IP パス MTU ディスカバリと DLSw

内容

概要

IBM のプロトコル スイート、DLSw、STUN および BSTUN は、あるルータから別のルータへの IP セッション パイプを確立します。TCP はその信頼性から、ルータ間の転送方式として一般に使用されます。このドキュメントは、フラグメンテーションを最小にして、効率を最大化する、セッション パイプで使用できる最大 MTU を動的に検出する TCP の機能について説明します。

はじめに

表記法

ドキュメント表記の詳細は、『シスコ テクニカル ティップスの表記法』を参照してください。

前提条件

このドキュメントに関しては個別の前提条件はありません。

使用するコンポーネント

このドキュメントの内容は、特定のソフトウェアやハードウェアのバージョンに限定されるものではありません。

このマニュアルの情報は、特定のラボ環境に置かれたデバイスに基づいて作成されました。このドキュメントで使用するすべてのデバイスは、初期（デフォルト）設定の状態から起動しています。実稼動中のネットワークで作業をしている場合、実際にコマンドを使用する前に、その潜在的な影響について理解しておく必要があります。

背景説明

Path MTU Discovery(PMTD)は、RFC 1191で説明されているように、IPパケットのデフォルトバイトサイズが576であることを指定します。フレームのIPおよびTCP部分は、データペイロードとして536バイトを残して40バイトになります。この領域は、最大セグメント サイズまたは MSS と呼ばれています。RFC 1191 のセクション 3.1 では、大きい MSS はネゴシエーションが必要とあり、シスコ ルータで **ip tcp path-mtu-discovery** コマンドを発行する理由はまさにここにあります。このコマンドを設定して TCP セッションを開始すると、ルータ内から送信される SYN パケットには大きい MSS を指定する TCP オプションが含まれます。この大きい MSS は、発信インターフェイスから 40 バイトを引いた MTU です。発信インターフェイスの MTU が 1500 バイトの場合、アドバタイズされる MSS は 1460 バイトになります。発信インターフェイスの MTU がそれよりも大きい場合、たとえば MTU が 4096 バイトのフレーム リレーであれば、MSS は 4096 バイトから IP 情報の 40 バイトを引いた値バイト数として **show tcp** コマンドの出力結果に表示されます (データ セグメントの最大バイト数は 4056 バイト) 。

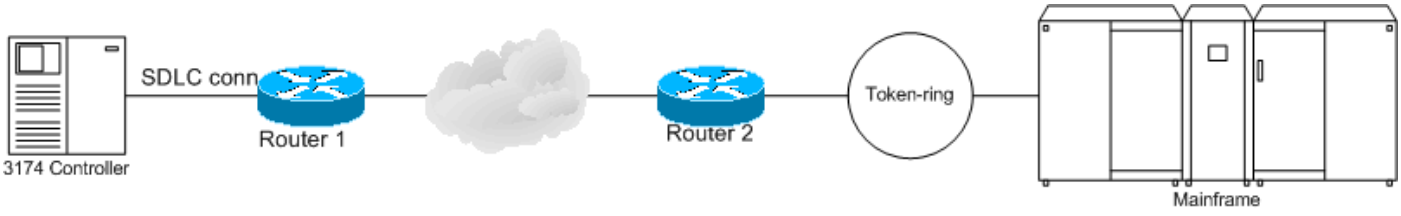
ルータに PMTD を設定しても、ルータ間ですでに確立された TCP セッションには影響ありません。PMTD は 11.3.5T IOS レベルと続く IOS リリースに実装され、オプションのコマンドになりました。IOS 11.3(5)T 以前ではデフォルトで使用できました。また、PMTD は IP アドレスが同じサブネットにあるときは、同じメディアに直接接続されていることから、自動実行されます。

PMTD をルータで正常稼働させるには、両ルータで設定が必要です。両ルータを設定すると、一方のルータからもう一方のルータに流れる SYN には大きい MSS をアドバタイズするオプションの TCP 値が含まれるようになります。返ってきた SYN は、大きい MSS 値をアドバタイズします。このようにして、両ルータは互いに大きな MSS を受け取れることをアドバタイズします。片方のルータ (ルータ 1) のみに **ip tcp path-mtu-discovery** コマンドがある場合、ルータ 1 は大きな MSS をアドバタイズし、ルータ 2 はルータ 1 へ 1460 バイト フレームを送信できるようになります。ルータ 2 は大きな MSS をアドバタイズしないので、ルータ 1 が送信できる値はデフォルト値に固定されます。

PMTD を使用した DLSw

IBM プロトコル スイートの DLSw、STUN、BSTUN では、ルータ間の TCP セッションに大容量のデータを乗せるよう設定できます。特に 11.2 および以前の IOS レベルではデフォルトで有効になっていたことを考えると、PMTD を実装することは重要で、かつ非常に有益なことです。RFC にあるとおり、デフォルトの最大フレームは 576 バイトで、TCP/IP カプセル化は 40 バイトが引かれた値になります。DLSw はカプセル化で別の 16 バイトを使用します。デフォルトの MSS を使用して転送される実際のデータは 520 バイトです。このほか、DLSw は 1 つの TCP フレームで異なる 2 つの論理リンク制御 2 (LLC2) パケットを運ぶことができます。2 台のワークステーションがそれぞれ LLC2 フレームを送信した場合、DLSw は両方の LLC2 フレームを 1 つのフレームとして DLSw リモート ピアに渡すことができます。TCP ドライバがこのようなピギーバック手法をとれるのは、MSS が大きいからです。次の主要な 3 つのシナリオに、**path-mtu-discovery** コマンドのメリットを示します。

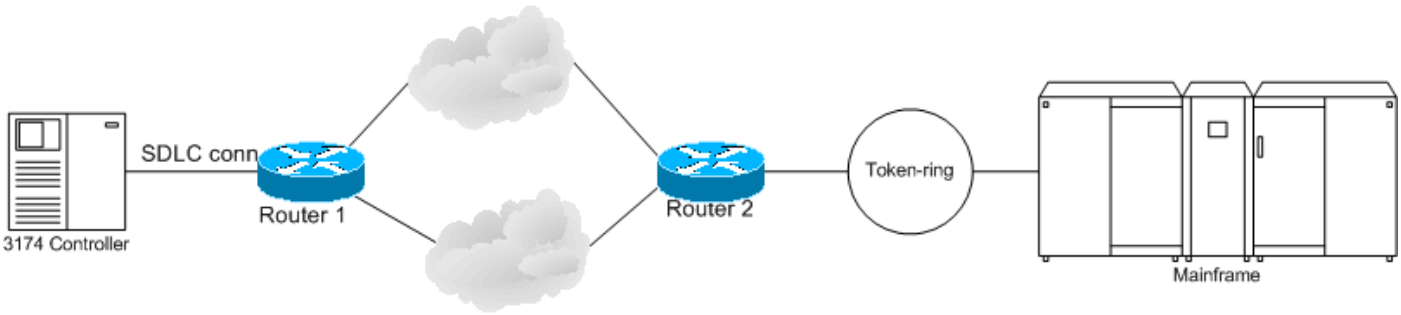
シナリオ 1 : 不要なオーバーヘッド



SDLC デバイスでは通常、各フレーム内のデータが最大 265 バイトまたは 521 バイトに設定されています。値が521で、3174がルータ1に521バイトのSDLCフレームを送信すると、ルータ1は2つのTCPフレームを作成して、これをDLSWピアのルータ2に転送します。最初のフレームには

520バイト、DLSw情報40バイト0バイトがが0合計576バイト。2 つめのパケットには、1 バイトのデータ、16 バイトの DLSw 情報、40 バイトの IP 情報が含まれています。PMTD が使用されており、1460 バイトの MSS を取得するのに 1500 バイトの MTU が想定されているとき、ルータ 2 はルータ 1 に対して 1460 バイトのデータを受信できると伝えます。これを受けて、ルータ 1 は 521 バイトの SDLC データすべてを、16 バイトの DLSw 情報と 40 バイトの IP 情報を含む 1 つのパケットとしてルータ 2 に送信します。DLSw はプロセス切り替え型イベントであることから、PMTD を使ってこの 1 つの SDLC フレームを処理するための CPU 使用率を半分にします。それに加えて、ルータ 2 は LLC2 フレームを構成するのに 2 つめのパケットを待つ必要がなくなります。PMTD が有効であれば、ルータ 2 はパケット全体を受信でき、パケットから IP と DLSw の情報を削除してから遅延なく 3745 コントローラへ送信できます。

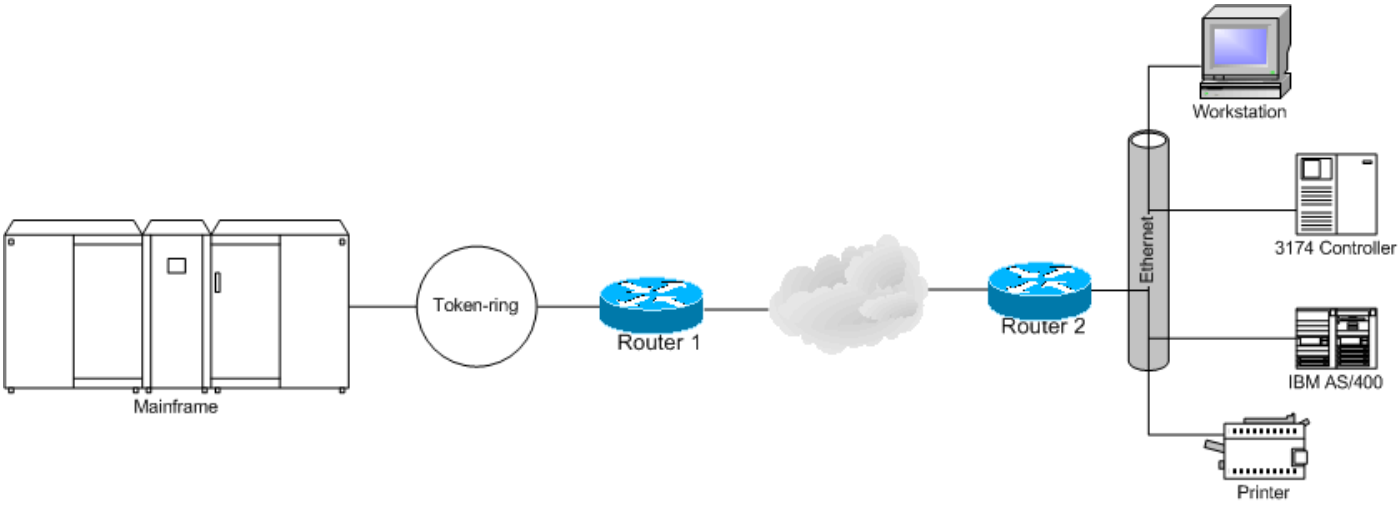
シナリオ 2：Out-of-Order パケットによる遅延



このシナリオでは、ロードバランシングまたは冗長性のいずれかのために同じメトリクスが設定された 2 つの IP クラウドがあります。このとき、PMTD が有効でないと DLSw は著しく遅くなります。PMTD が有効ではない場合、ルータ 1 は 521 バイトのフレームを 2 つの TCP パケット（1 つはデータが 520 バイトで、もう 1 つはデータが 1 バイト）を構成しなければなりません。1 つめのパケットが上の IP クラウドを通過し、1 つめのパケットが同じ処理性能を持つ下の IP クラウドを介して送信された場合、1 つめのパケットの到着は大幅に遅れる可能性があります。その結果、Out-of-Order パケットと呼ばれる減少が発生します。TCP/IP には、この問題を制御する機能が備わっています。Out-of-Order パケットはストリーム全体が到着するまでメモリに格納され、そのあとで再構成されます。Out-of-Order パケットはよくあることですが、メモリや CPU のリソースを消費することから極力最小限に抑えるようにするべきです。Out-of-Order パケットが大量にあると、TCP レベルで大幅な遅延が発生します。レイヤ 3/DLSw セッションが遅延すると、DLSw 上で転送される LLC2/SDLC セッション DLSw も続いて遅延します。このシナリオで PMTD が有効であれば、521 バイトのフレームは 1 つの TCP フレームとしていずれかの IP クラウドを介して送信されます。受信側のルータは 1 つの TCP フレームをバッファおよびカプセル化するだけで済みます。

PMTD は、SNA 環境でエンド ステーション間をアドバタイズされる最大フレームとは関係しません。これには、トークン リングのルーティング情報フィールド（RIF）内の Largest Frame（LF）が含まれます。PMTD は、1 つの TCP フレームにカプセル化できるデータの容量を厳しく指定します。LLC2 と SDLC にはパケットをフラグメント化する機能はありませんが、TCP/IP にはあります。大きな SNA フレームは、TCP にカプセル化されるため、セグメント化できます。このデータはリモート DLSw ルータでカプセル化が解除され、再びフラグメント化されていない SNA データになります。

シナリオ 3：LLC2 の接続およびスループットの高速化



このシナリオでは、3174 コントローラとワークステーションは 3745 TIC を介してメインフレームにセッションを張ります。このとき、両デバイスがホスト宛てにデータを送信すると、TCP 側で LLC2 フレームを 1 つのパケットにまとめることができます。しかし、PMTD が有効でないと、2 つのフレームが 521 バイト以上の場合、まとめることができません。その場合、TCP ソフトウェアはパケットをそれぞれ送信しなければなりません。たとえば 3174 コントローラとワークステーションがほぼ同じタイミングでフレームを送信し、そのパケットに 400 バイトのデータが含まれている場合、ルータはフレームそれぞれを受信してバッファします。続いて、ルータは 400 バイトのデータ ストリームをそれぞれ別の TCP パケットにカプセル化してから、ピアに転送することになります。

PMTD が有効で MSS が 1460 バイトに想定されていれば、ルータは 2 つの LLC2 パケットを受信しバッファします。その後、1 つのパケットにカプセル化することができます。1 つの TCP パケットには、40 バイトの IP 情報、16 バイトの DLSw 情報を含む最初の DLSw 回線ペア、400 バイトのデータ、16 バイトの DLSw 情報を含む 2 つめの DLSw 回線ペア、さらに 400 バイトのデータがまとめられます。そのため、今回のシナリオではデバイス 2 台と DLSw 回線 2 つを使用しています。PMTD があることで、DLSw はより効率的に DLSw 回線数を拡張できます。スポークとハブのネットワークの多くでは何百ものリモート サイトが必要で、サイトにはそれぞれ 1 台または 2 台の SNA デバイスがあり、中心となるサイトのルータにピアリングして OSA または FEP 経由でホスト アプリケーションに接続します。PMTD は、ルータの CPU やメモリを過剰に使用することなく、かつ高速転送を実現し、大規模な要件に対応できるよう TCP や DLSw を拡張することができます。

注：12.1(5)Tの後半に見つかったソフトウェアのバグが12.2(5)Tで解決されています。この場合、PMTDはバーチャルプライベートネットワーク(VPN)トンネルで動作していません。このソフトウェア欠陥の Cisco Bug ID は CSCdt49552（登録ユーザのみ）です。

DLSw の PMTD の確認

show tcp コマンドを発行します。

```
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
```

この出力結果では、TCP セッションの 1 つのポートが 2065 であるため、DLSw TCP セッションとして認識されています。出力結果の一番下近くに、最大データ セグメントが 536 バイトであると表示されています。この値は、10.1.1.1 のリモート DLSw ピア ルータに **ip tcp path-mtu-discovery** コマンドが設定されていないことを示しています。536 バイトの値は、IP フレーム内の 40 バイトの IP 情報がすでに含まれています。この 536 バイトの値には、SNA トラフィックを運ぶ TCP パケットに追加される 16 バイトの DLSw 情報は含まれていません。

ip tcp path-mtu-discovery コマンドを設定すると、最大データ セグメントは 1460 になります。さらに、**show tcp** コマンドの出力結果を見ると、**max data segment** の表記の直前に **path mtu capable** と表示されています。発信インターフェイスの MTU は 1500 バイトです。MTU の 1500 バイトから IP 情報の 40 バイトを引くと、1460 バイトになります。DLSw は別途 16 バイト使用します。つまり、1 つの TCP フレームで 1444 バイト フレームの LLC2 または SDLC を送信することになります。

```
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
```

関連情報

- 互換システムに関するテクニカル ノート：VPN における IP フラグメンテーションおよび MTU パス ディスカバリ
- テクニカルサポート - Cisco Systems