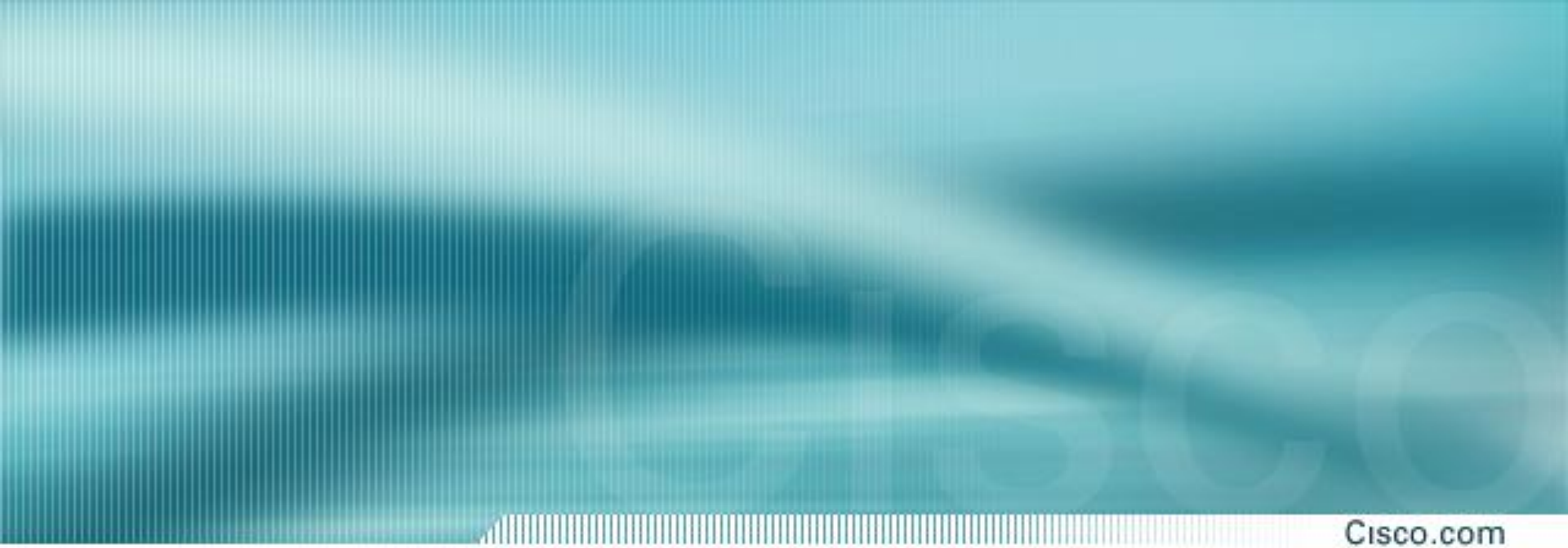


Understanding MPLS MTU Fragmentation & ICMP Unreachables

Prepared By Soumitra Mukherji

Advanced Network Services Consulting Engineer

Cisco Systems

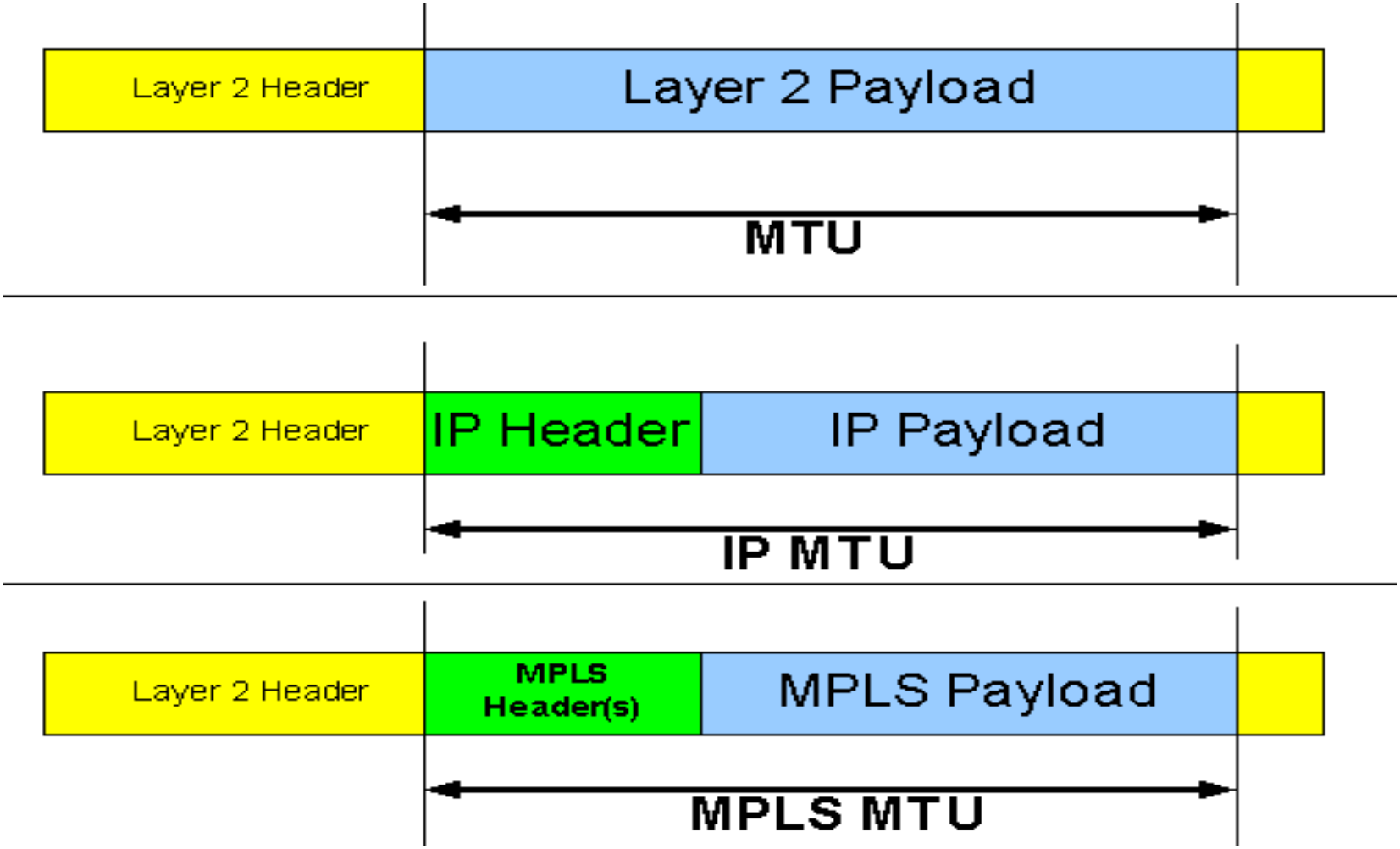
© 2003, Cisco Systems, Inc. All rights reserved.

Topics of Discussion

- MTU
- IP MTU
- MPLS MTU
- MPLS MTU & Fragmentation
- MPLS MTU & ICMP Unreachables

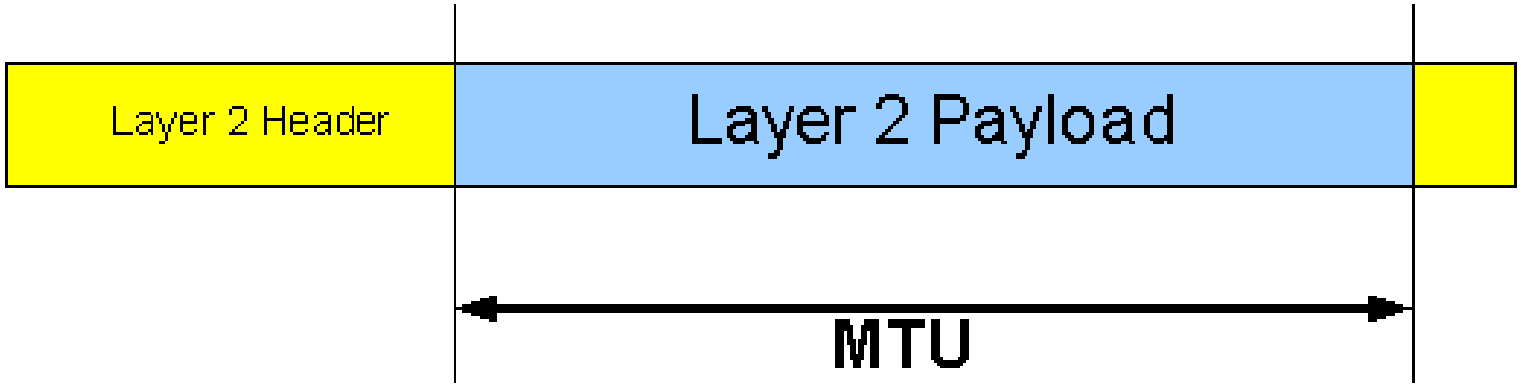
© 2003, Cisco Systems, Inc. All rights reserved.

MTU, IP MTU, MPLS MTU



© 2003, Cisco Systems, Inc. All rights reserved.

MTU



- **MTU setting sets a hard limit to the payload of the layer 2 encapsulation**
- **Default behavior is for IP MTU and MPLS MTU to follow MTU settings**
- **MTU setting determines the Maximum outgoing MTU of an interface**
- **Default setting of Ethernet Links is 1500 Bytes, max 9180 Bytes**
- **Default setting of Sonet based links is 4470 Bytes, max 18020 Bytes**

© 2003, Cisco Systems, Inc. All rights reserved.

MTU

- **Interface Command to change outgoing MTU on interface:**

mtu <x>

Example Changing Mtu in E2, Trident Line Card in GSR

R28-12410a(config-if)#mtu ?

<1500-9180> MTU size in bytes

- **Interface Command to view outgoing MTU on interface:**

sh int <interface type, number>

Example Viewing MTU

R28-12410a# sh int gigabitEthernet 3/1 | i MTU

MTU 1500 bytes, BW 1000000 Kbit, DLY 10 usec, rely 255/255, load 1/255

© 2003, Cisco Systems, Inc. All rights reserved.

MTU

Support for Jumbo Frames on Cisco 12000 Series 3-Port Gigabit Ethernet Line Cards

The Cisco 12000 series 3-port Gigabit Ethernet line cards that are numbered from 73-4775-02 revision E0 and later and from 800-06376-01 revision E0 and later are now enabled to handle incoming packets of sizes till 9180 bytes (also referred to as Jumbo Frames) on the first two ports of the line cards. Older revision line

cards can support maximum transmission unit (MTU) sizes up to 2450 bytes.

Documented by:

http://www.cisco.com/univercd/cc/td/doc/product/software/ios120/relnote/xprn120s/120snewf.htm

© 2003, Cisco Systems, Inc. All rights reserved.

MTU

- **Example of GigabitEthernet Trident LC on GSR**

R28-12410a#sh diag 3 | i PCA

PCA: 73-4775-02 rev E0 ver 2

R28-12410a(config-if)#int g 3/1

R28-12410a(config-if)#mtu ?

<1500-9180> MTU size in bytes

R28-12410a(config-if)#int g 3/1

R28-12410a(config-if)#mtu ?

<1500-9180> MTU size in bytes

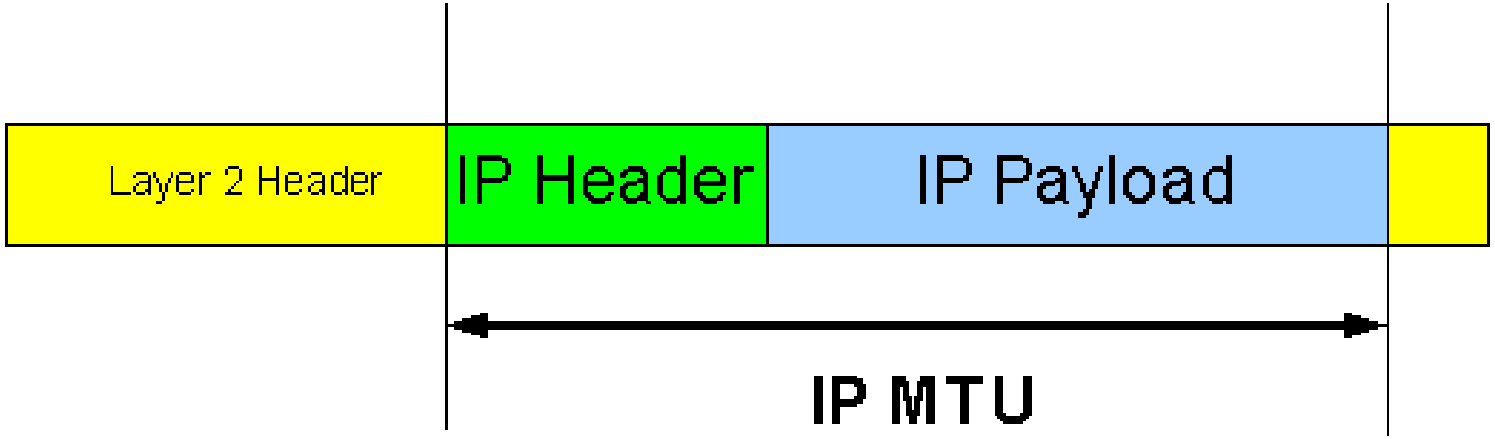
R28-12410a(config-if)#int g 3/2

R28-12410a(config-if)#mtu ?

<1500-1700> MTU size in bytes

© 2003, Cisco Systems, Inc. All rights reserved.

IP MTU



- IP MTU determines the maximum MTU of outgoing interface

KEY Concept:

- IP MTU follows the MTU setting by default
- IP MTU settings does not effect the MPLS MTU in any way
- MPLS MTU settings does not effect the IP MTU in any way

© 2003, Cisco Systems, Inc. All rights reserved.

IP MTU

- **Interface Command to change outgoing IP MTU on interface:**

ip mtu <x>

Example Changing IP Mtu in E2, Trident Line Card in GSR

```
R28-12410a(config)#int g 3/1
R28-12410a(config-if)#ip mtu ?
  <68-1000000>  MTU (bytes )
```

- **Interface Command to view outgoing IP MTU on interface:**

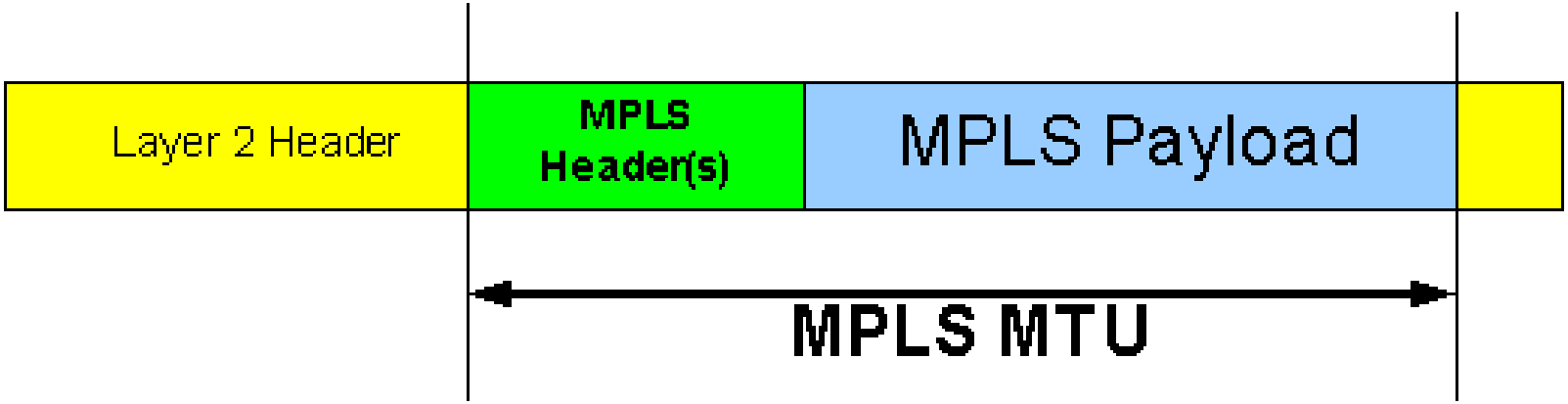
sh ip int <interface type, number>

Example Viewing IP MTU example

```
R28-12410a#sh ip int g3/1 | i MTU
  MTU is 9180 bytes
```

© 2003, Cisco Systems, Inc. All rights reserved.

MPLS MTU



- MPLS MTU determines the maximum MPLS MTU of outgoing interface
- MPLS MTU follows the MTU setting by default

KEY Concept:

- IP MTU settings does not effect the MPLS MTU in any way
- MPLS MTU settings does not effect the IP MTU in any way

© 2003, Cisco Systems, Inc. All rights reserved.

MPLS MTU

- **Interface Command to change outgoing IP MTU on interface:**

mpls mtu <x>

Example Changing IP Mtu in E2, Trident Line Card in GSR

```
R28-12410a(config)#int g 3/1
R28-12410a(config-if)#mpls mtu ?
  <64-65535>  MTU (bytes)
```

Interface Command to view outgoing IP MTU on interface:

sh mpls int <interface type, number> detail

Example Viewing IP MTU example

```
R28-12410a#sh mpls int g3/1 det | i MTU
MTU = 9180
```

© 2003, Cisco Systems, Inc. All rights reserved.

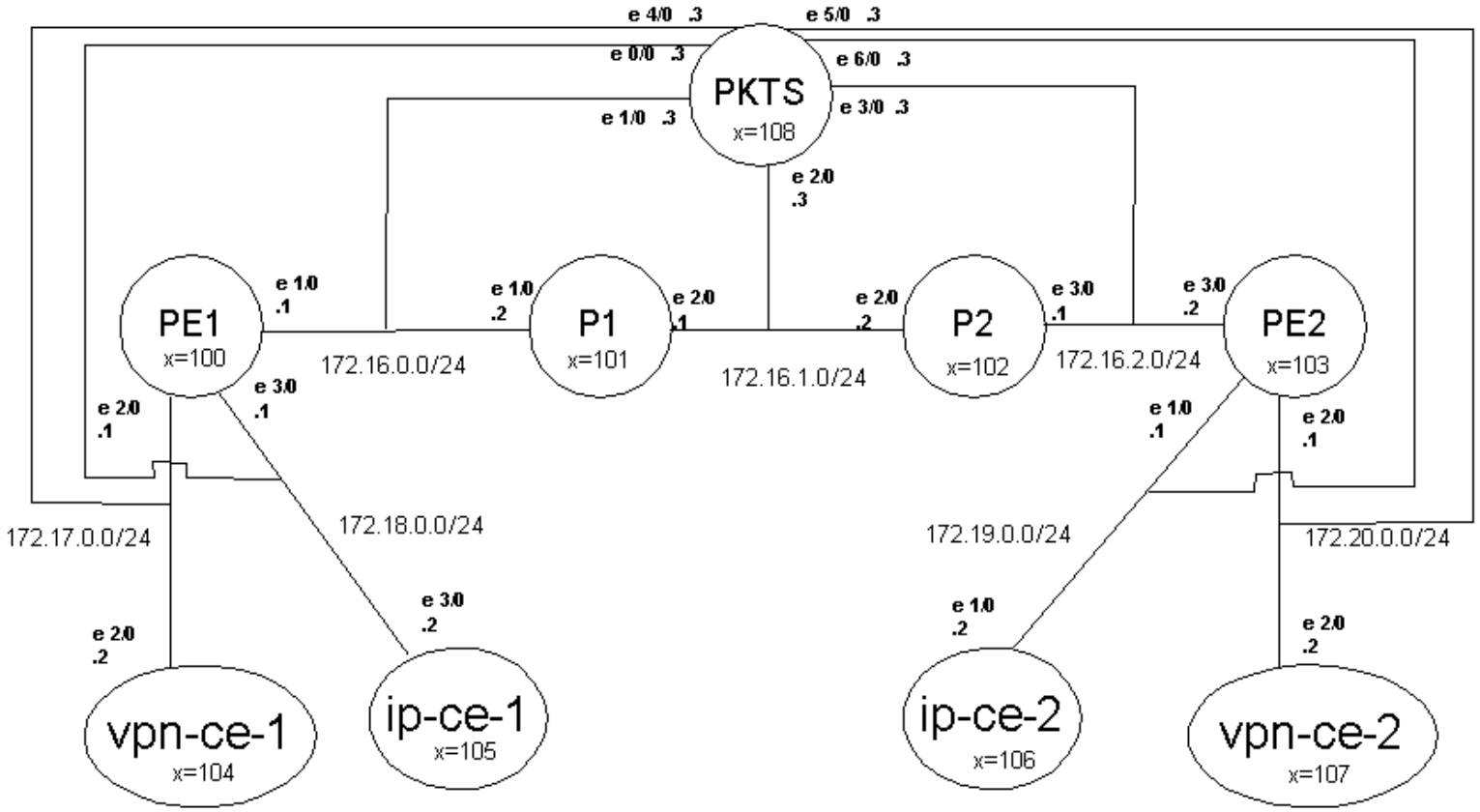
MTU and Fragmentation

- **Please see RFC3032 for more insight to MPLS MTU and Fragmentation**

http://www.faqs.org/rfcs/rfc3032.html

© 2003, Cisco Systems, Inc. All rights reserved.

MTU and Fragmentation, example, Simulation Lab results:

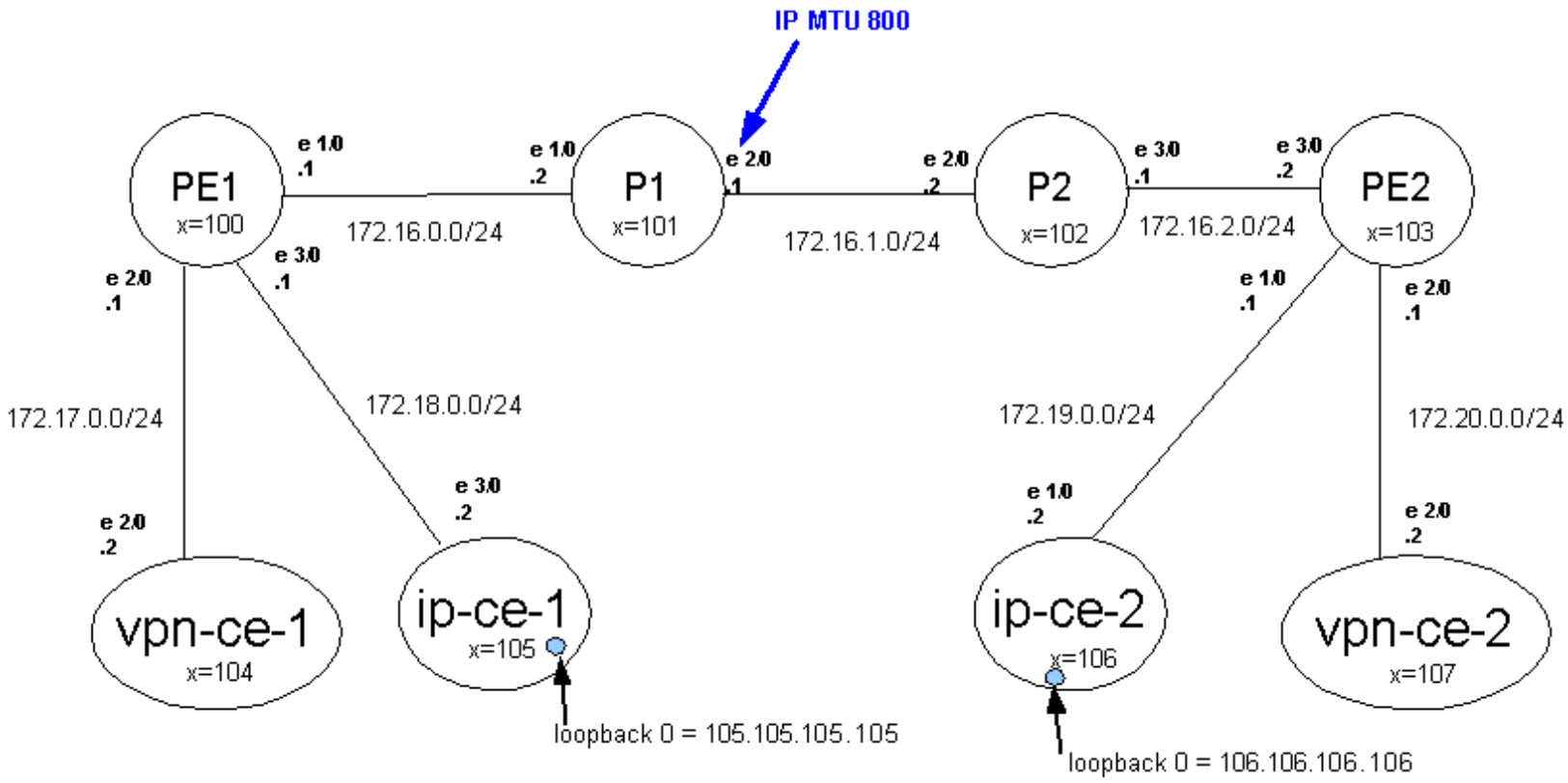


“PKTS” is a packet sniffer, LDP filtering is configured in P and PE routers so that packets between “ip-ce-1” and “ip-ce-2” are not mpls switched, loopbacks between “ip-ce-1” and “ip-ce2” are label switched

© 2003, Cisco Systems, Inc. All rights reserved.

MTU and Fragmentation, example
Changing "IP MTU"

Changing IP MTU on P1 Ethernet 2/0 to 800 Bytes:



© 2003, Cisco Systems, Inc. All rights reserved.

MTU and Fragmentation, example
Changing “IP MTU to 800 bytes”

From Show Commands on “P1”, below, it is verified that only “IP MTU” of e 2/0 is changed to 800 bytes

p1#sh int e 2/0 | i MTU

MTU 1500 bytes, BW 10000 Kbit, DLY 1000 usec, rely 255/255, load 1/255 ← Default Value, not changed, MTU

p1#sh ip int e 2/0 | i MTU

MTU is 800 bytes ← Changed, IP MTU to 800 bytes

p1#sh mpls int e 2/0 det | i MTU

MTU = 1500 ← Default Value, not changed, MPLS MTU

© 2003, Cisco Systems, Inc. All rights reserved.

MTU and Fragmentation, example

Changing “IP MTU to 800 bytes” & IP don’t fragment = 0

Pure IP Packet from Source to Destination:

Ping from “ip-ce-1” to “ip-ce-2” (note as previously mentioned, configurations are done such that no labels will be associated with the packets). Thus the packets are purely IP packets from source to destination.

One MPLS Label (4 bytes) :

Ping from loopback of “ip-ce-1” to loopback of “ip-ce-2” (note as previously mentioned, configurations are done such that one MPLS label stack will be associated with the packets).

Two MPLS Labels (4 bytes + 4 bytes = 8 bytes) :

Ping from “vpn-ce-1” to “vpn-ce-2”

© 2003, Cisco Systems, Inc. All rights reserved.

MTU and Fragmentation, example

Changing “IP MTU to 800 bytes” & IP don’t fragment = 0

Results:

No MPLS labels: Ping from “ip-ce-1” to “ip-ce-2”
800 byte IP packet goes through with no fragmentation

22:35:34.245 CST Sun Aug 17 2003 Relative Time: 1.119991

Packet 2 of 16 In: Ethernet2/0

Ethernet Packet: 814 bytes

Dest Addr: aabb.cc00.6602, Source Addr: aabb.cc00.6502

Protocol: 0x0800

IP Version: 0x4, HdrLen: 0x5, TOS: 0x00

Length: 800, ID: 0x0041, **Flags-Offset: 0x0000**

TTL: 253, Protocol: 1 (ICMP), Checksum: 0x6272 (OK)

Source: 172.18.0.2, Dest: 172.19.0.2

ICMP Type: 8, Code: 0 (Echo Request)

Checksum: 0x85E2 (OK)

Identifier: 042C, Sequence: 2032

© 2003, Cisco Systems, Inc. All rights reserved.

MTU and Fragmentation, example

Changing "IP MTU to 800 bytes" & IP don't fragment = 0

Results:

No MPLS labels: **Ping from "ip-ce-1" to "ip-ce-2"**

801 byte IP packet needs to be fragmented

22:42:54.113 CST Sun Aug 17 2003 Relative Time: 1.619987

Packet 4 of 22 In: Ethernet2/0

Ethernet Packet: 810 bytes

Dest Addr: aabb.cc00.6602, Source Addr: aabb.cc00.6502

Protocol: 0x0800

IP Version: 0x4, HdrLen: 0x5, TOS: 0x00

Length: 796, ID: 0x0046, **Flags-Offset: 0x2000 (more fragments)**

TTL: 253, Protocol: 1 (ICMP), Checksum: 0x4271 (OK)

Source: 172.18.0.2, Dest: 172.19.0.2

ICMP Type: 8, Code: 0 (Echo Request)

Checksum: 0xB8D5 ERROR: 579B

Identifier: 175A, Sequence: 23D6

1'st fragment of packet

© 2003, Cisco Systems, Inc. All rights reserved.

MTU and Fragmentation, example

Changing “IP MTU to 800 bytes” & IP don’t fragment = 0

Results:

One MPLS Label Stack: **Ping between loopbacks--“ip-ce-1” to “ip-ce-2”**

1496 byte IP packet can pass without fragmentation

(1496 bytes IP + 4 bytes label) = 1500 bytes MPLS MTU

22:51:49.065 CST Sun Aug 17 2003 Relative Time: 3.879970

Packet 7 of 19 In: Ethernet2/0

Ethernet Packet: 1514 bytes

Dest Addr: aabb.cc00.6602, Source Addr: aabb.cc00.6502

MPLS Label: 23, CoS: 0, Bottom: 1, TTL: 253

Protocol: 0x0800

IP Version: 0x4, HdrLen: 0x5, TOS: 0x00

Length: 1496, ID: 0x004B, **Flags-Offset: 0x0000**

TTL: 254, Protocol: 1 (ICMP), Checksum: 0x0F33 (OK)

Source: 105.105.105.105, Dest: 106.106.106.106

ICMP Type: 8, Code: 0 (Echo Request)

Checksum: 0xF934 (OK)

Identifier: 23A1, Sequence: 21E2

800 byte "ip mtu" has no effect

© 2003, Cisco Systems, Inc. All rights reserved.

MTU and Fragmentation, example

Changing "IP MTU to 800 bytes" & IP don't fragment = 0

Results:

One MPLS Label Stack: **Ping between loopbacks--"ip-ce-1" to "ip-ce-2"**

1497 byte IP packet needs to be fragmented

(1497 bytes IP + 4 bytes label) = 1501 bytes → exceeds MPLS MTU

23:07:09.093 CST Sun Aug 17 2003 Relative Time: 3.539972

Packet 6 of 29 In: Ethernet2/0

Ethernet Packet: 1510 bytes

Dest Addr: aabb.cc00.6602, Source Addr: aabb.cc00.6502

MPLS Label: 23, CoS: 0, Bottom: 1, TTL: 253

Protocol: 0x0800

IP Version: 0x4, HdrLen: 0x5, TOS: 0x00

Length: 1492, ID: 0x005A, **Flags-Offset: 0x2000 (more fragments)**

TTL: 254, Protocol: 1 (ICMP), Checksum: 0xEF27 (OK)

Source: 105.105.105.105, Dest: 106.106.106.106

ICMP Type: 8, Code: 0 (Echo Request)

Checksum: 0x183C ERROR: 579B

Identifier: 0DE4, Sequence: 0EB6

1'st fragment of packet

© 2003, Cisco Systems, Inc. All rights reserved.

MTU and Fragmentation, example

Changing “IP MTU to 800 bytes” & IP don’t fragment = 0

Results:

Two MPLS Label Stack: **Ping between “vpn-ce-1” to “vpn-ce-2”**

1492 byte IP packet can pass without fragmentation

(1492 bytes IP + 4 bytes + 4 bytes label) = 1500 bytes MPLS MTU

23:26:05.525 CST Sun Aug 17 2003 Relative Time: 1.599987

Packet 2 of 14 In: Ethernet2/0

Ethernet Packet: 1514 bytes

Dest Addr: aabb.cc00.6602, Source Addr: aabb.cc00.6502

MPLS Label: 20, CoS: 0, Bottom: 0, TTL: 253

MPLS Label: 23, CoS: 0, Bottom: 1, TTL: 254

Protocol: 0x0800

IP Version: 0x4, HdrLen: 0x5, TOS: 0x00

Length: 1492, ID: 0x05CC, **Flags-Offset: 0x0000**

TTL: 254, Protocol: 1 (ICMP), Checksum: 0x5933 (OK)

Source: 172.17.0.2, Dest: 172.20.0.2

ICMP Type: 8, Code: 0 (Echo Request)

Checksum: 0x13E5 (OK)

Identifier: 16A3, Sequence: 0BCC

800 byte "ip mtu" has no effect

© 2003, Cisco Systems, Inc. All rights reserved.

MTU and Fragmentation, example

Changing "IP MTU to 800 bytes" & IP don't fragment = 0

Results:

Two MPLS Label Stack: **Ping between "vpn-ce-1" to "vpn-ce-2"**

1493 byte IP packet needs fragmentation

(1493 bytes IP + 4 bytes + 4 bytes label) = 1501 bytes → exceeds MPLS MTU

23:31:52.125 CST Sun Aug 17 2003 Relative Time: 1.699986

Packet 3 of 23 In: Ethernet2/0

Ethernet Packet: 1514 bytes

Dest Addr: aabb.cc00.6602, Source Addr: aabb.cc00.6502

MPLS Label: 20, CoS: 0, Bottom: 0, TTL: 253

MPLS Label: 23, CoS: 0, Bottom: 1, TTL: 254

Protocol: 0x0800

IP Version: 0x4, HdrLen: 0x5, TOS: 0x00

Length: 1492, ID: 0x05D1, **Flags-Offset: 0x2000 (more fragments)**

TTL: 254, Protocol: 1 (ICMP), Checksum: 0x392E (OK)

Source: 172.17.0.2, Dest: 172.20.0.2

ICMP Type: 8, Code: 0 (Echo Request)

Checksum: 0xC5D4 (OK)

Identifier: 0D82, Sequence: 1918

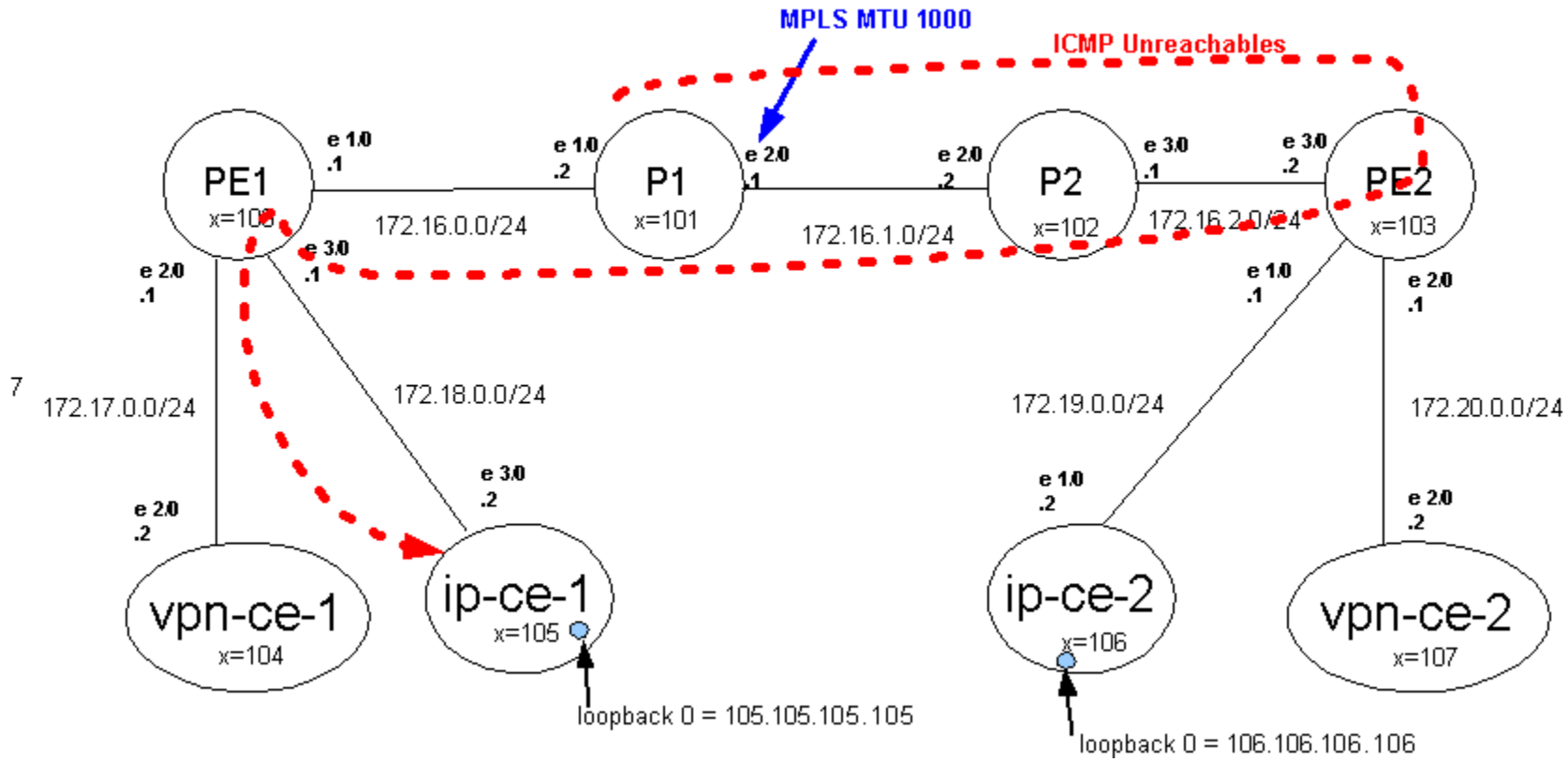
1'st fragment of packet

© 2003, Cisco Systems, Inc. All rights reserved.

MTU and icmp unreachables, example

Changing "MPLS MTU to 1,000 bytes" & IP don't fragment = 1

Cisco.com



For Single Labelled Packets, If IP Don't Fragment Bits are set, and MPLS MTU is too large, then, packet is dropped and ICMP unreachables are sent to the egress PE and is then U-turned back from the Egress PE to the source IP Address, as shown in diagram.

© 2003, Cisco Systems, Inc. All rights reserved.

MTU and icmp unreachables, example

Changing "MPLS MTU to 1,000 bytes" & IP don't fragment = 1

Results:

One MPLS Label Stack: Ping between loopbacks--"ip-ce-1" to "ip-ce-2"

987 byte IP packet needs to be fragmented, but IP Don't fragment set

(987 bytes IP + 4 bytes label) = 1,001 bytes → exceeds MPLS MTU of 1,000 bytes

00:13:44.533 CST Mon Aug 18 2003 Relative Time: 7.227944

Packet 42 of 62 In: Ethernet3/0

Ethernet Packet: 74 bytes

Dest Addr: aabb.cc00.6603, Source Addr: aabb.cc00.6703

MPLS Label: 22, CoS: 0, Bottom: 1, TTL: 252

Protocol: 0x0800

IP Version: 0x4, HdrLen: 0x5, TOS: 0xC0 (Prec=Internet Contrl)

Length: 56, ID: 0x161A, Flags-Offset: 0x0000

TTL: 255, Protocol: 1 (ICMP), Checksum: 0x2606 (OK)

Source: 172.16.0.2, Dest: 105.105.105.105

ICMP Type: 3, Code: 4 (Dest Unreachable)(Fragmentation Needed and DF set)

Checksum: 0x97D5 (OK)

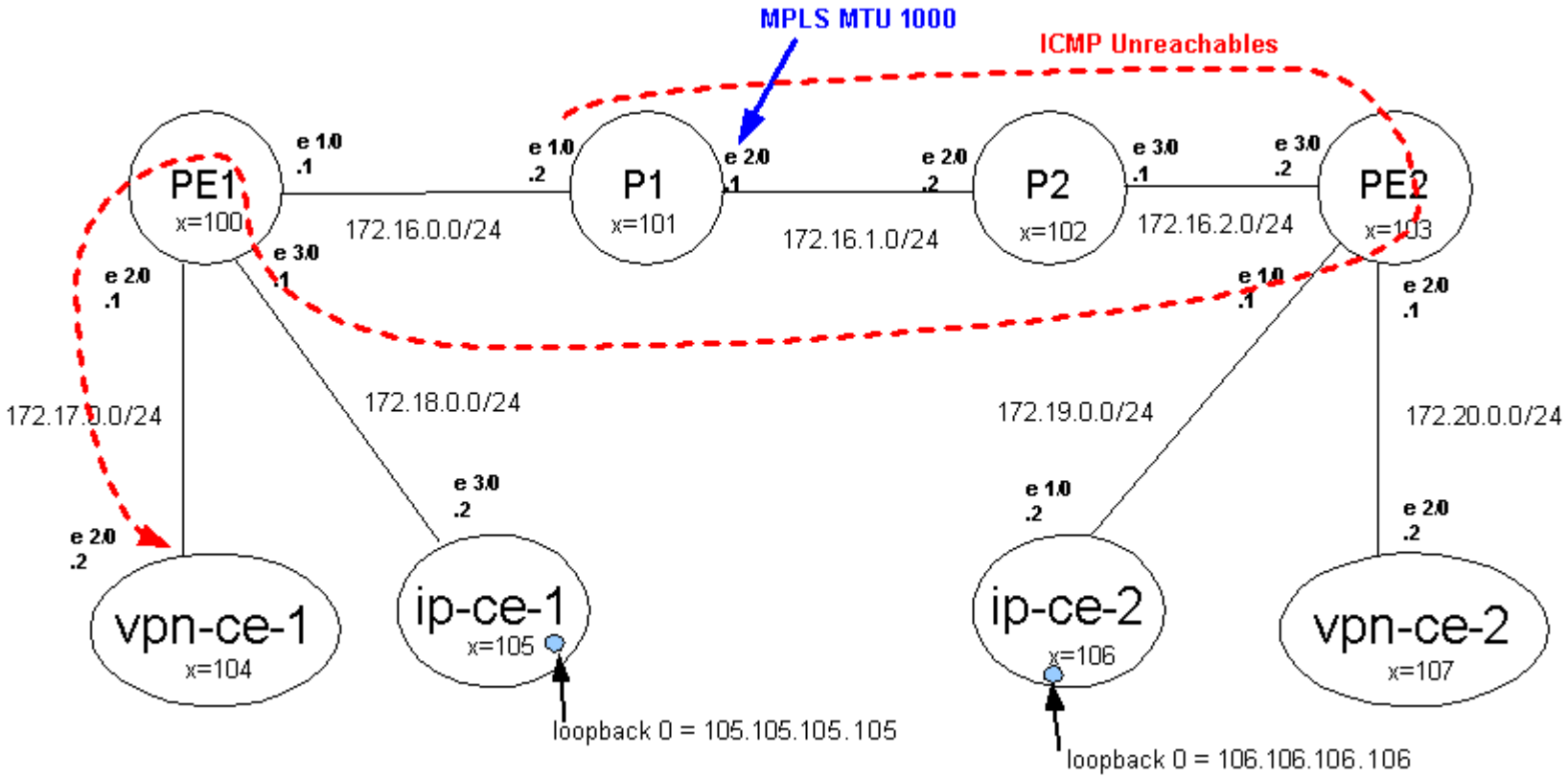
Next-Hop MTU: 996

icmp unreachable U-Turned back from PE2

© 2003, Cisco Systems, Inc. All rights reserved.

MTU and icmp unreachables, example

Changing "MPLS MTU to 1,000 bytes" & IP don't fragment = 1



For Single Labelled Packets, If IP Don't Fragment Bits are set, and MPLS MTU is too large, then, packet is dropped and ICMP unreachables are sent to the egress PE and is then U-turned back from the Egress PE to the source IP Address, as shown in diagram.

© 2003, Cisco Systems, Inc. All rights reserved.

MTU and icmp unreachables, example

Changing "MPLS MTU to 1,000 bytes" & IP don't fragment = 1

Results:

TWO MPLS Label Stack: **Ping between "vpn-ce-1" to "vpn-ce-2"**

987 byte IP packet needs to be fragmented, but IP Don't fragment set

(987 bytes IP + 4 bytes label) = 1,001 bytes → exceeds MPLS MTU of 1,000 bytes

00:32:01.745 CST Mon Aug 18 2003 Relative Time: 4.639964

Packet 14 of 55 In: Ethernet3/0

Ethernet Packet: 78 bytes

Dest Addr: aabb.cc00.6603, Source Addr: aabb.cc00.6703

MPLS Label: 18, CoS: 6, Bottom: 0, TTL: 251

MPLS Label: 22, CoS: 6, Bottom: 1, TTL: 251

Protocol: 0x0800

IP Version: 0x4, HdrLen: 0x5, TOS: 0xC0 **(Prec=Internet Contrl)**

Length: 56, ID: 0x1948, Flags-Offset: 0x0000

TTL: 251, Protocol: 1 (ICMP), Checksum: 0x4D97 (OK)

Source: 172.16.0.2, Dest: 172.17.0.2

ICMP Type: 3, Code: 4 (Dest Unreachable)(Fragmentation Needed and DF set)

Checksum: 0xA9B5 (OK)

Next-Hop MTU: 992

icmp unreachable U-Turned back from PE2

© 2003, Cisco Systems, Inc. All rights reserved.

© 2001, Cisco Systems, Inc. All rights reserved.