IP路径MTU发现和DLSw

目录

简介

IBM协议簇、DLSw、STUN和BSTUN建立从一台路由器到另一台路由器的IP会话管道。由于TCP的可靠性，它通常用作路由器之间的传输方法。本文档提供有关TCP动态发现会话管道上可使用的最大MTU的能力的信息，这可以最大限度地减少分段并提高效率。

开始使用前

规则

有关文档规则的详细信息，请参阅 Cisco 技术提示规则。

先决条件

本文档没有任何特定的前提条件。

使用的组件

本文档不限于特定的软件和硬件版本。

本文档中的信息都是基于特定实验室环境中的设备创建的。本文档中使用的所有设备最初均采用原始（默认）配置。如果您是在真实网络上操作，请确保您在使用任何命令前已经了解其潜在影响。

背景信息

路径MTU发现(PMTD)（如RFC 1191中所述）指定IP数据包的默认字节大小为576。帧的IP和TCP部分占用40字节，保留536字节作为数据负载。此空间称为最大数据段大小或MSS。RFC1191第3.1节指定可协商的更大MSS，这正是在Cisco路由器中发出**ip tcp path-mtu-discovery**命

令的作用。当配置此命令并启动TCP会话时，路由器外的SYN数据包包含指定更大MSS的TCP选项。此较大的MSS是出站接口的MTU减去40字节。如果出站接口的MTU为1500字节，则通告的MSS为1460。如果出站接口的MTU更大，例如帧中继的MTU为4096字节，则MSS将是4096字节减去40字节的IP信息，并显示在**show tcp**命令输出中（最大数据段为4056字节）。

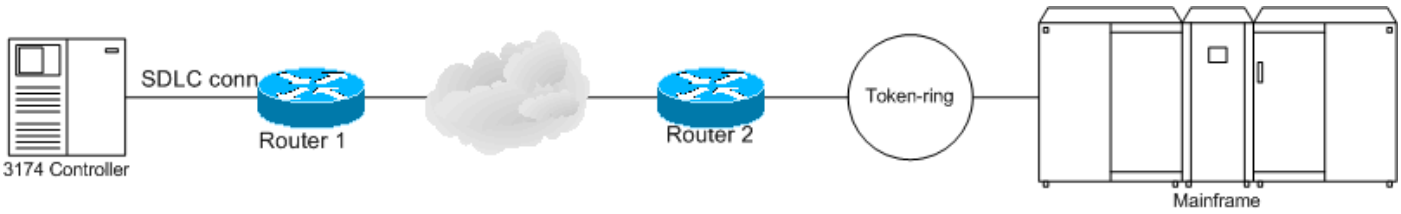
在路由器上配置PMTD对从路由器建立的现有TCP会话没有任何影响。PMTD被引入到11.3.5T IOS级别，在IOS的后续版本中，它成为可选命令。在IOS 11.3(5)T之前，默认情况下为打开状态。此外，当IP地址在同一子网中时，PMTD是自动的，表示它们直接连接在同一介质上。

必须配置两台路由器，PMTD才能正常工作。配置两台路由器后，从一台路由器到另一台路由器的SYN包含通告更高MSS的可选TCP值。返回的SYN随后通告更高的MSS值。因此，双方都向对方通告可以接受更大的MSS。如果只配置了一台路由器（路由器1）的**ip tcp path-mtu-discovery**命令，它将通告较大的MSS，因此路由器2可以向路由器1发送1460字节的帧。路由器2永远不会通告较大的MSS，因此路由器1被锁定以发送默认值。

带PMTD的DLSw

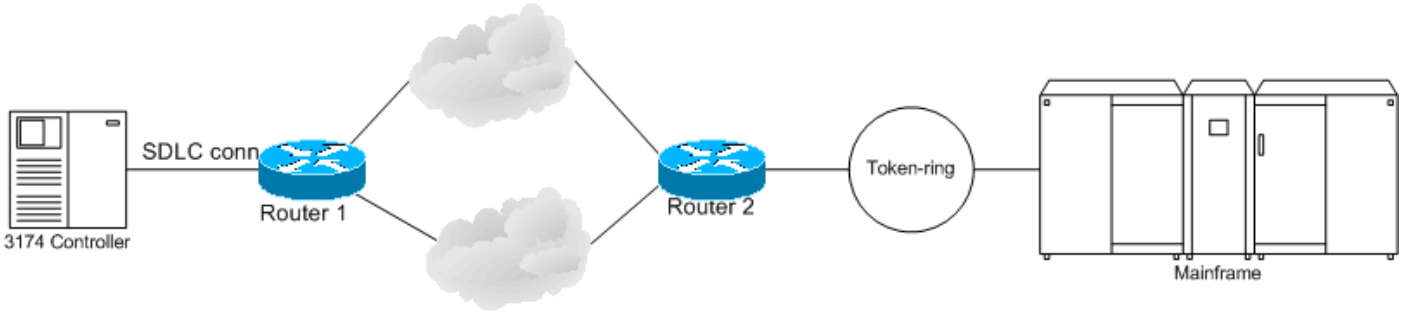
在IBM套件中，DLSw、STUN和BSTUN的任务是通过TCP会话从路由器传输大量数据。实施PMTD非常重要，而且非常有益，特别是考虑到它在11.2和之前的IOS级别默认启用。根据RFC，最大帧默认为576字节，IP/TCP封装的最大帧为减40字节。DLSw使用另外16个字节进行封装。使用默认MSS传输的实际数据为520字节。DLSw还能够将两个不同的逻辑链路控制2(LLC2)数据包传输到一个TCP帧中。如果两台工作站各自发送一个LLC2帧，DLSw可以在一个帧中将两个LLC2帧传送给DLSw远程对等体。通过拥有更大的MSS，TCP驱动程序可以适应这种捎带模式。以下三个主要场景用于说明**path-mtu-discovery命令的**值。

场景1 — 不需要的开销



SDLC设备通常配置为每个帧的最大数据为265或521字节。如果值为521,3174向路由器1发送一个521字节的SDLC帧，路由器1将制作两个TCP帧，将其传输到DLSW对等体路由器2。第一个帧将包含520字节的数据、16字节的DLSw信息和40字节的IP信息总共576字节。第二个数据包将包含1个字节的数据、16个字节的DLSw信息和40个字节的IP信息。当使用PMTD并假设1500字节的MTU获得1460 MSS时，路由器2告知路由器1，路由器2可以接收1460字节的数据。路由器1将在一个数据包中将所有521字节的SDLC数据发送到路由器2，其中包含16字节的DLSw信息和40字节的IP信息。由于DLSw是进程交换事件，因此通过使用PMTD，处理此SDLC帧的CPU利用率已减半。此外，路由器2不必等待第二个数据包组装LLC2帧。启用PMTD后，路由器2将接收整个数据包，然后可以从数据包中删除IP和DLSW信息，并无延迟地将其发送到3745。

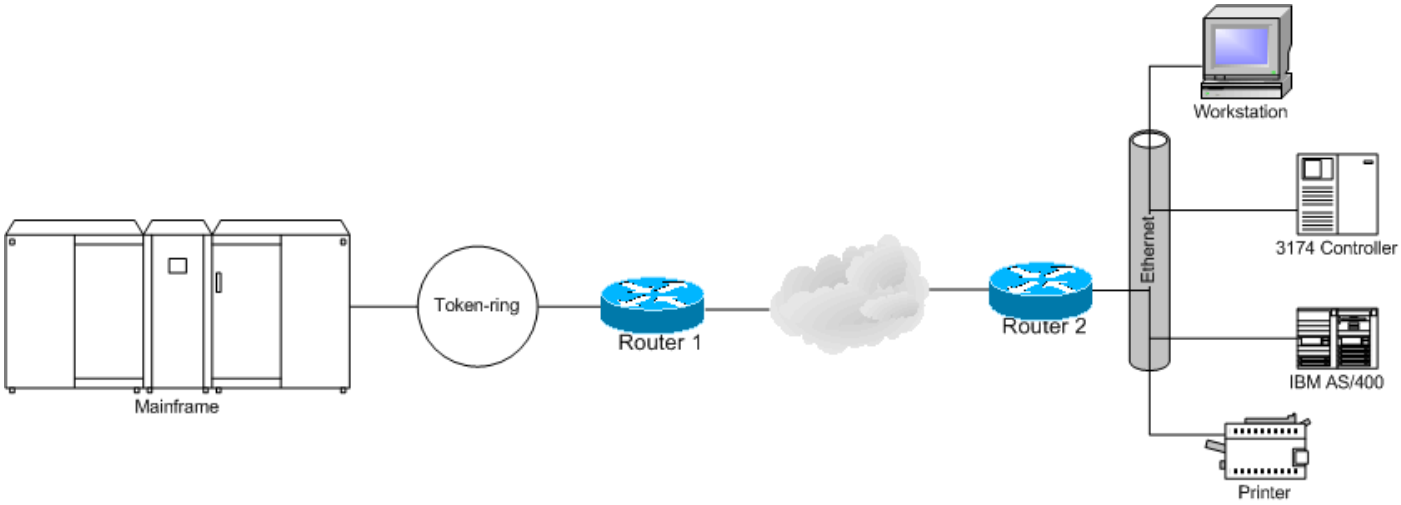
场景2 — 无序数据包的延迟



在此场景中，有两个IP云可用，它们的负载均衡或冗余度量相等。不启用PMTD会严重降低DLSw的速度。如果未启用PMTD，路由器1必须将521字节的帧组合为两个TCP数据包，一个包含520字节的数据，另一个包含1字节的数据。如果第一个数据包通过顶部IP云，则如果第二个数据包通过较低的等成本IP云发送，则很有可能在第二个数据包之后到达。这将生成称为无序数据包的数据包。IP/TCP协议的固有功能是能够管理此问题。无序数据包存储在等待整个流到达的内存中，然后重新组装。无序数据包并不罕见，但是，当此事件利用内存和CPU资源时，应进行所有尽量减少它们的尝试。大量的订单失序可能导致TCP级别出现严重延迟。如果第3层/DLSw会话延迟，则通过此DLSw承载的LLC2/SDLC会话随后将受到影响。如果在此场景中启用PMTD，则单个521字节的帧会通过任一IP云在一个TCP帧中发送。接收路由器只需要缓冲区，并解封一个TCP帧。

PMTD与SNA环境中通告给终端站的最大帧没有关系。这包括令牌环上路由信息字段(RIF)中的最大帧(LF)。PMTD严格规定可封装到一个TCP帧中的数据量。LLC2和SDLC不具有分段数据包的功能，但IP/TCP具有分段数据包的功能。大型SNA帧封装到TCP中时，可将其分段。此数据在远程DLSw路由器上解封，并再次显示为非分段SNA数据。

场景3 - LLC2连接和吞吐量更快



在此场景中，3174和工作站具有通过3745 TIC到大型机的会话，如果两台设备都发送发往主机的数据，则TCP可能会将两个LLC2帧放入一个数据包中。如果没有PMTD，则两个帧中的合并数据为521字节或更大，则不可能执行此操作。在这种情况下，TCP软件需要单独发送每个数据包。例如，如果3174和工作站大约同时发送帧，并且这些数据包中的每个数据包有400字节的数据，则路由器会接收并缓冲每个帧。现在，路由器必须将这400字节数据流中的每一个封装成单独的TCP数据包，以便传输到对等体。

启用PMTD并假设MSS为1460，路由器将接收并缓冲两个LLC2数据包。现在，它可以将两者封装到一个数据包中。此TCP数据包将包含40个字节的IP信息、第一个DLSw电路配对的16个字节的DLSw信息、400个字节的数据、第二个DLSw电路配对的另外16个字节的数据，以及其他400个字节的数据。此特定场景使用两台设备，因此使用两条DLSw电路。PMTD使DLSw能够更高效地扩展到更多DLSw电路。许多分支中心网络需要数百个远程站点，每个远程站点都有一个或两个SNA设备，它们对等到连接到OSA或FEP的中心站点路由器，以访问主机应用。PMTD使TCP和DLSw能够

扩展到更大的需求，而不会过度利用路由器CPU和内存资源，并提供更快的传输。

注意：在12.1(5)T后期发现并在12.2(5)T中解决了软件Bug，其中PMTD未通过虚拟专用网络(VPN)隧道工作。此软件缺陷为CSCdt49552(仅限注册客户)。

检验DLSW的PMTD

发出**show tcp** 命令。

```
havoc#show tcp

Stand-alone TCP connection to host 10.1.1.1
Connection state is ESTAB, I/O status: 1, unread input bytes: 0
Local host: 30.1.1.1, Local port: 11044
Foreign host: 10.1.1.1, Foreign port: 2065

Enqueued packets for retransmit: 0, input: 0  mis-ordered: 0 (0 bytes)

TCP driver queue size 0, flow controlled FALSE

Event Timers (current time is 0xA18A78):
Timer          Starts    Wakeups            Next
Retrans             3          0             0x0
TimeWait            0          0             0x0
AckHold             0          0             0x0
SendWnd             0          0             0x0
KeepAlive           0          0             0x0
GiveUp              2          0             0x0
PmtuAger            0          0             0x0
DeadWait            0          0             0x0

iss: 3215333571  snduna: 3215334045  sndnxt: 3215334045     sndwnd:  20007
irs: 3541505479  rcvnxt: 3541505480  rcvwnd:      20480  delrcvwnd:      0

SRTT: 99 ms, RTTO: 1539 ms, RTV: 1440 ms, KRTT: 0 ms
minRTT: 24 ms, maxRTT: 300 ms, ACK hold: 200 ms
Flags: higher precedence, retransmission timeout

Datagrams (max data segment is 536 bytes):
Rcvd: 30 (out of order: 0), with data: 0, total data bytes: 0
Sent: 4 (retransmit: 0, fastretransmit: 0), with data: 2, total data bytes: 473
```

此显示被标识为DLSw TCP会话，因为TCP会话中的一个端口是2065。输出底部附近的最大数据段为536字节。此值表示10.1.1.1的远程DLSw对等路由器未配置**ip tcp path-mtu-discovery**命令。536字节值已经占用了IP帧中40字节的IP信息。此536字节值不包括将添加到承载SNA流量的TCP数据包的16个字节的DLSw信息。

配置**ip tcp path-mtu-discovery**命令后，最大数据段现在为1460。此外，show **tcp命**令输出指示**路径mtu能**力紧**接在max data segment语**句前。出站接口的MTU为1500字节。MTU等于1500字节减去40字节的IP信息为1460字节。DLSw将再使用16个字节。因此，在一个TCP帧中最多可以传输1444字节的LLC2或SDLC帧。

```
havoc#show tcp

Stand-alone TCP connection to host 10.1.1.1
```

```
Connection state is ESTAB, I/O status: 1, unread input bytes: 0
Local host: 30.1.1.1, Local port: 11045
Foreign host: 10.1.1.1, Foreign port: 2065

Enqueued packets for retransmit: 0, input: 0  mis-ordered: 0 (0 bytes)

TCP driver queue size 0, flow controlled FALSE

Event Timers (current time is 0xA6DA58):
Timer          Starts    Wakeups            Next
Retrans             4          0             0x0
TimeWait            0          0             0x0
AckHold             1          0             0x0
SendWnd             0          0             0x0
KeepAlive           0          0             0x0
GiveUp              3          0             0x0
PmtuAger            0          0             0x0
DeadWait            0          0             0x0

iss: 3423657490  snduna: 3423657976  sndnxt: 3423657976     sndwnd:  19995
irs:  649085675  rcvnxt:  649085688  rcvwnd:      20468  delrcvwnd:     12

SRTT: 124 ms, RTTO: 1405 ms, RTV: 1281 ms, KRTT: 0 ms
minRTT: 24 ms, maxRTT: 300 ms, ACK hold: 200 ms
Flags: higher precedence, retransmission timeout, path mtu capable

Datagrams (max data segment is 1460 bytes):
Rcvd: 5 (out of order: 0), with data: 1, total data bytes: 12
Sent: 6 (retransmit: 0, fastretransmit: 0), with data: 3, total data bytes: 485
```

相关信息

- Compatible Systems技术说明：VPN的IP分段和MTU路径发现
- 技术支持 - Cisco Systems