IP路徑MTU探索和DLSw

目錄

簡介

IBM通訊協定套件、DLSw、STUN和BSTUN會建立路由器之間的IP作業階段管道。由於TCP的可靠性，它通常用作路由器之間的傳輸方法。本檔案將提供TCP動態探索作業階段管道上可使用的最大MTU的能力方面的資訊，此能力可以最小化分段並最大化效率。

開始之前

慣例

如需文件慣例的詳細資訊，請參閱思科技術提示慣例。

必要條件

本文件沒有特定先決條件。

採用元件

本文件所述內容不限於特定軟體和硬體版本。

本文中的資訊是根據特定實驗室環境內的裝置所建立。文中使用到的所有裝置皆從已清除（預設）的組態來啟動。如果您在即時網路中工作，請確保在使用任何命令之前瞭解其潛在影響。

背景資訊

如RFC 1191所述，路徑MTU發現(PMTD)指定IP資料包的預設位元組大小為576。幀的IP和TCP部分佔用40位元組，剩餘536位元組作為資料負載。此空間稱為最大段大小或MSS。RFC1191第3.1節指定可交涉的較大MSS，這正是發出**ip tcp path-mtu-discovery** 指令在Cisco路由器上執行的操

作。當配置此命令並啟動TCP會話時，路由器外部的SYN資料包包含指定更大的MSS的TCP選項。此較大的MSS是傳出介面的MTU減去40位元組。如果傳出介面的MTU為1500位元組，則通告的MSS為1460。如果傳出介面的MTU較大（例如具有4096位元組MTU的訊框中繼），則MSS將為4096位元組減去40位元組的IP資訊，並將顯示在**show tcp指令**輸出中（最大資料區段為4056位元組）。

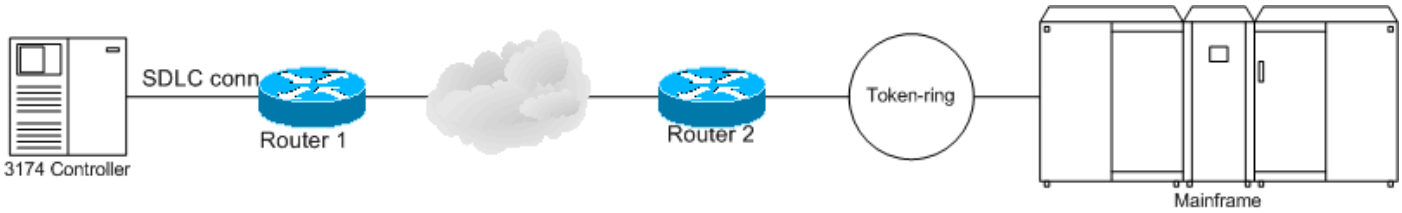
在路由器上配置PMTD對已建立自/至路由器的現有TCP會話沒有任何影響。PMTD已匯入11.3.5T IOS層級，在後續的IOS版本中，它成為一個可選命令。在IOS 11.3(5)T之前，它預設處於開啟狀態。此外，當IP位址位於同一子網中時，PMTD會自動執行，表示這些位址直接連線到同一媒體上。

兩台路由器都必須配置，PMTD才能正常工作。當兩台路由器都配置好時，從一台路由器到另一台路由器的SYN包含通告更高MSS的可選TCP值。傳回的SYN會通告更高的MSS值。因此，兩端向對方通告它們可以接受更大的MSS。如果只有一個路由器（路由器1）配置了**ip tcp path-mtu-discovery** 命令，它將通告更大的MSS，因此，路由器2可以向路由器1傳送1460位元組的幀。Router 2永遠不會通告較大的MSS，因此Router 1會鎖定以傳送預設值。

含PMTD的DLSw

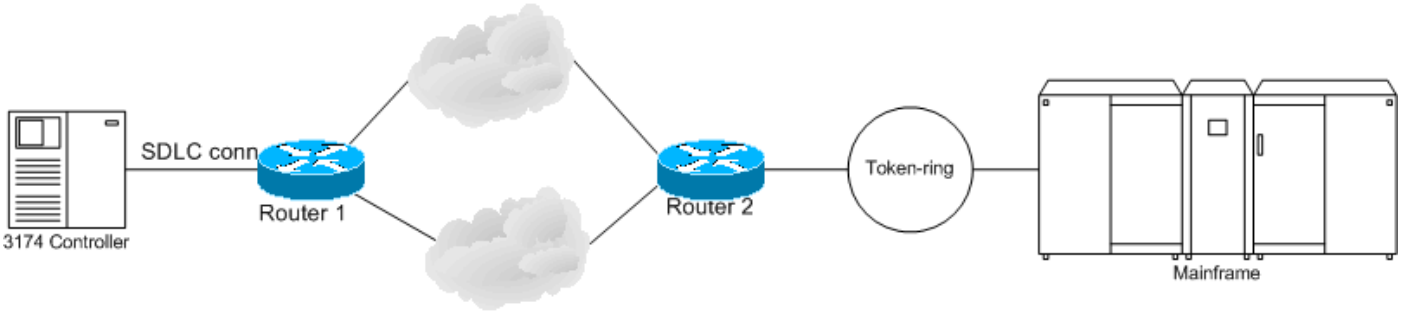
在IBM套件中，DLSw、STUN和BSTUN可以承擔在路由器之間通過TCP會話傳輸大量資料的任務。實作PMTD可能會非常重要和受益，尤其是考慮到在11.2和先前IOS層級中預設啟用了該功能。根據RFC，最大幀預設為576位元組，減去IP/TCP封裝的40位元組。DLSw使用另一個16位元組進行封裝。使用預設MSS傳輸的實際資料為520位元組。DLSw還能夠將兩個不同的邏輯連結控制2(LLC2)封包承載到一個TCP訊框內。如果兩個工作站分別傳送一個LLC2幀，則DLSw可以在一個幀中將兩個LLC2幀傳送到DLSw遠端對等路由器。TCP驅動程式通過具有更大的MSS可以適應這種傳送模式。以下三個主要案例說明**path-mtu-discovery** 命令的值。

場景1 — 不需要的開銷



SDLC裝置通常配置為每個幀中的資料最大為265或521位元組。如果值為521,3174將521位元組的SDLC幀傳送到路由器1，則路由器1將製作兩個TCP幀，以將此幀傳輸到DLSW對等路由器2。第一個幀將包含520位元組的資料、16位元組的DLSw資訊和40位元組的IP資訊，總計576位元組。第二個資料包將包含1位元組的資料、16位元組的DLSw資訊和40位元組的IP資訊。使用PMTD並假設1500位元組的MTU取得1460 MSS時，路由器2告知路由器1路由器2可接收1460位元組的資料。Router 1會將521位元組的SDLC資料以16位元組的DLSw資訊和40位元組的IP資訊傳送到路由器2。由於DLSw是進程交換事件，通過使用PMTD，處理此SDLC幀的CPU利用率減半。此外，路由器2不必等待第二個資料包組裝LLC2幀。啟用PMTD後，路由器2會收到整個封包，然後可以從封包中移除IP和DLSW資訊，並將其毫不延遲地傳送到3745。

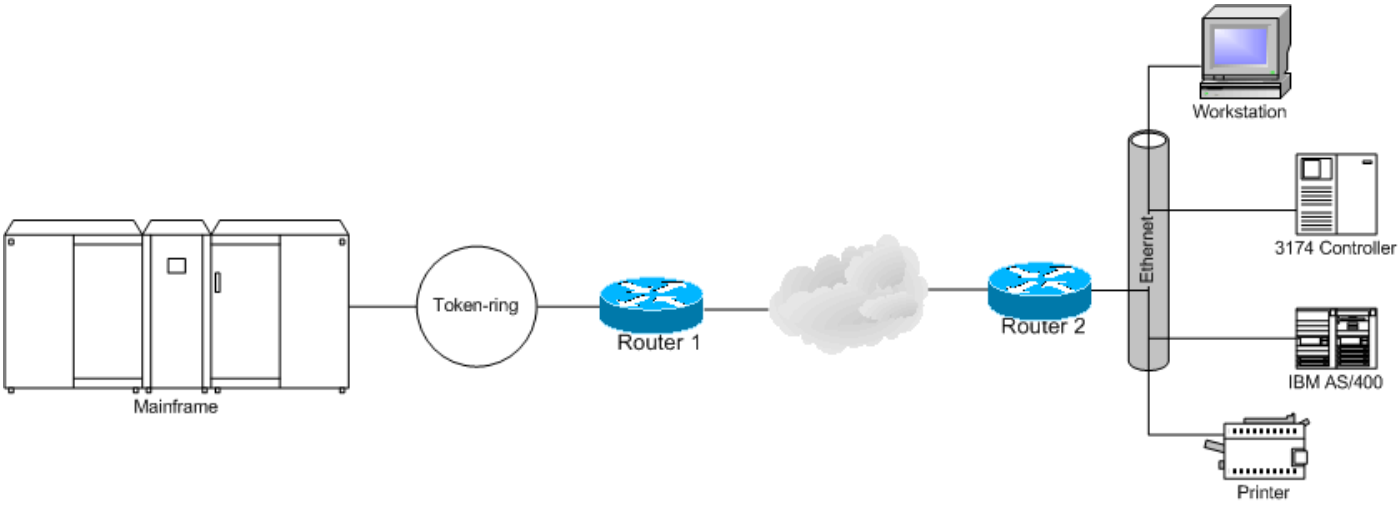
案例2 — 來自無序封包的延遲



在此方案中，有兩個IP雲可用於負載平衡或冗餘的同等度量。不啟用PMTD可能會嚴重減慢DLSw。如果不啟用PMTD，Router 1必須將521位元組的訊框組裝成兩個TCP封包 — 一個具有520位元組的資料，另一個具有1位元組的資料。如果第一個資料包經過頂級IP雲，那麼如果第二個資料包通過較低的、等價的IP雲傳送，則它將在第二個資料包之後到達。這樣會生成所謂的無序資料包。IP/TCP協定的固有功能是能夠管理此問題。亂序資料包儲存在記憶體中，等待整個資料流到達後再重新組裝。亂序資料包並不罕見，但是，應嘗試將其降至最低，因為此事件會利用記憶體和CPU資源。大量亂序可能導致在TCP級別出現大量延遲。如果第3層/DLSw作業階段延遲，則透過此DLSw傳輸的LLC2/SDLC作業階段隨後將受影響。如果在此案例中啟用PMTD，則會透過任一IP網雲在一個TCP訊框內傳送一個521位元組的訊框。接收路由器只需要緩衝和解封一個TCP幀。

PMTD與SNA環境中通告給終端站的最大訊框沒有關係。其中包括權杖環上的路由資訊欄位(RIF)中的最大訊框(LF)。PMTD嚴格規定可以封裝到一個TCP訊框中的資料量。LLC2和SDLC沒有功能分段資料包，但是IP/TCP有。大SNA幀在封裝到TCP中時可以分段。此資料在遠端DLSw路由器上解除封裝，並再次顯示為非分段SNA資料。

場景3 — 更快的LLC2連線和吞吐量



在此案例中，3174和工作站通過3745 TIC與大型機建立會話，如果兩台裝置都傳送了目標為主機的資料，則TCP可能可以將兩個LLC2幀放入一個資料包中。如果沒有PMTD，則當兩個訊框的合併資料為521位元組或以上時，則無法如此操作。在這種情況下，TCP軟體將需要單獨傳送每個封包。例如，如果3174和工作站大約在同一時間傳送一個幀，並且這些資料包中的每個資料包都有400位元組的資料，則路由器會接收並緩衝每個幀。路由器現在必須將這些400位元組資料流封裝成單獨的TCP封包，才能傳輸到對等路由器。

啟用PMTD並假設MSS為1460時，路由器會接收並緩衝兩個LLC2封包。現在能夠將兩者封裝到一個資料包中。這個TCP資料包將包含40位元組的IP資訊、用於第一個DLSw電路配對的16位元組的DLSw資訊、400位元組的資料、用於第二個DLSw電路配對的另外16位元組的資料以及其他400位元組的資料。此特定情況使用兩台裝置，因此使用兩條DLSw電路。PMTD允許DLSw更有效地擴展到更高數量的DLSw電路。許多分支中心網路需要數百個遠端站點（每個站點都有一或兩個SNA裝置），以對等方式連線到中央站點路由器，該路由器連線到OSA或FEP，以提供對主機應用程式的

訪問。PMTD使TCP和DLSw能夠擴充為較大的需求，而不會過度利用路由器CPU和記憶體資源，同時提供更快的傳輸。

註：在12.1(5)T後期發現並解決12.2(5)T中的軟體錯誤，其中PMTD無法通過虛擬專用網路(VPN)隧道工作。此軟體缺陷是CSCdt49552(僅限註冊客戶)。

檢驗DLSW的PMTD

發出**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
```

此顯示標識為DLSw TCP會話，因為TCP會話中的一個埠是2065。輸出底部附近的最大資料段為536位元組。此值表示10.1.1.1的遠端DLSw對等路由器未設定**ip tcp path-mtu-discovery** 指令。536位元組的值已佔IP訊框中的40位元組IP資訊。此536位元組值不考慮將新增到承載SNA流量的TCP封包中的16位元組的DLSw資訊。

配置**ip tcp path-mtu-discovery** 命令後，最大資料段現在為1460。此外，**show tcp**命令輸出會指示**path mtu capable**，緊接在**max data segment**語句之前。傳出介面的MTU為1500位元組。MTU等於1500位元組，減去40位元組的IP資訊為1460位元組。DLSw將使用另一個16位元組。因此，在一個TCP幀中最多可以傳輸LLC2或SDLC的1444位元組幀。

```
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
```

相關資訊

- 相容系統技術說明：使用VPN的IP分段和MTU路徑探索
- 技術支援 - Cisco Systems