MPLS

What is MPLS?

- Multi Protocol Label Switching
- Scalable, protocol agnostic, data-carrying mechanism
- MPLS was standardized in IETF in mid 90s, developed from Cisco’s “Tag Switching”

- A technology originally developed to switch (forward) a IP packet at a high speed at layer 2 using fixed length labels generated from layer 3 routing information
  - Small label lookup instead of longest prefix match
  - Roots in ATM - combines the connection oriented and fast forwarding algorithm used in ATM with IP

➢ But today’s routers does IP lookup in hardware at very high speeds

Current MPLS Advantages

- Label switching can be used for Traffic Engineering
  - MPLS TE provides efficient way of forwarding traffic through the network – not based on IGP but other policies (in order to achieve more efficient bandwidth utilization, etc.)
- Aggregating a class of traffic and treating it in a specific way
- Label switching can be used to support VPNs – Virtual Private Networks
  - In conjunction with MP-BGP
- Labels can be used to forward using other fields than destination address
  - The Generalized form of MPLS: G-MPLS can be used for optical networking such as management of wavelengths: ”lambdas”
- BGP-free core design
  - Core routers do not need to run BGP - simplified configuration, processing of data flows and troubleshooting

MPLS Concept

Ingress LER receives packet and adds (PUSH) label

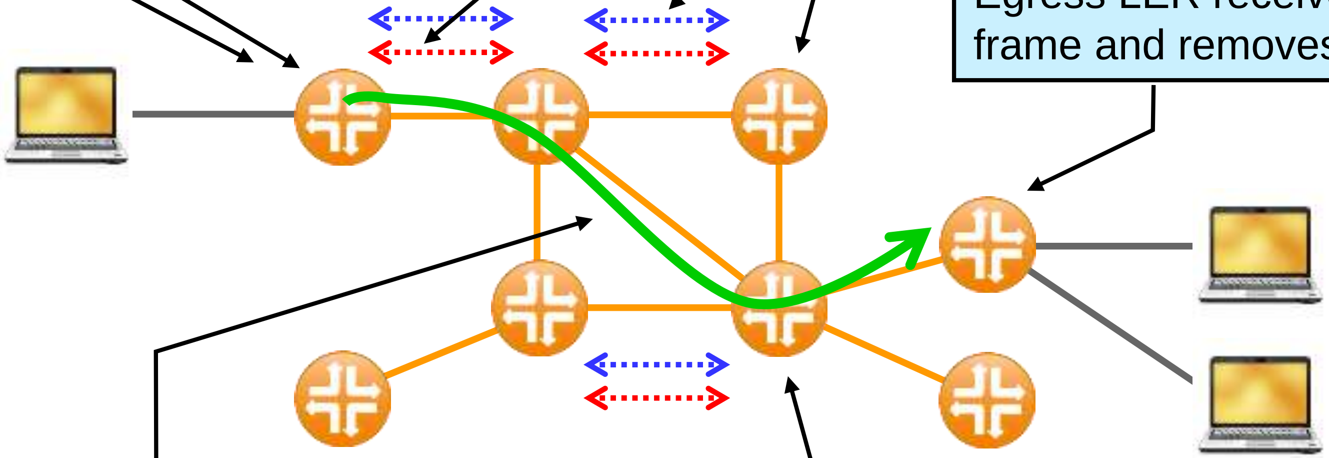
LER - Label Edge Router

Existing IGP establishes the reachability to the destination

Label Distribution Protocol (LDP) assignes a label to destination network, the label is locally significant, different label assigment on each router

LSR - Label Switch Router

Egress LER receives MPLS frame and removes (POP) label

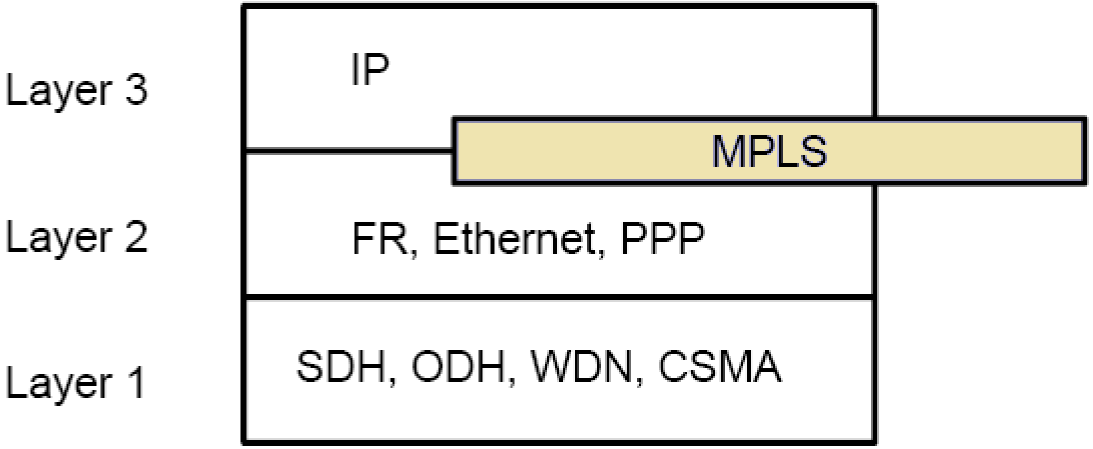


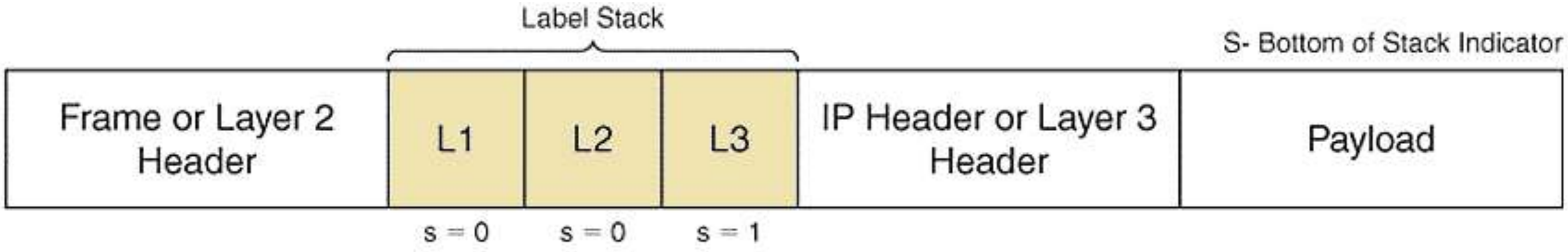
Label Switch Path (LSP), unidirectional connection

LSR switches packets using label swapping table (SWAP)

MPLS Encapsulation

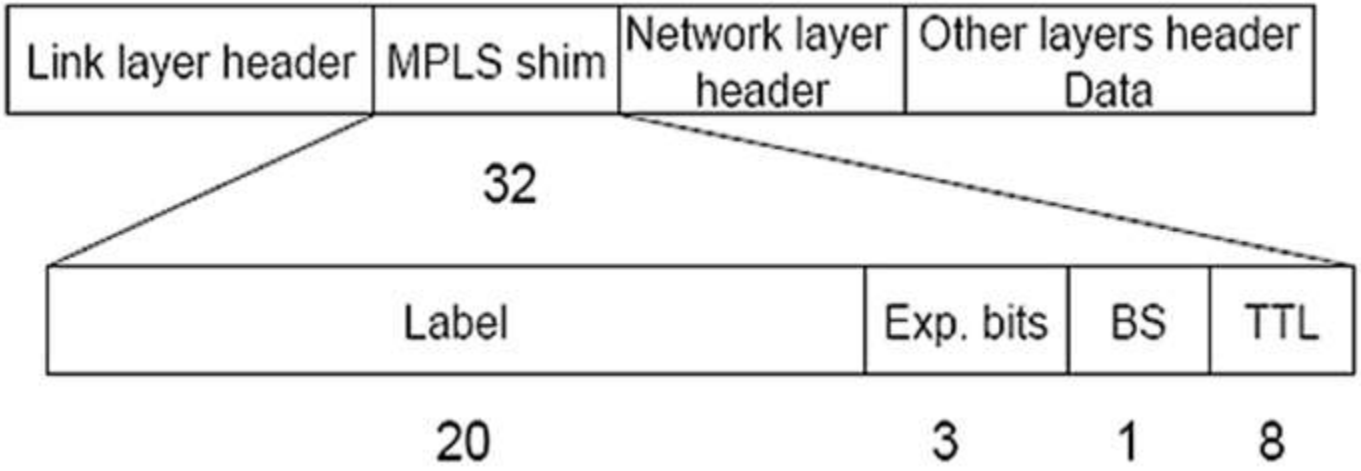
- MPLS header is present between Layer 2 and Layer 3
- MPLS header can contains one or more labels - this is called a Label Stack





MPLS Frame or Shim Header Format

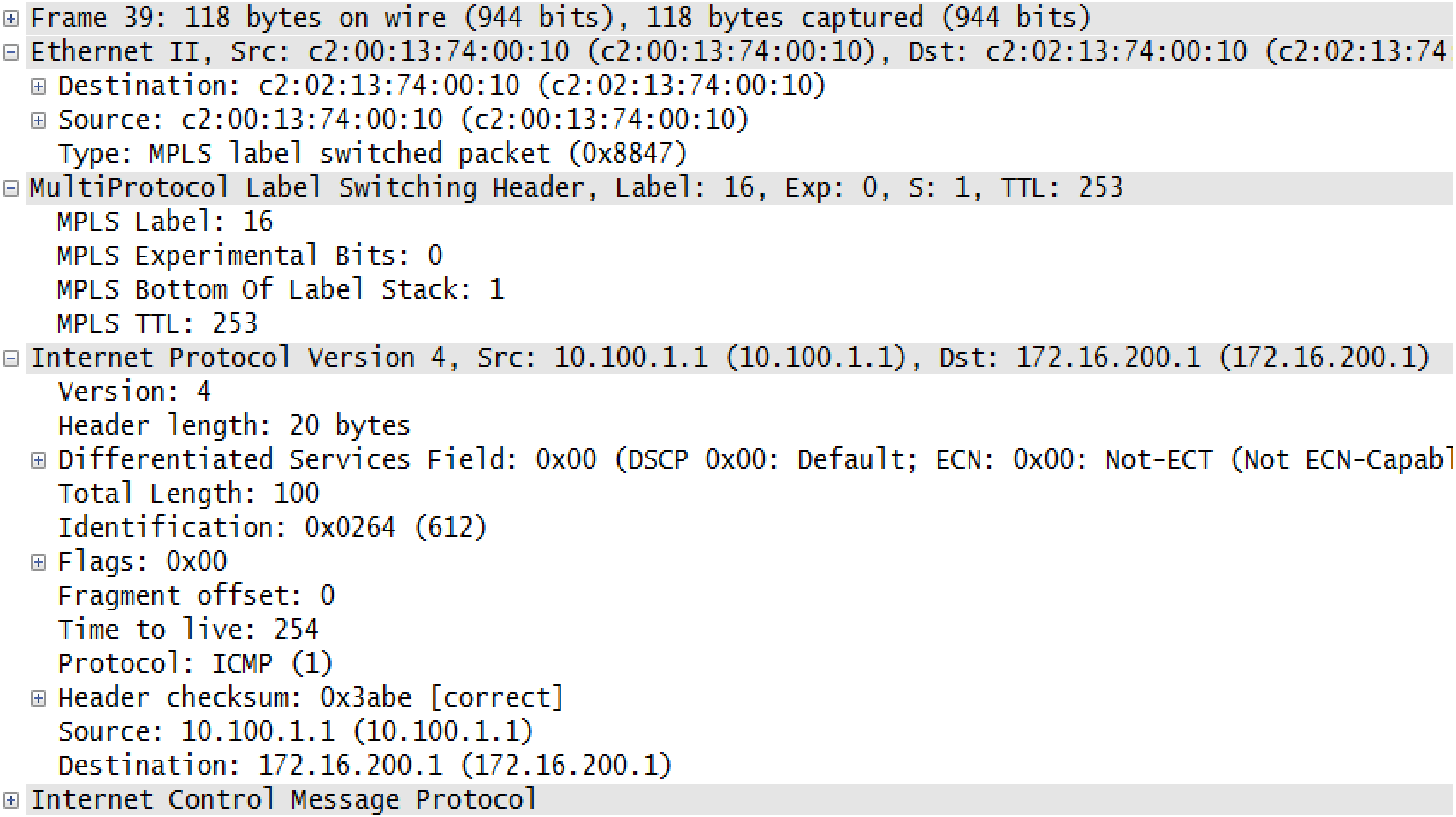
- MPLS uses a 32-bit shim header
  - Label - Value for table lookup in router
  - Exp - Traffic class field, can be used as class-of-service identification for QoS
  - Stack - Indicates that the bottom of a stack of labels has been reached
  - TTL - Time To Live (similar to IP TTL)



Label

- A label is an integer number identifying a flow (or a FEC - Forwarding equivalence class)
- Locally significant, cannot have globally or network based unique labels
    - Too complex to negotiate
    - Too large labels
    - Labels are unique only between two nodes
    - Labels change at each node as a packet traverses LSP
    - Labels assigned from the range 0-1048575.
    - 0-15 reserved by the IETF
    - Possible to set labels manually or to use some of label distribution protocols
        - LDP
        - RSVP-TE
        - MP-BGP

MPLS Frame



```
⊞ Frame 39: 118 bytes on wire (944 bits), 118 bytes captured (944 bits)
⊟ Ethernet II, Src: c2:00:13:74:00:10 (c2:00:13:74:00:10), Dst: c2:02:13:74:00:10 (c2:02:13:74
  ⊞ Destination: c2:02:13:74:00:10 (c2:02:13:74:00:10)
  ⊞ Source: c2:00:13:74:00:10 (c2:00:13:74:00:10)
    Type: MPLS label switched packet (0x8847)
⊟ MultiProtocol Label Switching Header, Label: 16, Exp: 0, S: 1, TTL: 253
    MPLS Label: 16
    MPLS Experimental Bits: 0
    MPLS Bottom Of Label Stack: 1
    MPLS TTL: 253
⊟ Internet Protocol Version 4, Src: 10.100.1.1 (10.100.1.1), Dst: 172.16.200.1 (172.16.200.1)
    Version: 4
    Header length: 20 bytes
  ⊞ Differentiated Services Field: 0x00 (DSCP 0x00: Default; ECN: 0x00: Not-ECT (Not ECN-Capabl
    Total Length: 100
    Identification: 0x0264 (612)
  ⊞ Flags: 0x00
    Fragment offset: 0
    Time to live: 254
    Protocol: ICMP (1)
  ⊞ Header checksum: 0x3abe [correct]
    Source: 10.100.1.1 (10.100.1.1)
    Destination: 172.16.200.1 (172.16.200.1)
⊞ Internet Control Message Protocol
```

Forwarding Equivalence Class (FEC)

- FEC means sorting packets into different classes

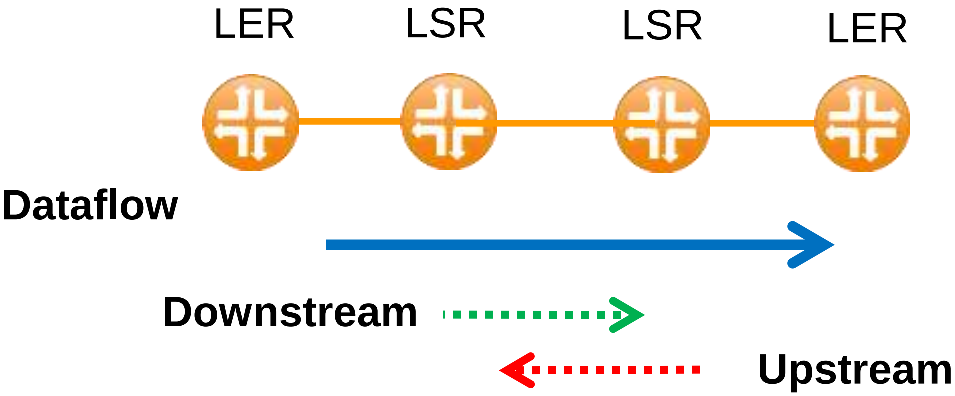
- Classification is a more general form of lookup
  - Examples:
    - Typically packets with Layer 3 destination IP addresses that belong to a set of prefixes having the same next hop
    - But can be also all UDP packets with the ToS field set to 0x42 from subnet 10.10.20.0/24

- MPLS binds labels to FECs
- The meaning of a FEC (the semantics) is added by the overlying application/protocol

Label Operations and Up/Down stream definition

- PUSH a label - typically at ingress
- SWAP a label - made by LSR
- POP a label - typically at egress or pen-ultimate (one hope before) LSR
- Label operations are interface-specific
  - Since labels are unique between LSRs

- Both downstream and upstream are defined with reference to the destination network: prefix or FEC



MPLS Label Switching

FEC	In Lbl	Out Lbl	Out If
192.168.10.0/24	-	L3	if1

PUSH

FEC	In Lbl	IP Lookup
192.168.10.0/24	L1	192.168.10.0/24

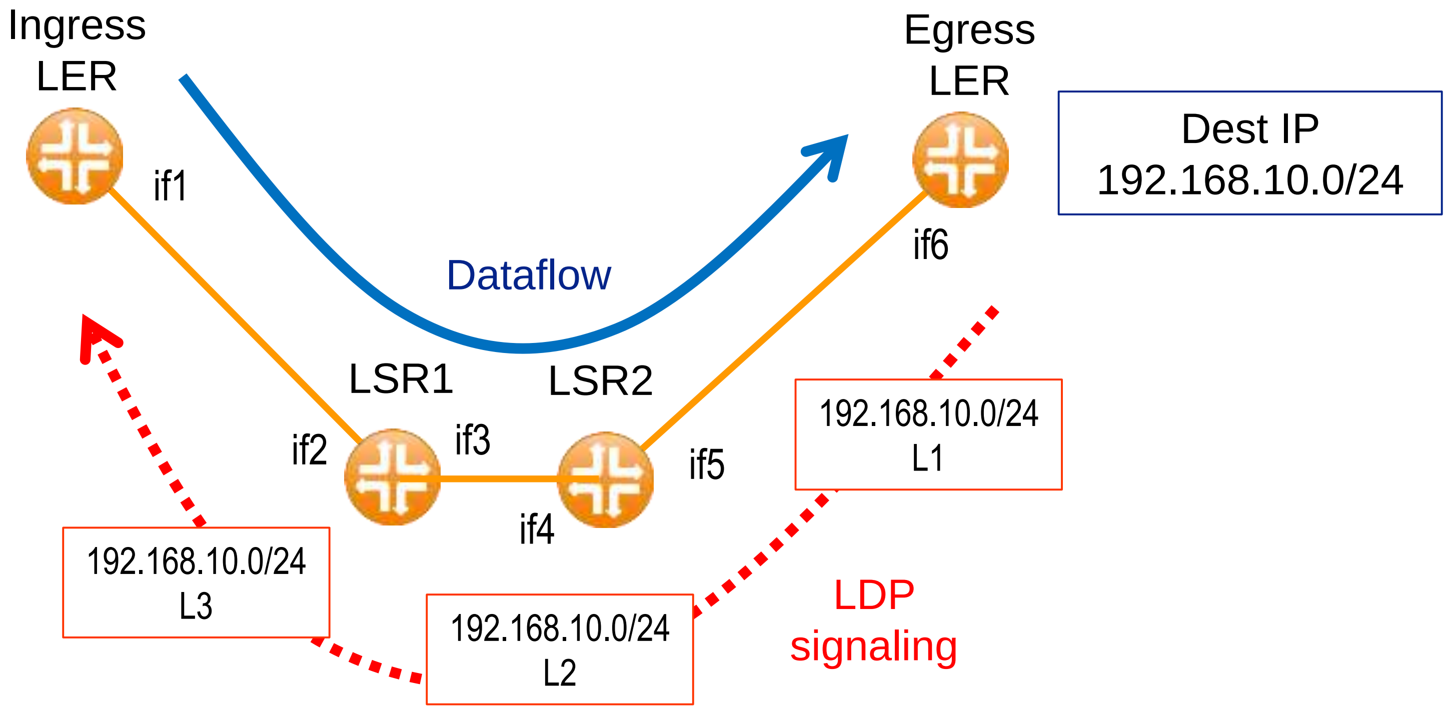
POP

FEC	In Lbl	Out Lbl	Out If
192.168.10.0/24	L3	L2	if3

SWAP

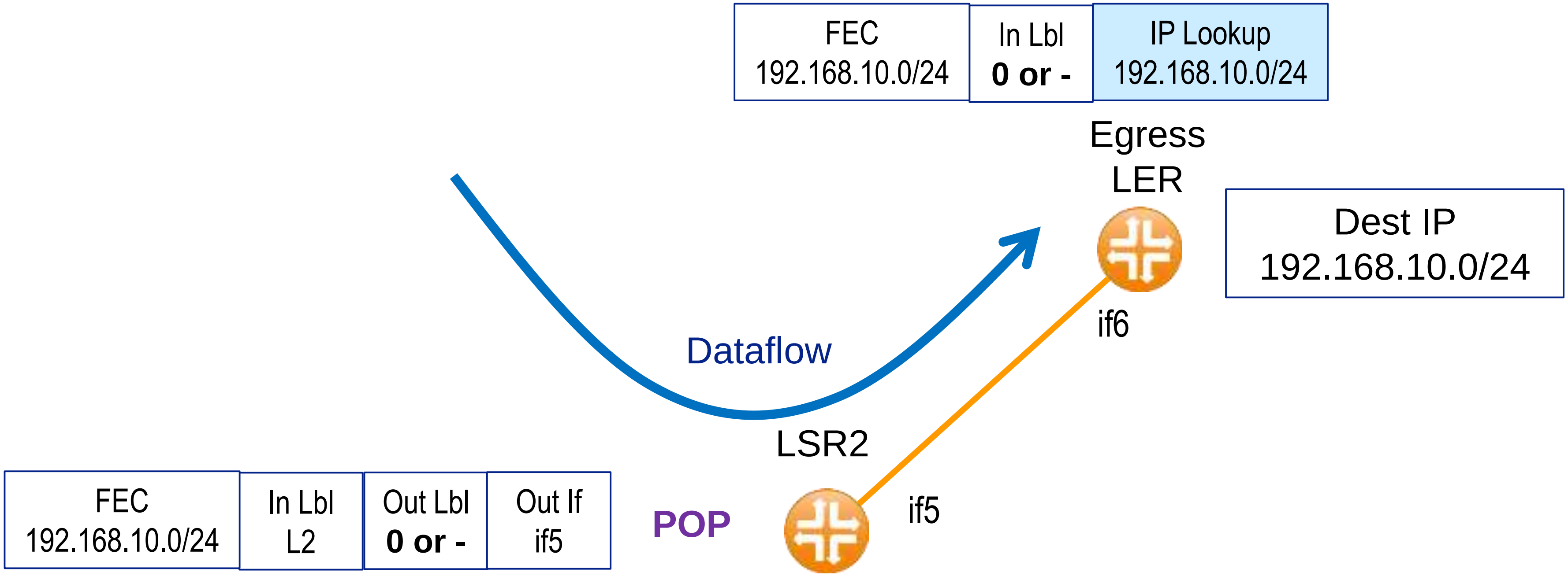
FEC	In Lbl	Out Lbl	Out If
192.168.10.0/24	L2	L1	if5

SWAP



Pen-ultimate Hop Popping or Explicit NULL

- Both MPLS and IP forwarding on egress router!
- To make it easier for the border router, pop the label on the previous router (pen-ultimate)
  - The pen-ultimate LSR does MPLS pop
  - The egress LER does only IP routing



FEC	In Lbl	IP Lookup
192.168.10.0/24	**0 or -**	192.168.10.0/24

FEC	In Lbl	Out Lbl	Out If
192.168.10.0/24	L2	**0 or -**	if5

Special Label Operations

0 - **IPv4 explicit NULL**

- Downstream/egress LER should pop label unconditionally, last LSR replaces incoming label with value 0
- Preserves EXP bits, used for QoS
- Popped packet is an IPv4 datagram

1 - **Router alert**

- Deliver to control plane – do not forward

2 - **IPv6 explicit-NULL**

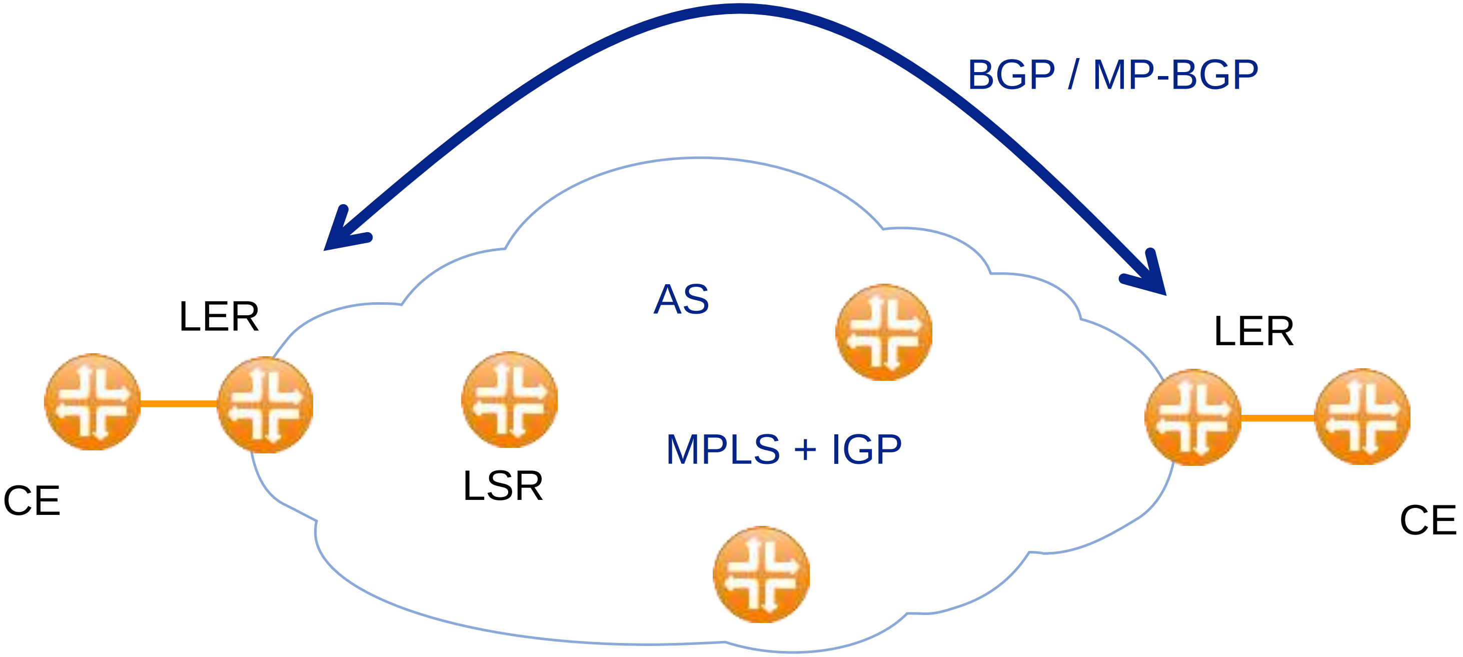
- Downstream LSR should pop unconditionally
- Popped packet is an IPv6 datagram

3 - **Implicit-NULL**

- Pop immediately and treat as IPv4 packet
- Does not appear on link, label removed
- *Pen-ultimate hop popping*

BGP Free Core, MPLS for Transit

- Setup LSPs between border routers using the IGP
- Send **transit traffic** via LSPs (src and dst outside the AS) using BGP next hops (transit src,dst not known in IGP)
- But still may send **internal traffic** via native IP
- External routes need not be distributed to non-border routers, so we do not need iBGP

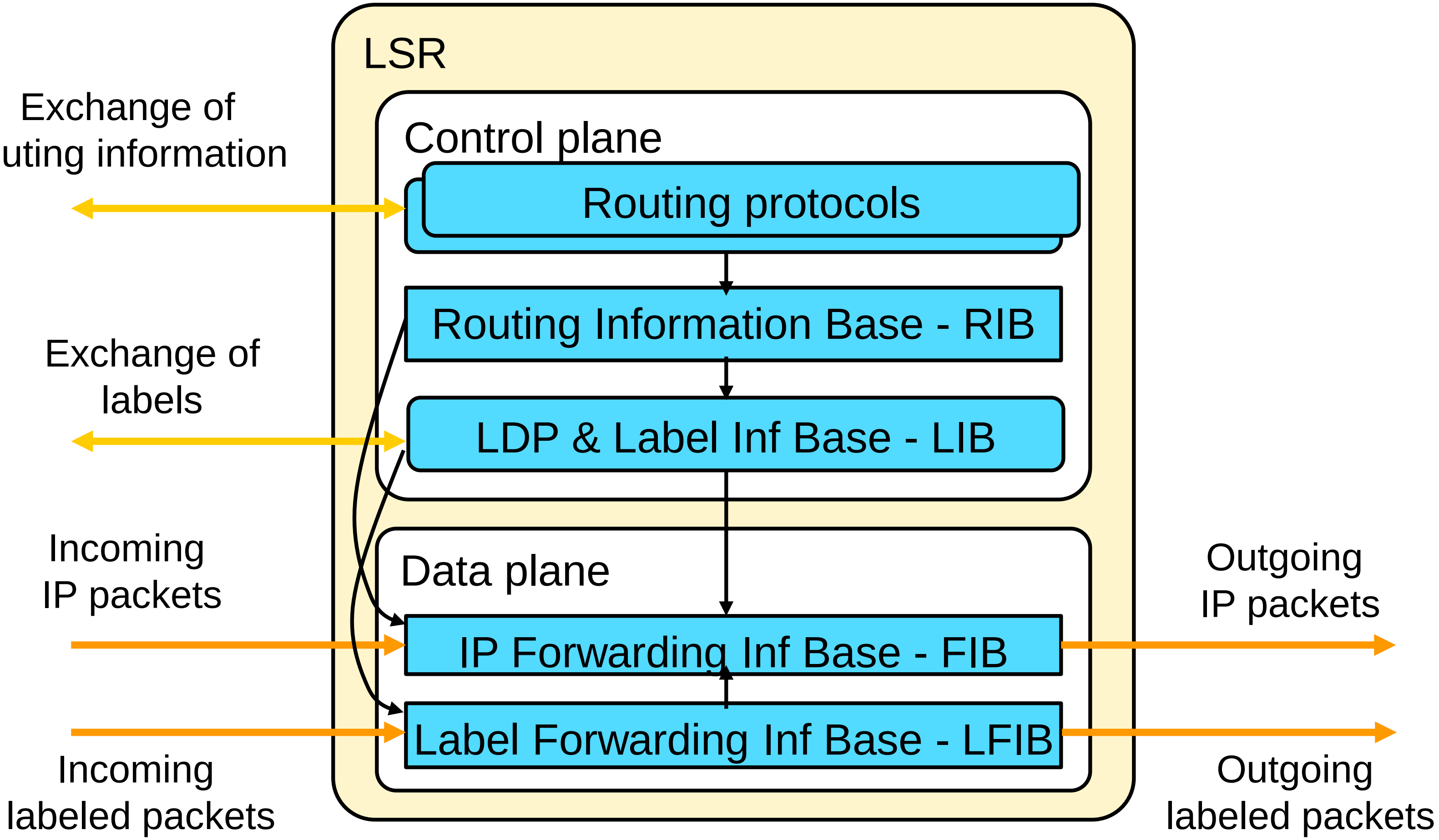


Label Distribution

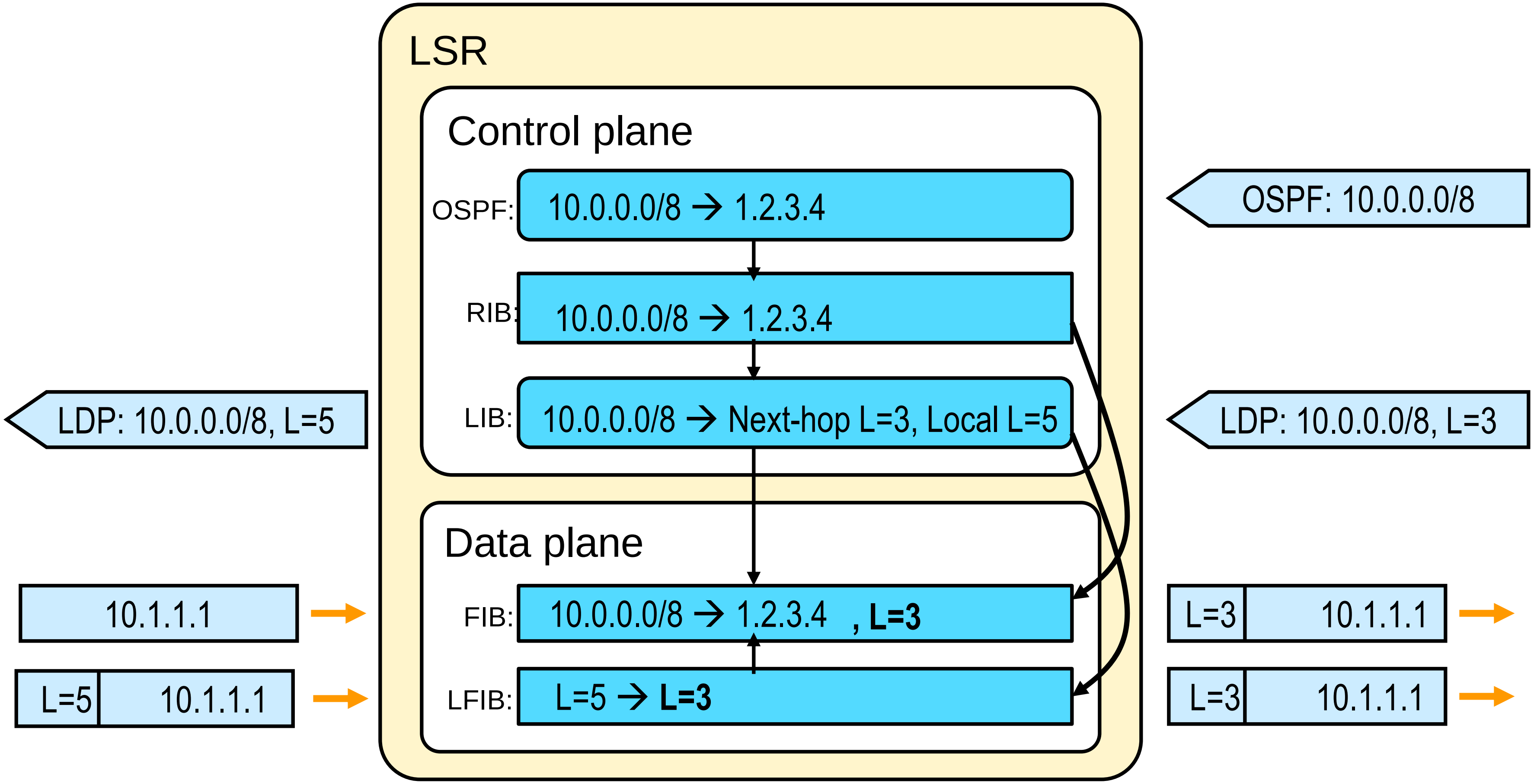
- Labels need to be assigned
- A signaling protocol distributes labels
  - Creates an LSP through an MPLS network
    - LDP, MP-BGP, RSVP-TE
    - These protocols all distributes label but they are somewhat different and can be combined to transfer different labels, eg BGP+RSVP, where BGP transfers inner labels and RSVP negotiate outer labels

- LDP - Label Distribution Protocol
  - Relies on IGP
  - Labels have link-local significance
  - Each LSR binds his own label mappings
  - Each LSR assign labels to his FECs
  - Labels are assigned and exchanged between adjacent neighboring LSR/LER in upstream direction

MPLS Routing Architecture


LSR
Exchange of routing information
Control plane
Routing protocols
Routing Information Base - RIB
Exchange of labels
LDP & Label Inf Base - LIB
Incoming IP packets
Data plane
Outgoing IP packets
IP Forwarding Inf Base - FIB
Label Forwarding Inf Base - LFIB
Incoming labeled packets
Outgoing labeled packets

MPLS Routing Architecture



LDP Session Establishment

- Defined in RFC 5036
- Neighbor discovery capability

- LDP uses a similar process to establish a session:
  - Hello messages are periodically sent on all interfaces enabled for MPLS
  - If there is another router on that interface it will respond by trying to establish a session with the source of the hello messages
- UDP is used for hello messages. It is targeted at “all routers on this subnet” multicast address (224.0.0.2).
- TCP is used to establish the session
- Both TCP and UDP use well-known LDP port number 646

LDP Messages

- Discovery/Hello messages (UDP)
  - Used to announce and maintain the presence of an LSR
- Session/Adjacency messages (TCP)
  - Used to establish, maintain and terminate sessions between LDP peers
- Advertisement messages (TCP)
  - Used to create, change, and delete label mappings
- Notification messages (TCP)
  - Used to provide advisory information and to signal error information

LDP Discovery

- A mechanism that enables an LSR to discover potential LDP peers
- Avoids unnecessary explicit configuration of LSR label switching peers

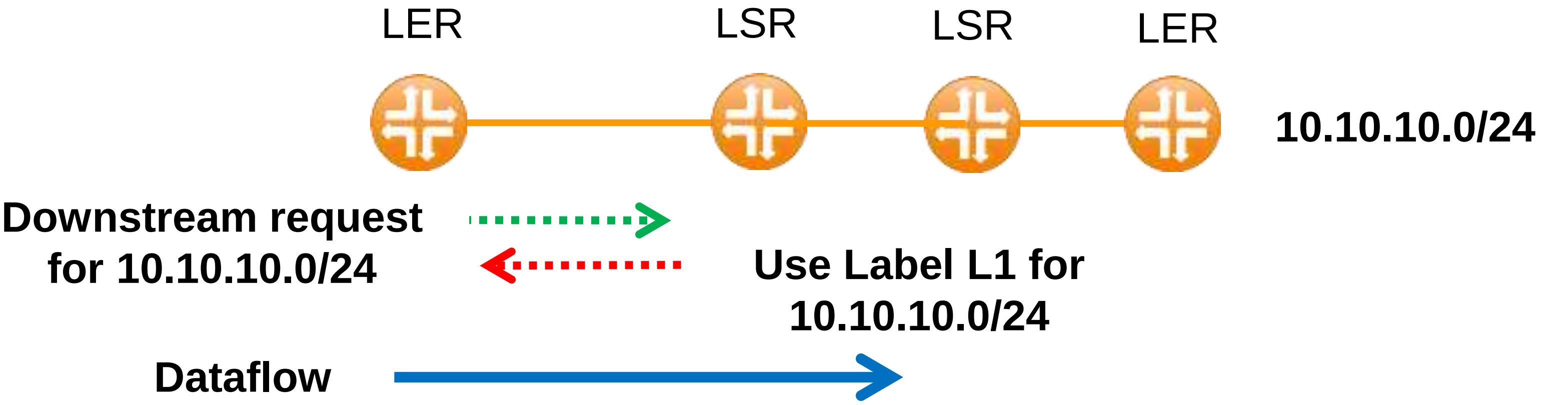
- Two variants of the discovery mechanism
  - Basic discovery mechanism: used to discover LSR neighbors that are directly connected at the link level
  - Extended discovery mechanism: used to locate LSRs that are not directly connected at the link level

Note: Traffic Engineering scenarios are examples of these applications that require a targeted LDP session between non directly connected routers

MPLS Label distribution modes

- Unsolicited Downstream
  - Each LSR automatically distributes its label bindings to all peers without waiting for a request message from those peers. The LSR receives a label binding for the FEC from all adjacent LSRs

- Downstream on-demand
  - In this mode each upstream LSR sends a label binding request to its downstream router only for specific FEC. Downstream LSR cannot automatically distributes


LER
LSR
LSR
LER
10.10.10.0/24
Downstream request for 10.10.10.0/24
Use Label L1 for 10.10.10.0/24
Dataflow

Label Retention Modes

- Liberal retention mode
  - LSR keeps/retains labels from all neighbors

    Improve convergence time, when next-hop is again available after IP convergence

    Require more memory and label space

- Conservative retention mode
  - LSR retains labels only if the sending LSR is the next hop downstream router for this specific FEC

    LSR discards all labels for FECs without next-hop

    Free memory and label space

Note: In Cisco IOS, the retention mode for label controlled ATM interfaces is the Conservative Label Retention mode. The Liberal Label Retention mode is used for all other types of interfaces.

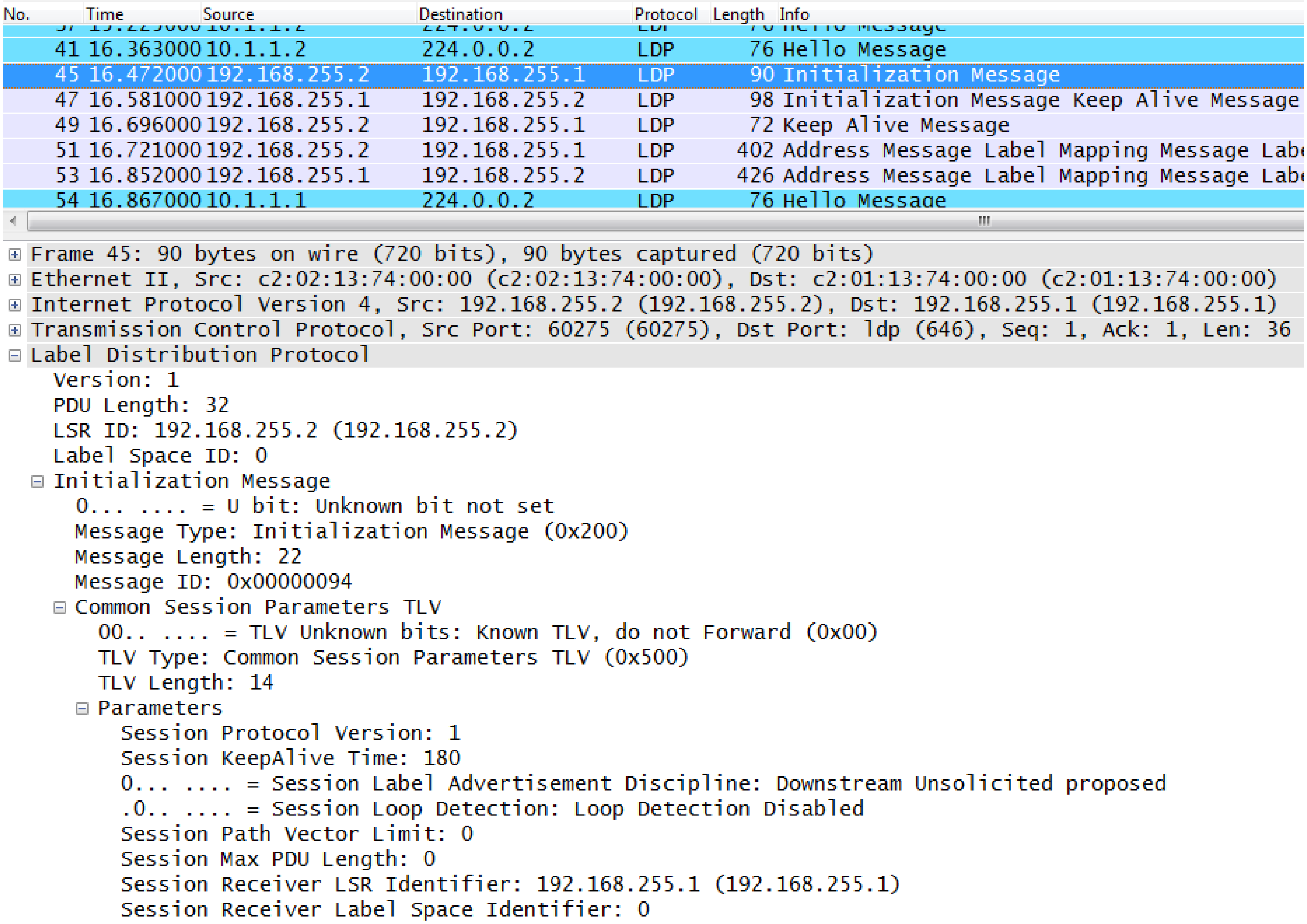
MPLS Label Control modes

- Independent label distribution control mode (default)
  - LSR assigns a local binding to a FEC as soon as it realizes its existence in the routing table. Does not wait to receive any labels from downstream LSR

- Ordered label distribution control mode
  - LSR assigns a local binding to a FEC only if it recognizes it is the egress LSR for that FEC (customer static/connected routes) or if it receives a label binding from the next hop downstream LSR (loopbacks)

- Note: By default labels are not assigned to BGP routes in the IP routing table. The BGP routes use the same label as the interior route toward the BGP next hop

LDP Session Establishment



```
No.   Time        Source         Destination    Protocol  Length  Info
   37 15.225000   10.1.1.2       224.0.0.2      LDP           76  Hello Message
   41 16.363000   10.1.1.2       224.0.0.2      LDP           76  Hello Message
   45 16.472000   192.168.255.2  192.168.255.1  LDP           90  Initialization Message
   47 16.581000   192.168.255.1  192.168.255.2  LDP           98  Initialization Message Keep Alive Message
   49 16.696000   192.168.255.2  192.168.255.1  LDP           72  Keep Alive Message
   51 16.721000   192.168.255.2  192.168.255.1  LDP          402  Address Message Label Mapping Message Labe
   53 16.852000   192.168.255.1  192.168.255.2  LDP          426  Address Message Label Mapping Message Labe
   54 16.867000   10.1.1.1       224.0.0.2      LDP           76  Hello Message
```

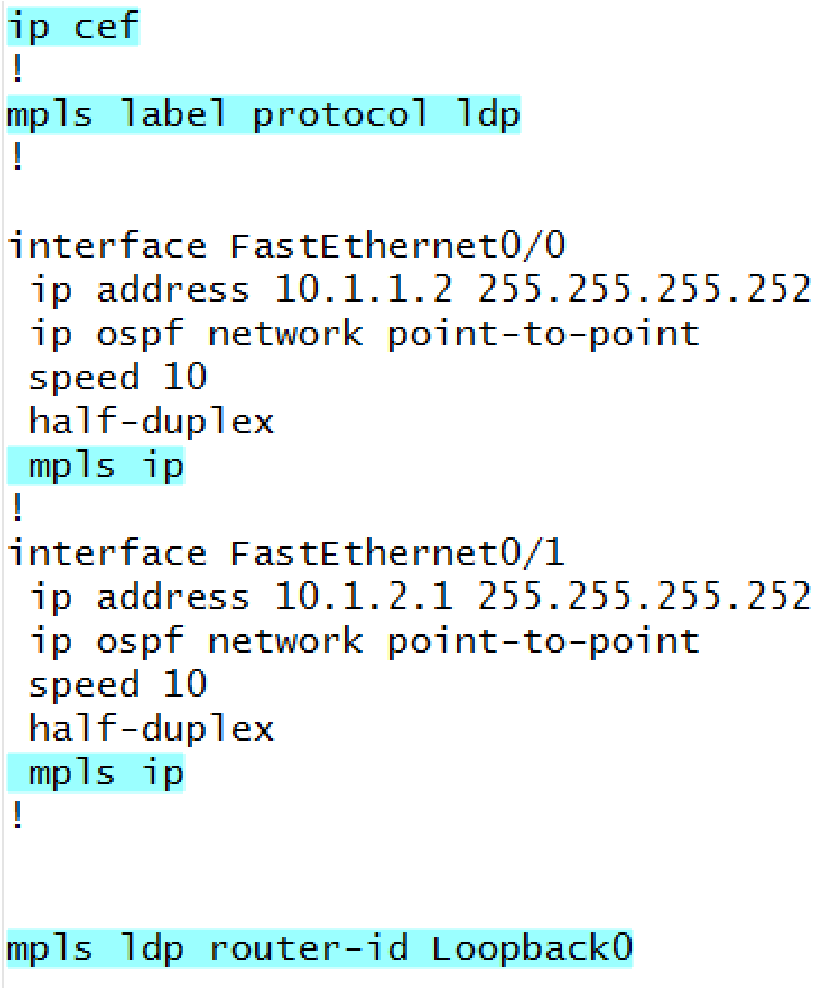
```
⊞ Frame 45: 90 bytes on wire (720 bits), 90 bytes captured (720 bits)
⊞ Ethernet II, Src: c2:02:13:74:00:00 (c2:02:13:74:00:00), Dst: c2:01:13:74:00:00 (c2:01:13:74:00:00)
⊞ Internet Protocol Version 4, Src: 192.168.255.2 (192.168.255.2), Dst: 192.168.255.1 (192.168.255.1)
⊞ Transmission Control Protocol, Src Port: 60275 (60275), Dst Port: ldp (646), Seq: 1, Ack: 1, Len: 36
⊟ Label Distribution Protocol
    Version: 1
    PDU Length: 32
    LSR ID: 192.168.255.2 (192.168.255.2)
    Label Space ID: 0
  ⊟ Initialization Message
      0... .... = U bit: Unknown bit not set
      Message Type: Initialization Message (0x200)
      Message Length: 22
      Message ID: 0x00000094
    ⊟ Common Session Parameters TLV
        00.. .... = TLV Unknown bits: Known TLV, do not Forward (0x00)
        TLV Type: Common Session Parameters TLV (0x500)
        TLV Length: 14
      ⊟ Parameters
          Session Protocol Version: 1
          Session KeepAlive Time: 180
          0... .... = Session Label Advertisement Discipline: Downstream Unsolicited proposed
          .0.. .... = Session Loop Detection: Loop Detection Disabled
          Session Path Vector Limit: 0
          Session Max PDU Length: 0
          Session Receiver LSR Identifier: 192.168.255.1 (192.168.255.1)
          Session Receiver Label Space Identifier: 0
```

LIB - Label Information Base

- LSR maintains learned labels in LIB
- This table associates each label pair with its corresponding FEC and the outbound interface
- When next hop changes for a FEC, routing table will receive the label for the new next hop from the LIB

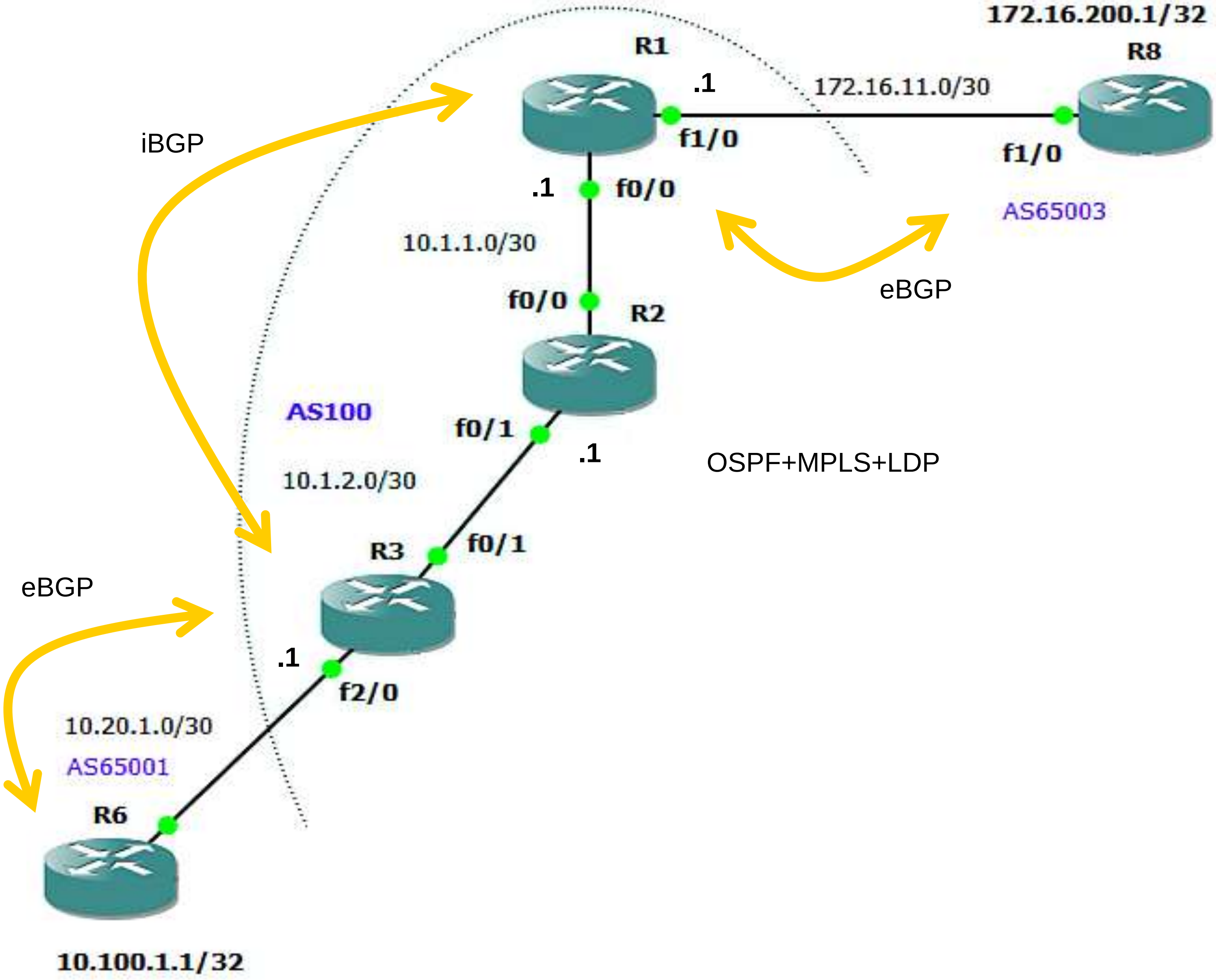
- Contents of the LIB
  - Address prefix
  - Incoming label
  - Outgoing label
  - Outgoing interface

MPLS & LDP configuration on LSR

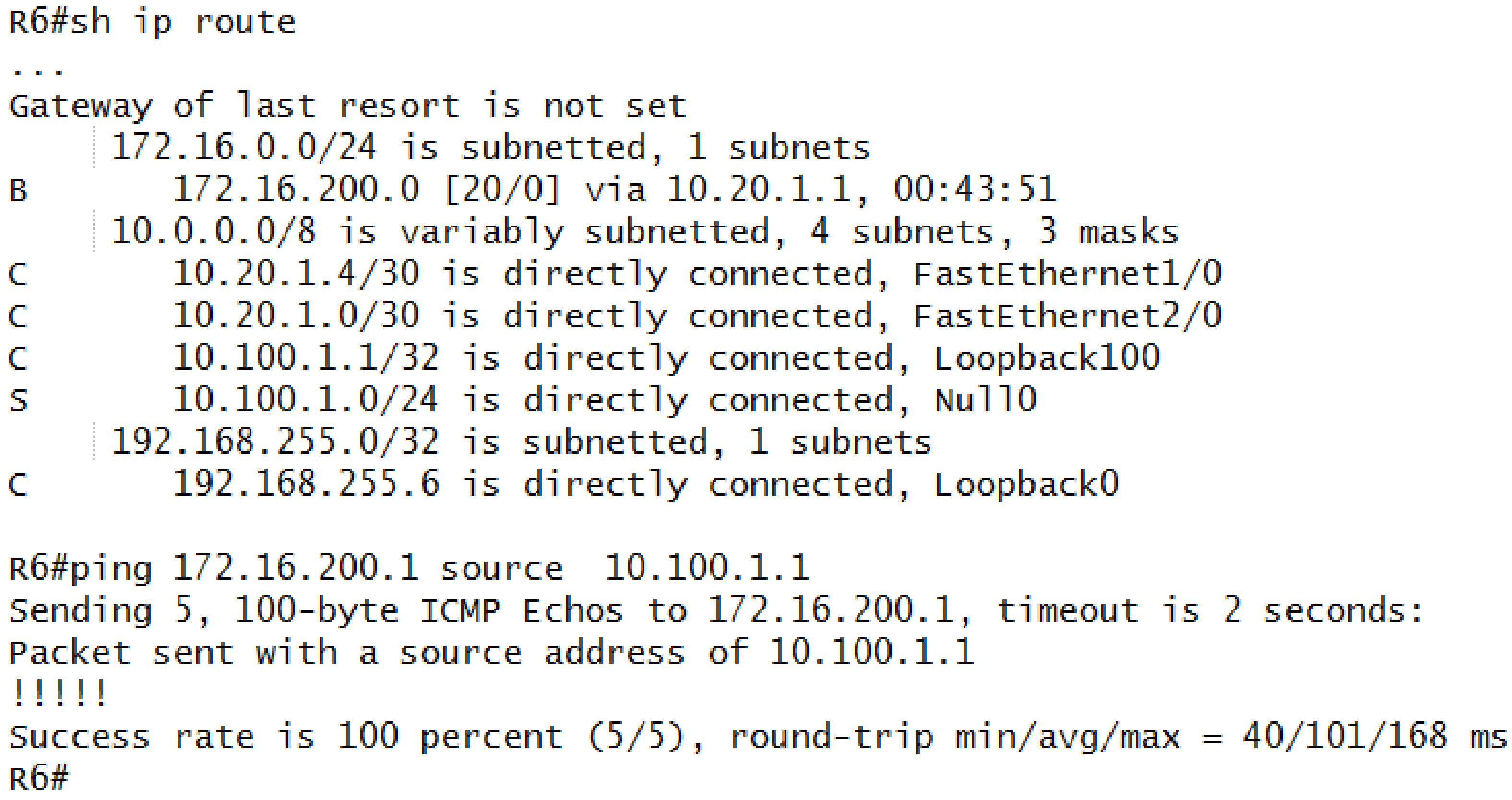


```
ip cef
!
mpls label protocol ldp
!

interface FastEthernet0/0
 ip address 10.1.1.2 255.255.255.252
 ip ospf network point-to-point
 speed 10
 half-duplex
 mpls ip
!
interface FastEthernet0/1
 ip address 10.1.2.1 255.255.255.252
 ip ospf network point-to-point
 speed 10
 half-duplex
 mpls ip
!



mpls ldp router-id Loopback0
```

MPLS Topology Example

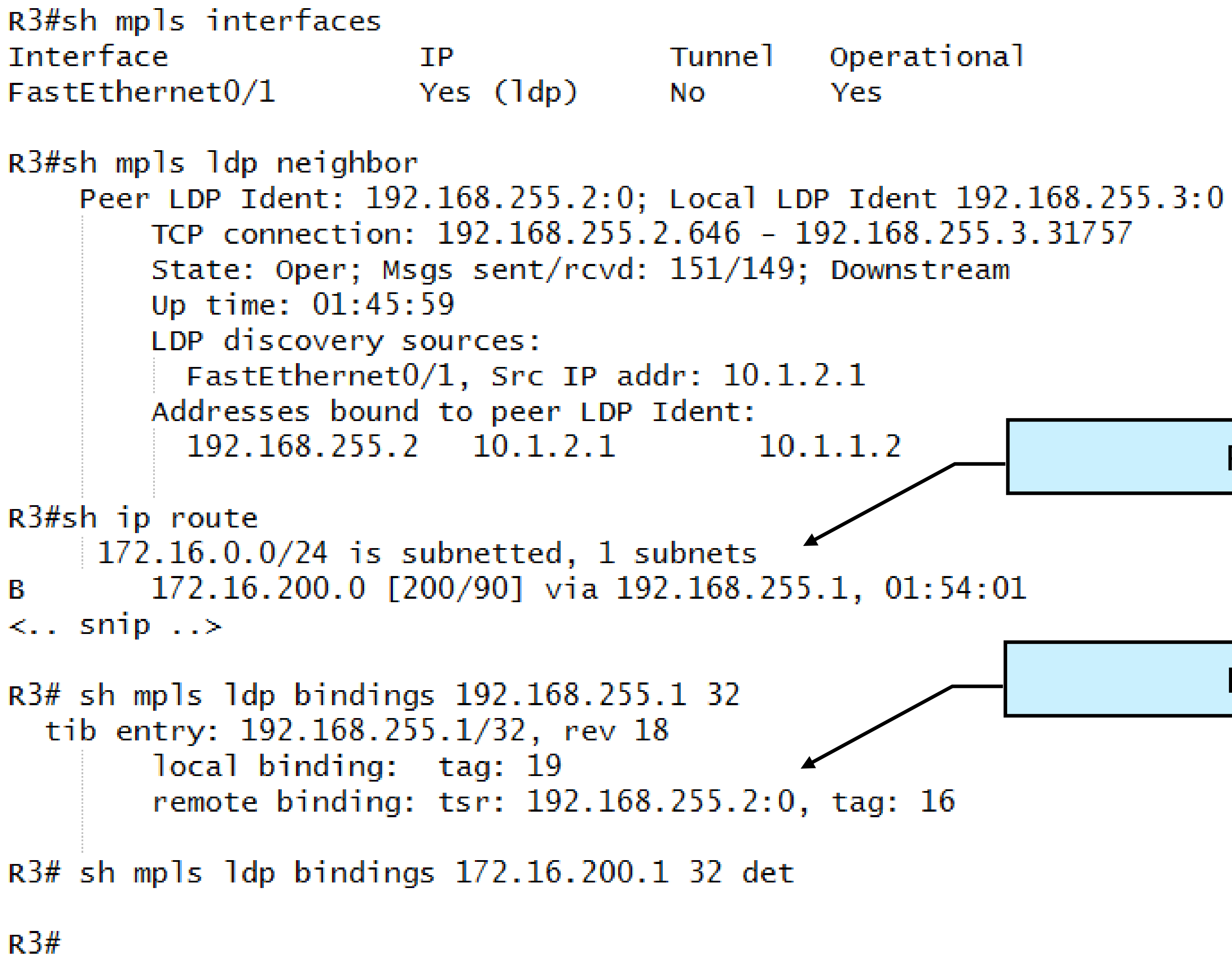


MPLS Monitoring
R6# - Customer Edge Router



```
R6#sh ip route
...
Gateway of last resort is not set
     172.16.0.0/24 is subnetted, 1 subnets
B       172.16.200.0 [20/0] via 10.20.1.1, 00:43:51
     10.0.0.0/8 is variably subnetted, 4 subnets, 3 masks
C       10.20.1.4/30 is directly connected, FastEthernet1/0
C       10.20.1.0/30 is directly connected, FastEthernet2/0
C       10.100.1.1/32 is directly connected, Loopback100
S       10.100.1.0/24 is directly connected, Null0
     192.168.255.0/32 is subnetted, 1 subnets
C       192.168.255.6 is directly connected, Loopback0

R6#ping 172.16.200.1 source  10.100.1.1
Sending 5, 100-byte ICMP Echos to 172.16.200.1, timeout is 2 seconds:
Packet sent with a source address of 10.100.1.1
!!!!!
Success rate is 100 percent (5/5), round-trip min/avg/max = 40/101/168 ms
R6#
```

MPLS Monitoring
R3# - Ingress Router



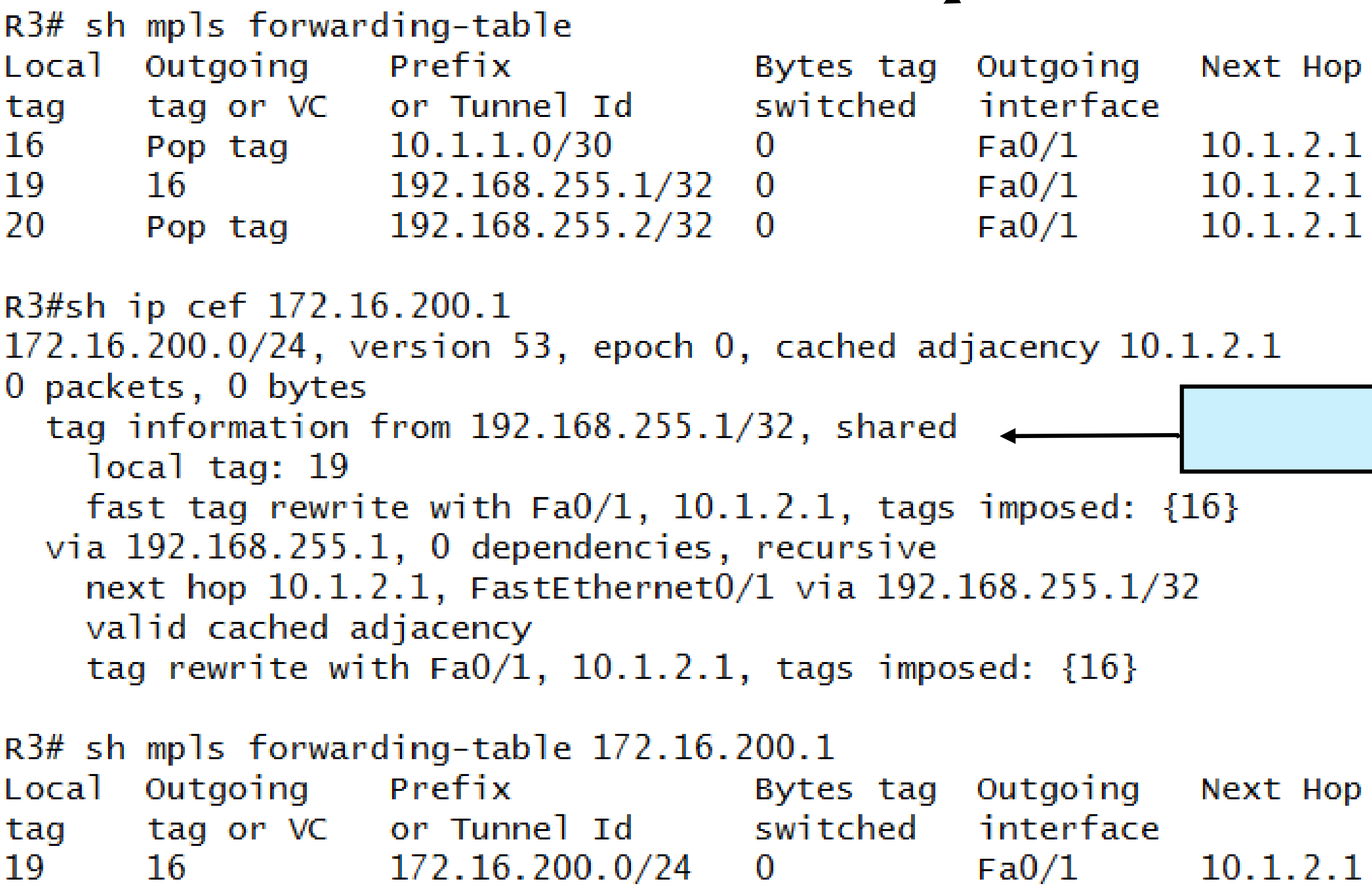
```
R3#sh mpls interfaces
Interface              IP              Tunnel   Operational
FastEthernet0/1        Yes (ldp)       No       Yes

R3#sh mpls ldp neighbor
    Peer LDP Ident: 192.168.255.2:0; Local LDP Ident 192.168.255.3:0
        TCP connection: 192.168.255.2.646 - 192.168.255.3.31757
        State: Oper; Msgs sent/rcvd: 151/149; Downstream
        Up time: 01:45:59
        LDP discovery sources:
          FastEthernet0/1, Src IP addr: 10.1.2.1
        Addresses bound to peer LDP Ident:
          192.168.255.2   10.1.2.1        10.1.1.2

R3#sh ip route
     172.16.0.0/24 is subnetted, 1 subnets
B       172.16.200.0 [200/90] via 192.168.255.1, 01:54:01
<.. snip ..>

R3# sh mpls ldp bindings 192.168.255.1 32
  tib entry: 192.168.255.1/32, rev 18
        local binding:  tag: 19
        remote binding: tsr: 192.168.255.2:0, tag: 16

R3# sh mpls ldp bindings 172.16.200.1 32 det

R3#
```

RIB

LIB

MPLS Monitoring
R3# - Recursive Lookup

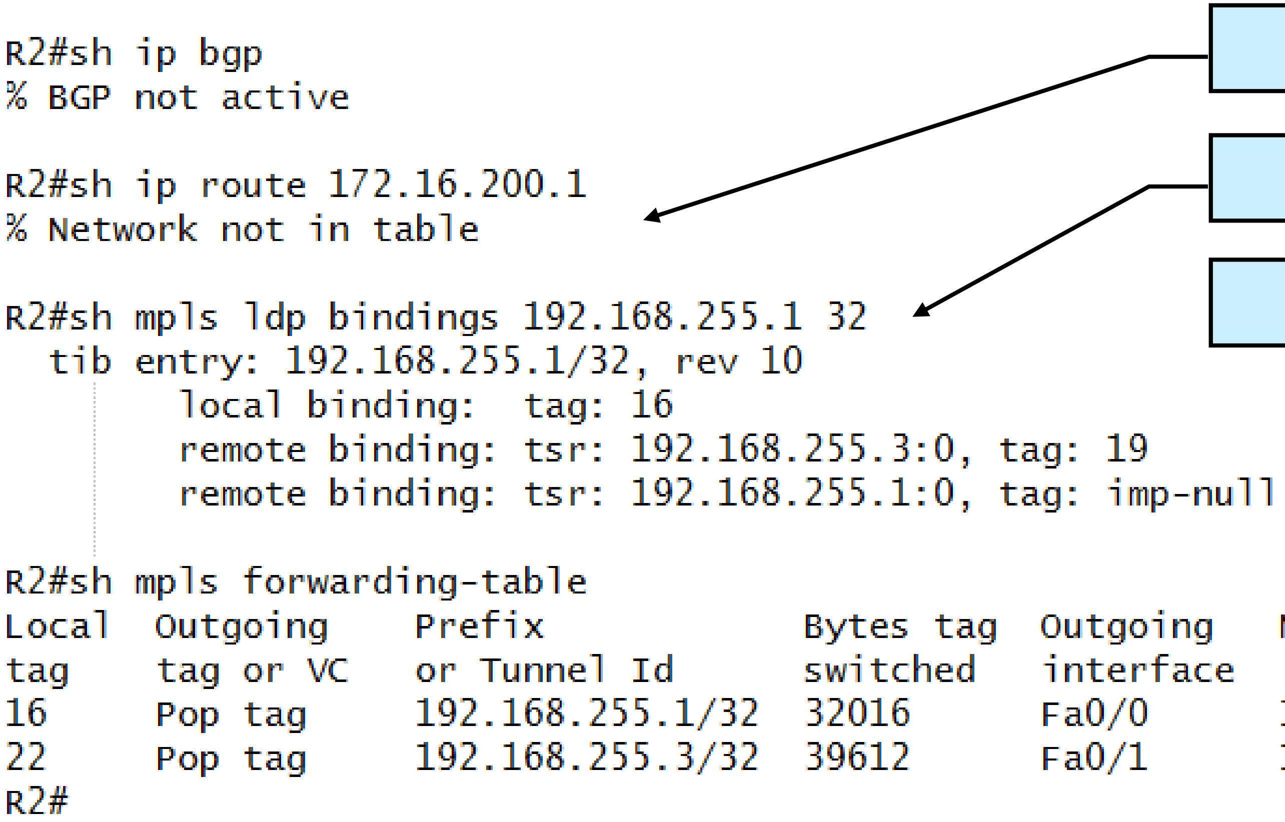
LFIB



```
R3# sh mpls forwarding-table
Local  Outgoing    Prefix            Bytes tag  Outgoing   Next Hop
tag    tag or VC   or Tunnel Id      switched   interface
16     Pop tag     10.1.1.0/30       0          Fa0/1      10.1.2.1
19     16          192.168.255.1/32  0          Fa0/1      10.1.2.1
20     Pop tag     192.168.255.2/32  0          Fa0/1      10.1.2.1

R3#sh ip cef 172.16.200.1
172.16.200.0/24, version 53, epoch 0, cached adjacency 10.1.2.1
0 packets, 0 bytes
  tag information from 192.168.255.1/32, shared
    local tag: 19
    fast tag rewrite with Fa0/1, 10.1.2.1, tags imposed: {16}
  via 192.168.255.1, 0 dependencies, recursive
    next hop 10.1.2.1, FastEthernet0/1 via 192.168.255.1/32
    valid cached adjacency
    tag rewrite with Fa0/1, 10.1.2.1, tags imposed: {16}

R3# sh mpls forwarding-table 172.16.200.1
Local  Outgoing    Prefix            Bytes tag  Outgoing   Next Hop
tag    tag or VC   or Tunnel Id      switched   interface
19     16          172.16.200.0/24   0          Fa0/1      10.1.2.1
```

FIB

MPLS Monitoring
R2# - LSR, BGP Free



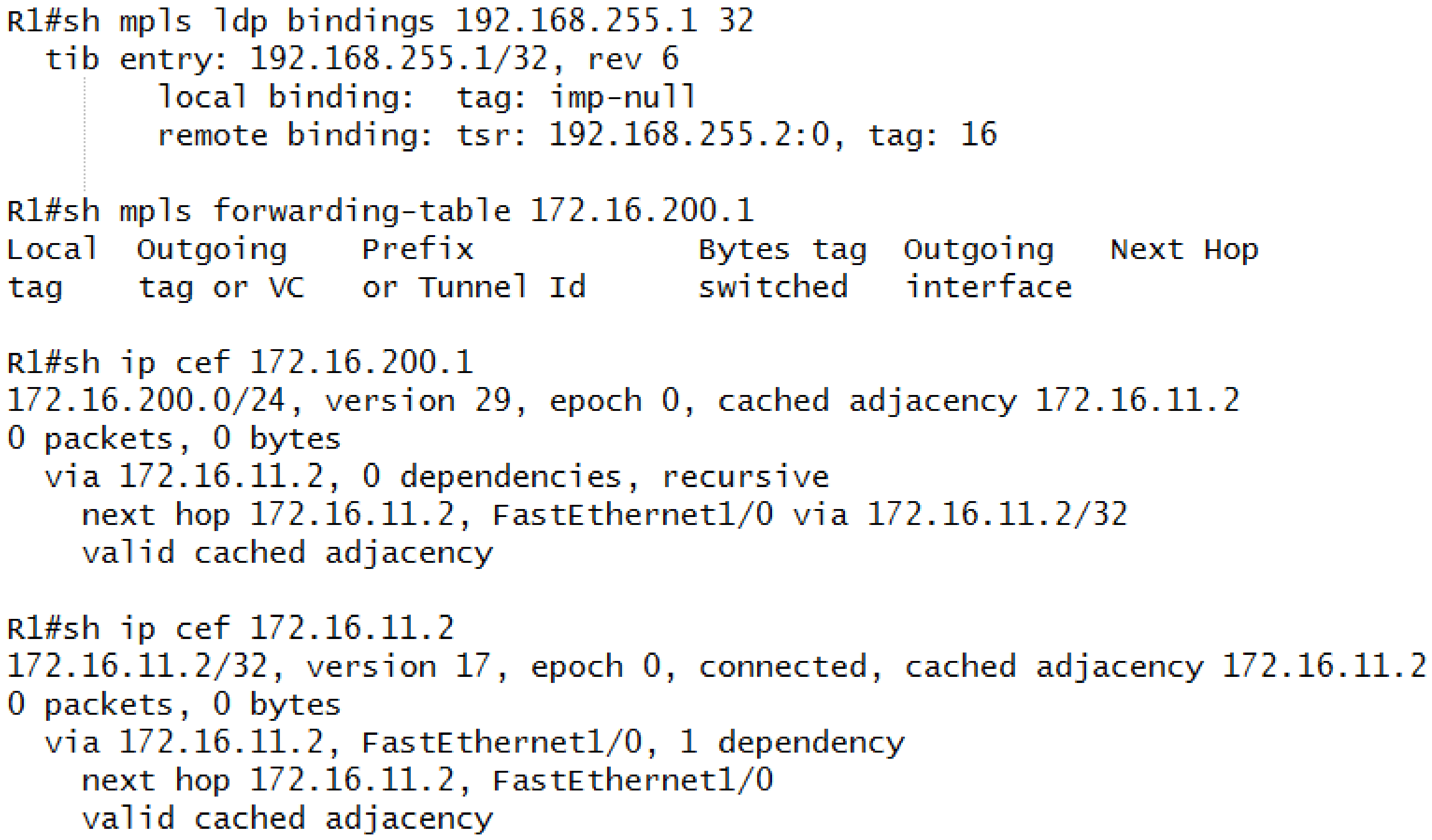
```
R2#sh ip bgp
% BGP not active

R2#sh ip route 172.16.200.1
% Network not in table

R2#sh mpls ldp bindings 192.168.255.1 32
  tib entry: 192.168.255.1/32, rev 10
        local binding:  tag: 16
        remote binding: tsr: 192.168.255.3:0, tag: 19
        remote binding: tsr: 192.168.255.1:0, tag: imp-null

R2#sh mpls forwarding-table
Local  Outgoing    Prefix            Bytes tag  Outgoing    Next Hop
tag    tag or VC   or Tunnel Id      switched   interface
16     Pop tag     192.168.255.1/32  32016      Fa0/0       10.1.1.1
22     Pop tag     192.168.255.3/32  39612      Fa0/1       10.1.2.2
R2#
```

RIB

LIB

LFIB

MPLS Monitoring
R1# - Egress Router



```
R1#sh mpls ldp bindings 192.168.255.1 32
  tib entry: 192.168.255.1/32, rev 6
        local binding:  tag: imp-null
        remote binding: tsr: 192.168.255.2:0, tag: 16

R1#sh mpls forwarding-table 172.16.200.1
Local  Outgoing    Prefix            Bytes tag  Outgoing   Next Hop
tag    tag or VC   or Tunnel Id      switched   interface

R1#sh ip cef 172.16.200.1
172.16.200.0/24, version 29, epoch 0, cached adjacency 172.16.11.2
0 packets, 0 bytes
  via 172.16.11.2, 0 dependencies, recursive
    next hop 172.16.11.2, FastEthernet1/0 via 172.16.11.2/32
    valid cached adjacency

R1#sh ip cef 172.16.11.2
172.16.11.2/32, version 17, epoch 0, connected, cached adjacency 172.16.11.2
0 packets, 0 bytes
  via 172.16.11.2, FastEthernet1/0, 1 dependency
    next hop 172.16.11.2, FastEthernet1/0
    valid cached adjacency
```

- "dependencies" refers to the number of routes that are resolvable via the current route
- "recursive" means that the prefix is itself resolvable via another route, meaning that it is not a connected route and has to be resolved via another route (172.16.11.2 in this case)

Ďakujem za pozornosť

roman dot kaloc at gmail dot com