Verify Path MTU Discovery on Cisco IOS XR and BGP

Contents

Introduction

Background Information

TCP PMTUD and TCP MSS

Scenarios – TCP PMTUD Disabled

- Use Default MTU Values
- Use Non-Default MTU Value – Active TCP Peer
- Use Non-Default MTU Value – Passive TCP Peer
- Use TCP Options – XR Active
- Use TCP Options – XR Passive
- TCP Peers not Directly Connected
- TCP Peers not Directly Connected – Use TCP Options (MD5)
- TCP Peers not Directly Connected – Path Segment has Lower IP MTU

Scenarios – TCP PMTUD Enabled

- Enable PMTUD
- PMTUD – Path Segment has Lower IP MTU
- PMTUD – TCP Options (MD5)
- PMTUD – Blackhole Detection

Introduction

This document describes the Transmission Control Protocol (TCP) Path Maximum Transmission Unit (MTU) Discovery (PMTUD) on Cisco IOS® XR devices.

Background Information

PMTUD mechanism attempts to determine the largest Internet Protocol (IP) packet size that does not require fragmentation anywhere along the path between two hosts. The value established is designated Path MTU and is equal to a minimum of the MTU values across each hop. If you consider Path MTU when you transmit information, it allows you to make the most out of network capacity and avoid fragmentation and transmission efficiency. PMTUD mechanics and implementation is introduced across a diverse set of scenarios with the use of Border Gateway Protocol (BGP) as client protocol which gradually reveals PMTUD behavior.

TCP PMTUD and TCP MSS

TCP leverages PMTUD outcome in order to influence local Maximum Segment Size (MSS), which means, it dynamically adapts to discovered Path MTU. Hence, before you move on to PMTUD, you can quickly review TCP Maximum Segment Size (MSS) and understand what it means and its purpose.

As per MSS original definition from RFC879: The definition of the MSS option can be stated: The maximum number of data octets that can be received by the sender of this TCP option in TCP segments with no TCP header options transmitted in IP datagrams with no IP header options.

To clarify some aspects and provide advice to implementers, RFC6691 highlights how MSS value must be calculated:

When you calculate the value to put in the TCP MSS option, the MTU value must be decreased by only the size of the fixed IP and TCP headers and must not be decreased to account for any possible IP or TCP options; conversely, the sender MUST reduce the TCP data length to account for any IP or TCP options that it is including in the packets that it sends.

A more elaborated definition of MSS can be extracted from the Routing Configuration Guide for Cisco ASR 9000 Series Routers, IOS XR Release 7.9.x:

MSS is the largest amount of data that a computer or a communication device can receive in a single, unfragmented TCP segment. All TCP sessions are bounded by a limit on the number of bytes that can be transported in a single packet; this limit is MSS. TCP breaks up packets into chunks in a transmit queue before passing packets down to the IP layer.

The TCP MSS value is dependent on the MTU of an interface, which is the maximum length of data that can be transmitted by a protocol at one instance. The maximum TCP packet length is determined by both the MTU of the outbound interface on the source device and the MSS announced by the destination device during the TCP setup process. The closer the MSS is to the MTU, the more efficient is the transfer of BGP messages. Each direction of data flow can use a different MSS value.

What would then be the value that TCP must consider for MSS on a given TCP session? And how is it calculated?

For the default values as per RFC879, you have: Hosts must not send datagrams larger than 576 octets unless they have specific knowledge that the destination host is prepared to accept larger datagrams. THE TCP MAXIMUM SEGMENT SIZE IS THE IP MAXIMUM DATAGRAM SIZE MINUS FORTY.

The default IP Maximum Datagram Size is 576.

The default TCP Maximum Segment Size is 536.

This takes into consideration an IP MTU value of 576 bytes. But if you ignore the actual IP MTU value then the TCP MSS calculation can be summarized as follows:

- Active Peer – calculates and sends initial MSS with SYN packet.

```
MSS = IPMTU - sizeof(minimum TCPHDR) - sizeof(minimum IPHDR)
Where,
sizeof(minimum TCPHDR) = 20 bytes.
sizeof(minimum IPHDR) = 20 bytes.
```

- Passive Peer – calculates initial MSS, compares with received MSS from Active Peer, and sends SYN, ACK with the lower of these MSS values.

```
MIN[IPMTU - sizeof(minimum TCPHDR) - sizeof(minimum IPHDR) , Received MSS value]
Where,
sizeof(minimum TCPHDR) = 20 bytes.
sizeof(minimum IPHDR) = 20 bytes.

Received MSS value = MSS value received with Active Peer TCP SYN.
```

There is no negotiation with regard to the MSS option value. Each node determines its own value and announces the same upon TCP session establishment. It becomes clear that if the IP MTU value considered for MSS calculation can be derived from PMTUD then the MSS value can be adapted to the most effective value for a given Path MTU. Cisco IOS XR behaviour has some specifics with regards to MSS calculation and PMTUD role summarized here.

PMTUD is disabled by default on Cisco IOS XR:

- Local initial MSS calculation considers IP MTU as per this:
  - If directly connected peers – consider egress interface IP MTU.
  - If non-directly connected peers – consider IP MTU of 1280 bytes.
  - MSS value is influenced by configured TCP options.

When PMTUD is enabled on Cisco IOS XR:

- Local initial MSS calculation considers IP MTU as per this:
  - Irrespective of directly/non-directly connected peers – consider egress interface IP MTU.
  - MSS value is influenced by configured TCP options.

There are additional details on PMTUD mechanics and implementation that need to be taken into account and which this document introduces through practical examples summarized in the next table. This table also presents Active and Passive TCP peers IP MTU as well as selected MSS values for each scenario considered.

PMTUD	Scenarios	ACTIVE IP MTU	PASSIVE IP MTU	MSS
Disabled	Using default MTU values	1500	1500	1460
	Using non-default MTU value – Active TCP peer	4460	1500	1460
	Using non-default MTU value – Passive TCP peer	1500	4460	1460
	Using TCP Options (MD5) – XR Active	1500	1500	1436
	Using TCP Options (MD5) – XR Passive	1500	1500	1460
	TCP peers not directly connected	1500	1500	1240
	TCP peers not directly connected – Using TCP Options (MD5)	1500	1500	1216
Enabled	Enabling TCP PMTUD	1500	1500	1460
	PMTUD in action – Path segment has lower MTU	1500	1500	1460
	PMTUD in action – TCP Options (MD5)	1500	1500	1436

Scenarios – TCP PMTUD Disabled

Use Default MTU Values

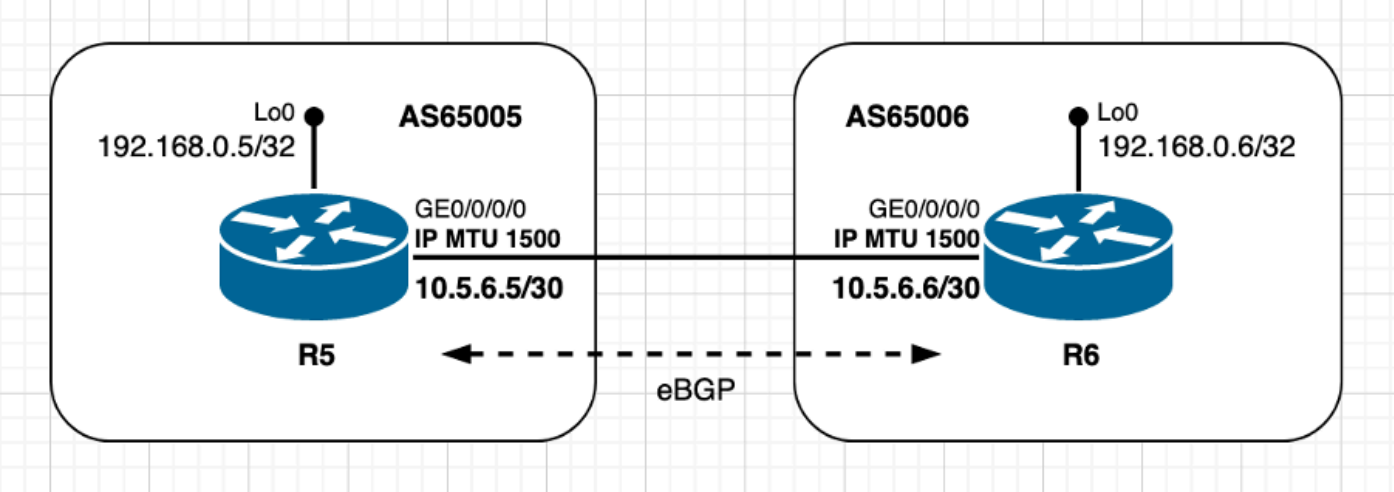


Image 2.1. With default MTU Values

In the case of the eBGP peers shown in Image 2.1 R6 manages TCP connection, this means, it plays the active role and starts the TCP session with R5 on destination port 179. Peers are directly connected, and both use the default IP MTU values on respective interfaces. Based on the information shared in the start of this document, the MSS calculation in this scenario can be summarized as follows:

- Both nodes use a default IP MTU of 1500 bytes
- TCP path MTU discovery is disabled by default
- TCP peers are directly connected
  - R6 manages the BGP connection
  - R6 sends SYN with MSS of 1460 bytes
    - 1500 (Interface IP MTU) - 20 (minTCP_H) - 20 (minIP_H)
  - R5 sends SYN, ACK with MSS of 1460 bytes
    - Sends the lower of [Received MSS; Local initial MSS]
    - Received MSS 1460 bytes; Local initial MSS 1460 bytes
  - The lowest MSS value is used on both peers

TCP session details as seen on R6 – ACTIVE:

<#root>

```
! - As seen on R6 - ACTIVE

RP/0/0/CPU0:R6#show interfaces gigabitEthernet 0/0/0/0
Fri Jan  8 09:35:48.553 UTC
GigabitEthernet0/0/0/0 is up, line protocol is up
Interface state transitions: 1
Hardware is GigabitEthernet, address is fa16.3e85.3dc2 (bia fa16.3e85.3dc2)
Internet address is 10.5.6.6/30

MTU 1514 bytes

, BW 1000000 Kbit (Max: 1000000 Kbit)
<snip>

RP/0/0/CPU0:R6#show tcp brief
Fri Jan  8 09:36:22.491 UTC
PCB      VRF-ID     Recv-Q Send-Q Local Address          Foreign Address        State
<snip>
0x121649fc 0x60000000      0      0  10.5.6.6:24454         10.5.6.5:179           ESTAB
<snip>
```

```
RP/0/0/CPU0:R6#show tcp detail pcb 0x121649fc
Fri Jan  8 09:37:00.888 UTC
==============================================================
Connection state is ESTAB, I/O status: 0, socket status: 0
Established at Fri Jan  8 09:28:28 2021

PCB 0x121649fc, SO 0x121561b8, TCPCB 0x12156f64, vrfid 0x60000000,
Pak Prio: Medium, TOS: 192, TTL: 1, Hash index: 78
Local host: 10.5.6.6, Local port: 24454 (Local App PID: 1011918)
Foreign host: 10.5.6.5, Foreign port: 179

Current send queue size in bytes: 0 (max 24576)
Current receive queue size in bytes: 0 (max 32768)  mis-ordered: 0 bytes
Current receive queue size in packets: 0 (max 0)

Timer          Starts    Wakeups         Next(msec)
Retrans            13          1                 0
SendWnd             0          0                 0
TimeWait            0          0                 0
AckHold            10          2                 0
KeepAlive           1          0                 0
PmtuAger            0          0                 0
GiveUp              0          0                 0
Throttle            0          0                 0

iss: 3757770712  snduna: 3757770960  sndnxt: 3757770960
sndmax: 3757770960  sndwnd: 32574       sndcwnd: 4380
irs: 1072103647  rcvnxt: 1072103895  rcvwnd: 32593   rcvadv: 1072136488

SRTT: 155 ms,  RTTO: 540 ms,  RTV: 385 ms,  KRTT: 0 ms
minRTT: 9 ms,  maxRTT: 229 ms

ACK hold time: 200 ms, Keepalive time: 0 sec, SYN waittime: 30 sec
Giveup time: 0 ms, Retransmission retries: 0, Retransmit forever: FALSE
Connect retries remaining: 30, connect retry interval: 50 secs

State flags: none
Feature flags: Win Scale, Nagle
Request flags: Win Scale


Datagrams (in bytes): MSS 1460, peer MSS 1460, min MSS 1460, max MSS 1460



Window scales: rcv 0, snd 0, request rcv 0, request snd 0
Timestamp option: recent 0, recent age 0, last ACK sent 0
Sack blocks {start, end}: none
Sack holes {start, end, dups, rxmit}: none

Socket options: SO_REUSEADDR, SO_REUSEPORT, SO_NBIO
Socket states: SS_ISCONNECTED, SS_PRIV
Socket receive buffer states: SB_DEL_WAKEUP
Socket send buffer states: SB_DEL_WAKEUP
Socket receive buffer: Low/High watermark 1/32768
Socket send buffer   : Low/High watermark 2048/24576, Notify threshold 0

PDU information:
#PDU's in buffer: 0
FIB Lookup Cache:  IFH: 0x40  PD ctx: size: 0  data:
Num Labels: 0  Label Stack:
```

```
RP/0/0/CPU0:R6
```

TCP session details as seen on R5 – PASSIVE:

<#root>

```
! - As seen on R5 - PASSIVE

RP/0/0/CPU0:R5#show interfaces gigabitEthernet 0/0/0/0
Fri Jan  8 09:33:04.564 UTC
GigabitEthernet0/0/0/0 is up, line protocol is up
Interface state transitions: 1
Hardware is GigabitEthernet, address is fa16.3ead.518f (bia fa16.3ead.518f)
Internet address is 10.5.6.5/30
```

MTU 1514 bytes

```
, BW 1000000 Kbit (Max: 1000000 Kbit)
<snip>

RP/0/0/CPU0:R5#show tcp brief
Fri Jan  8 09:33:53.221 UTC
PCB      VRF-ID     Recv-Q Send-Q Local Address          Foreign Address        State
<snip>
0x12155884 0x60000000      0      0  10.5.6.5:179           10.5.6.6:24454         ESTAB
<snip>

RP/0/0/CPU0:R5#show tcp detail pcb 0x12155884
Fri Jan  8 09:34:47.317 UTC
================================================================
Connection state is ESTAB, I/O status: 0, socket status: 0
Established at Fri Jan  8 09:28:29 2021

PCB 0x12155884, SO 0x1215568c, TCPCB 0x12155a54, vrfid 0x60000000,
Pak Prio: Medium, TOS: 192, TTL: 1, Hash index: 78
Local host: 10.5.6.5, Local port: 179 (Local App PID: 1044686)
Foreign host: 10.5.6.6, Foreign port: 24454

Current send queue size in bytes: 0 (max 24576)
Current receive queue size in bytes: 0 (max 32768)  mis-ordered: 0 bytes
Current receive queue size in packets: 0 (max 0)

Timer          Starts    Wakeups         Next(msec)
Retrans             9          0                0
SendWnd             0          0                0
TimeWait            0          0                0
AckHold             9          7                0
KeepAlive           1          0                0
PmtuAger            0          0                0
GiveUp              0          0                0
Throttle            0          0                0

iss: 1072103647  snduna: 1072103857  sndnxt: 1072103857
sndmax: 1072103857  sndwnd: 32631       sndcwnd: 4380
irs: 3757770712  rcvnxt: 3757770922  rcvwnd: 32612   rcvadv: 3757803534

SRTT: 47 ms,  RTTO: 300 ms,  RTV: 170 ms,  KRTT: 0 ms
minRTT: 19 ms,  maxRTT: 219 ms

ACK hold time: 200 ms, Keepalive time: 0 sec, SYN waittime: 30 sec
```

```
Giveup time: 0 ms, Retransmission retries: 0, Retransmit forever: FALSE
Connect retries remaining: 0, connect retry interval: 0 secs

State flags: none
Feature flags: Win Scale, Nagle
Request flags: Win Scale


Datagrams (in bytes): MSS 1460, peer MSS 1460, min MSS 1460, max MSS 1460



Window scales: rcv 0, snd 0, request rcv 0, request snd 0
Timestamp option: recent 0, recent age 0, last ACK sent 0
Sack blocks {start, end}: none
Sack holes {start, end, dups, rxmit}: none

Socket options: SO_REUSEADDR, SO_REUSEPORT, SO_NBIO
Socket states: SS_ISCONNECTED, SS_PRIV
Socket receive buffer states: SB_DEL_WAKEUP
Socket send buffer states: SB_DEL_WAKEUP
Socket receive buffer: Low/High watermark 1/32768
Socket send buffer   : Low/High watermark 2048/24576, Notify threshold 0

PDU information:
#PDU's in buffer: 0
FIB Lookup Cache:  IFH: 0x40  PD ctx: size: 0  data:
Num Labels: 0  Label Stack:

RP/0/0/CPU0:R5#
```

Use Non-Default MTU Value – Active TCP Peer

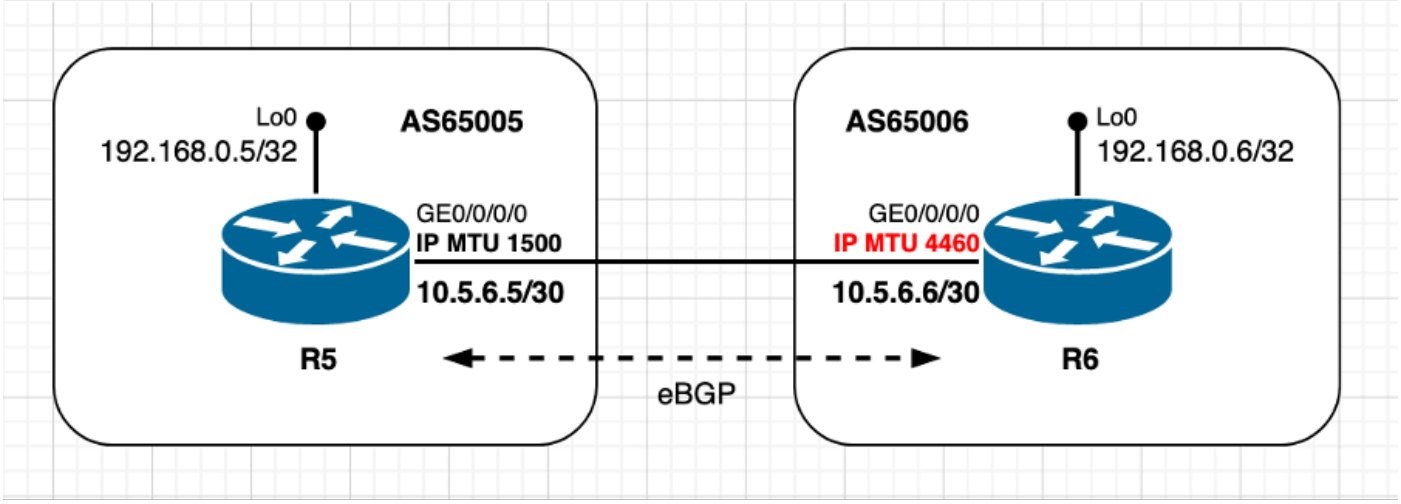


Image 2.2 – ACTIVE peer uses non-default MTU Value

This scenario is essentially the same as the previous, with the only difference that active TCP peer R6 now uses a non-default IP MTU value. Note how initial calculation and decision on MSS value is made by passive TCP peer R5. TCP MSS calculation in this scenario can be summarized as follows:

- R6 uses non-default IP MTU 4460 bytes
- TCP path MTU discovery is disabled by default
- TCP peers are directly connected

- R6 manages BGP connection
- R6 sends SYN with MSS of 4420 bytes
  - 4460 (Interface IP MTU) - 20 (minTCP_H) - 20 (minIP_H)
- R5 send SYN, ACK with MSS of 1460 bytes
  - sends the lower of [Received MSS; Local initial MSS]
  - Received MSS 4420 bytes; Local initial MSS 1460 bytes
- The lowest MSS value is used on both peers

TCP SYN sourced from R6:

<#root>

```
! – TCP SYN sourced from R6

140    1598.150521    10.5.6.6       10.5.6.5       TCP    62     35502 → 179 [SYN] Seq=0 Win=16384 Len=0

MSS=4420

 WS=1

Frame 140: 62 bytes on wire (496 bits), 62 bytes captured (496 bits) on interface 0
Ethernet II, Src: fa:16:3e:85:3d:c2 (fa:16:3e:85:3d:c2), Dst: fa:16:3e:ad:51:8f (fa:16:3e:ad:51:8f)
Internet Protocol Version 4, Src: 10.5.6.6, Dst: 10.5.6.5
Transmission Control Protocol, Src Port: 35502, Dst Port: 179, Seq: 0, Len: 0
    Source Port: 35502
    Destination Port: 179
    [Stream index: 6]
    [TCP Segment Len: 0]
    Sequence number: 0    (relative sequence number)
    Acknowledgment number: 0
    Header Length: 28 bytes
    Flags: 0x002 (SYN)
    Window size value: 16384
    [Calculated window size: 16384]
    Checksum: 0x219d [unverified]
    [Checksum Status: Unverified]
    Urgent pointer: 0
    Options: (8 bytes), Maximum segment size, Window scale, End of Option List (EOL)
        Maximum segment size: 4420 bytes
            Kind: Maximum Segment Size (2)
            Length: 4

MSS Value: 4420


        Window scale: 0 (multiply by 1)
        End of Option List (EOL)
```

TCP SYN, ACK sourced from R5:

<#root>

```
! – TCP SYN, ACK sourced from R5

141    1598.154866    10.5.6.5       10.5.6.6       TCP    62     179 → 35502 [SYN, ACK] Seq=0 Ack=1 Win=16

MSS=1460
```

```
 WS=1

Frame 141: 62 bytes on wire (496 bits), 62 bytes captured (496 bits) on interface 0
Ethernet II, Src: fa:16:3e:ad:51:8f (fa:16:3e:ad:51:8f), Dst: fa:16:3e:85:3d:c2 (fa:16:3e:85:3d:c2)
Internet Protocol Version 4, Src: 10.5.6.5, Dst: 10.5.6.6
Transmission Control Protocol, Src Port: 179, Dst Port: 35502, Seq: 0, Ack: 1, Len: 0
    Source Port: 179
    Destination Port: 35502
    [Stream index: 6]
    [TCP Segment Len: 0]
    Sequence number: 0    (relative sequence number)
    Acknowledgment number: 1    (relative ack number)
    Header Length: 28 bytes
    Flags: 0x012 (SYN, ACK)
    Window size value: 16384
    [Calculated window size: 16384]
    Checksum: 0xe2b4 [unverified]
    [Checksum Status: Unverified]
    Urgent pointer: 0
    Options: (8 bytes), Maximum segment size, Window scale, End of Option List (EOL)
        Maximum segment size: 1460 bytes
            Kind: Maximum Segment Size (2)
            Length: 4
```

MSS Value: 1460

```
        Window scale: 0 (multiply by 1)
        End of Option List (EOL)
```

TCP session details as seen on R6 – ACTIVE:

<#root>

```
! - as seen on R6 – Active

RP/0/0/CPU0:R6#show interfaces gigabitEthernet 0/0/0/0
Fri Jan  8 09:46:54.138 UTC
GigabitEthernet0/0/0/0 is up, line protocol is up
  Interface state transitions: 1
  Hardware is GigabitEthernet, address is fa16.3e85.3dc2 (bia fa16.3e85.3dc2)
  Internet address is 10.5.6.6/30
```

MTU 4474 bytes

```
, BW 1000000 Kbit (Max: 1000000 Kbit)
<snip>

RP/0/0/CPU0:R6#show tcp detail pcb 0x1215761c
Fri Jan  8 09:56:25.819 UTC
================================================================
Connection state is ESTAB, I/O status: 0, socket status: 0
Established at Fri Jan  8 09:51:46 2021

PCB 0x1215761c, SO 0x12156f64, TCPCB 0x1216419c, vrfid 0x60000000,
Pak Prio: Medium, TOS: 192, TTL: 1, Hash index: 886
Local host: 10.5.6.6, Local port: 35502 (Local App PID: 1011918)
Foreign host: 10.5.6.5, Foreign port: 179
```

```
Current send queue size in bytes: 0 (max 24576)
Current receive queue size in bytes: 0 (max 32768)  mis-ordered: 0 bytes
Current receive queue size in packets: 0 (max 0)

Timer          Starts    Wakeups         Next(msec)
Retrans             9          1                0
SendWnd             0          0                0
TimeWait            0          0                0
AckHold             6          5                0
KeepAlive           1          0                0
PmtuAger            0          0                0
GiveUp              0          0                0
Throttle            0          0                0

   iss: 764231407   snduna: 764231579   sndnxt: 764231579
sndmax: 764231579   sndwnd: 32650       sndcwnd: 4380
   irs: 2712512697  rcvnxt: 2712512869  rcvwnd: 32669   rcvadv: 2712545538

SRTT: 31 ms,  RTTO: 300 ms,  RTV: 130 ms,  KRTT: 0 ms
minRTT: 9 ms,  maxRTT: 239 ms

ACK hold time: 200 ms, Keepalive time: 0 sec, SYN waittime: 30 sec
Giveup time: 0 ms, Retransmission retries: 0, Retransmit forever: FALSE
Connect retries remaining: 30, connect retry interval: 50 secs

State flags: none
Feature flags: Win Scale, Nagle
Request flags: Win Scale


Datagrams (in bytes): MSS 1460, peer MSS 1460, min MSS 4420, max MSS 4420



Window scales: rcv 0, snd 0, request rcv 0, request snd 0
Timestamp option: recent 0, recent age 0, last ACK sent 0
Sack blocks {start, end}: none
Sack holes {start, end, dups, rxmit}: none

Socket options: SO_REUSEADDR, SO_REUSEPORT, SO_NBIO
Socket states: SS_ISCONNECTED, SS_PRIV
Socket receive buffer states: SB_DEL_WAKEUP
Socket send buffer states: SB_DEL_WAKEUP
Socket receive buffer: Low/High watermark 1/32768
Socket send buffer   : Low/High watermark 2048/24576, Notify threshold 0

PDU information:
 #PDU's in buffer: 0
FIB Lookup Cache:  IFH: 0x40  PD ctx: size: 0  data:
  Num Labels: 0  Label Stack:

RP/0/0/CPU0:R6#
```

TCP session details as seen on R5 – PASSIVE:

<#root>

```
! - as seen on R5 - Passive
```

```
RP/0/0/CPU0:R5#show tcp detail pcb 0x12155a98
Fri Jan  8 09:55:18.193 UTC
==============================================================
Connection state is ESTAB, I/O status: 0, socket status: 0
Established at Fri Jan  8 09:51:47 2021

PCB 0x12155a98, SO 0x12153ea0, TCPCB 0x12154e18, vrfid 0x60000000,
Pak Prio: Medium, TOS: 192, TTL: 1, Hash index: 886
Local host: 10.5.6.5, Local port: 179 (Local App PID: 1044686)
Foreign host: 10.5.6.6, Foreign port: 35502

Current send queue size in bytes: 0 (max 24576)
Current receive queue size in bytes: 0 (max 32768)  mis-ordered: 0 bytes
Current receive queue size in packets: 0 (max 0)

Timer          Starts    Wakeups         Next(msec)
Retrans             6          0                 0
SendWnd             0          0                 0
TimeWait            0          0                 0
AckHold             6          1                 0
KeepAlive           1          0                 0
PmtuAger            0          0                 0
GiveUp              0          0                 0
Throttle            0          0                 0

   iss: 2712512697  snduna: 2712512850  sndnxt: 2712512850
sndmax: 2712512850  sndwnd: 32688       sndcwnd: 4380
   irs: 764231407   rcvnxt: 764231560   rcvwnd: 32669   rcvadv: 764264229

SRTT: 107 ms,  RTTO: 538 ms,  RTV: 431 ms,  KRTT: 0 ms
minRTT: 29 ms,  maxRTT: 219 ms

ACK hold time: 200 ms, Keepalive time: 0 sec, SYN waittime: 30 sec
Giveup time: 0 ms, Retransmission retries: 0, Retransmit forever: FALSE
Connect retries remaining: 0, connect retry interval: 0 secs

State flags: none
Feature flags: Win Scale, Nagle
Request flags: Win Scale


Datagrams (in bytes): MSS 1460, peer MSS 4420, min MSS 1460, max MSS 1460



Window scales: rcv 0, snd 0, request rcv 0, request snd 0
Timestamp option: recent 0, recent age 0, last ACK sent 0
Sack blocks {start, end}: none
Sack holes {start, end, dups, rxmit}: none
Socket options: SO_REUSEADDR, SO_REUSEPORT, SO_NBIO
Socket states: SS_ISCONNECTED, SS_PRIV
Socket receive buffer states: SB_DEL_WAKEUP
Socket send buffer states: SB_DEL_WAKEUP
Socket receive buffer: Low/High watermark 1/32768
Socket send buffer   : Low/High watermark 2048/24576, Notify threshold 0

PDU information:
 #PDU's in buffer: 0
FIB Lookup Cache:  IFH: 0x40  PD ctx: size: 0  data:
  Num Labels: 0  Label Stack:

RP/0/0/CPU0:R5#
```

Use Non-Default MTU Value – Passive TCP Peer

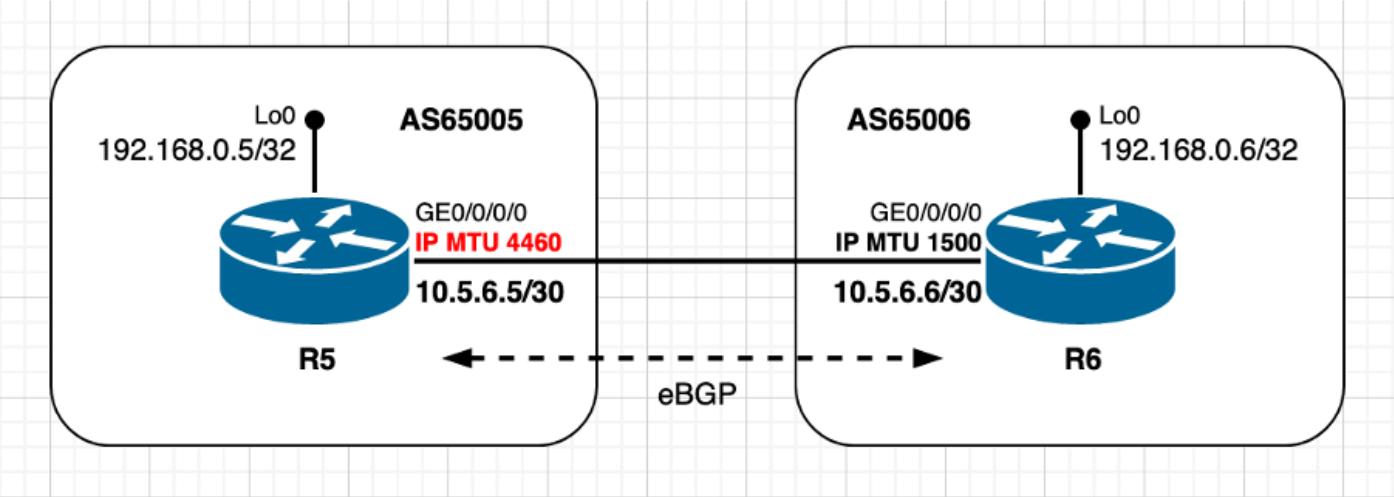


Image 2.3 – PASSIVE peer uses non-default MTU Value.

With still the same eBGP scenario, but now with passive TCP peer R5 configured with non-default IP MTU and active TCP peer R6 with default IP MTU value. As with the previous scenario, take note on how MSS value is selected by passive peer R5. TCP MSS calculation in this scenario can be summarized as follows:

- R5 uses non-default IP MTU 4460 bytes
- TCP path MTU discovery is disabled by default
- TCP peers are directly connected
  - R6 manages BGP connection
  - R6 sends SYN with MSS of 1460 bytes
    - 1500 (Interface IP MTU) - 20 (minTCP_H) - 20 (minIP_H)
  - R5 send SYN, ACK with MSS of 1460 bytes
    - sends the lower of [Received MSS; Local initial MSS]
    - Received MSS 1460 bytes; Local initial MSS 4420 bytes
  - The lowest MSS value is used on both peers

TCP SYN sourced from R6:

<#root>

```
! – TCP SYN sourced from R6

237    2696.666481    10.5.6.6       10.5.6.5       TCP    62     47007 → 179 [SYN] Seq=0 Win=16384 Len=0

MSS=1460

 WS=1

Frame 237: 62 bytes on wire (496 bits), 62 bytes captured (496 bits) on interface 0
Ethernet II, Src: fa:16:3e:85:3d:c2 (fa:16:3e:85:3d:c2), Dst: fa:16:3e:ad:51:8f (fa:16:3e:ad:51:8f)
Internet Protocol Version 4, Src: 10.5.6.6, Dst: 10.5.6.5
Transmission Control Protocol, Src Port: 47007, Dst Port: 179, Seq: 0, Len: 0
    Source Port: 47007
    Destination Port: 179
    [Stream index: 10]
    [TCP Segment Len: 0]
    Sequence number: 0    (relative sequence number)
    Acknowledgment number: 0
    Header Length: 28 bytes
```

```
    Flags: 0x002 (SYN)
    Window size value: 16384
    [Calculated window size: 16384]
    Checksum: 0x2025 [unverified]
    [Checksum Status: Unverified]
    Urgent pointer: 0
    Options: (8 bytes), Maximum segment size, Window scale, End of Option List (EOL)
        Maximum segment size: 1460 bytes
            Kind: Maximum Segment Size (2)
            Length: 4
```

MSS Value: 1460

```
        Window scale: 0 (multiply by 1)
        End of Option List (EOL)
```

TCP SYN, ACK sourced from R5:

<#root>

```
! – TCP SYN, ACK sourced from R5

238    2696.702792    10.5.6.5       10.5.6.6       TCP    62     179 → 47007 [SYN, ACK] Seq=0 Ack=1 Win=16
```

MSS=1460

```
 WS=1

Frame 238: 62 bytes on wire (496 bits), 62 bytes captured (496 bits) on interface 0
Ethernet II, Src: fa:16:3e:ad:51:8f (fa:16:3e:ad:51:8f), Dst: fa:16:3e:85:3d:c2 (fa:16:3e:85:3d:c2)
Internet Protocol Version 4, Src: 10.5.6.5, Dst: 10.5.6.6
Transmission Control Protocol, Src Port: 179, Dst Port: 47007, Seq: 0, Ack: 1, Len: 0
    Source Port: 179
    Destination Port: 47007
    [Stream index: 10]
    [TCP Segment Len: 0]
    Sequence number: 0    (relative sequence number)
    Acknowledgment number: 1    (relative ack number)
    Header Length: 28 bytes
    Flags: 0x012 (SYN, ACK)
    Window size value: 16384
    [Calculated window size: 16384]
    Checksum: 0x7078 [unverified]
    [Checksum Status: Unverified]
    Urgent pointer: 0
    Options: (8 bytes), Maximum segment size, Window scale, End of Option List (EOL)
        Maximum segment size: 1460 bytes
            Kind: Maximum Segment Size (2)
            Length: 4
```

MSS Value: 1460

```
        Window scale: 0 (multiply by 1)
        End of Option List (EOL)
```

TCP session details as seen on R6 – ACTIVE:

<#root>

```
! - as seen on R6 - Active

RP/0/0/CPU0:R6#show tcp detail pcb 0x1215761c
Fri Jan  8 10:15:20.351 UTC
==============================================================
Connection state is ESTAB, I/O status: 0, socket status: 0
Established at Fri Jan  8 10:10:04 2021

PCB 0x1215761c, SO 0x12162aac, TCPCB 0x12156f64, vrfid 0x60000000,
Pak Prio: Medium, TOS: 192, TTL: 1, Hash index: 103
Local host: 10.5.6.6, Local port: 47007 (Local App PID: 1011918)
Foreign host: 10.5.6.5, Foreign port: 179

Current send queue size in bytes: 0 (max 24576)
Current receive queue size in bytes: 0 (max 32768)  mis-ordered: 0 bytes
Current receive queue size in packets: 0 (max 0)

Timer          Starts    Wakeups         Next(msec)
Retrans            10          1                 0
SendWnd             0          0                 0
TimeWait            0          0                 0
AckHold             7          5                 0
KeepAlive           1          0                 0
PmtuAger            0          0                 0
GiveUp              0          0                 0
Throttle            0          0                 0

   iss: 3949093168  snduna: 3949093359  sndnxt: 3949093359
sndmax: 3949093359  sndwnd: 32631       sndcwnd: 4380
   irs: 54439005    rcvnxt: 54439196    rcvwnd: 32650   rcvadv: 54471846

SRTT: 75 ms,  RTTO: 459 ms,  RTV: 384 ms,  KRTT: 0 ms
minRTT: 9 ms,  maxRTT: 239 ms

ACK hold time: 200 ms, Keepalive time: 0 sec, SYN waittime: 30 sec
Giveup time: 0 ms, Retransmission retries: 0, Retransmit forever: FALSE
Connect retries remaining: 30, connect retry interval: 50 secs

State flags: none
Feature flags: Win Scale, Nagle
Request flags: Win Scale


Datagrams (in bytes): MSS 1460, peer MSS 1460, min MSS 1460, max MSS 1460



Window scales: rcv 0, snd 0, request rcv 0, request snd 0
Timestamp option: recent 0, recent age 0, last ACK sent 0
Sack blocks {start, end}: none
Sack holes {start, end, dups, rxmit}: none
Socket options: SO_REUSEADDR, SO_REUSEPORT, SO_NBIO
Socket states: SS_ISCONNECTED, SS_PRIV
Socket receive buffer states: SB_DEL_WAKEUP
Socket send buffer states: SB_DEL_WAKEUP
Socket receive buffer: Low/High watermark 1/32768
Socket send buffer   : Low/High watermark 2048/24576, Notify threshold 0
```

```
PDU information:
 #PDU's in buffer: 0
FIB Lookup Cache:  IFH: 0x40  PD ctx: size: 0  data:
  Num Labels: 0  Label Stack:

RP/0/0/CPU0:R6#
```

TCP session details as seen on R5 – PASSIVE:

<#root>

```
! - as seen on R5 - Passive

RP/0/0/CPU0:R5#show interfaces gigabitEthernet 0/0/0/0
Fri Jan  8 10:10:39.110 UTC
GigabitEthernet0/0/0/0 is up, line protocol is up
  Interface state transitions: 1
  Hardware is GigabitEthernet, address is fa16.3ead.518f (bia fa16.3ead.518f)
  Internet address is 10.5.6.5/30
```

MTU 4474 bytes

```
, BW 1000000 Kbit (Max: 1000000 Kbit)
<snip>

RP/0/0/CPU0:R5#show tcp detail pcb 0x121550fc
Fri Jan  8 10:14:20.105 UTC
================================================================
Connection state is ESTAB, I/O status: 0, socket status: 0
Established at Fri Jan  8 10:10:05 2021

PCB 0x121550fc, SO 0x12154e18, TCPCB 0x12154304, vrfid 0x60000000,
Pak Prio: Medium, TOS: 192, TTL: 1, Hash index: 103
Local host: 10.5.6.5, Local port: 179 (Local App PID: 1044686)
Foreign host: 10.5.6.6, Foreign port: 47007

Current send queue size in bytes: 0 (max 24576)
Current receive queue size in bytes: 0 (max 32768)  mis-ordered: 0 bytes
Current receive queue size in packets: 0 (max 0)

Timer          Starts    Wakeups         Next(msec)
Retrans             7          0                0
SendWnd             0          0                0
TimeWait            0          0                0
AckHold             7          2                0
KeepAlive           1          0                0
PmtuAger            0          0                0
GiveUp              0          0                0
Throttle            0          0                0

   iss: 54439005    snduna: 54439177    sndnxt: 54439177
sndmax: 54439177    sndwnd: 32669       sndcwnd: 4380
   irs: 3949093168  rcvnxt: 3949093340  rcvwnd: 32650   rcvadv: 3949125990

SRTT: 117 ms,  RTTO: 570 ms,  RTV: 453 ms,  KRTT: 0 ms
minRTT: 19 ms,  maxRTT: 229 ms

ACK hold time: 200 ms, Keepalive time: 0 sec, SYN waittime: 30 sec
```

```
Giveup time: 0 ms, Retransmission retries: 0, Retransmit forever: FALSE
Connect retries remaining: 0, connect retry interval: 0 secs

State flags: none
Feature flags: Win Scale, Nagle
Request flags: Win Scale


Datagrams (in bytes): MSS 1460, peer MSS 1460, min MSS 4420, max MSS 4420



Window scales: rcv 0, snd 0, request rcv 0, request snd 0
Timestamp option: recent 0, recent age 0, last ACK sent 0
Sack blocks {start, end}: none
Sack holes {start, end, dups, rxmit}: none

Socket options: SO_REUSEADDR, SO_REUSEPORT, SO_NBIO
Socket states: SS_ISCONNECTED, SS_PRIV
Socket receive buffer states: SB_DEL_WAKEUP
Socket send buffer states: SB_DEL_WAKEUP
Socket receive buffer: Low/High watermark 1/32768
Socket send buffer   : Low/High watermark 2048/24576, Notify threshold 0

PDU information:
 #PDU's in buffer: 0
FIB Lookup Cache:  IFH: 0x40  PD ctx: size: 0  data:
  Num Labels: 0  Label Stack:

RP/0/0/CPU0:R5#
```

Use TCP Options – XR Active

As mentioned earlier in this document, the use of TCP options (such as TCP MD5, TCP selective-ack, or TCP timestamps) influences the MSS calculation as these options lead to additional bytes to be accounted for in the MSS calculation.

This section as well as the next aims to illustrate the MSS calculation made by peers when in presence of TCP options. TCP MD5 authentication option is used as an example. Please refer to the reference scenario in Images 2.4 as shown in the image.

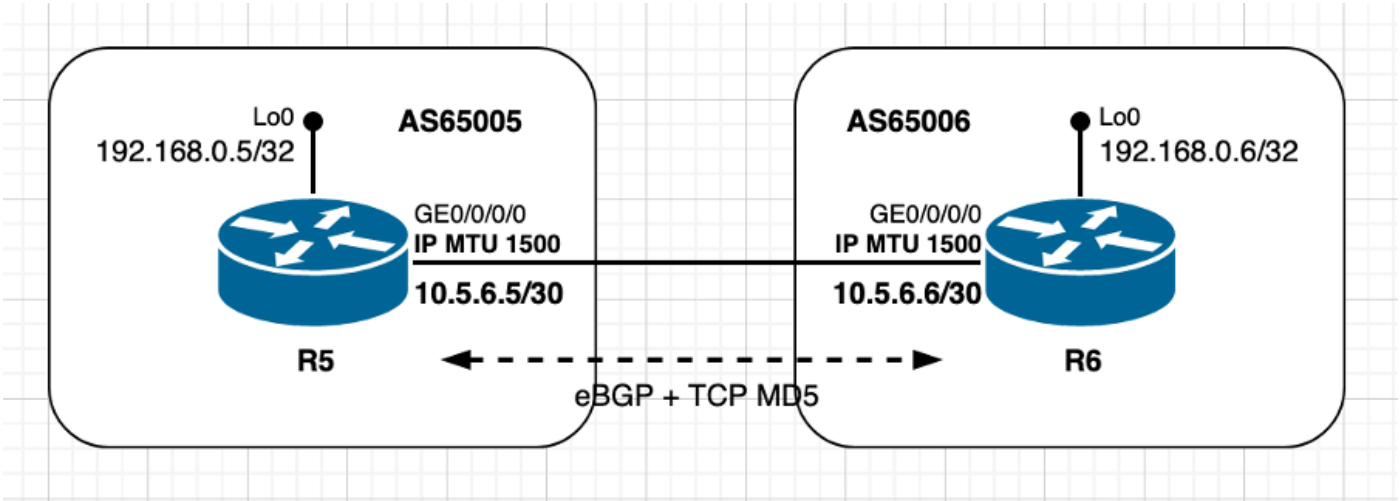


Image 2.4 – Use TCP Options (MD5) – XR Active.

In this scenario both peers use default IP MTU values, are directly connected, and peer R6 plays TCP active role. As already shared the configuration and use of TCP MD5 Authentication account for additional overhead. TCP MSS calculation in this particular scenario can be summarized as follows:

- Both nodes use a default IP MTU of 1500 bytes
- TCP path MTU discovery is disabled by default
- TCP peers are directly connected
- TCP MD5 authentication enabled on both nodes
    - R6 manages BGP connection
    - R6 sends SYN with MSS of 1436 bytes
        - 1500 (Interface IP MTU) - 20 (minTCP_H) - 20 (minIP_H) - 24 Bytes (IOS XR TCP Options Overhead)
    - R5 send SYN, ACK with MSS of 1436 bytes
        - sends the lower of [Received MSS; Local initial MSS]
        - Received MSS 1436 bytes; Local initial MSS 1460 bytes
    - The lowest MSS value is used on both peers

As seen from the summary, the way Cisco IOS XR behaves is not strictly as per RFC 879, and RFC 6691, which state that the TCP options must not be accounted on the MSS calculation.

The Cisco IOS XR account of an extra factor on **tcp header length** is further documented on Cisco bug ID CSCvf20166:

"(…)When XR is starting the BGP connection, BGP first creates the socket, then sets the socket options including **MD5**. This makes **tcp option header length = 24**. And hence the initial MSS becomes 1500 - 40 - 24 = 1436. This is sent to peer and peer uses min(1436, 1460) = 1436.(…)

TCP SYN sourced from R6:

<#root>

```
! - TCP SYN sourced from R6

430    5775.839420    10.5.6.6       10.5.6.5       TCP    82     24785 → 179 [SYN] Seq=0 Win=16384 Len=0

MSS=1436

 WS=1

Frame 430: 82 bytes on wire (656 bits), 82 bytes captured (656 bits) on interface 0
Ethernet II, Src: fa:16:3e:85:3d:c2 (fa:16:3e:85:3d:c2), Dst: fa:16:3e:ad:51:8f (fa:16:3e:ad:51:8f)
Internet Protocol Version 4, Src: 10.5.6.6, Dst: 10.5.6.5
Transmission Control Protocol, Src Port: 24785, Dst Port: 179, Seq: 0, Len: 0
    Source Port: 24785
    Destination Port: 179
    [Stream index: 14]
    [TCP Segment Len: 0]
    Sequence number: 0    (relative sequence number)
    Acknowledgment number: 0
    Header Length: 48 bytes
    Flags: 0x002 (SYN)
    Window size value: 16384
    [Calculated window size: 16384]
    Checksum: 0xd62b [unverified]
    [Checksum Status: Unverified]
    Urgent pointer: 0
    Options: (28 bytes), Maximum segment size, Window scale, No-Operation (NOP),
```

TCP MD5 signature

```
, End of Option List (EOL)
        Maximum segment size: 1436 bytes
            Kind: Maximum Segment Size (2)
            Length: 4
```

MSS Value: 1436

```
        Window scale: 0 (multiply by 1)
        No-Operation (NOP)
        TCP MD5 signature
        End of Option List (EOL)
```

TCP SYN, ACK sourced from R5:

<#root>

```
! – TCP SYN, ACK sourced from R5

431    5775.845744    10.5.6.5       10.5.6.6       TCP    82     179 → 24785 [SYN, ACK] Seq=0 Ack=1 Win=16
```

MSS=1436

```
 WS=1

Frame 431: 82 bytes on wire (656 bits), 82 bytes captured (656 bits) on interface 0
Ethernet II, Src: fa:16:3e:ad:51:8f (fa:16:3e:ad:51:8f), Dst: fa:16:3e:85:3d:c2 (fa:16:3e:85:3d:c2)
Internet Protocol Version 4, Src: 10.5.6.5, Dst: 10.5.6.6
Transmission Control Protocol, Src Port: 179, Dst Port: 24785, Seq: 0, Ack: 1, Len: 0
    Source Port: 179
    Destination Port: 24785
    [Stream index: 14]
    [TCP Segment Len: 0]
    Sequence number: 0    (relative sequence number)
    Acknowledgment number: 1    (relative ack number)
    Header Length: 48 bytes
    Flags: 0x012 (SYN, ACK)
    Window size value: 16384
    [Calculated window size: 16384]
    Checksum: 0xe83d [unverified]
    [Checksum Status: Unverified]
    Urgent pointer: 0
    Options: (28 bytes), Maximum segment size, Window scale, No-Operation (NOP),
```

TCP MD5 signature

```
, End of Option List (EOL)
        Maximum segment size: 1436 bytes
            Kind: Maximum Segment Size (2)
            Length: 4
```

MSS Value: 1436

```
        Window scale: 0 (multiply by 1)
        No-Operation (NOP)
        TCP MD5 signature
```

```
        End of Option List (EOL)
```

TCP session details as seen on R6 – ACTIVE:

<#root>

```
! - as seen on R6 - Active

RP/0/0/CPU0:R6#show tcp detail pcb 0x1215761c
Fri Jan  8 11:14:13.599 UTC
==============================================================
Connection state is ESTAB, I/O status: 0, socket status: 0
Established at Fri Jan  8 11:01:21 2021

PCB 0x1215761c, SO 0x1216419c, TCPCB 0x121649fc, vrfid 0x60000000,
Pak Prio: Medium, TOS: 192, TTL: 1, Hash index: 409
Local host: 10.5.6.6, Local port: 24785 (Local App PID: 1011918)
Foreign host: 10.5.6.5, Foreign port: 179

Current send queue size in bytes: 0 (max 24576)
Current receive queue size in bytes: 0 (max 32768)  mis-ordered: 0 bytes
Current receive queue size in packets: 0 (max 0)

Timer          Starts    Wakeups         Next(msec)
Retrans            17          1                 0
SendWnd             0          0                 0
TimeWait            0          0                 0
AckHold            14         13                 0
KeepAlive           1          0                 0
PmtuAger            0          0                 0
GiveUp              0          0                 0
Throttle            0          0                 0

   iss: 1379482495  snduna: 1379482819  sndnxt: 1379482819
sndmax: 1379482819  sndwnd: 32498       sndcwnd: 4308
   irs: 3750694052  rcvnxt: 3750694376  rcvwnd: 32517   rcvadv: 3750726893

SRTT: 55 ms,  RTTO: 300 ms,  RTV: 176 ms,  KRTT: 0 ms
minRTT: 9 ms,  maxRTT: 259 ms

ACK hold time: 200 ms, Keepalive time: 0 sec, SYN waittime: 30 sec
Giveup time: 0 ms, Retransmission retries: 0, Retransmit forever: FALSE
Connect retries remaining: 30, connect retry interval: 50 secs

State flags: none
Feature flags:
```

MD5

```
, Win Scale, Nagle
Request flags: Win Scale
```

Datagrams (in bytes): MSS 1436, peer MSS 1436, min MSS 1436, max MSS 1436

```
Window scales: rcv 0, snd 0, request rcv 0, request snd 0
Timestamp option: recent 0, recent age 0, last ACK sent 0
Sack blocks {start, end}: none
```

```
Sack holes {start, end, dups, rxmit}: none

Socket options: SO_REUSEADDR, SO_REUSEPORT, SO_NBIO
Socket states: SS_ISCONNECTED, SS_PRIV
Socket receive buffer states: SB_DEL_WAKEUP
Socket send buffer states: SB_DEL_WAKEUP
Socket receive buffer: Low/High watermark 1/32768
Socket send buffer   : Low/High watermark 2048/24576, Notify threshold 0

PDU information:
 #PDU's in buffer: 0
FIB Lookup Cache:  IFH: 0x40  PD ctx: size: 0  data:
  Num Labels: 0  Label Stack:

RP/0/0/CPU0:R6#
```

TCP session details as seen on R5 – PASSIVE:

<#root>

```
! - as seen on R5 - Passive

RP/0/0/CPU0:R5#show tcp detail pcb 0x12155d04
Fri Jan  8 11:12:51.984 UTC
================================================================
Connection state is ESTAB, I/O status: 0, socket status: 0
Established at Fri Jan  8 11:01:22 2021

PCB 0x12155d04, SO 0x12154e18, TCPCB 0x12154304, vrfid 0x60000000,
Pak Prio: Medium, TOS: 192, TTL: 1, Hash index: 409
Local host: 10.5.6.5, Local port: 179 (Local App PID: 1044686)
Foreign host: 10.5.6.6, Foreign port: 24785

Current send queue size in bytes: 0 (max 24576)
Current receive queue size in bytes: 0 (max 32768)  mis-ordered: 0 bytes
Current receive queue size in packets: 0 (max 0)

Timer          Starts    Wakeups         Next(msec)
Retrans            14          0                 0
SendWnd             0          0                 0
TimeWait            0          0                 0
AckHold            14          3                 0
KeepAlive           1          0                 0
PmtuAger            0          0                 0
GiveUp              0          0                 0
Throttle            0          0                 0

iss: 3750694052  snduna: 3750694357  sndnxt: 3750694357
sndmax: 3750694357  sndwnd: 32536       sndcwnd: 4308
irs: 1379482495  rcvnxt: 1379482800  rcvwnd: 32517   rcvadv: 1379515317
SRTT: 181 ms,  RTTO: 443 ms,  RTV: 262 ms,  KRTT: 0 ms
minRTT: 29 ms,  maxRTT: 219 ms

ACK hold time: 200 ms, Keepalive time: 0 sec, SYN waittime: 30 sec
Giveup time: 0 ms, Retransmission retries: 0, Retransmit forever: FALSE
Connect retries remaining: 0, connect retry interval: 0 secs

State flags: none
Feature flags:
```

MD5

```
, Win Scale, Nagle
Request flags: Win Scale
```

Datagrams (in bytes): MSS 1436, peer MSS 1436, min MSS 1460, max MSS 1460

```
Window scales: rcv 0, snd 0, request rcv 0, request snd 0
Timestamp option: recent 0, recent age 0, last ACK sent 0
Sack blocks {start, end}: none
Sack holes {start, end, dups, rxmit}: none

Socket options: SO_REUSEADDR, SO_REUSEPORT, SO_NBIO
Socket states: SS_ISCONNECTED, SS_PRIV
Socket receive buffer states: SB_DEL_WAKEUP
Socket send buffer states: SB_DEL_WAKEUP
Socket receive buffer: Low/High watermark 1/32768
Socket send buffer   : Low/High watermark 2048/24576, Notify threshold 0

PDU information:
#PDU's in buffer: 0
FIB Lookup Cache:  IFH: 0x40  PD ctx: size: 0  data:
Num Labels: 0  Label Stack:

RP/0/0/CPU0:R5#
```

Similar behavior can be observed with other TCP options which when configured account for additional overhead and influence MSS calculation in Cisco IOS XR. Consider the same scenario and these examples which document the MSS calculation when TCP timestamps and TCP selective-ack options are configured.

TCP session details as seen on R6 – ACTIVE – with TCP options timestamp and selective-ack options configured:

<#root>

```
! - as seen on R6 – Active
```

! -- tcp timestamp configured

```
! -- 12 bytes of additional overhead

RP/0/0/CPU0:R6#show tcp detail pcb 0x1539c844
<snip>
Feature flags: Timestamp, Win Scale, Nagle
Request flags: Timestamp, Win Scale
```

Datagrams (in bytes): MSS 1448, peer MSS 1448, min MSS 1448, max MSS 1448

```
<snip>

! - as seen on R6 – Active
```

! -- tcp selective-ack configured

```
! -- 36 bytes of additional overhead

RP/0/0/CPU0:R6#show tcp detail pcb 0x1539df38
<snip>
Feature flags: Sack, Win Scale, Nagle
Request flags: Sack, Win Scale


Datagrams (in bytes): MSS 1424, peer MSS 1424, min MSS 1424, max MSS 1424


<snip>

! - as seen on R6 – Active

! -- tcp selective-ack and tcp timestamp configured


! -- 40 bytes of additional overhead

RP/0/0/CPU0:R6#show tcp detail pcb 0x1539e130
<snip>
State flags: none
Feature flags: Sack, Timestamp, Win Scale, Nagle
Request flags: Sack, Timestamp, Win Scale


Datagrams (in bytes): MSS 1420, peer MSS 1420, min MSS 1420, max MSS 1420


<snip>

! - as seen on R6 – Active

! -- MD5 and tcp selective-ack configured


! -- 36 bytes of additional overhead

RP/0/0/CPU0:R6#show tcp detail pcb 0x1539b3cc
<snip>
Feature flags: Sack, MD5, Win Scale, Nagle
Request flags: Sack, Win Scale


Datagrams (in bytes): MSS 1424, peer MSS 1424, min MSS 1424, max MSS 1424


<snip>

! - as seen on R6 – Active

! -- MD5 and tcp timestamp configured


! -- 36 bytes of additional overhead

RP/0/0/CPU0:R6#show tcp detail pcb 0x15397b4c
<snip>

Feature flags: MD5, Timestamp, Win Scale, Nagle
Request flags: Timestamp, Win Scale


Datagrams (in bytes): MSS 1424, peer MSS 1424, min MSS 1424, max MSS 1424
```

```
<snip>

! - as seen on R6 – Active

! -- MD5, tcp timestamp, and tcp selective-ack configured


! -- 40 bytes of additional overhead

RP/0/0/CPU0:R6#show tcp detail pcb 0x1539a4cc
<snip>
State flags: none
Feature flags: MD5, Timestamp, Win Scale, Nagle
Request flags: Timestamp, Win Scale


Datagrams (in bytes): MSS 1420, peer MSS 1420, min MSS 1420, max MSS 1420


<snip>
```

Use TCP Options – XR Passive

From previous scenario you've probably noticed the distinct behavior of Cisco IOS XR node when in passive role with regards initial MSS calculation. Node does not account for **tcp option header length**. This scenario aims to highlight this distinct behavior which is also described by Cisco bug ID :

"(…) - When peer initiates the connection, it sends initial MSS as 1460. XR TCP creates socket, pcb etc then it takes these two actions in the given order:

- First, it computes initial MSS after subtracting **tcp option header length**. This is '0' as MD5 option is not yet inherited to this socket from listen socket.

- Then, it inherits the 'MD5' and other options and this makes 'option header bytes length' to 24.

So, in this case XR TCP sends initial MSS as 1460 and hence that is used by both. (…)"

In this scenario, although active TCP peer R8 is a Cisco IOS node, this fact does not introduce any difference or specifics on what the scenario aims to highlight. Nevertheless, and interestingly, note that differently than Cisco IOS XR as showed with previous section scenario, here the active TCP peer R8 does not consider TCP options on initial MSS calculation.

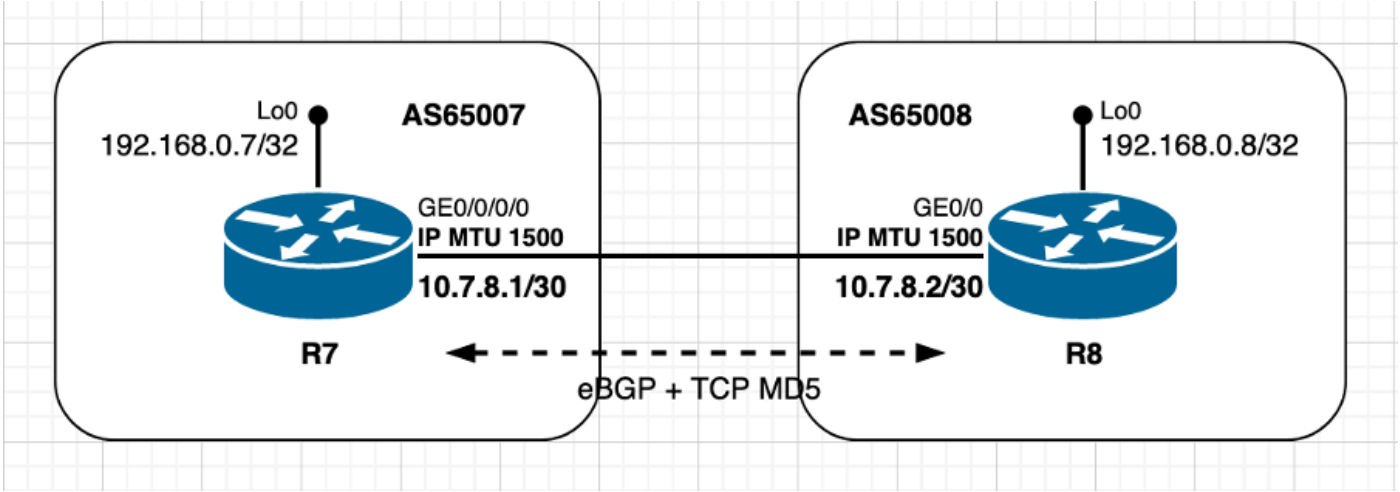


Image 2.5 – Use TCP Options (MD5) – XR Passive.

Both peers use default IP MTU values and are directly connected. Cisco IOS peer R8 plays active role. TCP MSS calculation in this scenario can be summarized as follows:

- Both nodes use a default IP MTU of 1500 bytes
- TCP path MTU discovery is disabled by default on Cisco IOS XR R7
- TCP path MTU discovery is enabled by default on Cisco IOS R8
- TCP Peers are directly connected
- TCP MD5 authentication enabled on both nodes
    - IOS R8 manages the BGP connection
    - IOS R8 sends SYN with MSS of 1460 bytes
        - 1500 (Interface IP MTU) - 20 (minTCP_H) - 20 (minIP_H)
    - IOS XR R7 sends SYN, ACK with MSS of 1460 bytes
        - sends the lower of [Received MSS; Local initial MSS]
        - Received MSS 1460 bytes; Local initial MSS 1460 bytes
    - The lowest MSS value is used on both peers

TCP SYN sourced from R8 – Cisco IOS:

<#root>

```
! – TCP SYN sourced from R8

96     5.907127       10.7.8.2       10.7.8.1       TCP    78     52975 → 179 [SYN] Seq=0 Win=16384 Len=0

MSS=1460



Frame 96: 78 bytes on wire (624 bits), 78 bytes captured (624 bits) on interface 0
Ethernet II, Src: fa:16:3e:58:21:ba (fa:16:3e:58:21:ba), Dst: fa:16:3e:68:d9:e5 (fa:16:3e:68:d9:e5)
Internet Protocol Version 4, Src: 10.7.8.2, Dst: 10.7.8.1
Transmission Control Protocol, Src Port: 52975, Dst Port: 179, Seq: 0, Len: 0
    Source Port: 52975
    Destination Port: 179
    [Stream index: 3]
    [TCP Segment Len: 0]
    Sequence number: 0    (relative sequence number)
    Acknowledgment number: 0
    Header Length: 44 bytes
    Flags: 0x002 (SYN)
    Window size value: 16384
    [Calculated window size: 16384]
    Checksum: 0xb612 [unverified]
    [Checksum Status: Unverified]
    Urgent pointer: 0
    Options: (24 bytes), Maximum segment size,

TCP MD5 signature

, End of Option List (EOL)
        Maximum segment size: 1460 bytes
            Kind: Maximum Segment Size (2)
            Length: 4


MSS Value: 1460
```

```
        TCP MD5 signature
        End of Option List (EOL)
```

TCP SYN, ACK sourced from R7 – Cisco IOS XR:

<#root>

```
! – TCP SYN,ACK sourced from R7

97     0.003446       10.7.8.1       10.7.8.2       TCP    78     179 → 52975 [SYN, ACK] Seq=0 Ack=1 Win=16
```

MSS=1460

```
Frame 97: 78 bytes on wire (624 bits), 78 bytes captured (624 bits) on interface 0
Ethernet II, Src: fa:16:3e:68:d9:e5 (fa:16:3e:68:d9:e5), Dst: fa:16:3e:58:21:ba (fa:16:3e:58:21:ba)
Internet Protocol Version 4, Src: 10.7.8.1, Dst: 10.7.8.2
Transmission Control Protocol, Src Port: 179, Dst Port: 52975, Seq: 0, Ack: 1, Len: 0
    Source Port: 179
    Destination Port: 52975
    [Stream index: 3]
    [TCP Segment Len: 0]
    Sequence number: 0    (relative sequence number)
    Acknowledgment number: 1    (relative ack number)
    Header Length: 44 bytes
    Flags: 0x012 (SYN, ACK)
    Window size value: 16384
    [Calculated window size: 16384]
    Checksum: 0xfb47 [unverified]
    [Checksum Status: Unverified]
    Urgent pointer: 0
    Options: (24 bytes), Maximum segment size,
```

TCP MD5 signature

```
, End of Option List (EOL)
        Maximum segment size: 1460 bytes
            Kind: Maximum Segment Size (2)
            Length: 4
```

MSS Value: 1460

```
        TCP MD5 signature
        End of Option List (EOL)
```

TCP session details as seen on R8 – Cisco IOS – ACTIVE:

<#root>

```
! – as seen from R8 - Cisco IOS

R8#show ip bgp neighbors
BGP neighbor is 10.7.8.1,  remote AS 65007, external link
  BGP version 4, remote router ID 192.168.0.7
  BGP state = Established, up for 00:06:12
```

```
 Last read 00:00:16, last write 00:00:16, hold time is 180, keepalive interval is 60 seconds
 Neighbor sessions:
   1 active, is not multisession capable (disabled)
 Neighbor capabilities:
   Route refresh: advertised and received(new)
   Four-octets ASN Capability: advertised and received
   Address family IPv4 Unicast: advertised and received
   Enhanced Refresh Capability: advertised
   Multisession Capability:
   Stateful switchover support enabled: NO for session 1
 Message statistics:
   InQ depth is 0
   OutQ depth is 0

                        Sent       Rcvd
   Opens:                  1          1
   Notifications:          0          0
   Updates:                1          1
   Keepalives:             7          7
   Route Refresh:          0          0
   Total:                  9          9
 Do log neighbor state changes (via global configuration)
 Default minimum time between advertisement runs is 30 seconds

For address family: IPv4 Unicast
 Session: 10.7.8.1
 BGP table version 1, neighbor version 1/0
 Output queue size : 0
 Index 6, Advertise bit 0
 6 update-group member
 Slow-peer detection is disabled
 Slow-peer split-update-group dynamic is disabled
                               Sent       Rcvd
 Prefix activity:              ----       ----
   Prefixes Current:              0          0
   Prefixes Total:                0          0
   Implicit Withdraw:             0          0
   Explicit Withdraw:             0          0
   Used as bestpath:            n/a          0
   Used as multipath:           n/a          0
   Used as secondary:           n/a          0

                                 Outbound    Inbound
 Local Policy Denied Prefixes:   --------    -------
   Total:                               0          0
 Number of NLRIs in the update sent: max 0, min 0
 Last detected as dynamic slow peer: never
 Dynamic slow peer recovered: never
 Refresh Epoch: 1
 Last Sent Refresh Start-of-rib: never
 Last Sent Refresh End-of-rib: never
 Last Received Refresh Start-of-rib: never
 Last Received Refresh End-of-rib: never
                                    Sent       Rcvd
       Refresh activity:            ----       ----
         Refresh Start-of-RIB         0          0
         Refresh End-of-RIB           0          0

 Address tracking is enabled, the RIB does have a route to 10.7.8.1
 Connections established 6; dropped 5
 Last reset 00:06:18, due to BGP Notification received of session 1, Administrative Reset
 External BGP neighbor configured for connected checks (single-hop no-disable-connected-check)
```

```
  Interface associated: GigabitEthernet0/1 (peering address in same link)
```

`Transport(tcp) path-mtu-discovery is enabled`

```
  Graceful-Restart is disabled
  SSO is disabled
Connection state is ESTAB, I/O status: 1, unread input bytes: 0
Connection is ECN Disabled, Mininum incoming TTL 0, Outgoing TTL 1
Local host: 10.7.8.2, Local port: 52975
Foreign host: 10.7.8.1, Foreign port: 179
Connection tableid (VRF): 0
Maximum output segment queue size: 50

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

Event Timers (current time is 0x15DD97):
Timer          Starts    Wakeups            Next
Retrans            10          0             0x0
TimeWait            0          0             0x0
AckHold             9          5             0x0
SendWnd             0          0             0x0
KeepAlive           0          0             0x0
GiveUp              0          0             0x0
PmtuAger            1          0        0x195465
DeadWait            0          0             0x0
Linger              0          0             0x0
ProcessQ            0          0             0x0

iss: 1154289541  snduna: 1154289755  sndnxt: 1154289755
irs: 2149897425  rcvnxt: 2149897635

sndwnd:  32612  scale:      0  maxrcvwnd:  16384
rcvwnd:  16175  scale:      0  delrcvwnd:    209

SRTT: 737 ms, RTTO: 2506 ms, RTV: 1769 ms, KRTT: 0 ms
minRTT: 7 ms, maxRTT: 1000 ms, ACK hold: 200 ms
uptime: 372981 ms, Sent idletime: 16648 ms, Receive idletime: 16431 ms
Status Flags: active open
Option Flags: nagle, path mtu capable,
```

`md5`

```
IP Precedence value : 6
```

`Datagrams (max data segment is 1460 bytes)`

```
:
Rcvd: 18 (out of order: 0), with data: 8, total data bytes: 209
Sent: 16 (retransmit: 0, fastretransmit: 0, partialack: 0, Second Congestion: 0), with data: 9, total d

 Packets received in fast path: 0, fast processed: 0, slow path: 0
 fast lock acquisition failures: 0, slow path: 0
TCP Semaphore      0x0FBFA8A4  FREE

R8#
```

TCP session details as seen on R7 – Cisco IOS XR – PASSIVE:

<#root>

```
! - as seen from R7 - Cisco IOS XR

RP/0/0/CPU0:R7#show tcp detail pcb 0x12152e48
Wed Jan 13 13:03:43.363 UTC
================================================================
Connection state is ESTAB, I/O status: 0, socket status: 0
Established at Wed Jan 13 12:58:16 2021

PCB 0x12152e48, SO 0x1213c130, TCPCB 0x12156060, vrfid 0x60000000,
Pak Prio: Medium, TOS: 192, TTL: 1, Hash index: 947
Local host: 10.7.8.1, Local port: 179 (Local App PID: 983244)
Foreign host: 10.7.8.2, Foreign port: 52975

Current send queue size in bytes: 0 (max 24576)
Current receive queue size in bytes: 0 (max 32768)  mis-ordered: 0 bytes
Current receive queue size in packets: 0 (max 0)

Timer          Starts    Wakeups         Next(msec)
Retrans             8          0                 0
SendWnd             0          0                 0
TimeWait            0          0                 0
AckHold             8          7                 0
KeepAlive           1          0                 0
PmtuAger            0          0                 0
GiveUp              0          0                 0
Throttle            0          0                 0

   iss: 2149897425  snduna: 2149897616  sndnxt: 2149897616
sndmax: 2149897616  sndwnd: 16194       sndcwnd: 4380
   irs: 1154289541  rcvnxt: 1154289736  rcvwnd: 32631   rcvadv: 1154322367

SRTT: 125 ms,  RTTO: 552 ms,  RTV: 427 ms,  KRTT: 0 ms
minRTT: 19 ms,  maxRTT: 229 ms

ACK hold time: 200 ms, Keepalive time: 0 sec, SYN waittime: 30 sec
Giveup time: 0 ms, Retransmission retries: 0, Retransmit forever: FALSE
Connect retries remaining: 0, connect retry interval: 0 secs

State flags: none
Feature flags:

MD5

, Nagle
Request flags: none


Datagrams (in bytes): MSS 1460, peer MSS 1460, min MSS 1460, max MSS 1460



Window scales: rcv 0, snd 0, request rcv 0, request snd 0
Timestamp option: recent 0, recent age 0, last ACK sent 0
Sack blocks {start, end}: none
Sack holes {start, end, dups, rxmit}: none

Socket options: SO_REUSEADDR, SO_REUSEPORT, SO_NBIO
Socket states: SS_ISCONNECTED, SS_PRIV
Socket receive buffer states: SB_DEL_WAKEUP
Socket send buffer states: SB_DEL_WAKEUP
Socket receive buffer: Low/High watermark 1/32768
Socket send buffer   : Low/High watermark 2048/24576, Notify threshold 0
```

```
PDU information:
 #PDU's in buffer: 0
FIB Lookup Cache:  IFH: 0x40  PD ctx: size: 0  data:
  Num Labels: 0  Label Stack:

RP/0/0/CPU0:R7#
```

TCP Peers not Directly Connected

When peers are not directly connected, the way TCP MSS initial calculation is done changes as previously described in the introductory section of this document. The scenario of an iBGP session with all peers configured with default IP MTU values is used to walk-through the MSS calculation.

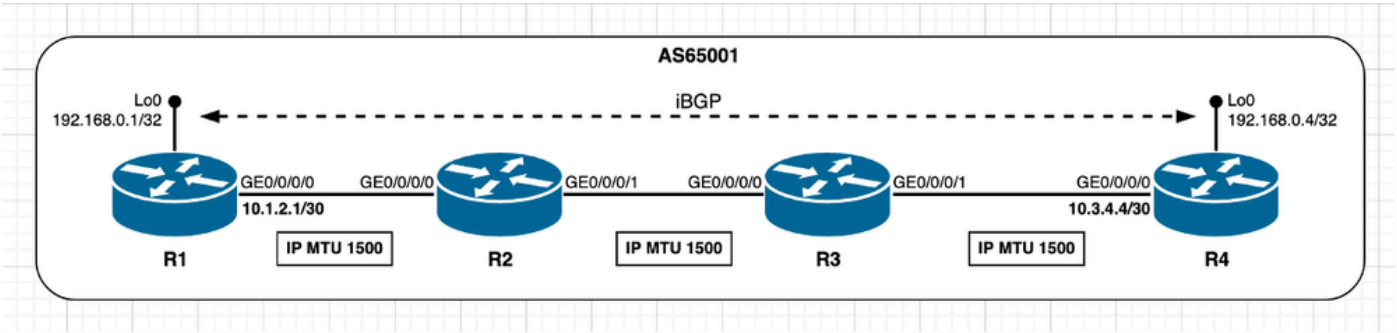


Image 2.6 – TCP Peers not directly connected – iBGP.

The important aspect to take note is that when TCP Path MTU Discovery is disabled, and peers are not directly connected, per design, Cisco IOS XR uses a fixed IP MTU value of 1280 bytes.

In the previous image R4 plays active role and manages TCP connection, R4 opens the TCP session with R1 on destination port 179. Both nodes use default IP MTU value on their interfaces. MSS calculation in this scenario can be summarized as follows:

- All nodes use a default IP MTU of 1500 bytes
- TCP path MTU discovery is disabled by default
- TCP peers are not directly connected
    - R4 manages the BGP connection
    - R4 sends SYN with MSS of 1240 bytes
        - Interface MTU is not considered when peers are not directly connected and TCP path MTU Discovery is disabled
        - As per Cisco IOS XR design, 1280 bytes is considered TCP_DEFAULT_MTU
        - 1280 (TCP_DEFAULT_MTU) - 20 (minTCP_H) - 20 (minIP_H)
    - R1 sends SYN, ACK with MSS of 1240 bytes
        - Sends the lower of [Received MSS; Local initial MSS]
        - Received MSS 1240 bytes; Local initial MSS 1240 bytes
    - The lowest MSS value is used on both peers

TCP SYN sourced from R4:

<#root>

```
! – TCP SYN sourced from R4

194    434.274181    192.168.0.4 192.168.0.1 TCP    62     37740 → 179 [SYN] Seq=0 Win=16384 Len=0
```

MSS=1240

WS=1

```
Frame 194: 62 bytes on wire (496 bits), 62 bytes captured (496 bits) on interface 0
Ethernet II, Src: fa:16:3e:d7:7e:f6 (fa:16:3e:d7:7e:f6), Dst: fa:16:3e:8f:8f:54 (fa:16:3e:8f:8f:54)
Internet Protocol Version 4, Src: 192.168.0.4, Dst: 192.168.0.1
Transmission Control Protocol, Src Port: 37740, Dst Port: 179, Seq: 0, Len: 0
    Source Port: 37740
    Destination Port: 179
    [Stream index: 7]
    [TCP Segment Len: 0]
    Sequence number: 0    (relative sequence number)
    Acknowledgment number: 0
    Header Length: 28 bytes
    Flags: 0x002 (SYN)
    Window size value: 16384
    [Calculated window size: 16384]
    Checksum: 0x8643 [unverified]
    [Checksum Status: Unverified]
    Urgent pointer: 0
    Options: (8 bytes), Maximum segment size, Window scale, End of Option List (EOL)
        Maximum segment size: 1240 bytes
            Kind: Maximum Segment Size (2)
            Length: 4
```

MSS Value: 1240

```
        Window scale: 0 (multiply by 1)
        End of Option List (EOL)
```

TCP SYN, ACK sourced from R1:

<#root>

```
! - TCP SYN,ACK sourced from R1

195    434.277985     192.168.0.1 192.168.0.4 TCP    62     179 → 37740 [SYN, ACK] Seq=0 Ack=1 Win=16384 Le
```

MSS=1240

WS=1

```
Frame 195: 62 bytes on wire (496 bits), 62 bytes captured (496 bits) on interface 0
Ethernet II, Src: fa:16:3e:8f:8f:54 (fa:16:3e:8f:8f:54), Dst: fa:16:3e:d7:7e:f6 (fa:16:3e:d7:7e:f6)
Internet Protocol Version 4, Src: 192.168.0.1, Dst: 192.168.0.4
Transmission Control Protocol, Src Port: 179, Dst Port: 37740, Seq: 0, Ack: 1, Len: 0
    Source Port: 179
    Destination Port: 37740
    [Stream index: 7]
    [TCP Segment Len: 0]
    Sequence number: 0    (relative sequence number)
    Acknowledgment number: 1    (relative ack number)
    Header Length: 28 bytes
    Flags: 0x012 (SYN, ACK)
    Window size value: 16384
    [Calculated window size: 16384]
    Checksum: 0xd8f7 [unverified]
```

```
    [Checksum Status: Unverified]
    Urgent pointer: 0
    Options: (8 bytes), Maximum segment size, Window scale, End of Option List (EOL)
        Maximum segment size: 1240 bytes
            Kind: Maximum Segment Size (2)
            Length: 4
```

MSS Value: 1240

```
        Window scale: 0 (multiply by 1)
        End of Option List (EOL)
```

TCP session details as seen on R4 – ACTIVE:

<#root>

```
! - as seen on R4 - Active

RP/0/0/CPU0:R4#show tcp detail pcb 0x12154d3c
Fri Jan  8 12:32:41.096 UTC
================================================================
Connection state is ESTAB, I/O status: 0, socket status: 0
Established at Fri Jan  8 12:17:46 2021

PCB 0x12154d3c, SO 0x12154460, TCPCB 0x1215486c, vrfid 0x60000000,
Pak Prio: Medium, TOS: 192, TTL: 255, Hash index: 1577
Local host: 192.168.0.4, Local port: 37740 (Local App PID: 1052958)
Foreign host: 192.168.0.1, Foreign port: 179

Current send queue size in bytes: 0 (max 24576)
Current receive queue size in bytes: 0 (max 32768)  mis-ordered: 0 bytes
Current receive queue size in packets: 0 (max 0)

Timer          Starts    Wakeups         Next(msec)
Retrans            19          1                0
SendWnd             0          0                0
TimeWait            0          0                0
AckHold            16         15                0
KeepAlive           1          0                0
PmtuAger            0          0                0
GiveUp              0          0                0
Throttle            0          0                0

   iss: 2075436506  snduna: 2075436868  sndnxt: 2075436868
sndmax: 2075436868  sndwnd: 32460       sndcwnd: 3720
   irs: 4238127261  rcvnxt: 4238127623  rcvwnd: 32479   rcvadv: 4238160102

SRTT: 65 ms,  RTTO: 300 ms,  RTV: 40 ms,  KRTT: 0 ms
minRTT: 9 ms,  maxRTT: 229 ms

ACK hold time: 200 ms, Keepalive time: 0 sec, SYN waittime: 30 sec
Giveup time: 0 ms, Retransmission retries: 0, Retransmit forever: FALSE
Connect retries remaining: 30, connect retry interval: 30 secs

State flags: none
Feature flags: Win Scale, Nagle
Request flags: Win Scale
```

Datagrams (in bytes): MSS 1240, peer MSS 1240, min MSS 1240, max MSS 1240

```
Window scales: rcv 0, snd 0, request rcv 0, request snd 0
Timestamp option: recent 0, recent age 0, last ACK sent 0
Sack blocks {start, end}: none
Sack holes {start, end, dups, rxmit}: none

Socket options: SO_REUSEADDR, SO_REUSEPORT, SO_NBIO
Socket states: SS_ISCONNECTED, SS_PRIV
Socket receive buffer states: SB_DEL_WAKEUP
Socket send buffer states: SB_DEL_WAKEUP
Socket receive buffer: Low/High watermark 1/32768
Socket send buffer   : Low/High watermark 2048/24576, Notify threshold 0

PDU information:
 #PDU's in buffer: 0
FIB Lookup Cache:  IFH: 0x40  PD ctx: size: 0  data:
  Num Labels: 0  Label Stack:

RP/0/0/CPU0:R4#
```

TCP session details as seen on R1 – PASSIVE:

<#root>

```
! - as seen on R1 - Passive

RP/0/0/CPU0:R1#show tcp detail pcb 0x12155390
Fri Jan  8 12:23:52.041 UTC
==============================================================
Connection state is ESTAB, I/O status: 0, socket status: 0
Established at Fri Jan  8 12:17:43 2021

PCB 0x12155390, SO 0x121573e4, TCPCB 0x12156948, vrfid 0x60000000,
Pak Prio: Medium, TOS: 192, TTL: 255, Hash index: 1577
Local host: 192.168.0.1, Local port: 179 (Local App PID: 983326)
Foreign host: 192.168.0.4, Foreign port: 37740

Current send queue size in bytes: 0 (max 24576)
Current receive queue size in bytes: 0 (max 32768)  mis-ordered: 0 bytes
Current receive queue size in packets: 0 (max 0)

Timer          Starts    Wakeups         Next(msec)
Retrans             9          0                 0
SendWnd             0          0                 0
TimeWait            0          0                 0
AckHold             9          1                 0
KeepAlive           1          0                 0
PmtuAger            0          0                 0
GiveUp              0          0                 0
Throttle            0          0                 0

iss: 4238127261  snduna: 4238127471  sndnxt: 4238127471
sndmax: 4238127471  sndwnd: 32631       sndcwnd: 3720
irs: 2075436506  rcvnxt: 2075436716  rcvwnd: 32612   rcvadv: 2075469328

SRTT: 144 ms,  RTTO: 578 ms,  RTV: 434 ms,  KRTT: 0 ms
```

```
minRTT: 19 ms,  maxRTT: 239 ms

ACK hold time: 200 ms, Keepalive time: 0 sec, SYN waittime: 30 sec
Giveup time: 0 ms, Retransmission retries: 0, Retransmit forever: FALSE
Connect retries remaining: 0, connect retry interval: 0 secs

State flags: none
Feature flags: Win Scale, Nagle
Request flags: Win Scale

Datagrams (in bytes): MSS 1240, peer MSS 1240, min MSS 1240, max MSS 1240


Window scales: rcv 0, snd 0, request rcv 0, request snd 0
Timestamp option: recent 0, recent age 0, last ACK sent 0
Sack blocks {start, end}: none
Sack holes {start, end, dups, rxmit}: none

Socket options: SO_REUSEADDR, SO_REUSEPORT, SO_NBIO
Socket states: SS_ISCONNECTED, SS_PRIV
Socket receive buffer states: SB_DEL_WAKEUP
Socket send buffer states: SB_DEL_WAKEUP
Socket receive buffer: Low/High watermark 1/32768
Socket send buffer   : Low/High watermark 2048/24576, Notify threshold 0

PDU information:
#PDU's in buffer: 0
FIB Lookup Cache:  IFH: 0x40  PD ctx: size: 0  data:
Num Labels: 0  Label Stack:

RP/0/0/CPU0:R1#
```

TCP Peers not Directly Connected – Use TCP Options (MD5)

For non-directly connected peer scenario and in use of TCP MD5 authentication, there is no fundamental difference from the previous test cases or scenarios already described. As seen previously with TCP MD5 authentication, Cisco IOS XR considers additional overhead and initial MSS value reflects the same. Please refer to previous sections Use TCP Options – XR Active and Use TCP Options – XR Passive for additional details on TCP options influence on TCP MSS calculation.

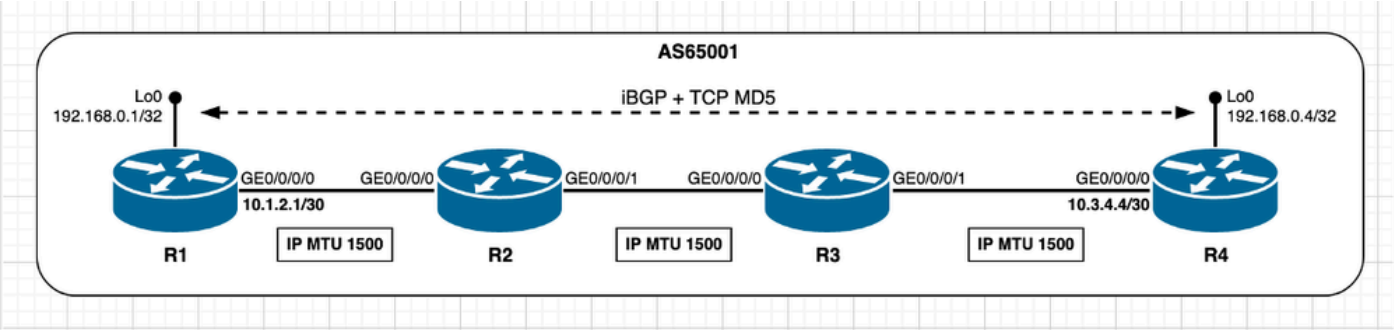


Image 2.7 – TCP Peers not directly connected – iBGP + TCP MD5.

TCP MSS calculation in this scenario can be summarized as follows:

- All nodes use a default IP MTU of 1500 bytes

- TCP path MTU discovery is disabled by default
- TCP Peers are not directly connected
    - R4 manages the BGP connection
    - Destination R1 is not directly connected
    - R4 sends SYN with MSS of 1216 bytes
        - Interface MTU is not considered when peers are not directly connected and TCP path MTU Discovery is disabled
        - As per design 1280 bytes is considered TCP_DEFAULT_MTU
        - 1280 (TCP_DEFAULT_MTU) - 20 (minTCP_H) - 20 (minIP_H) - 24 Bytes (IOS XR TCP Options Overhead)
    - R1 sends SYN, ACK with MSS of 1216 bytes
        - Sends the lower of [Received MSS; Local initial MSS]
        - Received MSS 1216 bytes; Local initial MSS 1240 bytes
    - The lowest MSS value is used on both peers

TCP SYN sourced from R4:

<#root>

```
! – TCP SYN sourced from R4

3425   3.691042       192.168.0.4 192.168.0.1 TCP    82     42135 → 179 [SYN] Seq=0 Win=16384 Len=0
```

MSS=1216

```
 WS=1

Frame 3425: 82 bytes on wire (656 bits), 82 bytes captured (656 bits) on interface 0
Ethernet II, Src: fa:16:3e:d7:7e:f6 (fa:16:3e:d7:7e:f6), Dst: fa:16:3e:8f:8f:54 (fa:16:3e:8f:8f:54)
Internet Protocol Version 4, Src: 192.168.0.4, Dst: 192.168.0.1
Transmission Control Protocol, Src Port: 42135, Dst Port: 179, Seq: 0, Len: 0
    Source Port: 42135
    Destination Port: 179
    [Stream index: 10]
    [TCP Segment Len: 0]
    Sequence number: 0    (relative sequence number)
    Acknowledgment number: 0
    Header Length: 48 bytes
    Flags: 0x002 (SYN)
    Window size value: 16384
    [Calculated window size: 16384]
    Checksum: 0xc503 [unverified]
    [Checksum Status: Unverified]
    Urgent pointer: 0
    Options: (28 bytes), Maximum segment size, Window scale, No-Operation (NOP),
```

TCP MD5 signature

```
, End of Option List (EOL)
        Maximum segment size: 1216 bytes
            Kind: Maximum Segment Size (2)
            Length: 4
```

MSS Value: 1216

```
        Window scale: 0 (multiply by 1)
        No-Operation (NOP)
        TCP MD5 signature
```

```
        End of Option List (EOL)
```

TCP SYN, ACK sourced from R1:

<#root>

```
! - TCP SYN,ACK sourced from R1

3426   0.004186       192.168.0.1 192.168.0.4 TCP    82     179 → 42135 [SYN, ACK] Seq=0 Ack=1 Win=16384 Le
```

MSS=1216

```
 WS=1

Frame 3426: 82 bytes on wire (656 bits), 82 bytes captured (656 bits) on interface 0
Ethernet II, Src: fa:16:3e:8f:8f:54 (fa:16:3e:8f:8f:54), Dst: fa:16:3e:d7:7e:f6 (fa:16:3e:d7:7e:f6)
Internet Protocol Version 4, Src: 192.168.0.1, Dst: 192.168.0.4
Transmission Control Protocol, Src Port: 179, Dst Port: 42135, Seq: 0, Ack: 1, Len: 0
    Source Port: 179
    Destination Port: 42135
    [Stream index: 10]
    [TCP Segment Len: 0]
    Sequence number: 0    (relative sequence number)
    Acknowledgment number: 1    (relative ack number)
    Header Length: 48 bytes
    Flags: 0x012 (SYN, ACK)
    Window size value: 16384
    [Calculated window size: 16384]
    Checksum: 0xbb05 [unverified]
    [Checksum Status: Unverified]
    Urgent pointer: 0
    Options: (28 bytes), Maximum segment size, Window scale, No-Operation (NOP),
```

TCP MD5 signature

```
, End of Option List (EOL)
        Maximum segment size: 1216 bytes
            Kind: Maximum Segment Size (2)
            Length: 4
```

MSS Value: 1216

```
        Window scale: 0 (multiply by 1)
        No-Operation (NOP)
        TCP MD5 signature
        End of Option List (EOL)
```

TCP session details as seen on R4 – ACTIVE:

<#root>

```
! - as seen from R4 - Active

RP/0/0/CPU0:R4#show tcp detail pcb 0x12154490
Tue Jan 12 14:37:32.097 UTC
==============================================================
```

```
Connection state is ESTAB, I/O status: 0, socket status: 0
Established at Tue Jan 12 14:27:42 2021

PCB 0x12154490, SO 0x12155014, TCPCB 0x12155a84, vrfid 0x60000000,
Pak Prio: Medium, TOS: 192, TTL: 255, Hash index: 1876
Local host: 192.168.0.4, Local port: 42135 (Local App PID: 1052958)
Foreign host: 192.168.0.1, Foreign port: 179

Current send queue size in bytes: 0 (max 24576)
Current receive queue size in bytes: 0 (max 32768)  mis-ordered: 0 bytes
Current receive queue size in packets: 0 (max 0)

Timer          Starts    Wakeups         Next(msec)
Retrans            14          1                 0
SendWnd             0          0                 0
TimeWait            0          0                 0
AckHold            11          9                 0
KeepAlive           1          0                 0
PmtuAger            0          0                 0
GiveUp              0          0                 0
Throttle            0          0                 0


   iss: 3124761989  snduna: 3124763317  sndnxt: 3124763317
sndmax: 3124763317  sndwnd: 32711       sndcwnd: 3648
   irs: 1090344992  rcvnxt: 1090346320  rcvwnd: 32730   rcvadv: 1090379050

SRTT: 28 ms,  RTTO: 300 ms,  RTV: 57 ms,  KRTT: 0 ms
minRTT: 9 ms,  maxRTT: 229 ms

ACK hold time: 200 ms, Keepalive time: 0 sec, SYN waittime: 30 sec
Giveup time: 0 ms, Retransmission retries: 0, Retransmit forever: FALSE
Connect retries remaining: 30, connect retry interval: 30 secs

State flags: none
Feature flags: MD5, Win Scale, Nagle
Request flags: Win Scale


Datagrams (in bytes): MSS 1216, peer MSS 1216, min MSS 1216, max MSS 1216



Window scales: rcv 0, snd 0, request rcv 0, request snd 0
Timestamp option: recent 0, recent age 0, last ACK sent 0
Sack blocks {start, end}: none
Sack holes {start, end, dups, rxmit}: none

Socket options: SO_REUSEADDR, SO_REUSEPORT, SO_NBIO
Socket states: SS_ISCONNECTED, SS_PRIV
Socket receive buffer states: SB_DEL_WAKEUP
Socket send buffer states: SB_DEL_WAKEUP
Socket receive buffer: Low/High watermark 1/32768
Socket send buffer   : Low/High watermark 2048/24576, Notify threshold 0

PDU information:
 #PDU's in buffer: 0
FIB Lookup Cache:  IFH: 0x40  PD ctx: size: 0  data:
  Num Labels: 0  Label Stack:

RP/0/0/CPU0:R4#
```

TCP session details as seen on R1 – PASSIVE:

<#root>

```
! - as seen from R1 - Passive

RP/0/0/CPU0:R1#show tcp detail pcb 0x12168df4
Tue Jan 12 14:36:38.860 UTC
================================================================
Connection state is ESTAB, I/O status: 0, socket status: 0
Established at Tue Jan 12 14:27:32 2021

PCB 0x12168df4, SO 0x12156bf8, TCPCB 0x12157a44, vrfid 0x60000000,
Pak Prio: Medium, TOS: 192, TTL: 255, Hash index: 1876
Local host: 192.168.0.1, Local port: 179 (Local App PID: 983326)
Foreign host: 192.168.0.4, Foreign port: 42135

Current send queue size in bytes: 0 (max 24576)
Current receive queue size in bytes: 0 (max 32768)  mis-ordered: 0 bytes
Current receive queue size in packets: 0 (max 0)

Timer          Starts    Wakeups         Next(msec)
Retrans            12          0                 0
SendWnd             0          0                 0
TimeWait            0          0                 0
AckHold            12          1                 0
KeepAlive           1          0                 0
PmtuAger            0          0                 0
GiveUp              0          0                 0
Throttle            0          0                 0

   iss: 1090344992  snduna: 1090346320  sndnxt: 1090346320
sndmax: 1090346320  sndwnd: 32730       sndcwnd: 3648
   irs: 3124761989  rcvnxt: 3124763317  rcvwnd: 32711   rcvadv: 3124796028

SRTT: 150 ms,  RTTO: 558 ms,  RTV: 408 ms,  KRTT: 0 ms
minRTT: 19 ms,  maxRTT: 239 ms

ACK hold time: 200 ms, Keepalive time: 0 sec, SYN waittime: 30 sec
Giveup time: 0 ms, Retransmission retries: 0, Retransmit forever: FALSE
Connect retries remaining: 0, connect retry interval: 0 secs

State flags: none
Feature flags: MD5, Win Scale, Nagle
Request flags: Win Scale


Datagrams (in bytes): MSS 1216, peer MSS 1216, min MSS 1240, max MSS 1240



Window scales: rcv 0, snd 0, request rcv 0, request snd 0
Timestamp option: recent 0, recent age 0, last ACK sent 0
Sack blocks {start, end}: none
Sack holes {start, end, dups, rxmit}: none

Socket options: SO_REUSEADDR, SO_REUSEPORT, SO_NBIO
Socket states: SS_ISCONNECTED, SS_PRIV
Socket receive buffer states: SB_DEL_WAKEUP
Socket send buffer states: SB_DEL_WAKEUP
Socket receive buffer: Low/High watermark 1/32768
```

```
Socket send buffer   : Low/High watermark 2048/24576, Notify threshold 0

PDU information:
 #PDU's in buffer: 0
FIB Lookup Cache:  IFH: 0x40  PD ctx: size: 0  data:
  Num Labels: 0  Label Stack:

RP/0/0/CPU0:R1#
```

TCP Peers not Directly Connected – Path Segment has Lower IP MTU

With the next scenario, the goal is to observe and conclude on what happens if there is an intermediate path segment with a lower IP MTU while in default condition, this means TCP PMTUD is disabled. Please refer to this image.

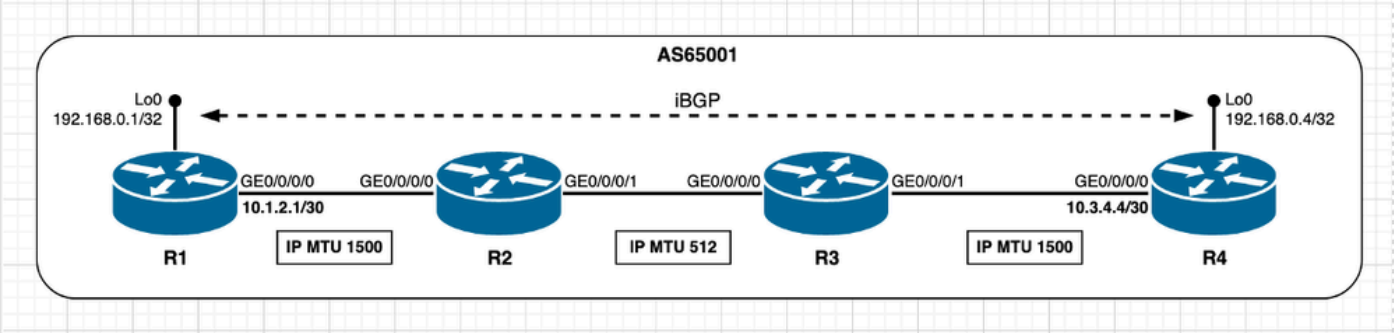


Image 2.8 – R2/R3 Path Segment has lower IP MTU.

As an initial scenario consider that BGP information is minimal, that is that whatever is required to be exchanged between BGP peers can be accomplished with IP packets that fit under the minimum path MTU of 512 bytes. With this assumption, MSS calculation happens as described in section **TCP Peers not Directly Connected**. Both R1 and R4 select an MSS value of 1240 bytes.

TCP session details as seen on R4 – ACTIVE:

<#root>

```
! - as seen from R4 - Active

RP/0/0/CPU0:R4#show tcp detail pcb 0x15390fe8
==============================================================
Connection state is ESTAB, I/O status: 0, socket status: 0
Established at Wed May 12 12:09:48 2021

PCB 0x15390fe8, SO 0x15391a7c, TCPCB 0x15391368, vrfid 0x60000000,
Pak Prio: Medium, TOS: 192, TTL: 255, Hash index: 835
Local host: 192.168.0.4, Local port: 39046 (Local App PID: 1196319)
Foreign host: 192.168.0.1, Foreign port: 179
(Local App PID/instance/SPL_APP_ID: 1196319/1/0)

Current send queue size in bytes: 0 (max 24576)
Current receive queue size in bytes: 0 (max 32768)  mis-ordered: 0 bytes
Current receive queue size in packets: 0 (max 0)

Timer          Starts    Wakeups         Next(msec)
Retrans          1267          1                  0
SendWnd             0          0                  0
```

```
TimeWait            0          0               0
AckHold          1280       1235               0
KeepAlive           1          0               0
PmtuAger            0          0               0
GiveUp              0          0               0
Throttle            0          0               0

   iss: 1991226354  snduna: 1991250450  sndnxt: 1991250450
sndmax: 1991250450  sndwnd: 32578       sndcwnd: 2480
   irs: 4276699304  rcvnxt: 4276746737  rcvwnd: 31568   rcvadv: 4276778305

SRTT: 213 ms,  RTTO: 300 ms,  RTV: 54 ms,  KRTT: 0 ms
minRTT: 9 ms,  maxRTT: 269 ms

ACK hold time: 200 ms, Keepalive time: 0 sec, SYN waittime: 30 sec
Giveup time: 0 ms, Retransmission retries: 0, Retransmit forever: FALSE
Connect retries remaining: 10, connect retry interval: 30 secs

State flags: none
Feature flags: Win Scale, Nagle
Request flags: Win Scale
```

Datagrams (in bytes): MSS 1240, peer MSS 1240, min MSS 1240, max MSS 1240

<snip>

TCP session details as seen on R1 – PASSIVE:

<#root>

```
! - as seen from R1 - Passive

RP/0/0/CPU0:R1#show tcp detail pcb 0x15393770
==============================================================
Connection state is ESTAB, I/O status: 0, socket status: 0
Established at Wed May 12 12:09:46 2021

PCB 0x15393770, SO 0x15392224, TCPCB 0x153928cc, vrfid 0x60000000,
Pak Prio: Medium, TOS: 192, TTL: 255, Hash index: 835
Local host: 192.168.0.1, Local port: 179 (Local App PID: 1192224)
Foreign host: 192.168.0.4, Foreign port: 39046
(Local App PID/instance/SPL_APP_ID: 1192224/1/0)

Current send queue size in bytes: 0 (max 24576)
Current receive queue size in bytes: 0 (max 32768)  mis-ordered: 0 bytes
Current receive queue size in packets: 0 (max 0)

Timer          Starts    Wakeups         Next(msec)
Retrans          1280          0               0
SendWnd             0          0               0
TimeWait            0          0               0
AckHold          1264       1213               0
KeepAlive           1          0               0
PmtuAger            0          0               0
GiveUp              0          0               0
Throttle            0          0               0

   iss: 4276699304  snduna: 4276746718  sndnxt: 4276746718
```

```
sndmax: 4276746718  sndwnd: 31587       sndcwnd: 3720
   irs: 1991226354  rcvnxt: 1991250431  rcvwnd: 32597   rcvadv: 1991283028

SRTT: 202 ms,  RTTO: 355 ms,  RTV: 153 ms,  KRTT: 0 ms
minRTT: 9 ms,  maxRTT: 309 ms

ACK hold time: 200 ms, Keepalive time: 0 sec, SYN waittime: 30 sec
Giveup time: 0 ms, Retransmission retries: 0, Retransmit forever: FALSE
Connect retries remaining: 0, connect retry interval: 0 secs

State flags: none
Feature flags: Win Scale, Nagle
Request flags: Win Scale
```

Datagrams (in bytes): MSS 1240, peer MSS 1240, min MSS 1240, max MSS 1240

<snip>

With the BGP session established now, consider that a BGP Update message with a size higher than the minimum path MTU of 512 bytes is triggered. As can be observed from the outputs, Cisco IOS XR does not set the df-bit with the BGP update message which means that BGP information is transmitted at the expense of packet fragmentation on intermediate nodes.

BGP Update sourced by R1 – PASSIVE:

<#root>

```
! - as seen from R1 - Passive - BGP UPDATE
! - Note Total Length of 1097 bytes higher than the IP MTU value of 512 bytes at R2-R3 path segment

23     3.450878       192.168.0.1 192.168.0.4 BGP    1111   UPDATE Message

Frame 23: 1111 bytes on wire (8888 bits), 1111 bytes captured (8888 bits) on interface 0
Ethernet II, Src: fa:16:3e:42:18:05 (fa:16:3e:42:18:05), Dst: fa:16:3e:5c:f1:80 (fa:16:3e:5c:f1:80)
Internet Protocol Version 4, Src: 192.168.0.1, Dst: 192.168.0.4
    0100 .... = Version: 4
    .... 0101 = Header Length: 20 bytes (5)
    Differentiated Services Field: 0xc0 (DSCP: CS6, ECN: Not-ECT)
```

Total Length: 1097

```
    Identification: 0x5841 (22593)
    Flags: 0x00
        0... .... = Reserved bit: Not set
        .0.. .... = Don't fragment: Not set
        ..0. .... = More fragments: Not set
    Fragment offset: 0
    Time to live: 255
    Protocol: TCP (6)
    Header checksum: 0x54a4 [validation disabled]
    [Header checksum status: Unverified]
    Source: 192.168.0.1
    Destination: 192.168.0.4
    [Source GeoIP: Unknown]
    [Destination GeoIP: Unknown]
Transmission Control Protocol, Src Port: 179, Dst Port: 39046, Seq: 20, Ack: 20, Len: 1057
```

```
Border Gateway Protocol - UPDATE Message
    Marker: ffffffffffffffffffffffffffffffff
    Length: 1057
    Type: UPDATE Message (2)
    Withdrawn Routes Length: 0
    Total Path Attribute Length: 1034
    Path attributes
        Path Attribute - MP_REACH_NLRI
        Path Attribute - ORIGIN: INCOMPLETE
        Path Attribute - AS_PATH: empty
        Path Attribute - MULTI_EXIT_DISC: 0
        Path Attribute - LOCAL_PREF: 100
```

Fragmentation of BGP Update message sourced by node R1 takes place at node R2 as can be observed by traffic capture done at R2 interface GE0/0/0/1.

IP fragmentation at node R2:

<#root>

```
! - as seen from R2 - GE0/0/0/1
! - Node R2 fragments original packet in three distinct packets

4      1.334852       192.168.0.1 192.168.0.4 BGP    522    UPDATE Message
5      0.000289       192.168.0.1 192.168.0.4 IPv4   522    Fragmented IP protocol (proto=TCP 6, off=48
6      0.000122       192.168.0.1 192.168.0.4 IPv4   135    Fragmented IP protocol (proto=TCP 6, off=97

! - Captured frame details

Frame 4: 522 bytes on wire (4176 bits), 522 bytes captured (4176 bits) on interface 0
Ethernet II, Src: fa:16:3e:61:25:f0 (fa:16:3e:61:25:f0), Dst: fa:16:3e:23:ab:27 (fa:16:3e:23:ab:27)
Internet Protocol Version 4, Src: 192.168.0.1, Dst: 192.168.0.4
    0100 .... = Version: 4
    .... 0101 = Header Length: 20 bytes (5)
    Differentiated Services Field: 0xc0 (DSCP: CS6, ECN: Not-ECT)

Total Length: 508

Identification: 0x7b41 (31553)

    Flags: 0x01 (More Fragments)
        0... .... = Reserved bit: Not set
        .0.. .... = Don't fragment: Not set

..1. .... = More fragments: Set

Fragment offset: 0

    Time to live: 254
    Protocol: TCP (6)
    Header checksum: 0x14f1 [validation disabled]
```

```
    [Header checksum status: Unverified]
    Source: 192.168.0.1
    Destination: 192.168.0.4
    [Source GeoIP: Unknown]
    [Destination GeoIP: Unknown]
Transmission Control Protocol, Src Port: 179, Dst Port: 39046, Seq: 4276759681, Ack: 1991250830
Border Gateway Protocol - UPDATE Message
<snip>

Frame 5: 522 bytes on wire (4176 bits), 522 bytes captured (4176 bits) on interface 0
Ethernet II, Src: fa:16:3e:61:25:f0 (fa:16:3e:61:25:f0), Dst: fa:16:3e:23:ab:27 (fa:16:3e:23:ab:27)
Internet Protocol Version 4, Src: 192.168.0.1, Dst: 192.168.0.4
    0100 .... = Version: 4
    .... 0101 = Header Length: 20 bytes (5)
    Differentiated Services Field: 0xc0 (DSCP: CS6, ECN: Not-ECT)


Total Length: 508



Identification: 0x7b41 (31553)


    Flags: 0x01 (More Fragments)
        0... .... = Reserved bit: Not set
        .0.. .... = Don't fragment: Not set


..1. .... = More fragments: Set



Fragment offset: 488


    Time to live: 254
    Protocol: TCP (6)
    Header checksum: 0x14b4 [validation disabled]
    [Header checksum status: Unverified]
    Source: 192.168.0.1
    Destination: 192.168.0.4
    [Source GeoIP: Unknown]
    [Destination GeoIP: Unknown]
Data (488 bytes)
<snip>

Frame 6: 135 bytes on wire (1080 bits), 135 bytes captured (1080 bits) on interface 0
Ethernet II, Src: fa:16:3e:61:25:f0 (fa:16:3e:61:25:f0), Dst: fa:16:3e:23:ab:27 (fa:16:3e:23:ab:27)
Internet Protocol Version 4, Src: 192.168.0.1, Dst: 192.168.0.4
    0100 .... = Version: 4
    .... 0101 = Header Length: 20 bytes (5)
    Differentiated Services Field: 0xc0 (DSCP: CS6, ECN: Not-ECT)


Total Length: 121



Identification: 0x7b41 (31553)
```

```
    Flags: 0x00
        0... .... = Reserved bit: Not set
        .0.. .... = Don't fragment: Not set

..0. .... = More fragments: Not set



Fragment offset: 976


    Time to live: 254
    Protocol: TCP (6)
    Header checksum: 0x35fa [validation disabled]
    [Header checksum status: Unverified]
    Source: 192.168.0.1
    Destination: 192.168.0.4
    [Source GeoIP: Unknown]
    [Destination GeoIP: Unknown]
Data (101 bytes)
<snip>
```

Scenarios – TCP PMTUD Enabled

Enable PMTUD

Once PMTUD is enabled, then irrespective if peers are directly or non-directly connected, MSS initial calculation always take into consideration the egress interface IP MTU.

This scenario provides insight on the expected behavior when PMTUD is enabled. Here, Cisco IOS XR node R4 plays the active role, manages the TCP connection, and opens the TCP session with Cisco IOS XR node R1 on destination port 179. Both nodes use the default IP MTU values on their interfaces.

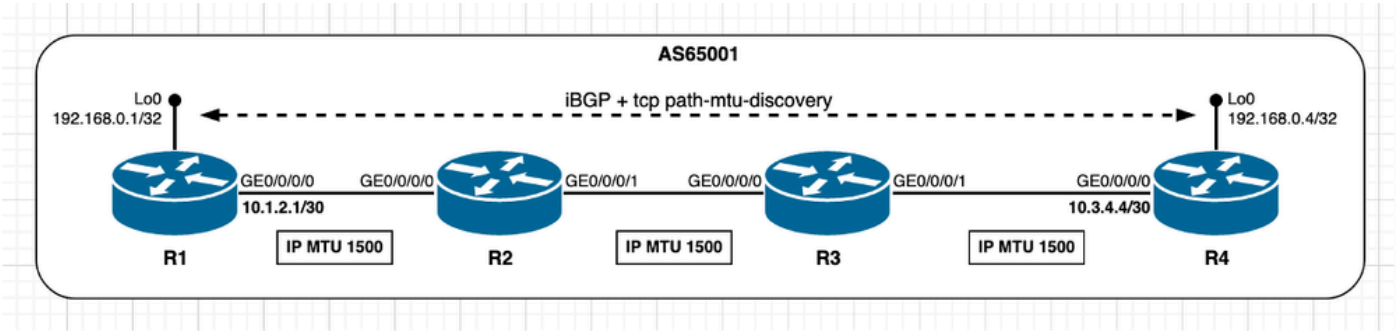


Image 3.1 – TCP PMTUD Enabled.

MSS calculation in this scenario can be summarized as follows:

- All nodes use a default IP MTU of 1500 bytes
- TCP Path MTU discovery is enabled
- TCP peers are not directly connected
    - R4 manages the BGP connection
    - R4 sends SYN with MSS of 1460 bytes
        - 1500 (Interface IP MTU) - 20 (minTCP_H) - 20 (minIP_H)
    - R1 sends SYN, ACK with MSS of 1460 bytes

    - Sends the lower of [Received MSS; Local initial MSS]
    - Received MSS 1460 bytes; Local initial MSS 1460 bytes
  - The lowest MSS value is used on both peers

In order to highlight the behavior change introduced from enable PMTUD, the next outputs illustrate the sequence of events:

1. The initial state of established TCP session in default scenario of PMTUD disabled;
2. PMTUD is configured and enabled on both TCP peers R4 and R1;
3. The TCP session is restarted, MSS calculation takes place, and is influenced by TCP PMTUD.

As seen on R4 – ACTIVE – TCP PMTUD disabled (default):

<#root>

```
! - as seen on R4 - Active
! - TCP path mtu discovery disabled (default)
! - TCP session initial state

RP/0/0/CPU0:R4#show tcp detail pcb 0x121536c8
Fri Jan  8 16:06:30.237 UTC
==========================================================
Connection state is ESTAB, I/O status: 0, socket status: 0
Established at Fri Jan  8 16:05:15 2021

PCB 0x121536c8, SO 0x12155370, TCPCB 0x12154f64, vrfid 0x60000000,
Pak Prio: Medium, TOS: 192, TTL: 255, Hash index: 376
Local host: 192.168.0.4, Local port: 20155 (Local App PID: 1052958)
Foreign host: 192.168.0.1, Foreign port: 179

Current send queue size in bytes: 0 (max 24576)
Current receive queue size in bytes: 0 (max 32768)  mis-ordered: 0 bytes
Current receive queue size in packets: 0 (max 0)

Timer          Starts    Wakeups         Next(msec)
Retrans             6          1                  0
SendWnd             0          0                  0
TimeWait            0          0                  0
AckHold             3          2                  0
KeepAlive           1          0                  0
PmtuAger            0          0                  0
GiveUp              0          0                  0
Throttle            0          0                  0

   iss: 357400981   snduna: 357401257   sndnxt: 357401257
sndmax: 357401257   sndwnd: 32546       sndcwnd: 3720
   irs: 524019443   rcvnxt: 524019719   rcvwnd: 32565   rcvadv: 524052284

SRTT: 72 ms,  RTTO: 416 ms,  RTV: 344 ms,  KRTT: 0 ms
minRTT: 19 ms,  maxRTT: 229 ms

ACK hold time: 200 ms, Keepalive time: 0 sec, SYN waittime: 30 sec
Giveup time: 0 ms, Retransmission retries: 0, Retransmit forever: FALSE
Connect retries remaining: 30, connect retry interval: 30 secs

State flags: none
Feature flags: Win Scale, Nagle
Request flags: Win Scale
```

Datagrams (in bytes): MSS 1240, peer MSS 1240, min MSS 1240, max MSS 1240

```
Window scales: rcv 0, snd 0, request rcv 0, request snd 0
Timestamp option: recent 0, recent age 0, last ACK sent 0
Sack blocks {start, end}: none
Sack holes {start, end, dups, rxmit}: none

Socket options: SO_REUSEADDR, SO_REUSEPORT, SO_NBIO
Socket states: SS_ISCONNECTED, SS_PRIV
Socket receive buffer states: SB_DEL_WAKEUP
Socket send buffer states: SB_DEL_WAKEUP
Socket receive buffer: Low/High watermark 1/32768
Socket send buffer   : Low/High watermark 2048/24576, Notify threshold 0

PDU information:
 #PDU's in buffer: 0
FIB Lookup Cache:  IFH: 0x40  PD ctx: size: 0  data:
  Num Labels: 0  Label Stack:

RP/0/0/CPU0:R4#
```

As seen on R1 – PASSIVE – TCP PMTUD disabled (default):

<#root>

```
! - as seen on R1 - Passive
! - TCP path mtu discovery disabled (default)
! - TCP session initial state

RP/0/0/CPU0:R1#show tcp detail pcb 0x12157020
Fri Jan  8 16:05:52.868 UTC
==============================================================
Connection state is ESTAB, I/O status: 0, socket status: 0
Established at Fri Jan  8 16:05:12 2021

PCB 0x12157020, SO 0x121565ac, TCPCB 0x121560ec, vrfid 0x60000000,
Pak Prio: Medium, TOS: 192, TTL: 255, Hash index: 376
Local host: 192.168.0.1, Local port: 179 (Local App PID: 983326)
Foreign host: 192.168.0.4, Foreign port: 20155

Current send queue size in bytes: 0 (max 24576)
Current receive queue size in bytes: 0 (max 32768)  mis-ordered: 0 bytes
Current receive queue size in packets: 0 (max 0)

Timer          Starts    Wakeups         Next(msec)
Retrans             3          0                 0
SendWnd             0          0                 0
TimeWait            0          0                 0
AckHold             3          1                 0
KeepAlive           1          0                 0
PmtuAger            0          0                 0
GiveUp              0          0                 0
Throttle            0          0                 0

   iss: 524019443   snduna: 524019700   sndnxt: 524019700
sndmax: 524019700   sndwnd: 32584       sndcwnd: 3720
   irs: 357400981   rcvnxt: 357401238   rcvwnd: 32565   rcvadv: 357433803
```

```
SRTT: 46 ms,  RTTO: 300 ms,  RTV: 249 ms,  KRTT: 0 ms
minRTT: 19 ms,  maxRTT: 239 ms

ACK hold time: 200 ms, Keepalive time: 0 sec, SYN waittime: 30 sec
Giveup time: 0 ms, Retransmission retries: 0, Retransmit forever: FALSE
Connect retries remaining: 0, connect retry interval: 0 secs

State flags: none
Feature flags: Win Scale, Nagle
Request flags: Win Scale
```

Datagrams (in bytes): MSS 1240, peer MSS 1240, min MSS 1240, max MSS 1240

```
Window scales: rcv 0, snd 0, request rcv 0, request snd 0
Timestamp option: recent 0, recent age 0, last ACK sent 0
Sack blocks {start, end}: none
Sack holes {start, end, dups, rxmit}: none

Socket options: SO_REUSEADDR, SO_REUSEPORT, SO_NBIO
Socket states: SS_ISCONNECTED, SS_PRIV
Socket receive buffer states: SB_DEL_WAKEUP
Socket send buffer states: SB_DEL_WAKEUP
Socket receive buffer: Low/High watermark 1/32768
Socket send buffer   : Low/High watermark 2048/24576, Notify threshold 0

PDU information:
 #PDU's in buffer: 0
FIB Lookup Cache:  IFH: 0x40  PD ctx: size: 0  data:
  Num Labels: 0  Label Stack:

RP/0/0/CPU0:R1#
```

As seen on R4 – ACTIVE – TCP PMTUD enabled:

<#root>

```
! - 'debug tcp pmtud' output on R4
! - tcp path mtu discovery enabled and uses default Path MTU aging timer (10 min / 600000 msec)

RP/0/0/CPU0:Jan  8 16:09:28.285 : tcp[399]: [t21] Try to enable path MTU discovery(neww age timer: 10 m
RP/0/0/CPU0:Jan  8 16:09:28.285 : tcp[399]: [t21] Path mtu is ON (age-timer: 10)

! - as seen on R4 - Active
! - TCP PMTUD is enabled

RP/0/0/CPU0:R4#show tcp detail pcb 0x121536c8
Fri Jan  8 16:11:00.138 UTC
================================================================
Connection state is ESTAB, I/O status: 0, socket status: 0
Established at Fri Jan  8 16:05:15 2021

PCB 0x121536c8, SO 0x12155370, TCPCB 0x12154f64, vrfid 0x60000000,
Pak Prio: Medium, TOS: 192, TTL: 255, Hash index: 376
Local host: 192.168.0.4, Local port: 20155 (Local App PID: 1052958)
Foreign host: 192.168.0.1, Foreign port: 179
```

```
Current send queue size in bytes: 0 (max 24576)
Current receive queue size in bytes: 0 (max 32768)  mis-ordered: 0 bytes
Current receive queue size in packets: 0 (max 0)

Timer          Starts    Wakeups         Next(msec)
Retrans            10          1                0
SendWnd             0          0                0
TimeWait            0          0                0
AckHold             7          4                0
KeepAlive           1          0                0

PmtuAger            1          0           508096


GiveUp              0          0                0
Throttle            0          0                0

   iss: 357400981   snduna: 357401333   sndnxt: 357401333
sndmax: 357401333   sndwnd: 32470       sndcwnd: 3720
   irs: 524019443   rcvnxt: 524019795   rcvwnd: 32489   rcvadv: 524052284

SRTT: 116 ms,  RTTO: 578 ms,  RTV: 462 ms,  KRTT: 0 ms
minRTT: 9 ms,  maxRTT: 229 ms

ACK hold time: 200 ms, Keepalive time: 0 sec, SYN waittime: 30 sec
Giveup time: 0 ms, Retransmission retries: 0, Retransmit forever: FALSE
Connect retries remaining: 30, connect retry interval: 30 secs

State flags: PMTU ager
Feature flags: Win Scale, Nagle,

Path MTU


Request flags: Win Scale


Datagrams (in bytes): MSS 1240, peer MSS 1240, min MSS 1240, max MSS 1240



Window scales: rcv 0, snd 0, request rcv 0, request snd 0
Timestamp option: recent 0, recent age 0, last ACK sent 0
Sack blocks {start, end}: none
Sack holes {start, end, dups, rxmit}: none

Socket options: SO_REUSEADDR, SO_REUSEPORT, SO_NBIO
Socket states: SS_ISCONNECTED, SS_PRIV
Socket receive buffer states: SB_DEL_WAKEUP
Socket send buffer states: SB_DEL_WAKEUP
Socket receive buffer: Low/High watermark 1/32768
Socket send buffer   : Low/High watermark 2048/24576, Notify threshold 0

PDU information:
 #PDU's in buffer: 0
FIB Lookup Cache:  IFH: 0x40  PD ctx: size: 0  data:
  Num Labels: 0  Label Stack:

RP/0/0/CPU0:R4#
```

As seen on R1 – PASSIVE – TCP PMTUD enabled:

<#root>

```
! - 'debug tcp pmtud' output on R1
! - tcp path mtu discovery is enabled and uses default Path MTU aging timer (10 min / 600000 msec)

RP/0/0/CPU0:Jan  8 16:09:25.214 : tcp[399]: [t21] Try to enable path MTU discovery(neww age timer: 10 m
RP/0/0/CPU0:Jan  8 16:09:25.214 : tcp[399]: [t21] Path mtu is ON (age-timer: 10)

! - as seen on R1 - Passive
! - TCP PMTUD is enabled

RP/0/0/CPU0:R1#show tcp detail pcb 0x12157020
Fri Jan  8 16:10:03.101 UTC
================================================================
Connection state is ESTAB, I/O status: 0, socket status: 0
Established at Fri Jan  8 16:05:12 2021

PCB 0x12157020, SO 0x121565ac, TCPCB 0x121560ec, vrfid 0x60000000,
Pak Prio: Medium, TOS: 192, TTL: 255, Hash index: 376
Local host: 192.168.0.1, Local port: 179 (Local App PID: 983326)
Foreign host: 192.168.0.4, Foreign port: 20155

Current send queue size in bytes: 0 (max 24576)
Current receive queue size in bytes: 0 (max 32768)  mis-ordered: 0 bytes
Current receive queue size in packets: 0 (max 0)

Timer          Starts    Wakeups         Next(msec)
Retrans             7          0                0
SendWnd             0          0                0
TimeWait            0          0                0
AckHold             7          4                0
KeepAlive           1          0                0

PmtuAger            1          0           562042


GiveUp              0          0                0
Throttle            0          0                0

   iss: 524019443   snduna: 524019776   sndnxt: 524019776
sndmax: 524019776   sndwnd: 32508       sndcwnd: 3720
   irs: 357400981   rcvnxt: 357401314   rcvwnd: 32489   rcvadv: 357433803

SRTT: 95 ms,  RTTO: 528 ms,  RTV: 433 ms,  KRTT: 0 ms
minRTT: 19 ms,  maxRTT: 239 ms

ACK hold time: 200 ms, Keepalive time: 0 sec, SYN waittime: 30 sec
Giveup time: 0 ms, Retransmission retries: 0, Retransmit forever: FALSE
Connect retries remaining: 0, connect retry interval: 0 secs

State flags: PMTU ager
Feature flags: Win Scale, Nagle,

Path MTU


Request flags: Win Scale


Datagrams (in bytes): MSS 1240, peer MSS 1240, min MSS 1240, max MSS 1240



Window scales: rcv 0, snd 0, request rcv 0, request snd 0
```

```
Timestamp option: recent 0, recent age 0, last ACK sent 0
Sack blocks {start, end}: none
Sack holes {start, end, dups, rxmit}: none

Socket options: SO_REUSEADDR, SO_REUSEPORT, SO_NBIO
Socket states: SS_ISCONNECTED, SS_PRIV
Socket receive buffer states: SB_DEL_WAKEUP
Socket send buffer states: SB_DEL_WAKEUP
Socket receive buffer: Low/High watermark 1/32768
Socket send buffer   : Low/High watermark 2048/24576, Notify threshold 0

PDU information:
 #PDU's in buffer: 0
FIB Lookup Cache:  IFH: 0x40  PD ctx: size: 0  data:
  Num Labels: 0  Label Stack:

RP/0/0/CPU0:R1#
```

Note the PMTU ager timer behavior:

```
! - Note PmtuAger timer initial value is 10min
! - but after initial interval expires then it expires every 2min
! - As seen from 'debug tcp pmtud' output
! - TCP PMTUD is enabled

RP/0/0/CPU0:Jan  8 16:09:25.214 : tcp[399]: [t21] Try to enable path MTU discovery(neww age timer: 10 m
RP/0/0/CPU0:Jan  8 16:09:25.214 : tcp[399]: [t21] Path mtu is ON (age-timer: 10)
RP/0/0/CPU0:Jan  8 16:19:25.233 : tcp[399]: [t21] PCB 0x12157020: Trying next higher MTU: 1240
RP/0/0/CPU0:Jan  8 16:21:25.245 : tcp[399]: [t21] PCB 0x12157020: Trying next higher MTU: 1240
RP/0/0/CPU0:Jan  8 16:23:25.256 : tcp[399]: [t21] PCB 0x12157020: Trying next higher MTU: 1240
```

As seen on R4 – ACTIVE – BGP Session restart – TCP SYN:

<#root>

```
! - Once BGP session is cleared
! - TCP SYN sourced from R4 - Active
! - MSS calculation takes place and is influenced by TCP PMTUD

2734   4.810311       192.168.0.4 192.168.0.1 TCP    62     32077 → 179 [SYN] Seq=0 Win=16384 Len=0
```

MSS=1460

```
 WS=1

Frame 2734: 62 bytes on wire (496 bits), 62 bytes captured (496 bits) on interface 0
Ethernet II, Src: fa:16:3e:d7:7e:f6 (fa:16:3e:d7:7e:f6), Dst: fa:16:3e:8f:8f:54 (fa:16:3e:8f:8f:54)
Internet Protocol Version 4, Src: 192.168.0.4, Dst: 192.168.0.1
Transmission Control Protocol, Src Port: 32077, Dst Port: 179, Seq: 0, Len: 0
    Source Port: 32077
    Destination Port: 179
    [Stream index: 25]
    [TCP Segment Len: 0]
    Sequence number: 0    (relative sequence number)
    Acknowledgment number: 0
    Header Length: 28 bytes
```

```
    Flags: 0x002 (SYN)
    Window size value: 16384
    [Calculated window size: 16384]
    Checksum: 0x6398 [unverified]
    [Checksum Status: Unverified]
    Urgent pointer: 0
    Options: (8 bytes), Maximum segment size, Window scale, End of Option List (EOL)
        Maximum segment size: 1460 bytes
            Kind: Maximum Segment Size (2)
            Length: 4
```

MSS Value: 1460

```
        Window scale: 0 (multiply by 1)
        End of Option List (EOL)
```

As seen on R1 – PASSIVE – BGP Session restart – TCP SYN, ACK.

<#root>

```
! - Once BGP session is cleared
! - TCP SYN,ACK sourced from R1 - Passive
! - MSS calculation takes place and is influenced by TCP PMTUD

2735   0.003879       192.168.0.1 192.168.0.4 TCP    62     179 → 32077 [SYN, ACK] Seq=0 Ack=1 Win=16384 Le
```

MSS=1460

```
 WS=1

Frame 2735: 62 bytes on wire (496 bits), 62 bytes captured (496 bits) on interface 0
Ethernet II, Src: fa:16:3e:8f:8f:54 (fa:16:3e:8f:8f:54), Dst: fa:16:3e:d7:7e:f6 (fa:16:3e:d7:7e:f6)
Internet Protocol Version 4, Src: 192.168.0.1, Dst: 192.168.0.4
Transmission Control Protocol, Src Port: 179, Dst Port: 32077, Seq: 0, Ack: 1, Len: 0
    Source Port: 179
    Destination Port: 32077
    [Stream index: 25]
    [TCP Segment Len: 0]
    Sequence number: 0    (relative sequence number)
    Acknowledgment number: 1    (relative ack number)
    Header Length: 28 bytes
    Flags: 0x012 (SYN, ACK)
    Window size value: 16384
    [Calculated window size: 16384]
    Checksum: 0xbf77 [unverified]
    [Checksum Status: Unverified]
    Urgent pointer: 0
    Options: (8 bytes), Maximum segment size, Window scale, End of Option List (EOL)
        Maximum segment size: 1460 bytes
            Kind: Maximum Segment Size (2)
            Length: 4
```

MSS Value: 1460

```
        Window scale: 0 (multiply by 1)
        End of Option List (EOL)
```

TCP session details as seen on R4 – ACTIVE – after TCP PMTUD is enabled and the BGP session is cleared:

<#root>

```
! - BGP session re-established
! - as seen on R4 - Active

RP/0/0/CPU0:R4#show tcp detail pcb 0x121567f4
Fri Jan  8 16:45:13.928 UTC
================================================================
Connection state is ESTAB, I/O status: 0, socket status: 0
Established at Fri Jan  8 16:41:49 2021

PCB 0x121567f4, SO 0x12154460, TCPCB 0x12156190, vrfid 0x60000000,
Pak Prio: Medium, TOS: 192, TTL: 255, Hash index: 10
Local host: 192.168.0.4, Local port: 32077 (Local App PID: 1052958)
Foreign host: 192.168.0.1, Foreign port: 179

Current send queue size in bytes: 0 (max 24576)
Current receive queue size in bytes: 0 (max 32768)  mis-ordered: 0 bytes
Current receive queue size in packets: 0 (max 0)

Timer          Starts    Wakeups         Next(msec)
Retrans             8          1                 0
SendWnd             0          0                 0
TimeWait            0          0                 0
AckHold             5          3                 0
KeepAlive           1          0                 0
PmtuAger            0          0                 0
GiveUp              0          0                 0
Throttle            0          0                 0

   iss: 1254100669  snduna: 1254100983  sndnxt: 1254100983
sndmax: 1254100983  sndwnd: 32508       sndcwnd: 4380
   irs: 839938559   rcvnxt: 839938873   rcvwnd: 32527   rcvadv: 839971400

SRTT: 79 ms,  RTTO: 485 ms,  RTV: 406 ms,  KRTT: 0 ms
minRTT: 9 ms,  maxRTT: 229 ms

ACK hold time: 200 ms, Keepalive time: 0 sec, SYN waittime: 30 sec
Giveup time: 0 ms, Retransmission retries: 0, Retransmit forever: FALSE
Connect retries remaining: 30, connect retry interval: 30 secs

State flags: none
Feature flags: Win Scale, Nagle,
```

Path MTU

```
Request flags: Win Scale
```

Datagrams (in bytes): MSS 1460, peer MSS 1460, min MSS 1460, max MSS 1460

```
Window scales: rcv 0, snd 0, request rcv 0, request snd 0
Timestamp option: recent 0, recent age 0, last ACK sent 0
Sack blocks {start, end}: none
Sack holes {start, end, dups, rxmit}: none
```

```
Socket options: SO_REUSEADDR, SO_REUSEPORT, SO_NBIO
Socket states: SS_ISCONNECTED, SS_PRIV
Socket receive buffer states: SB_DEL_WAKEUP
Socket send buffer states: SB_DEL_WAKEUP
Socket receive buffer: Low/High watermark 1/32768
Socket send buffer   : Low/High watermark 2048/24576, Notify threshold 0

PDU information:
 #PDU's in buffer: 0
FIB Lookup Cache:  IFH: 0x40  PD ctx: size: 0  data:
  Num Labels: 0  Label Stack:

RP/0/0/CPU0:R4#
```

TCP session details as seen on R1 – PASSIVE – after TCP PMTUD is enabled and BGP session is cleared.

<#root>

```
! - BGP session re-established
! - as seen on R1 - Passive

RP/0/0/CPU0:R1#show tcp detail pcb 0x121558cc
Fri Jan  8 16:44:59.448 UTC
==============================================================
Connection state is ESTAB, I/O status: 0, socket status: 0
Established at Fri Jan  8 16:41:46 2021

PCB 0x121558cc, SO 0x121556d4, TCPCB 0x121575bc, vrfid 0x60000000,
Pak Prio: Medium, TOS: 192, TTL: 255, Hash index: 10
Local host: 192.168.0.1, Local port: 179 (Local App PID: 983326)
Foreign host: 192.168.0.4, Foreign port: 32077

Current send queue size in bytes: 0 (max 24576)
Current receive queue size in bytes: 0 (max 32768)  mis-ordered: 0 bytes
Current receive queue size in packets: 0 (max 0)

Timer          Starts    Wakeups         Next(msec)
Retrans             6          0                 0
SendWnd             0          0                 0
TimeWait            0          0                 0
AckHold             6          3                 0
KeepAlive           1          0                 0
PmtuAger            0          0                 0
GiveUp              0          0                 0
Throttle            0          0                 0

   iss: 839938559   snduna: 839938873   sndnxt: 839938873
sndmax: 839938873   sndwnd: 32527       sndcwnd: 4380
   irs: 1254100669  rcvnxt: 1254100983  rcvwnd: 32508   rcvadv: 1254133491

SRTT: 76 ms,  RTTO: 454 ms,  RTV: 378 ms,  KRTT: 0 ms
minRTT: 19 ms,  maxRTT: 219 ms

ACK hold time: 200 ms, Keepalive time: 0 sec, SYN waittime: 30 sec
Giveup time: 0 ms, Retransmission retries: 0, Retransmit forever: FALSE
Connect retries remaining: 0, connect retry interval: 0 secs

State flags: none
```

```
Feature flags: Win Scale, Nagle,
```

Path MTU

```
Request flags: Win Scale
```

Datagrams (in bytes): MSS 1460, peer MSS 1460, min MSS 1460, max MSS 1460

```
Window scales: rcv 0, snd 0, request rcv 0, request snd 0
Timestamp option: recent 0, recent age 0, last ACK sent 0
Sack blocks {start, end}: none
Sack holes {start, end, dups, rxmit}: none

Socket options: SO_REUSEADDR, SO_REUSEPORT, SO_NBIO
Socket states: SS_ISCONNECTED, SS_PRIV
Socket receive buffer states: SB_DEL_WAKEUP
Socket send buffer states: SB_DEL_WAKEUP
Socket receive buffer: Low/High watermark 1/32768
Socket send buffer   : Low/High watermark 2048/24576, Notify threshold 0

PDU information:
 #PDU's in buffer: 0
FIB Lookup Cache:  IFH: 0x40  PD ctx: size: 0  data:
  Num Labels: 0  Label Stack:

RP/0/0/CPU0:R1#
```

PMTUD – Path Segment has Lower IP MTU

The previous scenario helped to understand what happens upon initial TCP session establishment with PMTUD enabled. This scenario builds on top and helps to understand how TCP PMTUD works and the influence it has on established TCP sessions.

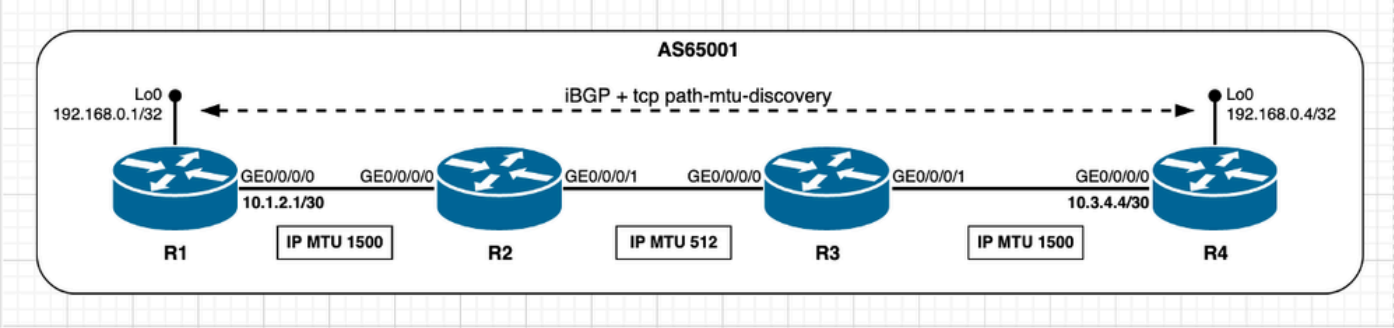


Image 3.2 – PMTUD enabled, and path segment has lower IP MTU.

Consider the previous image as a reference, assume that the BGP session is established, and R1 sends the BGP Update message carried by an IP packet with a size larger than 512 Bytes. With PMTUD enabled, the DF-bit (Don't Fragment) is now set. Hence, node R2 drops the IP packet and send an ICMP (Internet Control Message Protocol) message (Destination Unreachable – type 3; Fragmentation needed – code 4) back to R1. At node R1 after the ICMP message is received, PMTUD is triggered and attempts to establish the path lowest IP MTU. It does so by use of the next lower value from a set of well-defined plateau levels, that is considers a new TCP session MSS value. TCP then retransmits the original BGP update with the new MSS value and this process is repeated as many times as required till the ICMP message (Destination

Unreachable – type 3; Fragmentation needed – Code 4) is no longer received. This means till the MSS value in use is such that every packet sent falls under the lowest path segment IP MTU. As time passes, the PMTUD ruled by the PmtuAger timer walks through the plateau levels in the reverse direction and raises the MSS back to its maximum value. At any given time, if an ICMP message (Destination Unreachable – type 3 ; Fragmentation needed – Code 4) is again received then PMTUD acts as previously described.

The next outputs walk through the PMTUD behavior just described and starts from the scenario of an established TCP session. Here, Cisco IOS XR node R4 plays an active role hence manages TCP connection and opens the TCP session with R1 on destination port 179. Both nodes use the default IP MTU values on their interfaces. The initial MSS calculation in this scenario can be summarized as follows:

- The intermediate segment between R2 and R3 nodes uses non-default IP MTU 512 bytes.
- R1 and R4 use default MTU values on their interfaces.
- TCP Path MTU discovery is enabled.
- TCP Peers are not directly connected.
    - R4 manages BGP connection.
    - R4 sends SYN with MSS of 1460 bytes.
        - 1500 (Interface IP MTU) – 20 (minTCP_H) – 20 (minIP_H).
    - R1 sends SYN, ACK with MSS of 1460 bytes.
        - Sends the lower of [Received MSS ; Local initial MSS].
        - Received MSS 1460 bytes; Local initial MSS 1460 bytes.
    - The lowest MSS value is used on both peers.

TCP SYN sourced from R4:

<#root>

```
! – Initial TCP session establishment
! – TCP SYN sourced from R4

392    6.752774       192.168.0.4 192.168.0.1 TCP    62     32449 → 179 [SYN] Seq=0 Win=16384 Len=0

MSS=1460

 WS=1

Frame 392: 62 bytes on wire (496 bits), 62 bytes captured (496 bits) on interface 0
Ethernet II, Src: fa:16:3e:5c:f1:80 (fa:16:3e:5c:f1:80), Dst: fa:16:3e:42:18:05 (fa:16:3e:42:18:05)
Internet Protocol Version 4, Src: 192.168.0.4, Dst: 192.168.0.1
Transmission Control Protocol, Src Port: 32449, Dst Port: 179, Seq: 0, Len: 0
    Source Port: 32449
    Destination Port: 179
    [Stream index: 10]
    [TCP Segment Len: 0]
    Sequence number: 0    (relative sequence number)
    Acknowledgment number: 0
    Header Length: 28 bytes
    Flags: 0x002 (SYN)
    Window size value: 16384
    [Calculated window size: 16384]
    Checksum: 0x6858 [unverified]
    [Checksum Status: Unverified]
    Urgent pointer: 0
    Options: (8 bytes), Maximum segment size, Window scale, End of Option List (EOL)
        Maximum segment size: 1460 bytes
            Kind: Maximum Segment Size (2)
            Length: 4
```

MSS Value: 1460

```
        Window scale: 0 (multiply by 1)
        End of Option List (EOL)
```

TCP SYN, ACK sourced from R1:

<#root>

```
! – Initial TCP session establishment
! – TCP SYN,ACK sourced from R1

393    0.003628      192.168.0.1 192.168.0.4 TCP    62     179 → 32449 [SYN, ACK] Seq=0 Ack=1 Win=16384 Len
```

MSS=1460

```
 WS=1

Frame 393: 62 bytes on wire (496 bits), 62 bytes captured (496 bits) on interface 0
Ethernet II, Src: fa:16:3e:42:18:05 (fa:16:3e:42:18:05), Dst: fa:16:3e:5c:f1:80 (fa:16:3e:5c:f1:80)
Internet Protocol Version 4, Src: 192.168.0.1, Dst: 192.168.0.4
Transmission Control Protocol, Src Port: 179, Dst Port: 32449, Seq: 0, Ack: 1, Len: 0
    Source Port: 179
    Destination Port: 32449
    [Stream index: 10]
    [TCP Segment Len: 0]
    Sequence number: 0    (relative sequence number)
    Acknowledgment number: 1    (relative ack number)
    Header Length: 28 bytes
    Flags: 0x012 (SYN, ACK)
    Window size value: 16384
    [Calculated window size: 16384]
    Checksum: 0x509e [unverified]
    [Checksum Status: Unverified]
    Urgent pointer: 0
    Options: (8 bytes), Maximum segment size, Window scale, End of Option List (EOL)
        Maximum segment size: 1460 bytes
            Kind: Maximum Segment Size (2)
            Length: 4
```

MSS Value: 1460

```
        Window scale: 0 (multiply by 1)
        End of Option List (EOL)
```

With the BGP session established, node R1 sends the BGP Update message and receives the ICMP message (Destination Unreachable – type 3 ; Fragmentation needed – Code 4) in return sourced from node R2.

This occurs because the IP packet that carries the BGP update message has the DF-bit set and the IP MTU of 512 bytes used at the R2/R3 segment is lower than the IP packet size of 1116 bytes. As explained earlier, the reception of the ICMP message triggers PMTUD.

At R1 ICMP, Type 3/Code 4 message is received:

<#root>

```
! - as seen from R1 - Passive
! - After session is established R1 sends BGP Update message with IP length of 1116 Bytes
! - note IP Header Flags shows DF bit set

528    5.893055       192.168.0.1 192.168.0.4 BGP
```

1130

```
   UPDATE Message, KEEPALIVE Message

Frame 528: 1130 bytes on wire (9040 bits), 1130 bytes captured (9040 bits) on interface 0
Ethernet II, Src: fa:16:3e:42:18:05 (fa:16:3e:42:18:05), Dst: fa:16:3e:5c:f1:80 (fa:16:3e:5c:f1:80)
Internet Protocol Version 4, Src: 192.168.0.1, Dst: 192.168.0.4
    0100 .... = Version: 4
    .... 0101 = Header Length: 20 bytes (5)
    Differentiated Services Field: 0xc0 (DSCP: CS6, ECN: Not-ECT)
```

Total Length: 1116

```
    Identification: 0x8c37 (35895)
```

Flags: 0x02 (Don't Fragment)

```
    Fragment offset: 0
    Time to live: 255
    Protocol: TCP (6)
    Header checksum: 0xe09a [validation disabled]
    [Header checksum status: Unverified]
    Source: 192.168.0.1
    Destination: 192.168.0.4
    [Source GeoIP: Unknown]
    [Destination GeoIP: Unknown]
Transmission Control Protocol, Src Port: 179, Dst Port: 32449, Seq: 318, Ack: 251, Len: 1076
Border Gateway Protocol - UPDATE Message
Border Gateway Protocol - KEEPALIVE Message
<snip>

! - as seen from R1 - Passive
! - IP MTU on R2/R3 is lower than IP packet length and DF bit is set
! - R1 receives ICMP error message from R2
! - note R2 ICMP error message carries Next-Hop MTU
! - “The size in octets of the largest datagram that could be forwarded, along the path of
!   the original datagram, without being fragmented at this router. The size includes the
!   IP header and IP data, and does not include any lower-level headers.”

529    0.002423       10.2.3.1       192.168.0.1 ICMP   110
```

Destination unreachable (Fragmentation needed)

```
Frame 529: 110 bytes on wire (880 bits), 110 bytes captured (880 bits) on interface 0
Ethernet II, Src: fa:16:3e:5c:f1:80 (fa:16:3e:5c:f1:80), Dst: fa:16:3e:42:18:05 (fa:16:3e:42:18:05)
Internet Protocol Version 4, Src: 10.2.3.1, Dst: 192.168.0.1
    0100 .... = Version: 4
    .... 0101 = Header Length: 20 bytes (5)
    Differentiated Services Field: 0x00 (DSCP: CS0, ECN: Not-ECT)
    Total Length: 96
    Identification: 0x0001 (1)
```

```
    Flags: 0x00
    Fragment offset: 0
    Time to live: 255
    Protocol: ICMP (1)

    Header checksum: 0xac97 [validation disabled]
    [Header checksum status: Unverified]
    Source: 10.2.3.1
    Destination: 192.168.0.1
    [Source GeoIP: Unknown]
    [Destination GeoIP: Unknown]
Internet Control Message Protocol
    Type: 3 (Destination unreachable)

    Code: 4 (Fragmentation needed)

    Checksum: 0x2d52 [correct]
    [Checksum Status: Good]
    Length: 17
    [Length of original datagram: 68]
    Unused: 0011

  MTU of next hop: 512

    Internet Protocol Version 4, Src: 192.168.0.1, Dst: 192.168.0.4
        0100 .... = Version: 4
        .... 0101 = Header Length: 20 bytes (5)
        Differentiated Services Field: 0xc0 (DSCP: CS6, ECN: Not-ECT)
        Total Length: 1116
        Identification: 0x8c37 (35895)
        Flags: 0x02 (Don't Fragment)
        Fragment offset: 0
        Time to live: 254
        Protocol: TCP (6)
        Header checksum: 0xe19a [validation disabled]
        [Header checksum status: Unverified]
        Source: 192.168.0.1
        Destination: 192.168.0.4
        [Source GeoIP: Unknown]
        [Destination GeoIP: Unknown]
    Transmission Control Protocol, Src Port: 179, Dst Port: 32449, Seq: 2847698730, Ack: 2130367817
    Border Gateway Protocol - UPDATE Message
[Packet size limited during capture: IPv4 truncated]
```

At node R1, triggered by ICMP message, TCP PMTUD attempts to establish the end-to-end lowest IP MTU by use of the next lower value from a set of well-defined plateau (IP MTU) levels. These plateau levels are documented on RFC1191 – Path MTU discovery.

```
MTU plateaus from RFC 1191
- values include both TCP and IP headers
65535
32000
```

```
17914
8166
4352
2002
1492
1006
508
296
68
```

But since ICMP (Destination Unreachable – type 3; Fragmentation needed – Code 4) message received by node R1 conveys the **MTU of next hop** then as exhibited next, node R1 uses this value, which in our example is 512 bytes, and adjust the TCP session MSS value. Take notice that the original TCP segment length was 1076 bytes, hence three packets are required to retransmit the original TCP segment.

As seen on R1 – PASSIVE – PMTUD operation:

<#root>

```
! - As seen from R1 - Passive
! - Hint is provided by ICMP unreachable message MTU of next-hop field: 512 bytes
! - R1 then considers this value and retransmits BGP Update split in three distinct packets
! - Sum of TCP length = 472 + 472 + 132 = 1076 bytes

530    0.007497       192.168.0.1 192.168.0.4 TCP    526    [TCP Out-Of-Order] 179 → 32449 [ACK] Seq=318 A

Len=472


532    0.015374       192.168.0.1 192.168.0.4 TCP    526    [TCP Retransmission] 179 → 32449 [ACK] Seq=79

Len=472


533    0.004129       192.168.0.1 192.168.0.4 TCP    186    [TCP Retransmission] 179 → 32449 [PSH, ACK] S

Len=132
```

As stated earlier once all packets have been transmitted over time, PMTUD walks through the plateau levels in the reverse direction ruled by the PmtuAger timer and attempt to raise the MSS to its maximum value as per the scenario in place.

As seen on R1 – PMTUD across defined plateaus:

```
! - As seen from R1 - Passive - 'debug tcp pmtud' and 'debug icmp' active
! - TCP PMTUD is triggered once ICMP unreachable received

RP/0/0/CPU0:May 12 09:09:22.763 UTC: ipv4_io[266]: IPv4 ICMP: Received ICMP too big from 192.168.0.1 ab
RP/0/0/CPU0:May 12 09:09:22.763 UTC: ipv4_io[266]: ipv4_icmp_unreachable_rcvd ICMP unreach recvd: sendi
RP/0/0/CPU0:May 12 09:09:22.763 UTC: ipv4_io[266]: ip_icmp_lib_ipv4_receive: sending pak(0xb0c07d8f) to
RP/0/0/CPU0:May 12 09:09:22.763 UTC: tcp[399]: [t4] PCB 0x15393770: Process ICMP Dest-unreach (next hop

! - attempt new MSS 472 = MTU of next-hop(512) - TCP_H(20) - IP_H(20)

RP/0/0/CPU0:May 12 09:09:22.763 UTC: tcp[399]: [t4] PCB 0x15393770: Process ICMP Dest-unreach (next hop
```

```
RP/0/0/CPU0:May 12 09:09:22.763 UTC: tcp[399]: [t4] PCB 0x15393770: Try to use new MSS: 472
RP/0/0/CPU0:May 12 09:09:22.763 UTC: tcp[399]: [t4] PCB 0x15393770, New path MTU decided to use: 472  c

! - over time PMTUD attempts to raise MSS as per egress interface configured MTU

RP/0/0/CPU0:May 12 09:19:22.782 UTC: tcp[399]: [t23] PCB 0x15393770: Trying next higher MTU: 966
RP/0/0/CPU0:May 12 09:21:22.793 UTC: tcp[399]: [t23] PCB 0x15393770: Trying next higher MTU: 1452
RP/0/0/CPU0:May 12 09:23:22.805 UTC: tcp[399]: [t23] PCB 0x15393770: Trying next higher MTU: 1460
```

The final state can be observed on these outputs. Note in particular the min and max MSS values exhibited by node R1, which highlights and signals that PMTUD was triggered.

TCP session details as seen on R4 – ACTIVE:

<#root>

```
! - Final stage as seen from R4 - Active

RP/0/0/CPU0:R4#show tcp detail pcb 0x153913b8
Wed May 12 10:09:43.246 UTC
==========================================================
Connection state is ESTAB, I/O status: 0, socket status: 0
Established at Wed May 12 09:02:07 2021

PCB 0x153913b8, SO 0x153917f0, TCPCB 0x1538fb58, vrfid 0x60000000,
Pak Prio: Medium, TOS: 192, TTL: 255, Hash index: 382
Local host: 192.168.0.4, Local port: 32449 (Local App PID: 1196319)
Foreign host: 192.168.0.1, Foreign port: 179
(Local App PID/instance/SPL_APP_ID: 1196319/1/0)

Current send queue size in bytes: 0 (max 24576)
Current receive queue size in bytes: 0 (max 32768)  mis-ordered: 0 bytes
Current receive queue size in packets: 0 (max 0)

Timer          Starts    Wakeups         Next(msec)
Retrans            72          1                 0
SendWnd             0          0                 0
TimeWait            0          0                 0
AckHold            71         69                 0
KeepAlive           1          0                 0
PmtuAger            0          0                 0
GiveUp              0          0                 0
Throttle            0          0                 0

   iss: 2130367566  snduna: 2130368957  sndnxt: 2130368957
sndmax: 2130368957  sndwnd: 31453       sndcwnd: 2920
   irs: 2847698412  rcvnxt: 2847700946  rcvwnd: 31799   rcvadv: 2847732745

SRTT: 220 ms,  RTTO: 300 ms,  RTV: 12 ms,  KRTT: 0 ms
minRTT: 9 ms,  maxRTT: 239 ms

ACK hold time: 200 ms, Keepalive time: 0 sec, SYN waittime: 30 sec
Giveup time: 0 ms, Retransmission retries: 0, Retransmit forever: FALSE
Connect retries remaining: 10, connect retry interval: 30 secs

State flags: none
Feature flags: Win Scale, Nagle,
```

Path MTU

```
Request flags: Win Scale
```

Datagrams (in bytes): MSS 1460, peer MSS 1460, min MSS 1460, max MSS 1460

```
Window scales: rcv 0, snd 0, request rcv 0, request snd 0
Timestamp option: recent 0, recent age 0, last ACK sent 0
Sack blocks {start, end}: none
Sack holes {start, end, dups, rxmit}: none

Socket options: SO_REUSEADDR, SO_REUSEPORT, SO_NBIO
Socket states: SS_ISCONNECTED, SS_PRIV
Socket receive buffer states: SB_DEL_WAKEUP
Socket send buffer states: SB_DEL_WAKEUP
Socket receive buffer: Low/High watermark 1/32768
Socket send buffer   : Low/High watermark 2048/24576, Notify threshold 0
Socket misc info     : Rcv data size (sb_cc) 0, so_qlen 0,
                       so_q0len 0, so_qlimit 0, so_error 0
                       so_auto_rearm 1

PDU information:
 #PDU's in buffer: 0
FIB Lookup Cache:  IFH: 0x40  PD ctx: size: 0  data:
  Num Labels: 0  Label Stack:
Num of peers with authentication info: 0

RP/0/0/CPU0:R4#
```

TCP session details as seen on R1 – PASSIVE:

<#root>

```
! - Final stage as seen from R1 - Passive

RP/0/0/CPU0:R1#show tcp detail pcb 0x15393770
Wed May 12 10:12:41.432 UTC
==============================================================
Connection state is ESTAB, I/O status: 240, socket status: 0
Established at Wed May 12 09:02:05 2021

PCB 0x15393770, SO 0x15394ea0, TCPCB 0x15391c0c, vrfid 0x60000000,
Pak Prio: Medium, TOS: 192, TTL: 255, Hash index: 382
Local host: 192.168.0.1, Local port: 179 (Local App PID: 1192224)
Foreign host: 192.168.0.4, Foreign port: 32449
(Local App PID/instance/SPL_APP_ID: 1192224/1/0)

Current send queue size in bytes: 0 (max 24576)
Current receive queue size in bytes: 0 (max 32768)  mis-ordered: 0 bytes
Current receive queue size in packets: 0 (max 0)

Timer          Starts    Wakeups         Next(msec)
Retrans            75          0                  0
SendWnd             0          0                  0
TimeWait            0          0                  0
AckHold            73         71                  0
KeepAlive           1          0                  0
```

```
PmtuAger           28         27            41595


GiveUp              0          0               0
Throttle            0          0               0

   iss: 2847698412  snduna: 2847701003  sndnxt: 2847701003
sndmax: 2847701003  sndwnd: 31742       sndcwnd: 4380
   irs: 2130367566  rcvnxt: 2130369014  rcvwnd: 31396   rcvadv: 2130400410

SRTT: 224 ms,  RTTO: 300 ms,  RTV: 23 ms,  KRTT: 0 ms
minRTT: 9 ms,  maxRTT: 259 ms

ACK hold time: 200 ms, Keepalive time: 0 sec, SYN waittime: 30 sec
Giveup time: 0 ms, Retransmission retries: 0, Retransmit forever: FALSE
Connect retries remaining: 0, connect retry interval: 0 secs

State flags: PMTU ager
Feature flags: Win Scale, Nagle,

Path MTU


Request flags: Win Scale


Datagrams (in bytes): MSS 1460, peer MSS 1460, min MSS 472, max MSS 1460



Window scales: rcv 0, snd 0, request rcv 0, request snd 0
Timestamp option: recent 0, recent age 0, last ACK sent 0
Sack blocks {start, end}: none
Sack holes {start, end, dups, rxmit}: none

Socket options: SO_REUSEADDR, SO_REUSEPORT, SO_NBIO
Socket states: SS_ISCONNECTED, SS_PRIV
Socket receive buffer states: SB_DEL_WAKEUP
Socket send buffer states: SB_DEL_WAKEUP
Socket receive buffer: Low/High watermark 1/32768
Socket send buffer   : Low/High watermark 2048/24576, Notify threshold 0
Socket misc info     : Rcv data size (sb_cc) 0, so_qlen 0,
                       so_q0len 0, so_qlimit 0, so_error 0
                       so_auto_rearm 1

PDU information:
 #PDU's in buffer: 0
FIB Lookup Cache:  IFH: 0x20  PD ctx: size: 0  data:
  Num Labels: 0  Label Stack:
Num of peers with authentication info: 0

RP/0/0/CPU0:R1#
```

Lastly, if at any given time an ICMP (Destination Unreachable – type 3 ; Fragmentation needed – Code 4) message is again received then PMTUD once more acts as previously described.

As seen from R1 – PASSIVE – PMTUD has triggered again:

```
! - As seen from R1 - Passive
! - TCP PMTUD is again triggered upon new ICMP unreachable received
! - Behavior can be triggered via clearing redistributed, network and aggregate routes originated

RP/0/0/CPU0:R1#clear bgp ipv4 all self-originated
Wed May 12 10:19:06.836 UTC
RP/0/0/CPU0:R1#

! - New BGP update message is sourced from R1 after clear bgp command

1707   1.712657       192.168.0.1 192.168.0.4 BGP    1121   UPDATE Message

Frame 1707: 1121 bytes on wire (8968 bits), 1121 bytes captured (8968 bits) on interface 0
Ethernet II, Src: fa:16:3e:42:18:05 (fa:16:3e:42:18:05), Dst: fa:16:3e:5c:f1:80 (fa:16:3e:5c:f1:80)
Internet Protocol Version 4, Src: 192.168.0.1, Dst: 192.168.0.4
    0100 .... = Version: 4
    .... 0101 = Header Length: 20 bytes (5)
    Differentiated Services Field: 0xc0 (DSCP: CS6, ECN: Not-ECT)
    Total Length: 1107
    Identification: 0x1a38 (6712)
    Flags: 0x02 (Don't Fragment)
    Fragment offset: 0
    Time to live: 255
    Protocol: TCP (6)
    Header checksum: 0x52a3 [validation disabled]
    [Header checksum status: Unverified]
    Source: 192.168.0.1
    Destination: 192.168.0.4
    [Source GeoIP: Unknown]
    [Destination GeoIP: Unknown]
Transmission Control Protocol, Src Port: 179, Dst Port: 32449, Seq: 2705, Ack: 1562, Len: 1067
Border Gateway Protocol - UPDATE Message

! - ICMP Destination Unreachable / Fragmentation needed is received and triggers PMTUD

1708   0.001614       10.2.3.1       192.168.0.1 ICMP   110
```

Destination unreachable (Fragmentation needed)

```
Frame 1708: 110 bytes on wire (880 bits), 110 bytes captured (880 bits) on interface 0
Ethernet II, Src: fa:16:3e:5c:f1:80 (fa:16:3e:5c:f1:80), Dst: fa:16:3e:42:18:05 (fa:16:3e:42:18:05)
Internet Protocol Version 4, Src: 10.2.3.1, Dst: 192.168.0.1
    0100 .... = Version: 4
    .... 0101 = Header Length: 20 bytes (5)
    Differentiated Services Field: 0x00 (DSCP: CS0, ECN: Not-ECT)
    Total Length: 96
    Identification: 0x0002 (2)
    Flags: 0x00
    Fragment offset: 0
    Time to live: 255
```

Protocol: ICMP (1)

```
    Header checksum: 0xac96 [validation disabled]
    [Header checksum status: Unverified]
    Source: 10.2.3.1
    Destination: 192.168.0.1
    [Source GeoIP: Unknown]
    [Destination GeoIP: Unknown]
```

```
Internet Control Message Protocol
```

Type: 3 (Destination unreachable)

Code: 4 (Fragmentation needed)

```
    Checksum: 0x3b73 [correct]
    [Checksum Status: Good]
    Length: 17
    [Length of original datagram: 68]
    Unused: 0011
```

MTU of next hop: 512

```
    Internet Protocol Version 4, Src: 192.168.0.1, Dst: 192.168.0.4
        0100 .... = Version: 4
        .... 0101 = Header Length: 20 bytes (5)
        Differentiated Services Field: 0xc0 (DSCP: CS6, ECN: Not-ECT)
        Total Length: 1107
        Identification: 0x1a38 (6712)
        Flags: 0x02 (Don't Fragment)
        Fragment offset: 0
        Time to live: 254
        Protocol: TCP (6)
        Header checksum: 0x53a3 [validation disabled]
        [Header checksum status: Unverified]
        Source: 192.168.0.1
        Destination: 192.168.0.4
        [Source GeoIP: Unknown]
        [Destination GeoIP: Unknown]
    Transmission Control Protocol, Src Port: 179, Dst Port: 32449, Seq: 2847701117, Ack: 2130369128
    Border Gateway Protocol - UPDATE Message

! - Note new/updated MSS value and PmtuAger
! - MSS 472 ; Aligned with “MTU of next hop” value contained in ICMP message

RP/0/0/CPU0:R1#show tcp detail pcb 0x15393770
Wed May 12 10:19:31.494 UTC
================================================================
Connection state is ESTAB, I/O status: 240, socket status: 0
Established at Wed May 12 09:02:05 2021

PCB 0x15393770, SO 0x15394ea0, TCPCB 0x15391c0c, vrfid 0x60000000,
Pak Prio: Medium, TOS: 192, TTL: 255, Hash index: 382
Local host: 192.168.0.1, Local port: 179 (Local App PID: 1192224)
Foreign host: 192.168.0.4, Foreign port: 32449
(Local App PID/instance/SPL_APP_ID: 1192224/1/0)

Current send queue size in bytes: 0 (max 24576)
Current receive queue size in bytes: 0 (max 32768)  mis-ordered: 0 bytes
Current receive queue size in packets: 0 (max 0)

Timer          Starts    Wakeups         Next(msec)
Retrans            83          0                 0
SendWnd             0          0                 0
TimeWait            0          0                 0
AckHold            80         77                 0
KeepAlive           1          0                 0
```

```
PmtuAger          32         30          575401


GiveUp             0          0               0
Throttle           0          0               0

   iss: 2847698412  snduna: 2847702184  sndnxt: 2847702184
sndmax: 2847702184  sndwnd: 32173       sndcwnd: 944
   irs: 2130367566  rcvnxt: 2130369147  rcvwnd: 32730   rcvadv: 2130401877

SRTT: 221 ms,  RTTO: 300 ms,  RTV: 16 ms,  KRTT: 0 ms
minRTT: 9 ms,  maxRTT: 259 ms

ACK hold time: 200 ms, Keepalive time: 0 sec, SYN waittime: 30 sec
Giveup time: 0 ms, Retransmission retries: 0, Retransmit forever: FALSE
Connect retries remaining: 0, connect retry interval: 0 secs

State flags: PMTU ager
Feature flags: Win Scale, Nagle,

Path MTU


Request flags: Win Scale


Datagrams (in bytes): MSS 472, peer MSS 1460, min MSS 472, max MSS 1460



Window scales: rcv 0, snd 0, request rcv 0, request snd 0
Timestamp option: recent 0, recent age 0, last ACK sent 0
Sack blocks {start, end}: none
Sack holes {start, end, dups, rxmit}: none

Socket options: SO_REUSEADDR, SO_REUSEPORT, SO_NBIO
Socket states: SS_ISCONNECTED, SS_PRIV
Socket receive buffer states: SB_DEL_WAKEUP
Socket send buffer states: SB_DEL_WAKEUP
Socket receive buffer: Low/High watermark 1/32768
Socket send buffer   : Low/High watermark 2048/24576, Notify threshold 0
Socket misc info     : Rcv data size (sb_cc) 0, so_qlen 0,
                       so_qOlen 0, so_qlimit 0, so_error 0
                       so_auto_rearm 1

PDU information:
 #PDU's in buffer: 0
FIB Lookup Cache:  IFH: 0x20  PD ctx: size: 0  data:
  Num Labels: 0  Label Stack:
Num of peers with authentication info: 0

RP/0/0/CPU0:R1#
```

On Cisco IOS XR releases impacted by Cisco bug ID CSCvf10395, the next-hop contained in the ICMP error message is ignored and the node attempts to establish the end-to-end lowest IP MTU by use of the next lower value from the set of well-defined plateau (IP MTU) levels mentioned previously and documented by RFC1191 – Path MTU discovery. These attempts occur till successful transmission, this means till ICMP (Destination Unreachable – type 3 ; Fragmentation needed – Code 4) messages are no longer received.

As seen from a node with Cisco IOS XR release impacted by Cisco bug ID CSCvf10395:

```
! - As seen from IOX XR node with a release impacted by Cisco bug ID CSCvf10395
! - Node ignores "MTU of next hop" and tries next lower plateau
! - This is observed till ICMP error messages are no longer received
! - Practical consequence is extra retransmissions occurrence

RP/0/0/CPU0:Feb 23 17:05:32.929 : tcp[399]: [t4] PCB 0x12152adc: Process ICMP Dest-unreach (next hop mtu
RP/0/0/CPU0:Feb 23 17:05:32.929 : tcp[399]: [t4] PCB 0x12152adc: Invalid next hop mtu (33554432), ignore
RP/0/0/CPU0:Feb 23 17:05:34.649 : tcp[399]: [t27] PCB 0x12152adc: Trying next lower MTU: 1452 <<<<<<<< H

RP/0/0/CPU0:Feb 23 17:05:35.519 : tcp[399]: [t4] PCB 0x12152adc: Process ICMP Dest-unreach (next hop mtu
RP/0/0/CPU0:Feb 23 17:05:35.519 : tcp[399]: [t4] PCB 0x12152adc: Invalid next hop mtu (33554432), ignore
RP/0/0/CPU0:Feb 23 17:05:37.239 : tcp[399]: [t27] PCB 0x12152adc: Trying next lower MTU: 966 <<<<<<<< HE

RP/0/0/CPU0:Feb 23 17:05:38.109 : tcp[399]: [t4] PCB 0x12152adc: Process ICMP Dest-unreach (next hop mtu
RP/0/0/CPU0:Feb 23 17:05:38.109 : tcp[399]: [t4] PCB 0x12152adc: Invalid next hop mtu (33554432), ignore
RP/0/0/CPU0:Feb 23 17:05:39.829 : tcp[399]: [t27] PCB 0x12152adc: Trying next lower MTU: 468 <<<<<<<< HE
```

As next-step consider same scenario but with Label Distribution Protocol (LDP) across all interfaces. The goal here is to understand what differences can be observed from previous scenarios in a MPLS- enabled environment.

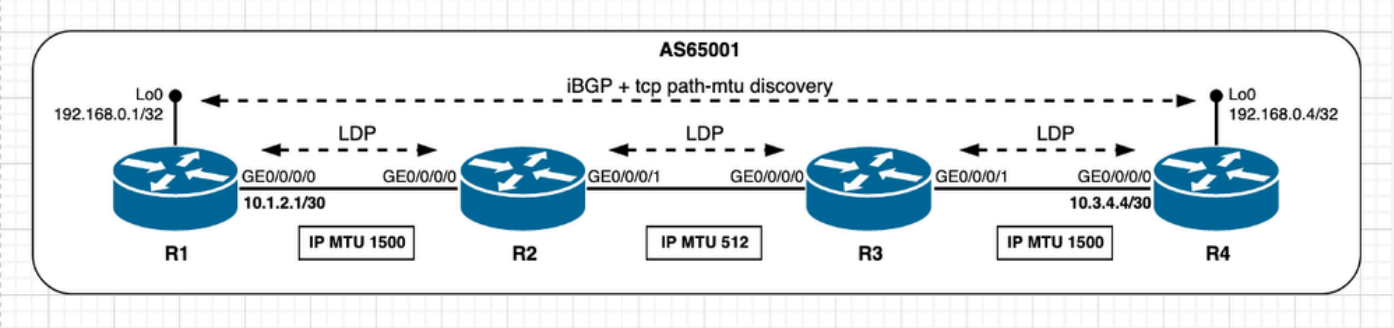


Image 3.3 – PMTUD enabled, and path segment has lower IP MTU – MPLS scenario.

First, consider the initial stage of the BGP session established prior to PMTUD trigger as shown here.

TCP (BGP) initial state as seen on R4 – ACTIVE – MPLS-enabled scenario:

<#root>

```
! - as seen on R4 - Active
! - TCP path MTU discovery enabled
! - MPLS LDP enabled
! - TCP session initial state

RP/0/0/CPU0:R4#show tcp detail pcb 0x153bdaf0
Mon May 17 08:32:16.673 UTC
================================================================
Connection state is ESTAB, I/O status: 0, socket status: 0
Established at Mon May 17 08:31:57 2021

PCB 0x153bdaf0, SO 0x153acc80, TCPCB 0x153acea8, vrfid 0x60000000,
Pak Prio: Medium, TOS: 192, TTL: 255, Hash index: 757
Local host: 192.168.0.4, Local port: 57400 (Local App PID: 1196319)
Foreign host: 192.168.0.1, Foreign port: 179
(Local App PID/instance/SPL_APP_ID: 1196319/1/0)
```

```
Current send queue size in bytes: 0 (max 24576)
Current receive queue size in bytes: 0 (max 32768)  mis-ordered: 0 bytes
Current receive queue size in packets: 0 (max 0)

Timer          Starts    Wakeups         Next(msec)
Retrans             5          1                 0
SendWnd             0          0                 0
TimeWait            0          0                 0
AckHold             2          1                 0
KeepAlive           1          0                 0
PmtuAger            0          0                 0
GiveUp              0          0                 0
Throttle            0          0                 0

   iss: 1386459919  snduna: 1386460037  sndnxt: 1386460037
sndmax: 1386460037  sndwnd: 32726       sndcwnd: 4380
   irs: 3874414679  rcvnxt: 3874414864  rcvwnd: 32678   rcvadv: 3874447542

SRTT: 48 ms,  RTTO: 300 ms,  RTV: 228 ms,  KRTT: 0 ms
minRTT: 9 ms,  maxRTT: 229 ms

ACK hold time: 200 ms, Keepalive time: 0 sec, SYN waittime: 30 sec
Giveup time: 0 ms, Retransmission retries: 0, Retransmit forever: FALSE
Connect retries remaining: 10, connect retry interval: 30 secs

State flags: none
Feature flags: Win Scale, Nagle,
```

Path MTU

```
Request flags: Win Scale
```

Datagrams (in bytes): MSS 1460, peer MSS 1460, min MSS 1460, max MSS 1460

```
Window scales: rcv 0, snd 0, request rcv 0, request snd 0
Timestamp option: recent 0, recent age 0, last ACK sent 0
Sack blocks {start, end}: none
Sack holes {start, end, dups, rxmit}: none

Socket options: SO_REUSEADDR, SO_REUSEPORT, SO_NBIO
Socket states: SS_ISCONNECTED, SS_PRIV
Socket receive buffer states: SB_DEL_WAKEUP
Socket send buffer states: SB_DEL_WAKEUP
Socket receive buffer: Low/High watermark 1/32768
Socket send buffer   : Low/High watermark 2048/24576, Notify threshold 0
Socket misc info     : Rcv data size (sb_cc) 0, so_qlen 0,
                       so_q0len 0, so_qlimit 0, so_error 0
                       so_auto_rearm 1

PDU information:
 #PDU's in buffer: 0
FIB Lookup Cache:  IFH: 0x40  PD ctx: size: 0  data:
  Num Labels: 1  Label Stack:  0x5dc2
Num of peers with authentication info: 0

RP/0/0/CPU0:R4#
```

TCP (BGP) initial state as seen on R1 – PASSIVE – MPLS-enabled scenario:

<#root>

```
! - as seen on R1 - Passive
! - TCP path MTU discovery enabled
! - MPLS LDP enabled
! - TCP session initial state

RP/0/0/CPU0:R1#show tcp detail pcb 0x153acc8c
Mon May 17 08:32:56.618 UTC
================================================================
Connection state is ESTAB, I/O status: 0, socket status: 0
Established at Mon May 17 08:31:55 2021

PCB 0x153acc8c, SO 0x153adad4, TCPCB 0x153adcfc, vrfid 0x60000000,
Pak Prio: Medium, TOS: 192, TTL: 255, Hash index: 757
Local host: 192.168.0.1, Local port: 179 (Local App PID: 1192224)
Foreign host: 192.168.0.4, Foreign port: 57400
(Local App PID/instance/SPL_APP_ID: 1192224/1/0)

Current send queue size in bytes: 0 (max 24576)
Current receive queue size in bytes: 0 (max 32768)  mis-ordered: 0 bytes
Current receive queue size in packets: 0 (max 0)

Timer          Starts    Wakeups         Next(msec)
Retrans             3          0                 0
SendWnd             0          0                 0
TimeWait            0          0                 0
AckHold             3          1                 0
KeepAlive           1          0                 0
PmtuAger            0          0                 0
GiveUp              0          0                 0
Throttle            0          0                 0

   iss: 3874414679  snduna: 3874414864  sndnxt: 3874414864
sndmax: 3874414864  sndwnd: 32678       sndcwnd: 4380
   irs: 1386459919  rcvnxt: 1386460037  rcvwnd: 32726   rcvadv: 1386492763

SRTT: 45 ms,  RTTO: 300 ms,  RTV: 239 ms,  KRTT: 0 ms
minRTT: 19 ms,  maxRTT: 229 ms

ACK hold time: 200 ms, Keepalive time: 0 sec, SYN waittime: 30 sec
Giveup time: 0 ms, Retransmission retries: 0, Retransmit forever: FALSE
Connect retries remaining: 0, connect retry interval: 0 secs

State flags: none
Feature flags: Win Scale, Nagle,
```

Path MTU

```
Request flags: Win Scale
```

Datagrams (in bytes): MSS 1460, peer MSS 1460, min MSS 1460, max MSS 1460

```
Window scales: rcv 0, snd 0, request rcv 0, request snd 0
Timestamp option: recent 0, recent age 0, last ACK sent 0
Sack blocks {start, end}: none
```

```
Sack holes {start, end, dups, rxmit}: none

Socket options: SO_REUSEADDR, SO_REUSEPORT, SO_NBIO
Socket states: SS_ISCONNECTED, SS_PRIV
Socket receive buffer states: SB_DEL_WAKEUP
Socket send buffer states: SB_DEL_WAKEUP
Socket receive buffer: Low/High watermark 1/32768
Socket send buffer   : Low/High watermark 2048/24576, Notify threshold 0
Socket misc info     : Rcv data size (sb_cc) 0, so_qlen 0,
                       so_q0len 0, so_qlimit 0, so_error 0
                       so_auto_rearm 1

PDU information:
 #PDU's in buffer: 0
FIB Lookup Cache:  IFH: 0x20  PD ctx: size: 0  data:
  Num Labels: 1  Label Stack:  0x5dc3
Num of peers with authentication info: 0

RP/0/0/CPU0:R1#
```

In this MPLS-enabled scenario, it is observed that the details for the TCP (LDP) sessions were established. Do note that all described previously with regards to the MSS calculation for TCP (BGP) sessions, applies to TCP (LDP) sessions as well. As an example, nodes R3 and R2 TCP (LDP) session MSS calculation can be summarized as follows:

- Both R2 and R3 use non-default IP MTU of 512 bytes.
- Path MTU discovery is enabled.
- TCP peers are not directly connected (TCP session is established between Loopback interfaces).
    - R3 manages LDP connection.
    - R3 sends SYN with MSS of 472 bytes.
        - 512 (Interface IP MTU) - 20 (minTCP_H) - 20 (minIP_H).
    - R2 sends SYN, ACK with an MSS of 472 bytes.
        - Sends the lower of [Received MSS; Local initial MSS].
        - Received MSS 472 bytes; Local initial MSS 472 bytes.
    - The lowest MSS value is used on both peers.

TCP (LDP) session details as seen on R3 – ACTIVE – MPLS-enabled scenario:

<#root>

```
! - as seen on R3 - Active
! - TCP path MTU discovery enabled
! - MPLS LDP enabled
! - TCP session initial state

RP/0/0/CPU0:R3#show tcp detail pcb 0x15393fbc
Mon May 17 08:33:30.627 UTC
================================================================
Connection state is ESTAB, I/O status: 0, socket status: 0
Established at Mon May 17 08:30:04 2021

PCB 0x15393fbc, SO 0x15393d94, TCPCB 0x153941b4, vrfid 0x60000000,
Pak Prio: Medium, TOS: 192, TTL: 255, Hash index: 970
Local host: 192.168.0.3, Local port: 57146 (Local App PID: 1151216)
Foreign host: 192.168.0.2, Foreign port: 646
(Local App PID/instance/SPL_APP_ID: 1151216/0/0)
```

```
Current send queue size in bytes: 0 (max 16384)
Current receive queue size in bytes: 0 (max 16384)  mis-ordered: 0 bytes
Current receive queue size in packets: 0 (max 60)

Timer          Starts    Wakeups         Next(msec)
Retrans             8          1                 0
SendWnd             0          0                 0
TimeWait            0          0                 0
AckHold             6          4                 0
KeepAlive           1          0                 0
PmtuAger            0          0                 0
GiveUp              0          0                 0
Throttle            0          0                 0

   iss: 2917752466  snduna: 2917752838  sndnxt: 2917752838
sndmax: 2917752838  sndwnd: 16013       sndcwnd: 944
   irs: 228184383   rcvnxt: 228184763   rcvwnd: 16005   rcvadv: 228200768

SRTT: 103 ms,  RTTO: 580 ms,  RTV: 477 ms,  KRTT: 0 ms
minRTT: 9 ms,  maxRTT: 279 ms

ACK hold time: 200 ms, Keepalive time: 0 sec, SYN waittime: 30 sec
Giveup time: 0 ms, Retransmission retries: 0, Retransmit forever: FALSE
Connect retries remaining: 1, connect retry interval: 3 secs

State flags: none
Feature flags: Win Scale, Nagle,
```

Path MTU

```
Request flags: Win Scale
```

Datagrams (in bytes): MSS 472, peer MSS 472, min MSS 472, max MSS 472

```
Window scales: rcv 0, snd 0, request rcv 0, request snd 0
Timestamp option: recent 0, recent age 0, last ACK sent 0
Sack blocks {start, end}: none
Sack holes {start, end, dups, rxmit}: none

Socket options: SO_REUSEADDR, SO_REUSEPORT, SO_NBIO
Socket states: SS_ISCONNECTED, SS_PRIV
Socket receive buffer states: SB_SEL, SB_DEL_WAKEUP
Socket send buffer states: SB_DEL_WAKEUP
Socket receive buffer: Low/High watermark 1/16384
Socket send buffer   : Low/High watermark 2048/16384, Notify threshold 0
Socket misc info     : Rcv data size (sb_cc) 0, so_qlen 0,
                       so_q0len 0, so_qlimit 0, so_error 0
                       so_auto_rearm 1

PDU information:
 #PDU's in buffer: 0
FIB Lookup Cache:  IFH: 0x40  PD ctx: size: 0  data:
  Num Labels: 1  Label Stack:  0x5dc2
Num of peers with authentication info: 0

RP/0/0/CPU0:R3#
```

TCP (LDP) session details as seen on R2 – PASSIVE – MPLS-enabled scenario:

<#root>

```
! - as seen on R2 - Passive
! - TCP path MTU discovery enabled
! - MPLS LDP enabled
! - TCP session initial state

RP/0/0/CPU0:R2#show tcp detail pcb 0x153a1f44
Mon May 17 08:34:28.843 UTC
================================================================
Connection state is ESTAB, I/O status: 0, socket status: 0
Established at Mon May 17 08:30:31 2021

PCB 0x153a1f44, SO 0x153a1d1c, TCPCB 0x153a213c, vrfid 0x60000000,
Pak Prio: Medium, TOS: 192, TTL: 255, Hash index: 970
Local host: 192.168.0.2, Local port: 646 (Local App PID: 1151216)
Foreign host: 192.168.0.3, Foreign port: 57146
(Local App PID/instance/SPL_APP_ID: 1151216/0/0)

Current send queue size in bytes: 0 (max 16384)
Current receive queue size in bytes: 0 (max 16384)  mis-ordered: 0 bytes
Current receive queue size in packets: 0 (max 60)

Timer          Starts    Wakeups         Next(msec)
Retrans             7          0                0
SendWnd             0          0                0
TimeWait            0          0                0
AckHold             7          5                0
KeepAlive           1          0                0
PmtuAger            0          0                0
GiveUp              0          0                0
Throttle            0          0                0

   iss: 228184383   snduna: 228184763   sndnxt: 228184763
sndmax: 228184763   sndwnd: 16005       sndcwnd: 944
   irs: 2917752466  rcvnxt: 2917752856  rcvwnd: 15995   rcvadv: 2917768851

SRTT: 95 ms,  RTTO: 561 ms,  RTV: 466 ms,  KRTT: 0 ms
minRTT: 0 ms,  maxRTT: 219 ms

ACK hold time: 200 ms, Keepalive time: 0 sec, SYN waittime: 30 sec
Giveup time: 0 ms, Retransmission retries: 0, Retransmit forever: FALSE
Connect retries remaining: 0, connect retry interval: 0 secs

State flags: none
Feature flags: Win Scale, Nagle,
```

Path MTU

```
Request flags: Win Scale
```

Datagrams (in bytes): MSS 472, peer MSS 472, min MSS 472, max MSS 472

```
Window scales: rcv 0, snd 0, request rcv 0, request snd 0
Timestamp option: recent 0, recent age 0, last ACK sent 0
Sack blocks {start, end}: none
```

```
Sack holes {start, end, dups, rxmit}: none

Socket options: SO_REUSEADDR, SO_REUSEPORT, SO_NBIO
Socket states: SS_ISCONNECTED, SS_PRIV
Socket receive buffer states: SB_SEL, SB_DEL_WAKEUP
Socket send buffer states: SB_DEL_WAKEUP
Socket receive buffer: Low/High watermark 1/16384
Socket send buffer   : Low/High watermark 2048/16384, Notify threshold 0
Socket misc info     : Rcv data size (sb_cc) 0, so_qlen 0,
                       so_q0len 0, so_qlimit 0, so_error 0
                       so_auto_rearm 1

PDU information:
 #PDU's in buffer: 0
FIB Lookup Cache:  IFH: 0x60  PD ctx: size: 0  data:
  Num Labels: 1  Label Stack:  0x5dc1
Num of peers with authentication info: 0

RP/0/0/CPU0:R2#
```

After the BGP session is established, R1 sends the BGP Update message and receives ICMP message (Destination Unreachable – type 3; Fragmentation needed – Code 4) in return sourced from node R2 which triggers TCP PMTUD at node R1. This occurs because the IP packet that carries the BGP update message has the DF-bit set and the IP MTU of 512 bytes used at the R2/R3 segment is lower than the IP packet size of 1116 bytes. As before, the reception of this ICMP message triggers PMTUD. The difference in the MPLS-enabled scenario when compared with the previous non-MPLS scenarios is with regards to the **MTU of next hop** value included on node R2 ICMP message (Destination Unreachable – type 3; Fragmentation needed – Code 4). In this MPLS-enabled scenario, the **MTU of next hop** value accounts for the additional MPLS overhead of 4 bytes, which means it accounts for the egress MPLS label stack at R2, as seen in these outputs.

TCP path MTU discovery in action as seen on R1 – PASSIVE – MPLS-enabled scenario:

<#root>

```
! - as seen from R1 - Passive
! - R1 sends BGP Update message with IP length of 1116 Bytes
! - Note MPLS Header as packet is to be label-switched (single label ; IGP label)
! - note IP Header Flags shows DF bit set

455    0.044859       192.168.0.1 192.168.0.4 BGP    1134   UPDATE Message, KEEPALIVE Message

Frame 455: 1134 bytes on wire (9072 bits), 1134 bytes captured (9072 bits) on interface 0
Ethernet II, Src: fa:16:3e:42:18:05 (fa:16:3e:42:18:05), Dst: fa:16:3e:5c:f1:80 (fa:16:3e:5c:f1:80)

MultiProtocol Label Switching Header

, Label: 24002, Exp: 6, S: 1, TTL: 255
Internet Protocol Version 4, Src: 192.168.0.1, Dst: 192.168.0.4
    0100 .... = Version: 4
    .... 0101 = Header Length: 20 bytes (5)
    Differentiated Services Field: 0xc0 (DSCP: CS6, ECN: Not-ECT)


Total Length: 1116


    Identification: 0xc6dd (50909)
    Flags: 0x02 (Don't Fragment)
```

```
        0... .... = Reserved bit: Not set
```

.1.. .... = Don't fragment: Set

```
        ..0. .... = More fragments: Not set
    Fragment offset: 0
    Time to live: 255
    Protocol: TCP (6)
    Header checksum: 0xa5f4 [validation disabled]
    [Header checksum status: Unverified]
    Source: 192.168.0.1
    Destination: 192.168.0.4
    [Source GeoIP: Unknown]
    [Destination GeoIP: Unknown]
Transmission Control Protocol, Src Port: 179, Dst Port: 57400, Seq: 242, Ack: 175, Len: 1076
Border Gateway Protocol - UPDATE Message
Border Gateway Protocol - KEEPALIVE Message
<snip>

! - as seen from R1 - Passive
! - IP MTU on R2/R3 of 512 bytes is lower than IP packet length and DF bit is set
! - R1 receives ICMP error message from R2
! - note R2 ICMP error message carries Next-Hop MTU
! - “The size in octets of the largest datagram that could be forwarded, along the path of
!   the original datagram, without being fragmented at this router. The size includes the
!   IP header and IP data, and does not include any lower-level headers.”
! - In present MPLS-enabled scenario Next-Hop MTU value is 508 bytes
! - In previous non-MPLS scenario Next-Hop MTU value was 512 bytes

456    0.014117       10.2.3.1       192.168.0.1 ICMP   182
```

Destination unreachable (Fragmentation needed)

```
Frame 456: 182 bytes on wire (1456 bits), 182 bytes captured (1456 bits) on interface 0
Ethernet II, Src: fa:16:3e:5c:f1:80 (fa:16:3e:5c:f1:80), Dst: fa:16:3e:42:18:05 (fa:16:3e:42:18:05)
Internet Protocol Version 4, Src: 10.2.3.1, Dst: 192.168.0.1
    0100 .... = Version: 4
    .... 0101 = Header Length: 20 bytes (5)
    Differentiated Services Field: 0x00 (DSCP: CS0, ECN: Not-ECT)
    Total Length: 168
    Identification: 0x001f (31)
    Flags: 0x00
        0... .... = Reserved bit: Not set
        .0.. .... = Don't fragment: Not set
        ..0. .... = More fragments: Not se
    Fragment offset: 0
    Time to live: 251
```

**   Protocol: ICMP (1)**

```
    Header checksum: 0xb031 [validation disabled]
    [Header checksum status: Unverified]
    Source: 10.2.3.1
    Destination: 192.168.0.1
    [Source GeoIP: Unknown]
    [Destination GeoIP: Unknown]
Internet Control Message Protocol
```

**   Type: 3 (Destination unreachable)**

Code: 4 (Fragmentation needed)

Checksum: 0x5199 [correct]
[Checksum Status: Good]
Length: 17
[Length of original datagram: 68]
Unused: 0011

MTU of next hop: 508

Internet Protocol Version 4, Src: 192.168.0.1, Dst: 192.168.0.4
Transmission Control Protocol, Src Port: 179, Dst Port: 57400, Seq: 3874414921, Ack: 1386460094
Border Gateway Protocol - UPDATE Message

```
! - As seen from R1 - Passive
! - Hint is provided by ICMP unreachable message MTU of next-hop field: 508 bytes
! - R1 then considers this value and retransmits BGP Update split in three distinct packets
! - Sum of TCP length = 468 + 468 + 140 = 1076 bytes

457    0.006689       192.168.0.1 192.168.0.4 TCP    526    [TCP Retransmission] 179 → 57400 [ACK] Seq=24
```

Len=468

```
460    0.004001       192.168.0.1 192.168.0.4 TCP    526    [TCP Retransmission] 179 → 57400 [ACK] Seq=71
```

Len=468

```
461    0.001788       192.168.0.1 192.168.0.4 TCP    198    [TCP Retransmission] 179 → 57400 [PSH, ACK] S
```

Len=140

```
463    0.056695       192.168.0.4 192.168.0.1 TCP    54     57400 → 179 [ACK] Seq=175 Ack=1318 Win=31545 Le

! - As seen from R1 - Passive - 'debug tcp pmtud' and 'debug icmp' active
! - TCP PMTUD is triggered once ICMP unreachable received

RP/0/0/CPU0:May 17 08:29:56.131 UTC: tcp[399]: [t1] Try to enable path MTU discovery(neww age timer: 10
RP/0/0/CPU0:May 17 08:29:56.131 UTC: tcp[399]: [t1] Path mtu is ON (age-timer: 10)
RP/0/0/CPU0:May 17 08:35:51.726 UTC: ipv4_io[266]: ip_icmp_lib_ipv4_receive: Receiving pak(0xb0c07d8f)
RP/0/0/CPU0:May 17 08:35:51.726 UTC: ipv4_io[266]: Entering ipv4_mtu_update_cb
RP/0/0/CPU0:May 17 08:35:51.726 UTC: ipv4_io[266]: IPv4 ICMP: Received ICMP too big from 192.168.0.1 ab
RP/0/0/CPU0:May 17 08:35:51.726 UTC: ipv4_io[266]: ipv4_icmp_unreachable_rcvd ICMP unreach recvd: sendi
RP/0/0/CPU0:May 17 08:35:51.726 UTC: ipv4_io[266]: ip_icmp_lib_ipv4_receive: sending pak(0xb0c07d8f) to
RP/0/0/CPU0:May 17 08:35:51.726 UTC: tcp[399]: [t4] PCB 0x153acc8c: Process ICMP Dest-unreach (next hop

! - attempt new MSS 468 = MTU of next-hop(508) - TCP_H(20) - IP_H(20)

RP/0/0/CPU0:May 17 08:35:51.726 UTC: tcp[399]: [t4] PCB 0x153acc8c: Try to use new MSS: 468
RP/0/0/CPU0:May 17 08:35:51.726 UTC: tcp[399]: [t4] PCB 0x153acc8c, New path MTU decided to use: 468  c

! - over time PMTUD attempts to raise MSS as per egress interface configured MTU

RP/0/0/CPU0:May 17 08:45:51.745 UTC: tcp[399]: [t29] PCB 0x153acc8c: Trying next higher MTU: 966
RP/0/0/CPU0:May 17 08:47:51.757 UTC: tcp[399]: [t29] PCB 0x153acc8c: Trying next higher MTU: 1452
RP/0/0/CPU0:May 17 08:49:51.769 UTC: tcp[399]: [t29] PCB 0x153acc8c: Trying next higher MTU: 1460
```

As seen from R1 – PASSIVE – TCP PMTUD triggered – MPLS-enabled scenario:

<#root>

```
! - as seen on R1 - Passive
! - R1 session details after TCP PMTUD trigger

RP/0/0/CPU0:R1#show tcp detail pcb 0x153acc8c
Mon May 17 08:43:07.077 UTC
================================================================
Connection state is ESTAB, I/O status: 240, socket status: 0
Established at Mon May 17 08:31:55 2021

PCB 0x153acc8c, SO 0x153adad4, TCPCB 0x153adcfc, vrfid 0x60000000,
Pak Prio: Medium, TOS: 192, TTL: 255, Hash index: 757
Local host: 192.168.0.1, Local port: 179 (Local App PID: 1192224)
Foreign host: 192.168.0.4, Foreign port: 57400
(Local App PID/instance/SPL_APP_ID: 1192224/1/0)

Current send queue size in bytes: 0 (max 24576)
Current receive queue size in bytes: 0 (max 32768)  mis-ordered: 0 bytes
Current receive queue size in packets: 0 (max 0)

Timer          Starts    Wakeups         Next(msec)
Retrans            15          0                0
SendWnd             0          0                0
TimeWait            0          0                0
AckHold            14          9                0
KeepAlive           1          0                0

PmtuAger            1          0           164599


GiveUp              0          0                0
Throttle            0          0                0

   iss: 3874414679  snduna: 3874416130  sndnxt: 3874416130
sndmax: 3874416130  sndwnd: 31412       sndcwnd: 936
   irs: 1386459919  rcvnxt: 1386460246  rcvwnd: 32517   rcvadv: 1386492763

SRTT: 180 ms,  RTTO: 509 ms,  RTV: 329 ms,  KRTT: 0 ms
minRTT: 19 ms,  maxRTT: 239 ms

ACK hold time: 200 ms, Keepalive time: 0 sec, SYN waittime: 30 sec
Giveup time: 0 ms, Retransmission retries: 0, Retransmit forever: FALSE
Connect retries remaining: 0, connect retry interval: 0 secs

State flags: PMTU ager
Feature flags: Win Scale, Nagle,

Path MTU


Request flags: Win Scale


Datagrams (in bytes): MSS 468, peer MSS 1460, min MSS 468, max MSS 1460



Window scales: rcv 0, snd 0, request rcv 0, request snd 0
Timestamp option: recent 0, recent age 0, last ACK sent 0
```

```
Sack blocks {start, end}: none
Sack holes {start, end, dups, rxmit}: none

Socket options: SO_REUSEADDR, SO_REUSEPORT, SO_NBIO
Socket states: SS_ISCONNECTED, SS_PRIV
Socket receive buffer states: SB_DEL_WAKEUP
Socket send buffer states: SB_DEL_WAKEUP
Socket receive buffer: Low/High watermark 1/32768
Socket send buffer   : Low/High watermark 2048/24576, Notify threshold 0
Socket misc info     : Rcv data size (sb_cc) 0, so_qlen 0,
                       so_q0len 0, so_qlimit 0, so_error 0
                       so_auto_rearm 1

PDU information:
 #PDU's in buffer: 0
FIB Lookup Cache:  IFH: 0x20  PD ctx: size: 0  data:
  Num Labels: 1  Label Stack:  0x5dc3
Num of peers with authentication info: 0

RP/0/0/CPU0:R1#
```

Do note that in the MPLS-enabled scenario the value of the **MTU of next hop** included on node R2 ICMP message accounts for the egress MPLS label stack. To further reinforce this aspect, consider the next example. If the IP packet filtered at R2 is associated with an L3VPN service, it means that the Ethernet frame now carries two labels (IGP label and VPN label). Then the **MTU of next hop** reflects the required label stack size. Refer to these outputs.

As seen on R1 – PASSIVE – L3 VPN service packet:

<#root>

```
! - as seen from R1 - Passive
! - L3 VPN service packet is sourced by node R1 and destined to node R4
! - Note presence of MPLS label stack - both IGP and VPN label are present
! - Note IP Total Length of 610 bytes higher than the IP MTU on R2/R3 segment
! - note IP Header Flags shows DF bit set

2024   0.302370       10.1.14.1      10.1.14.14     TELNET 632    Telnet Data ...

Frame 2024: 632 bytes on wire (5056 bits), 632 bytes captured (5056 bits) on interface 0
Ethernet II, Src: fa:16:3e:42:18:05 (fa:16:3e:42:18:05), Dst: fa:16:3e:5c:f1:80 (fa:16:3e:5c:f1:80)

MultiProtocol Label Switching Header, Label: 24002, Exp: 0, S: 0, TTL: 255


    0000 0101 1101 1100 0010 .... .... .... = MPLS Label: 24002
    .... .... .... .... .... 000. .... .... = MPLS Experimental Bits: 0
    .... .... .... .... .... ...0 .... .... = MPLS Bottom Of Label Stack: 0
    .... .... .... .... .... .... 1111 1111 = MPLS TTL: 255

MultiProtocol Label Switching Header, Label: 24005, Exp: 0, S: 1, TTL: 255


    0000 0101 1101 1100 0101 .... .... .... = MPLS Label: 24005
    .... .... .... .... .... 000. .... .... = MPLS Experimental Bits: 0
    .... .... .... .... .... ...1 .... .... = MPLS Bottom Of Label Stack: 1
    .... .... .... .... .... .... 1111 1111 = MPLS TTL: 255
Internet Protocol Version 4, Src: 10.1.14.1, Dst: 10.1.14.14
    0100 .... = Version: 4
```

```
    .... 0101 = Header Length: 20 bytes (5)
    Differentiated Services Field: 0x00 (DSCP: CS0, ECN: Not-ECT)
    Total Length: 610
    Identification: 0x7c9f (31903)
    Flags: 0x02 (Don't Fragment)
        0... .... = Reserved bit: Not set


.1.. .... = Don't fragment: Set


        ..0. .... = More fragments: Not set
    Fragment offset: 0
    Time to live: 255
    Protocol: TCP (6)
    Header checksum: 0xcce5 [validation disabled]
    [Header checksum status: Unverified]
    Source: 10.1.14.1
    Destination: 10.1.14.14
    [Source GeoIP: Unknown]
    [Destination GeoIP: Unknown]
Transmission Control Protocol, Src Port: 22008, Dst Port: 23, Seq: 34755, Ack: 93250, Len: 570
```

As seen on R1 – PASSIVE – L3 VPN service – ICMP Type 3/Code 4:

<#root>

```
! - as seen from R1 - Passive
! - IP MTU on R2/R3 of 512 bytes is lower than IP packet length and DF bit is set
! - R1 receives ICMP error message from R2
! - note R2 ICMP error message carries Next-Hop MTU
! - “The size in octets of the largest datagram that could be forwarded, along the path of
!   the original datagram, without being fragmented at this router. The size includes the
!   IP header and IP data, and does not include any lower-level headers.”
! - In present L3VPN MPLS-enabled scenario (dual-label) Next-Hop MTU value is 504 bytes
! - In previous MPLS scenario (single-label) Next-Hop MTU value was 508 bytes

2030   0.020299       10.2.3.1       10.1.14.1      ICMP   190

Destination unreachable (Fragmentation needed)


Frame 2030: 190 bytes on wire (1520 bits), 190 bytes captured (1520 bits) on interface 0
Ethernet II, Src: fa:16:3e:5c:f1:80 (fa:16:3e:5c:f1:80), Dst: fa:16:3e:42:18:05 (fa:16:3e:42:18:05)
MultiProtocol Label Switching Header, Label: 24005, Exp: 0, S: 1, TTL: 251
    0000 0101 1101 1100 0101 .... .... .... = MPLS Label: 24005
    .... .... .... .... .... 000. .... .... = MPLS Experimental Bits: 0
    .... .... .... .... .... ...1 .... .... = MPLS Bottom Of Label Stack: 1
    .... .... .... .... .... .... 1111 1011 = MPLS TTL: 251
Internet Protocol Version 4, Src: 10.2.3.1, Dst: 10.1.14.1
    0100 .... = Version: 4
    .... 0101 = Header Length: 20 bytes (5)
    Differentiated Services Field: 0x00 (DSCP: CS0, ECN: Not-ECT)
    Total Length: 172
    Identification: 0x002b (43)
    Flags: 0x00
        0... .... = Reserved bit: Not set
        .0.. .... = Don't fragment: Not set
        ..0. .... = More fragments: Not set
```

```
    Fragment offset: 0
    Time to live: 253


Protocol: ICMP (1)


    Header checksum: 0x9821 [validation disabled]
    [Header checksum status: Unverified]
    Source: 10.2.3.1
    Destination: 10.1.14.1
    [Source GeoIP: Unknown]
    [Destination GeoIP: Unknown]
Internet Control Message Protocol
    Type: 3 (Destination unreachable)


    Code: 4 (Fragmentation needed)


    Checksum: 0xbbac [correct]
    [Checksum Status: Good]
    Length: 17
    [Length of original datagram: 68]
    Unused: 0011


MTU of next hop: 504


    Internet Protocol Version 4, Src: 10.1.14.1, Dst: 10.1.14.14
        0100 .... = Version: 4
        .... 0101 = Header Length: 20 bytes (5)
        Differentiated Services Field: 0x00 (DSCP: CS0, ECN: Not-ECT)
        Total Length: 610
        Identification: 0x7c9f (31903)
        Flags: 0x02 (Don't Fragment)
            0... .... = Reserved bit: Not set
            .1.. .... = Don't fragment: Set
            ..0. .... = More fragments: Not set
        Fragment offset: 0
        Time to live: 255
        Protocol: TCP (6)
        Header checksum: 0xcce5 [validation disabled]
        [Header checksum status: Unverified]
        Source: 10.1.14.1
        Destination: 10.1.14.14
        [Source GeoIP: Unknown]
        [Destination GeoIP: Unknown]
    Transmission Control Protocol, Src Port: 22008, Dst Port: 23, Seq: 586828435, Ack: 754580617
```

PMTUD – TCP Options (MD5)

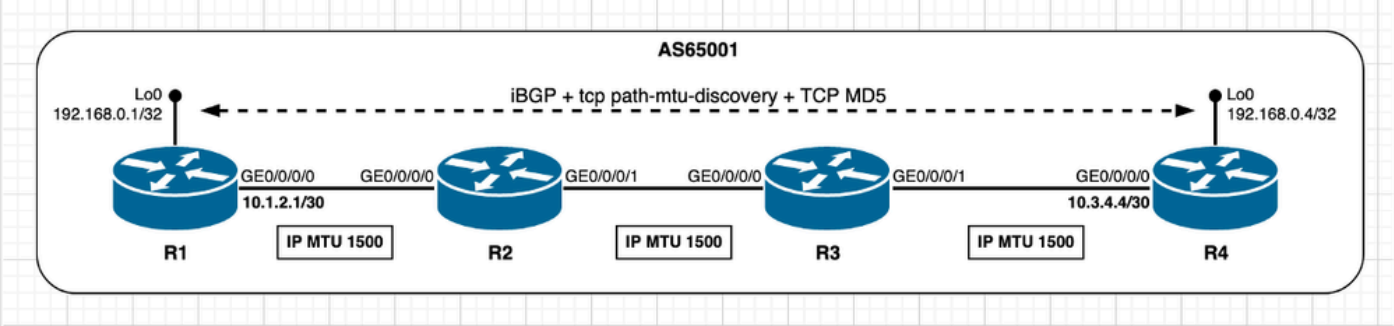


Image 3.4 – PMTUD enabled and TCP MD5 Authentication.

No distinction with regards to the PMTUD behavior is introduced from what has already been described in the previous scenarios with TCP MD5 authentication enabled. As previously shared with TCP MD5 authentication in use, Cisco IOS XR considers additional overhead and active TCP peer initial MSS value reflects same. Please refer to previous sections **Use TCP Options – XR Active** and **Use TCP Options – XR Passive** for additional details on the impact of TCP options use. TCP MSS calculation in this scenario can be summarized as follows:

- All nodes use default IP MTU of 1500 bytes.
- TCP Path MTU discovery is enabled.
- TCP Peers are not directly connected.
- TCP MD5 authentication enabled on both R1 and R4.
    - R4 manages BGP connection.
    - R4 sends SYN with MSS of 1436 bytes.
        - 1500 (Interface IP MTU) - 20 (minTCP_H) - 20 (minIP_H) - 24 Bytes (IOS XR TCP Options Overhead).
    - R1 send SYN, ACK with MSS of 1436 bytes.
        - sends the lower of [Received MSS ; Local initial MSS].
        - Received MSS 1436 bytes; Local initial MSS 1460 bytes.
    - Lowest MSS value is used on both peers.

TCP SYN sourced from R4:

<#root>

```
! – TCP SYN sourced from R4

2408   5.695076       192.168.0.4 192.168.0.1 TCP    82     59050 → 179 [SYN] Seq=0 Win=16384 Len=0
MSS=1436
 WS=1

Frame 2408: 82 bytes on wire (656 bits), 82 bytes captured (656 bits) on interface 0
Ethernet II, Src: fa:16:3e:d7:7e:f6 (fa:16:3e:d7:7e:f6), Dst: fa:16:3e:8f:8f:54 (fa:16:3e:8f:8f:54)
Internet Protocol Version 4, Src: 192.168.0.4, Dst: 192.168.0.1
Transmission Control Protocol, Src Port: 59050, Dst Port: 179, Seq: 0, Len: 0
    Source Port: 59050
    Destination Port: 179
    [Stream index: 8]
    [TCP Segment Len: 0]
    Sequence number: 0    (relative sequence number)
    Acknowledgment number: 0
    Header Length: 48 bytes
    Flags: 0x002 (SYN)
    Window size value: 16384
```

```
    [Calculated window size: 16384]
    Checksum: 0x20d7 [unverified]
    [Checksum Status: Unverified]
    Urgent pointer: 0
    Options: (28 bytes), Maximum segment size, Window scale, No-Operation (NOP),
```

TCP MD5 signature

```
, End of Option List (EOL)
        Maximum segment size: 1436 bytes
            Kind: Maximum Segment Size (2)
            Length: 4
```

MSS Value: 1436

```
        Window scale: 0 (multiply by 1)
        No-Operation (NOP)
        TCP MD5 signature
        End of Option List (EOL)
```

TCP SYN, ACK sourced from R1:

<#root>

```
! – TCP SYN,ACK sourced from R1

2409   0.004352       192.168.0.1 192.168.0.4 TCP    82     179 → 59050 [SYN, ACK] Seq=0 Ack=1 Win=16384 Le
```

MSS=1436

```
 WS=1

Frame 2409: 82 bytes on wire (656 bits), 82 bytes captured (656 bits) on interface 0
Ethernet II, Src: fa:16:3e:8f:8f:54 (fa:16:3e:8f:8f:54), Dst: fa:16:3e:d7:7e:f6 (fa:16:3e:d7:7e:f6)
Internet Protocol Version 4, Src: 192.168.0.1, Dst: 192.168.0.4
Transmission Control Protocol, Src Port: 179, Dst Port: 59050, Seq: 0, Ack: 1, Len: 0
    Source Port: 179
    Destination Port: 59050
    [Stream index: 8]
    [TCP Segment Len: 0]
    Sequence number: 0    (relative sequence number)
    Acknowledgment number: 1    (relative ack number)
    Header Length: 48 bytes
    Flags: 0x012 (SYN, ACK)
    Window size value: 16384
    [Calculated window size: 16384]
    Checksum: 0xcbf8 [unverified]
    [Checksum Status: Unverified]
    Urgent pointer: 0
    Options: (28 bytes), Maximum segment size, Window scale, No-Operation (NOP),
```

TCP MD5 signature

```
, End of Option List (EOL)
        Maximum segment size: 1436 bytes
            Kind: Maximum Segment Size (2)
            Length: 4
```

MSS Value: 1436

```
        Window scale: 0 (multiply by 1)
        No-Operation (NOP)
        TCP MD5 signature
        End of Option List (EOL)
```

TCP session details as seen on R4 – ACTIVE:

<#root>

```
! - as seen from R4 - Active

RP/0/0/CPU0:R4#show tcp detail pcb 0x121542c0
Tue Jan 12 13:27:23.526 UTC
==========================================================
Connection state is ESTAB, I/O status: 0, socket status: 0
Established at Tue Jan 12 13:25:41 2021

PCB 0x121542c0, SO 0x1213c0e4, TCPCB 0x12156010, vrfid 0x60000000,
Pak Prio: Medium, TOS: 192, TTL: 255, Hash index: 359
Local host: 192.168.0.4, Local port: 59050 (Local App PID: 1052958)
Foreign host: 192.168.0.1, Foreign port: 179

Current send queue size in bytes: 0 (max 24576)
Current receive queue size in bytes: 0 (max 32768)  mis-ordered: 0 bytes
Current receive queue size in packets: 0 (max 0)

Timer          Starts    Wakeups         Next(msec)
Retrans             6          1                 0
SendWnd             0          0                 0
TimeWait            0          0                 0
AckHold             3          2                 0
KeepAlive           1          0                 0
PmtuAger            0          0                 0
GiveUp              0          0                 0
Throttle            0          0                 0

   iss: 3299472269  snduna: 3299473445  sndnxt: 3299473445
sndmax: 3299473445  sndwnd: 31646       sndcwnd: 4308
   irs: 3225544359  rcvnxt: 3225545535  rcvwnd: 31665   rcvadv: 3225577200

SRTT: 89 ms,  RTTO: 530 ms,  RTV: 441 ms,  KRTT: 0 ms
minRTT: 19 ms,  maxRTT: 239 ms

ACK hold time: 200 ms, Keepalive time: 0 sec, SYN waittime: 30 sec
Giveup time: 0 ms, Retransmission retries: 0, Retransmit forever: FALSE
Connect retries remaining: 30, connect retry interval: 30 secs

State flags: none
Feature flags:
```

MD5

```
, Win Scale, Nagle,
```

Path MTU

```

Request flags: Win Scale
```

Datagrams (in bytes): MSS 1436, peer MSS 1436, min MSS 1436, max MSS 1436

```
Window scales: rcv 0, snd 0, request rcv 0, request snd 0
Timestamp option: recent 0, recent age 0, last ACK sent 0
Sack blocks {start, end}: none
Sack holes {start, end, dups, rxmit}: none

Socket options: SO_REUSEADDR, SO_REUSEPORT, SO_NBIO
Socket states: SS_ISCONNECTED, SS_PRIV
Socket receive buffer states: SB_DEL_WAKEUP
Socket send buffer states: SB_DEL_WAKEUP
Socket receive buffer: Low/High watermark 1/32768
Socket send buffer   : Low/High watermark 2048/24576, Notify threshold 0

PDU information:
 #PDU's in buffer: 0
FIB Lookup Cache:  IFH: 0x40  PD ctx: size: 0  data:
  Num Labels: 0  Label Stack:

RP/0/0/CPU0:R4#
```

TCP session details as seen on R1 – PASSIVE:

<#root>

```
! - as seen from R1 - Passive

RP/0/0/CPU0:R1#show tcp detail pcb 0x121560ec
Tue Jan 12 13:25:59.310 UTC
==============================================================
Connection state is ESTAB, I/O status: 0, socket status: 0
Established at Tue Jan 12 13:25:31 2021

PCB 0x121560ec, SO 0x121556d4, TCPCB 0x121575bc, vrfid 0x60000000,
Pak Prio: Medium, TOS: 192, TTL: 255, Hash index: 359
Local host: 192.168.0.1, Local port: 179 (Local App PID: 983326)
Foreign host: 192.168.0.4, Foreign port: 59050

Current send queue size in bytes: 0 (max 24576)
Current receive queue size in bytes: 0 (max 32768)  mis-ordered: 0 bytes
Current receive queue size in packets: 0 (max 0)

Timer          Starts    Wakeups         Next(msec)
Retrans             3          0                 0
SendWnd             0          0                 0
TimeWait            0          0                 0
AckHold             3          2                 0
KeepAlive           1          0                 0
PmtuAger            0          0                 0
GiveUp              0          0                 0
Throttle            0          0                 0

   iss: 3225544359  snduna: 3225545516  sndnxt: 3225545516
sndmax: 3225545516  sndwnd: 31684       sndcwnd: 4308
   irs: 3299472269  rcvnxt: 3299473426  rcvwnd: 31665   rcvadv: 3299505091

SRTT: 37 ms,  RTTO: 300 ms,  RTV: 244 ms,  KRTT: 0 ms
```

```
minRTT: 9 ms,  maxRTT: 239 ms

ACK hold time: 200 ms, Keepalive time: 0 sec, SYN waittime: 30 sec
Giveup time: 0 ms, Retransmission retries: 0, Retransmit forever: FALSE
Connect retries remaining: 0, connect retry interval: 0 secs

State flags: none
Feature flags:
```

MD5

```
, Win Scale, Nagle,
```

Path MTU

```
Request flags: Win Scale
```

Datagrams (in bytes): MSS 1436, peer MSS 1436, min MSS 1460, max MSS 1460

```
Window scales: rcv 0, snd 0, request rcv 0, request snd 0
Timestamp option: recent 0, recent age 0, last ACK sent 0
Sack blocks {start, end}: none
Sack holes {start, end, dups, rxmit}: none

Socket options: SO_REUSEADDR, SO_REUSEPORT, SO_NBIO
Socket states: SS_ISCONNECTED, SS_PRIV
Socket receive buffer states: SB_DEL_WAKEUP
Socket send buffer states: SB_DEL_WAKEUP
Socket receive buffer: Low/High watermark 1/32768
Socket send buffer   : Low/High watermark 2048/24576, Notify threshold 0

PDU information:
 #PDU's in buffer: 0
FIB Lookup Cache:  IFH: 0x40  PD ctx: size: 0  data:
  Num Labels: 0  Label Stack:

RP/0/0/CPU0:R1#
```

PMTUD – Blackhole Detection

As previously explained in section **PMTUD – Path Segment has Lower IP MTU**, the TCP PMTUD when enabled is triggered by the reception of an ICMP (Destination Unreachable – type 3; Fragmentation needed – Code 4) message. It can be the case that these messages are for some reason not received which results that PMTUD is not triggered. In that case, the lowest IP MTU of the path between the TCP peers is not learnt. Such a scenario would introduce a potential blackhole if IP packets have the DF-bit set and if they have a size higher than the lowest IP MTU path segment. Those packets would be silently dropped.

This section aims to highlight how Cisco IOS XR detects and acts on such potential blackhole scenario. For this purpose IPv4 unreachables feature is disabled at the R2 interface GE0/0/0/0 as depicted in the next image and CLI output.

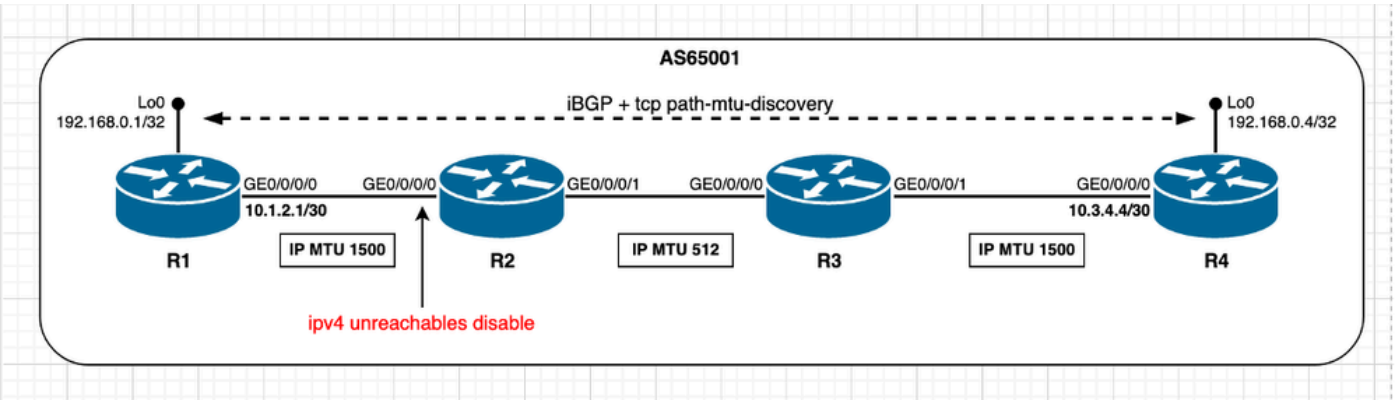


Image 3.5 – PMTUD enabled on R1/R4 and R2 IPv4 unreachables disabled.

IPv4 unreachables disabled at R2:

<#root>

```
!- R2 - IP unreachables is disabled

RP/0/0/CPU0:R2#show run interface gigabitEthernet 0/0/0/0
Thu May 13 12:09:45.483 UTC
interface GigabitEthernet0/0/0/0
 ipv4 address 10.1.2.2 255.255.255.252
ipv4 unreachables disable


!

RP/0/0/CPU0:R2#show ipv4 interface gigabitEthernet 0/0/0/0
Thu May 13 12:10:04.112 UTC
GigabitEthernet0/0/0/0 is Up, ipv4 protocol is Up
  Vrf is default (vrfid 0x60000000)
  Internet address is 10.1.2.2/30
  MTU is 1514 (1500 is available to IP)
  Helper address is not set
  Multicast reserved groups joined: 224.0.0.2 224.0.0.1 224.0.0.5
      224.0.0.6
  Directed broadcast forwarding is disabled
  Outgoing access list is not set
  Inbound  common access list is not set, access list is not set
  Proxy ARP is disabled
  ICMP redirects are never sent


ICMP unreachables are never sent


  ICMP mask replies are never sent
  Table Id is 0xe0000000
```

The way Cisco IOS XR deals with this blackhole scenario is to retransmit the same packet twice and if still unsuccessful, that is the expected TCP ACK is not received, then retry but use the next lower well-defined plateau value as documented in RFC1191 – Path MTU discovery (refer to section **PMTUD – Path Segment has Lower IP MTU** for the list of plateaus). In summary, Cisco IOS XR assumes that it can be the case that packets are dropped somewhere within the path to their destination due to their size and attempts to work

around it via packet retransmission. This behavior can be observed with the next example from a packet capture taken at node R1 interface and the output from **debug tcp pmtud**.

IOS-XR blackhole detection at R1:

<#root>

```
! - at R1
! - Original BGP Update message is sent
! - Note IP Total Length of 1116 bytes and TCP Segment Length of 1076 bytes
! - R2 filters such packet and send and ICMP error message towards R1 which triggers PMTUD
! - But because IPv4 unreachables are disabled at R2 GE0/0/0/0 ICMP message is not sent
! - Hence BGP message is silently filtered at R2

562    7.638774       192.168.0.1 192.168.0.4 BGP    1130   UPDATE Message, KEEPALIVE Message

Frame 562: 1130 bytes on wire (9040 bits), 1130 bytes captured (9040 bits) on interface 0
Ethernet II, Src: fa:16:3e:42:18:05 (fa:16:3e:42:18:05), Dst: fa:16:3e:5c:f1:80 (fa:16:3e:5c:f1:80)
Internet Protocol Version 4, Src: 192.168.0.1, Dst: 192.168.0.4
    0100 .... = Version: 4
    .... 0101 = Header Length: 20 bytes (5)
    Differentiated Services Field: 0xc0 (DSCP: CS6, ECN: Not-ECT)


Total Length: 1116


    Identification: 0x4a37 (18999)
    Flags: 0x02 (Don't Fragment)
        0... .... = Reserved bit: Not set


.1.. .... = Don't fragment: Set


        ..0. .... = More fragments: Not set
    Fragment offset: 0
    Time to live: 255
    Protocol: TCP (6)
    Header checksum: 0x229b [validation disabled]
    [Header checksum status: Unverified]
    Source: 192.168.0.1
    Destination: 192.168.0.4
    [Source GeoIP: Unknown]
    [Destination GeoIP: Unknown]
Transmission Control Protocol, Src Port: 179, Dst Port: 57082, Seq: 318, Ack: 251, Len: 1076
Border Gateway Protocol - UPDATE Message
Border Gateway Protocol - KEEPALIVE Message
<snip>

! - at R1
! - No TCP ACK is received
! - Packet retransmission is attempted (2 attempts)
! - Note initial MSS value is of 1460 bytes

563    0.560058       192.168.0.1 192.168.0.4 TCP    1130   [TCP Retransmission] 179 → 57082 [PSH, ACK] S
564    1.101367       192.168.0.1 192.168.0.4 TCP    1130   [TCP Retransmission] 179 → 57082 [PSH, ACK] S

! - at R1
! - Still no TCP ACK received; previous retransmissions failed
! - Next lower plateau value is attempted - 1492 bytes
! - Packet retransmission is attempted (2 attempts)
```

```
RP/0/0/CPU0:May 13 10:20:44.251 UTC: tcp[399]: [t1] PCB 0x15392224: Trying next lower MTU: 1452

567    1.850294       192.168.0.1 192.168.0.4 TCP    1130   [TCP Retransmission] 179 → 57082 [PSH, ACK] S
568    1.111361       192.168.0.1 192.168.0.4 TCP    1130   [TCP Retransmission] 179 → 57082 [PSH, ACK] S

! - at R1
! - Still no TCP ACK received; previous retransmissions failed
! - Next lower plateau value is attempted - 1006 bytes
! - Packet retransmission is attempted (2 attempts)

RP/0/0/CPU0:May 13 10:20:47.560 UTC: tcp[399]: [t1] PCB 0x15392224: Trying next lower MTU: 966

569    2.198327       192.168.0.1 192.168.0.4 TCP    1020   [TCP Retransmission] 179 → 57082 [ACK] Seq=31
570    1.109602       192.168.0.1 192.168.0.4 TCP    1020   [TCP Retransmission] 179 → 57082 [ACK] Seq=31

! - at R1
! - Still no TCP ACK received; previous retransmissions failed
! - Next lower plateau value is attempted - 508 bytes
! - Original information (TCP Length of 1076 bytes) is split in three distinct packets
! - TCP Segment Lengths 468 + 468 + 140 = 1076
! - TCP ACK is received from peer R4

RP/0/0/CPU0:May 13 10:20:50.870 UTC: tcp[399]: [t1] PCB 0x15392224: Trying next lower MTU: 468

571    2.205552       192.168.0.1 192.168.0.4 TCP    522    [TCP Retransmission] 179 → 57082 [ACK] Seq=31

Len=468


573    0.004254       192.168.0.1 192.168.0.4 TCP    522    [TCP Retransmission] 179 → 57082 [ACK] Seq=78

Len=468


574    0.002724       192.168.0.1 192.168.0.4 TCP    194    [TCP Retransmission] 179 → 57082 [PSH, ACK] S

Len=140



! - Peer R4 TCP ACK is received

575    0.223172       192.168.0.4 192.168.0.1 TCP    54     57082 → 179 [ACK] Seq=251 Ack=1394 Win=31469 Le
```