储转发交换机。

ACI和CRC:查找有故障的接口

- 如果枝叶在下行链路端口上看到CRC错误，则主要问题是下行SFP或外部设备/网络上的组件问题。
- 如果主干看到CRC错误，则主要是本地端口、SFP、光纤或邻居SFP出现问题。来自枝叶下行链路的CRC故障数据包不会存储到主干。如同其报头可读一样，它采用VXLAN封装，并将计算新的CRC。如果帧损坏导致报头不可读，数据包将被丢弃。
- 如果枝叶在交换矩阵链路上看到CRC错误，则可能是： 本地光纤/SFP对、主干的入口光纤或SFP对出现问题。一个从布料中穿过的窄框架。

踩踏：排除停机故障

- 查找交换矩阵上存在FCS错误的接口。由于FCS发生在端口本地，因此很可能是在任一端使用光纤或SFP。
- “show interface”输出上的CRC错误反映了总FCS+Stomp值。\

看一个例子：

使用命令检查端口

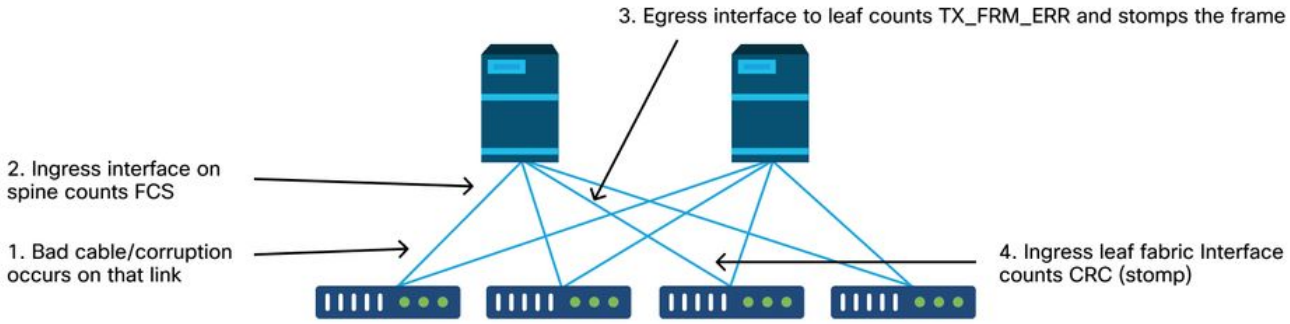
```
vsh_lc: 'show platform internal counter port <X>'
```

在此命令中，3个值非常重要：

- RX_FCS_ERR - FCS故障。
- RX_CRCERR — 收到存储的CRC错误帧。
- TX_FRM_ERROR — 传输的CRC错误帧。

```
module-1# show platform internal counters port 1 | egrep ERR
       RX_FCS_ERR           0      ---- Real error local between the devices and its direct
neighbor
       RX_CRCERR            0      ---- Stomped frame --- so likely stomped by underlying devices
and generated further down the network
       TX_FRM_ERROR         0      ---- Packet received from another interface that was stomp on
Tx direction
```

CRC堆栈故障排除场景



如果损坏的链路生成大量损坏的帧，则这些帧可能会泛洪到所有其他枝叶节点，并且非常有可能在交换矩阵中大多数枝叶节点的交换矩阵上行链路入口找到CRC。这些可能都来自一个损坏的链路。

关于此翻译
思科采用人工翻译与机器翻译相结合的方式将此文档翻译成不同语言，希望全球的用户都能通过各自的语言得到支持性的内容。

请注意：即使是最好的机器翻译，其准确度也不及专业翻译人员的水平。

Cisco Systems, Inc. 对于翻译的准确性不承担任何责任，并建议您总是参考英文原始文档（已提供链接）。