MPLS

Matěj Grégr
mgregr@fit.vut.cz

Agenda

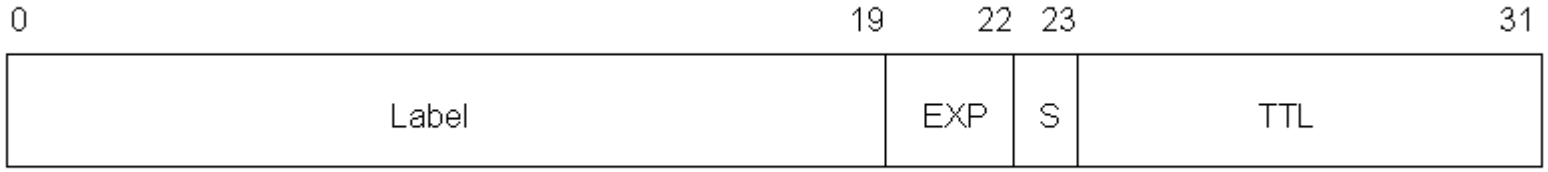
- Why MPLS – the BGP-free core
- MPLS basics: labels, LSPs, LDP
- MPLS L3VPN
- MPLS L2VPN: pseudowires, VPLS, EVPN
- Segment Routing, traffic engineering, fast reroute
- Operations

MPLS – MultiProtocol Label Switching

- Why?
  - Necessity to have protocol-independent transport
  - Speed? – historically yes; on modern hardware label and IP lookups are equally fast
  - Services: L3VPN, L2VPN (VPWS, VPLS, EVPN)
  - QoS (TC bits)
  - BGP free core
  - Traffic engineering and fast reroute (RSVP-TE, Segment Routing)

MPLS – MultiProtocol Label Switching

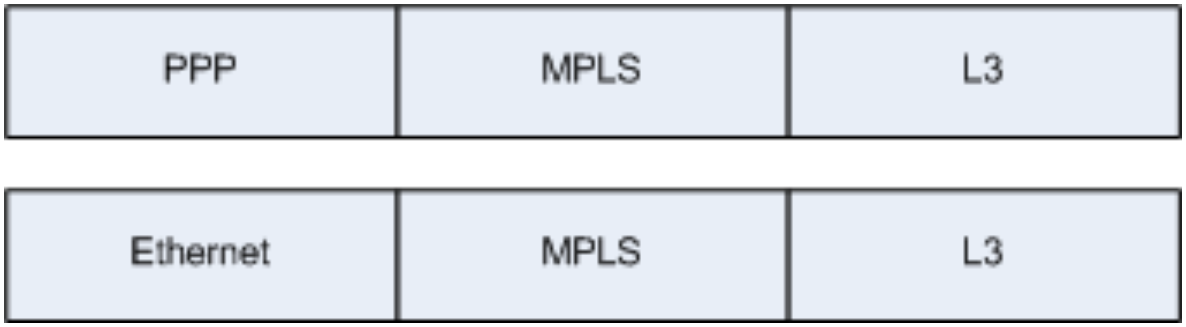
- Use labels for packet forwarding

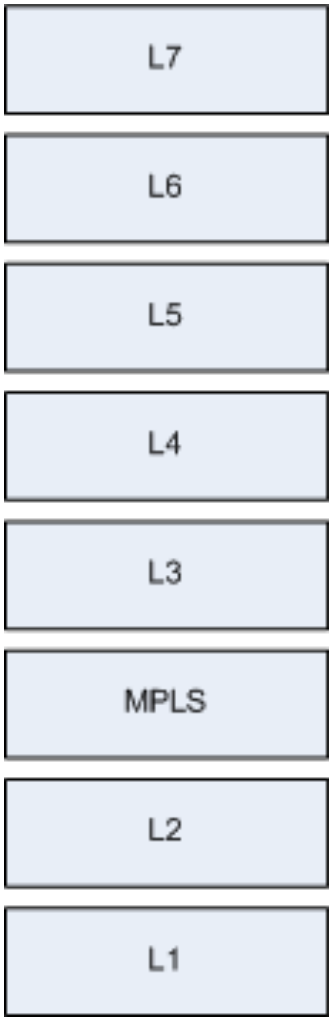


- Label
  - labels 0 – 15 reserved
- TC – Traffic Class (formerly EXP, RFC 5462) = QoS
- S – Bottom of stack – the last label
- TTL – like IP TTL; ingress copies the IP TTL into the label
  - no mpls ip propagate-ttl hides the core from traceroute

MPLS – label

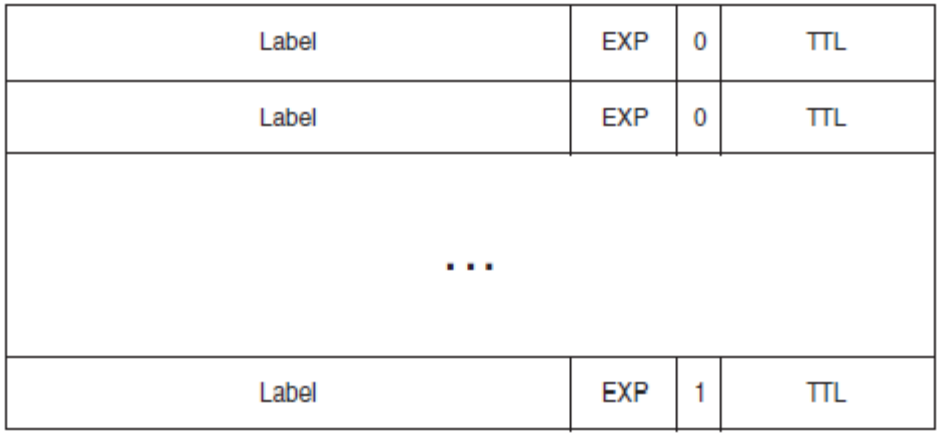
- MPLS label is inserted between L2/L3 layer
- Eliminates dependency on L2 technology
  - “Shim” header, EtherType 0x8847
  - No “next protocol” field in the header





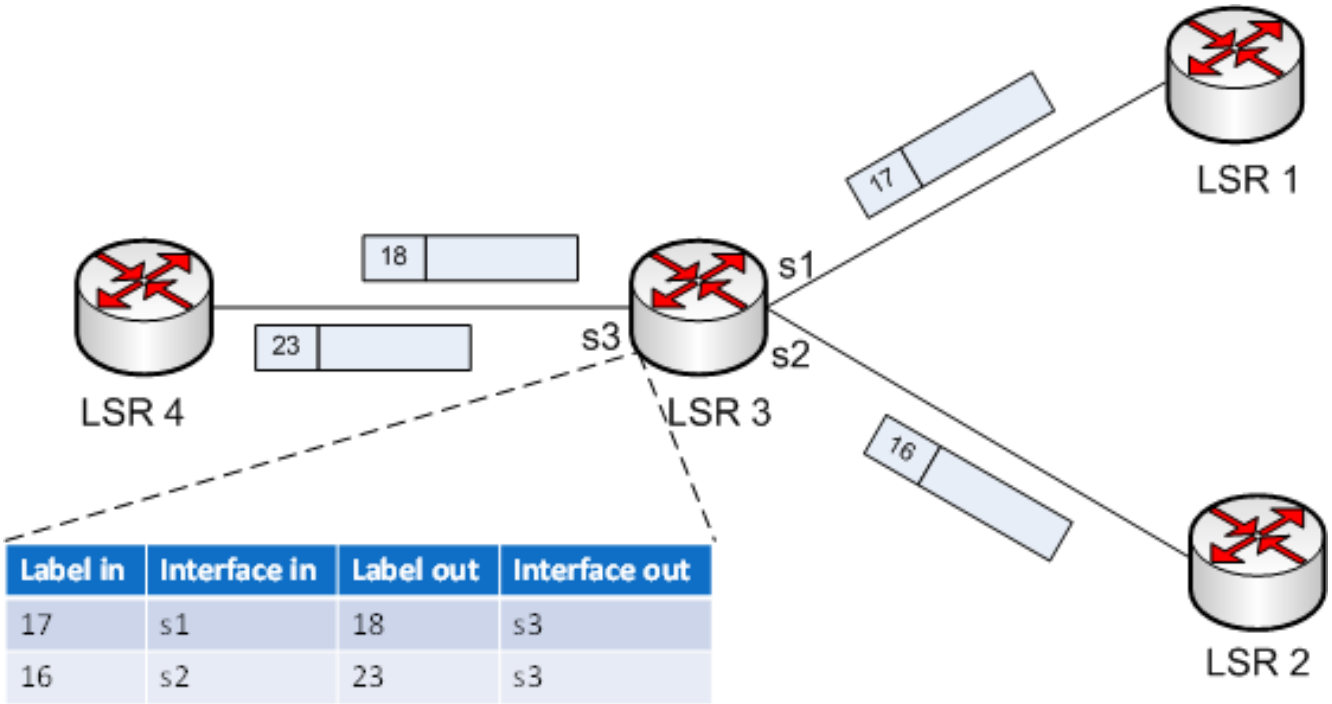
Label stack

- Several labels can be stacked
- MPLS switches operate with the topmost label only
- Last label in the stack is marked with the S bit



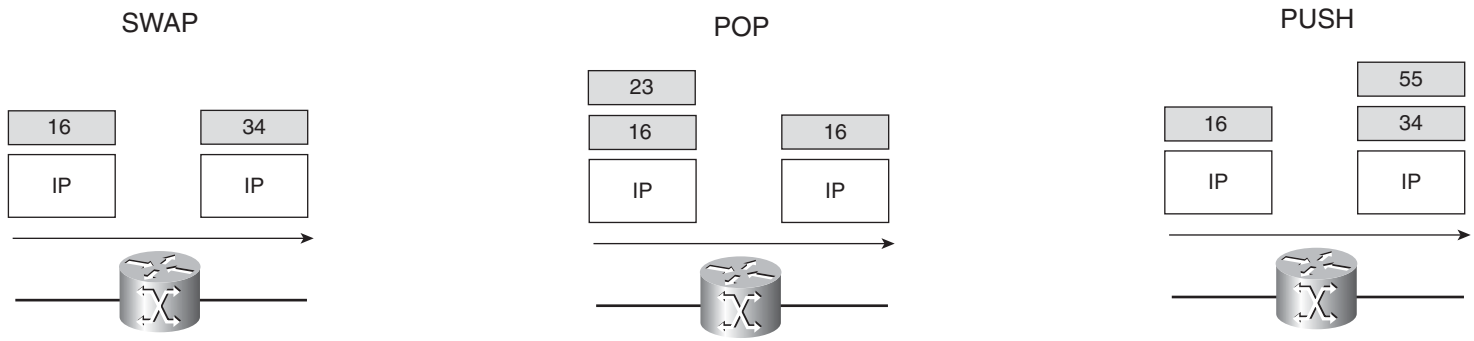
Label switching

- Labels are **locally** significant



Label in	Interface in	Label out	Interface out
17	s1	18	s3
16	s2	23	s3

Label stack – possible actions



PUSH example: incoming label 16 is swapped to 34, then label 55 is pushed on top (swap + push).

Figures: L. De Ghein, MPLS Fundamentals, Cisco Press

Reserved labels

- Implicit NULL
  - Value 3
  - Only signalled by LDP – never carried in a packet
  - Egress LSR requests the upstream LSR to perform a pop operation
    - Called PHP - Penultimate Hop Popping
  - The result - egress LSR receives an IP packet and only needs to perform an IP lookup to be able to forward the packet

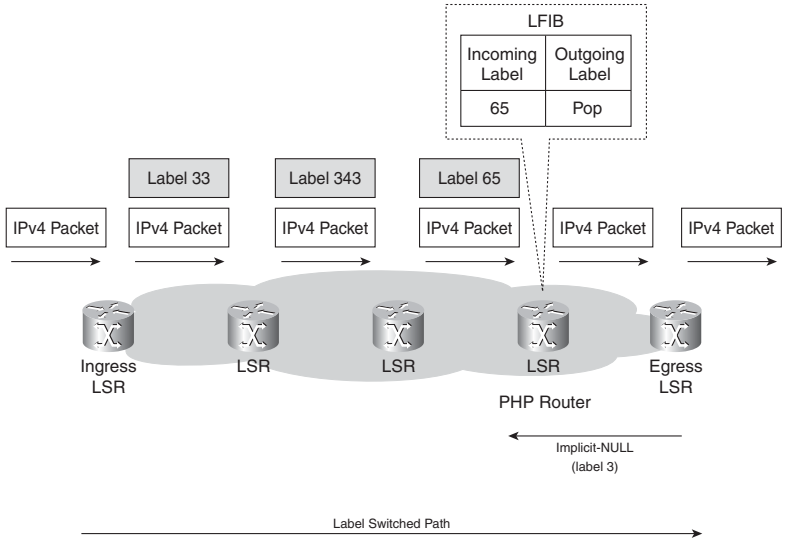


Figure: L. De Ghein, MPLS Fundamentals, Cisco Press

Reserved labels

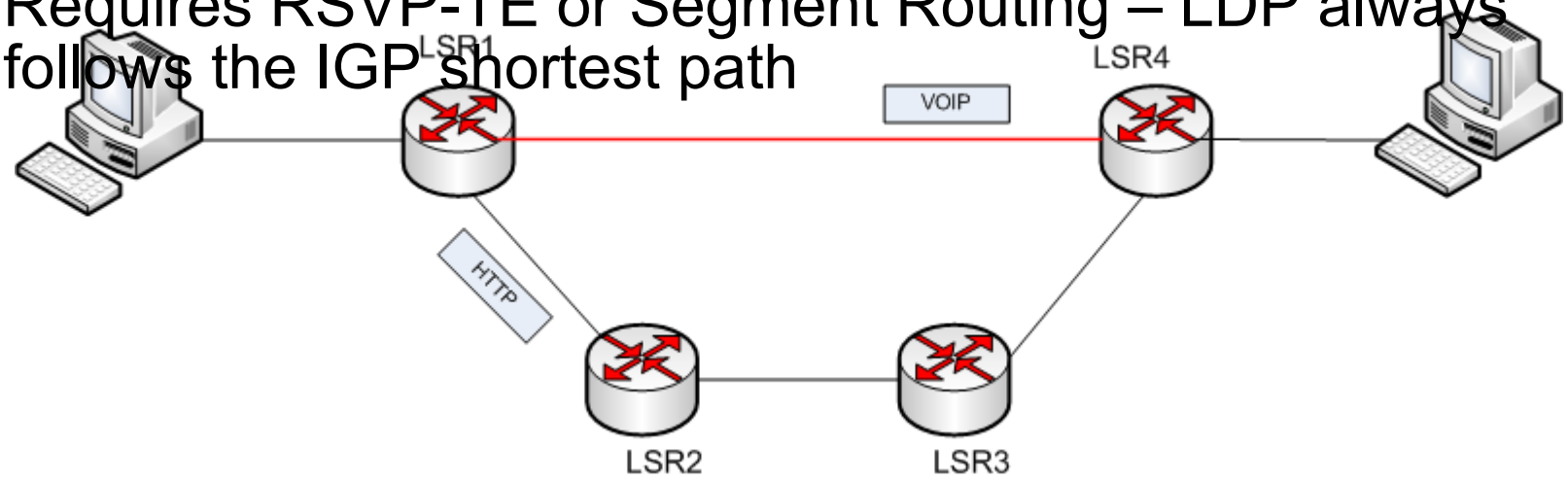
- Implicit NULL label removes the whole label – the EXP (QoS) bits are thus removed as well
- Explicit NULL label
  - Value 0 (2 for IPv6)
  - Egress LSR signals the explicit NULL label to the penultimate hop router
    - receives labeled packets with a label of value 0 (2) as the top label
  - Pops the label without an LFIB lookup, then IP lookup
  - QoS bits are preserved
  - Used with the DiffServ pipe / short-pipe models (RFC 3270)

FEC

- Flow of packets that are forwarded along the same path and are treated the same
- All packets belonging to the same FEC have the same label
- Usually:
  - Packets with Layer 3 destination IP addresses matching a certain prefix
  - BGP routes with the same next hop share the label of that next hop
- With LDP, a FEC is always an IGP prefix (usually a /32 loopback)

FEC example – traffic steering

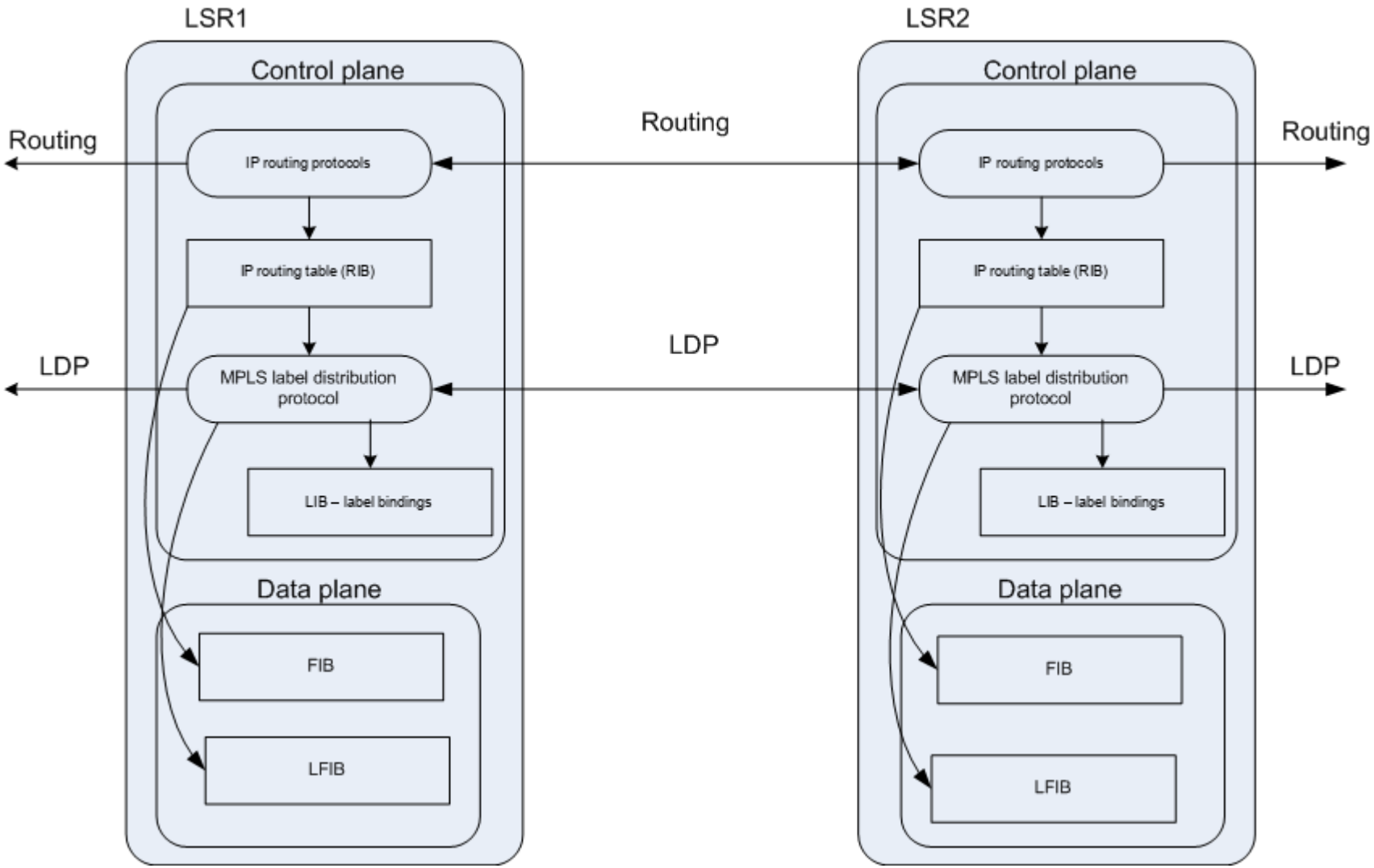
- LSP Paths
  - LSP1 = (LSR1, LSR4), LSP2 = (LSR1, LSR2, LSR3, LSR4)
  - LSP1 has lower latency than LSP2
  - FEC1 – all VoIP data, LSR1 adds label so all FEC1 packets go through LSP1
  - FEC2 – rest of the traffic use LSP2
- Requires RSVP-TE or Segment Routing – LDP always follows the IGP shortest path



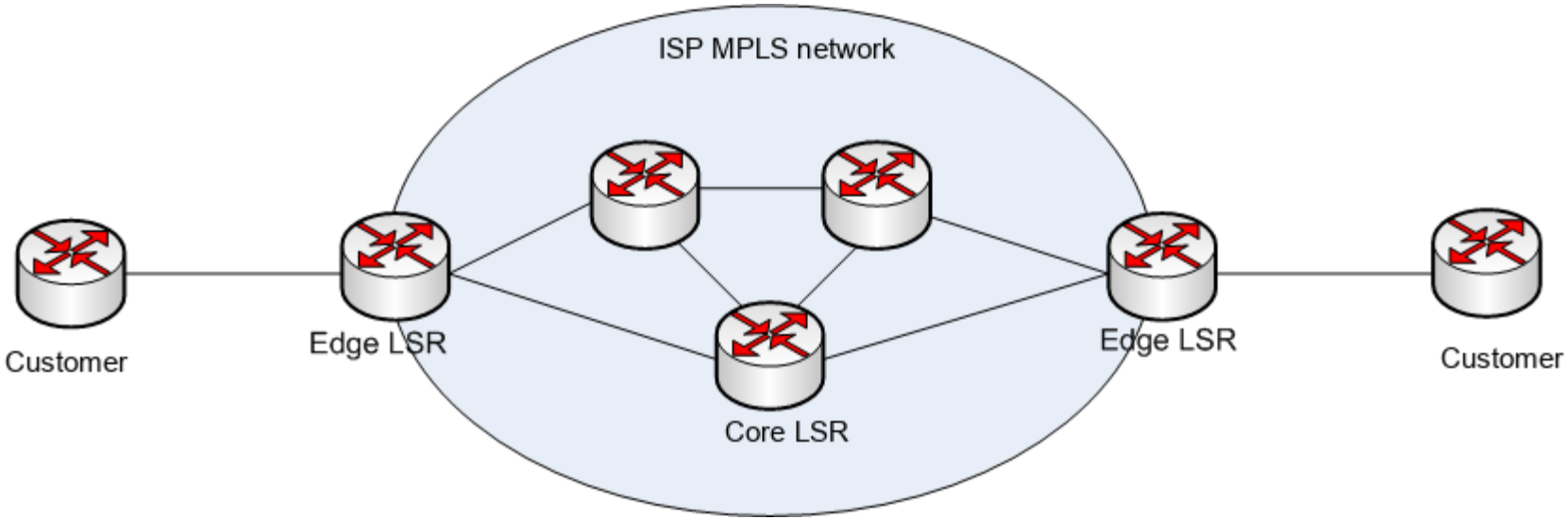
RIB, FIB, LIB and LFIB

- RIB – routes from the IGP (and BGP)
- FIB (CEF) – precomputed copy of the RIB for forwarding, not a cache
- LIB – all label bindings from all LDP neighbours (control plane)
- LFIB – only the label of the current IGP next hop (data plane)
- Liberal label retention: non-best labels stay in the LIB
  - after an IGP change the new next hop's label is already known → fast switchover

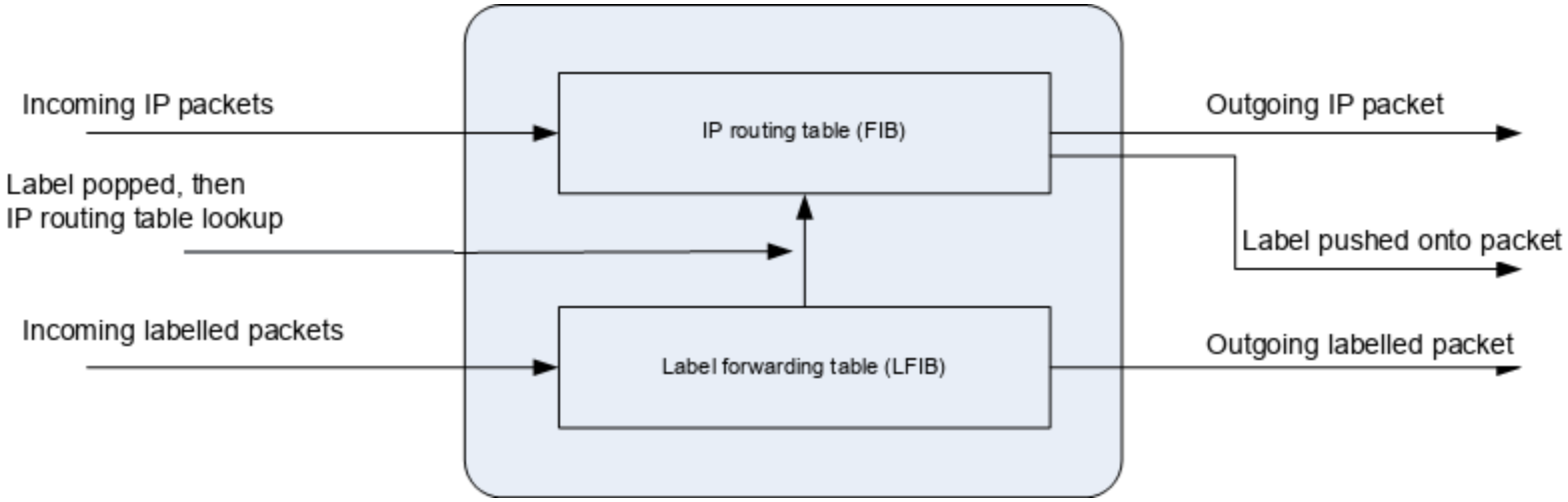
Router scheme



MPLS - network



Edge LSR

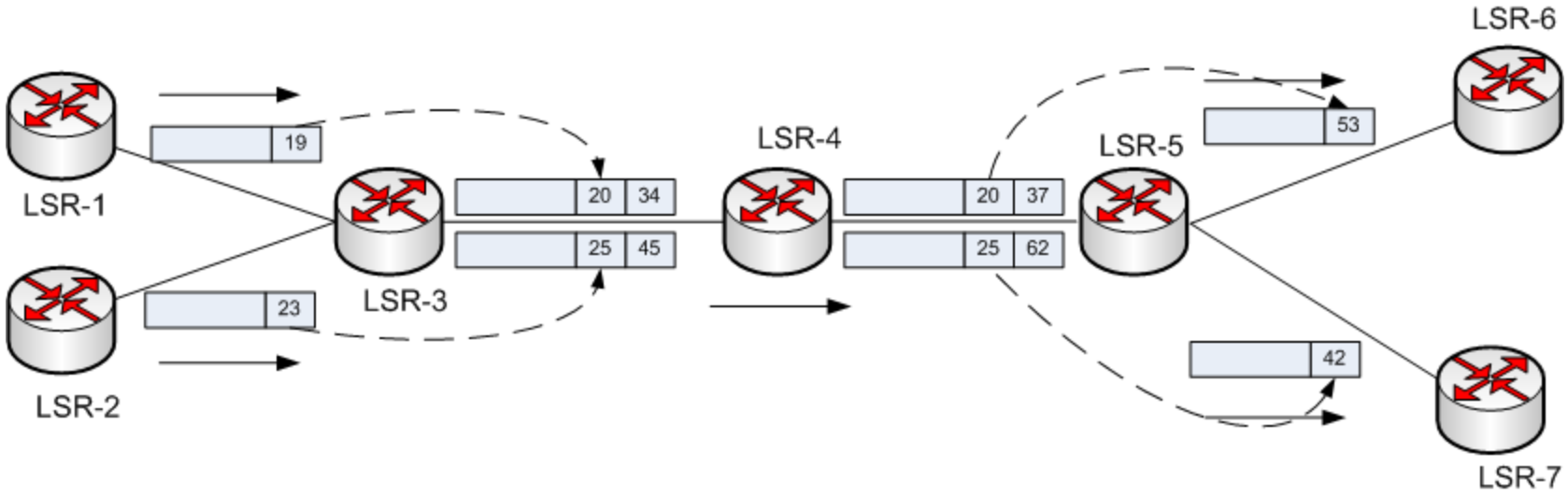


Ingress edge LSR: IP lookup → push label. Egress edge LSR: pop label → IP lookup.

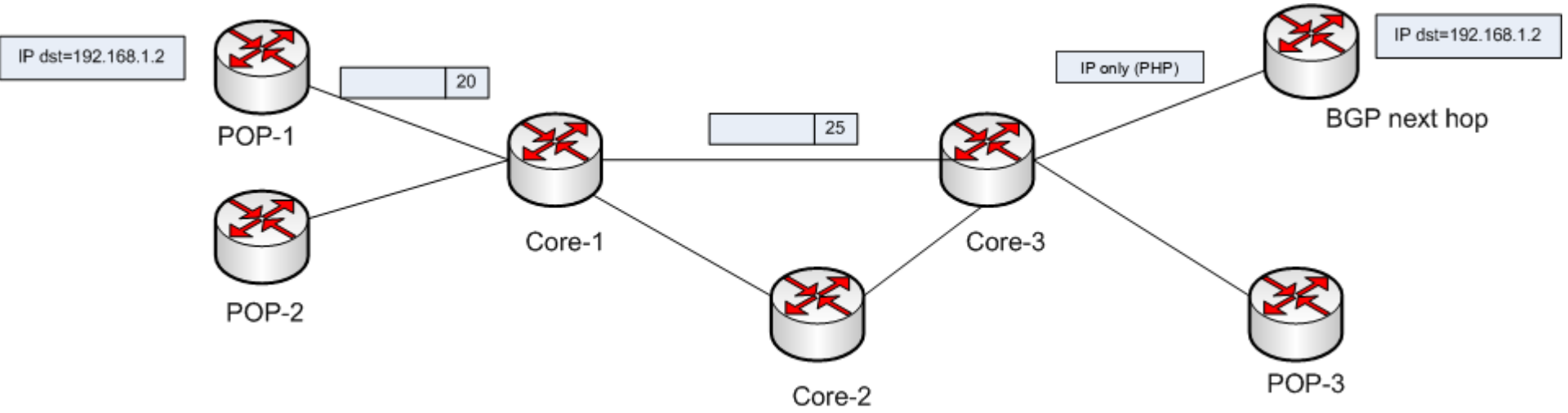
LSP – label switched path

- LSP – path of LSRs in MPLS network
- First LSR in LSP path – ingress LSR, last – egress LSR
- LSP is unidirectional
- Ingress LSR does not have to be the first router that adds label – labels can be stacked

LSP – tunnel


LSR-1
LSR-2
LSR-3
LSR-4
LSR-5
LSR-6
LSR-7
19
23
20
34
25
45
20
37
25
62
53
42

BGP free core



- POP-1 receives a packet for 192.168.1.2 – BGP route, next hop = egress loopback /32
- Recursive lookup: next hop resolves via IGP → POP-1 pushes the LDP label for that /32
- Core-1 – Core-3 only swap labels: no BGP, no full Internet table
- Core-3 pops the label (PHP); the BGP next-hop router does a normal IP lookup

LDP – Label Distribution protocol

- Labels are locally significant
- LDP protocol is used to distribute labels between routers

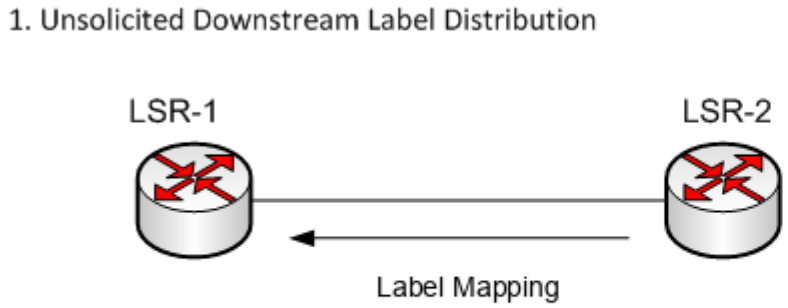
- Discovery messages – discover other LSR in network, UDP multicast
- Session messages – connection between LDP nodes
- Advertisement messages – changes, creation, deletion labels for FEC
- Notification messages – information messages and management signalization

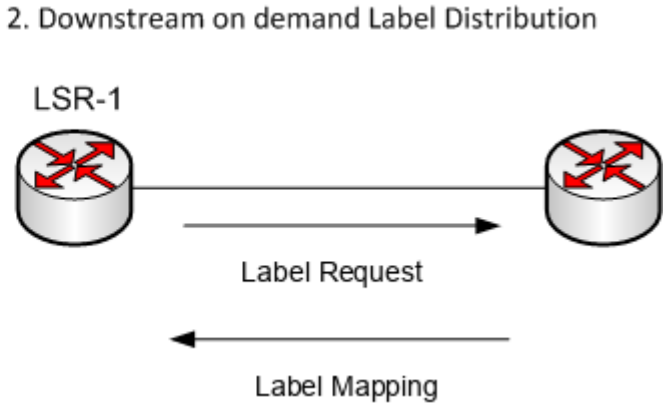
LDP – discovery

- UDP packet to all-routers multicast (224.0.0.2), or UDP unicast port (646), to a specific IP address
- After discovery – LDP session over TCP 646 between transport addresses (usually loopbacks)
- Transport address must be reachable via the IGP – most common lab failure
- LDP router-ID = highest loopback; pin it: mpls ldp router-id Loopback1 force
- Best practice: LDP–IGP synchronization, LDP session protection

LDP – labels distribution ①

1. Unsolicited Downstream Label Distribution
2. Downstream on Demand Label Distribution





LDP – label distribution ②

- Downstream Unsolicited (DU)
  - LSR-2 advertises a label for every FEC to all neighbours
  - Default for LDP on Cisco and Juniper
- Downstream on Demand (DoD)
  - LSR-1 requests a label from its next hop
  - Historical (ATM); the model used by RSVP-TE

Label distribution – control and retention

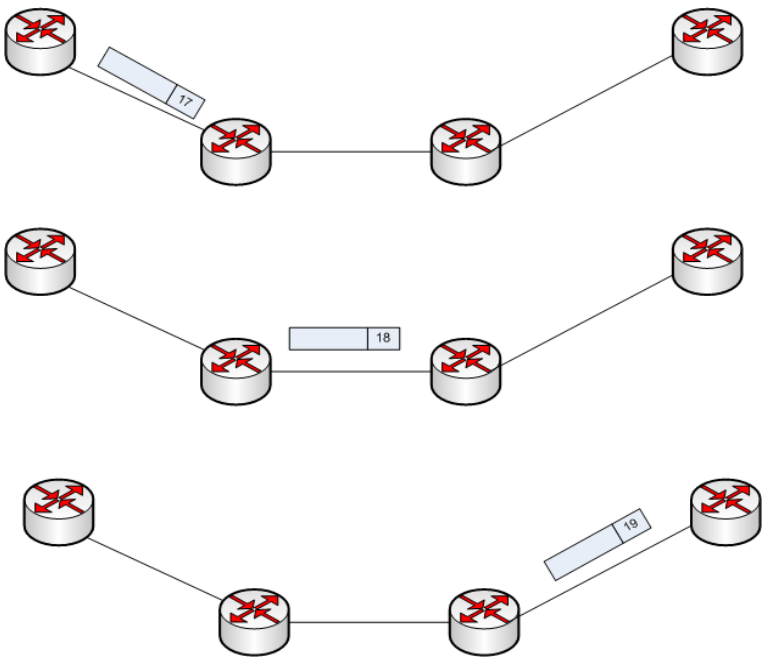
- Control mode – when is a local binding created?
  - Independent: as soon as the prefix is in the RIB (Cisco default)
  - Ordered: only on the egress LSR, or after the binding from the next hop arrives (Junos default)
- Retention mode – which received labels are kept?
  - Liberal: labels from all neighbours (default) → fast convergence
  - Conservative: only the next hop's label → less memory

Label distribution protocols

- LDP – labels for IGP prefixes, follows the IGP shortest path
- RSVP-TE – explicit paths with bandwidth reservation (traffic engineering)
- MP-BGP – VPN / service labels, BGP labelled unicast (RFC 8277)
- IS-IS / OSPF – Segment Routing: labels carried in the IGP itself
  - Current best practice for new networks – replaces LDP and RSVP-TE

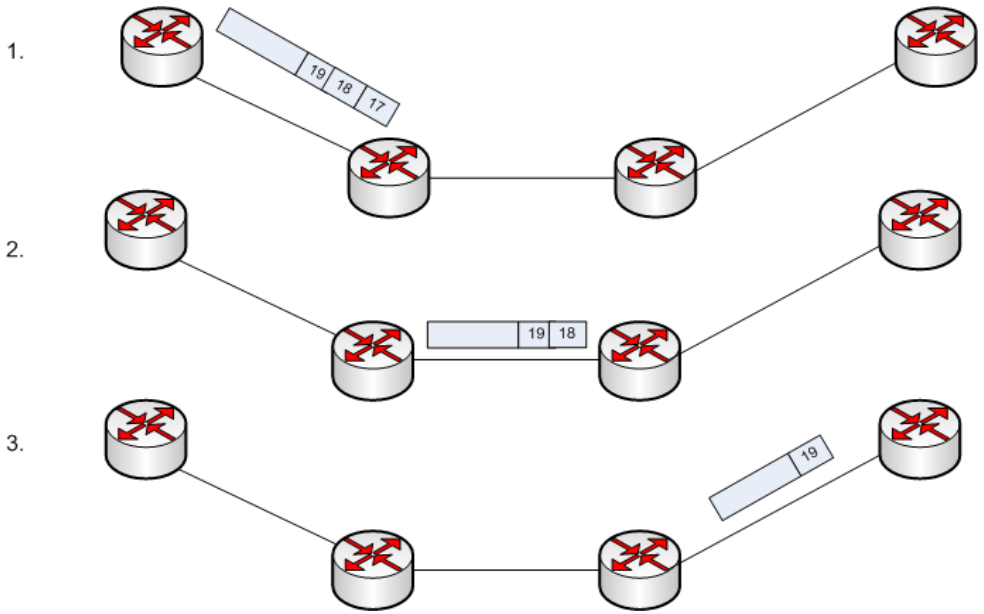
Hop-by-hop routing

- Ingress pushes, transit LSRs swap, penultimate hop pops (PHP)
- LSP – the same paths as classic routing



Explicit routing (source routing)

- Pre-defined path: ingress pushes the whole path as a label stack
  - Each hop pops its own label – this is how Segment Routing (SR-TE) works
  - RSVP-TE alternative: one label per LSP, swapped hop by hop, path signalled in advance



MPLS L3VPN

MPLS – VPN

- The most popular use case for MPLS
- Any-to-any by default; hub-and-spoke or extranet via RT policy
- ISP infrastructure is hidden for customer
- Provider Edge routers can be shared between several customers (cheaper)

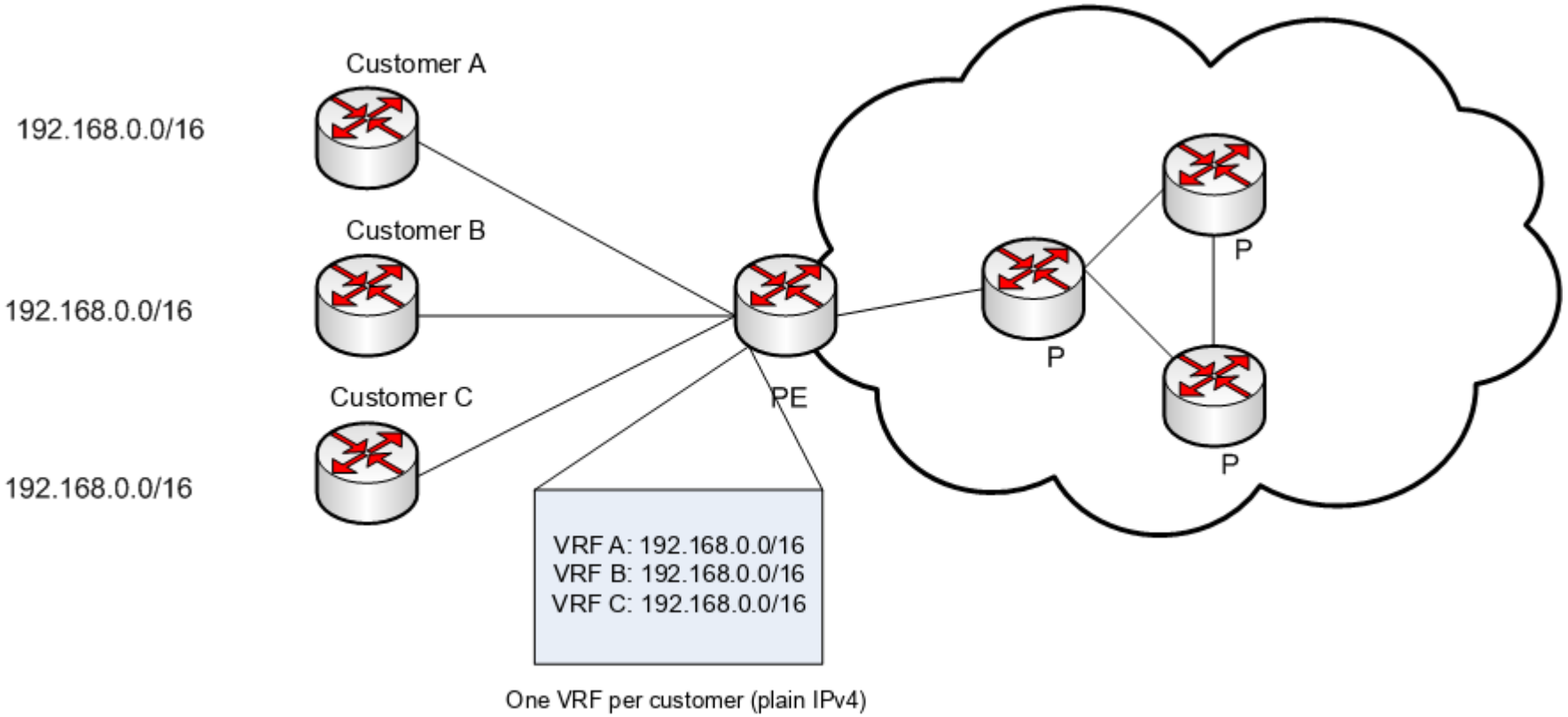
MPLS – VPN basic terms

- CE – Customer edge router
- PE – Provider edge router
- P – Provider core router (label switching only)
- VRF – Virtual routing and forwarding table (one per customer on a PE)
- RD – Route distinguisher (makes customer prefixes unique in BGP)
- RT – Route target (extended community: which VRFs import a route)
- MP-BGP – Multiprotocol BGP carrying VPNv4 / VPNv6 routes between PEs

CE and PE

- CE
  - Router in customer‘s network
  - PE–CE routing: eBGP (most common), static, OSPF
  - Not aware of MPLS network
- PE
  - ISP‘s PoP – can connect several customers
  - Customers are separated by VRF – each PE interface belongs to one VRF
  - Own VRF per customer → overlapping private addresses are fine

PE architecture



RD is added only when the route is exported to MP-BGP: 65300:1:192.168.0.0/16 (VPNv4)

RD – Route Distinguisher

- Customers can use same private address space (RFC 1918)
- IPv4 address + RD = unique identifier
- RD (64 bit) + IPv4 (32 bit) = 96-bit VPNv4 prefix; VPNv6 = RD + IPv6
- VPNv4 / VPNv6 routes are distributed between PEs by MP-BGP
- Format ASN:nn or IP:nn – only makes prefixes unique, carries no policy
- Best practice: unique RD per VRF per PE (route reflectors then keep all paths)

RT – Route Target

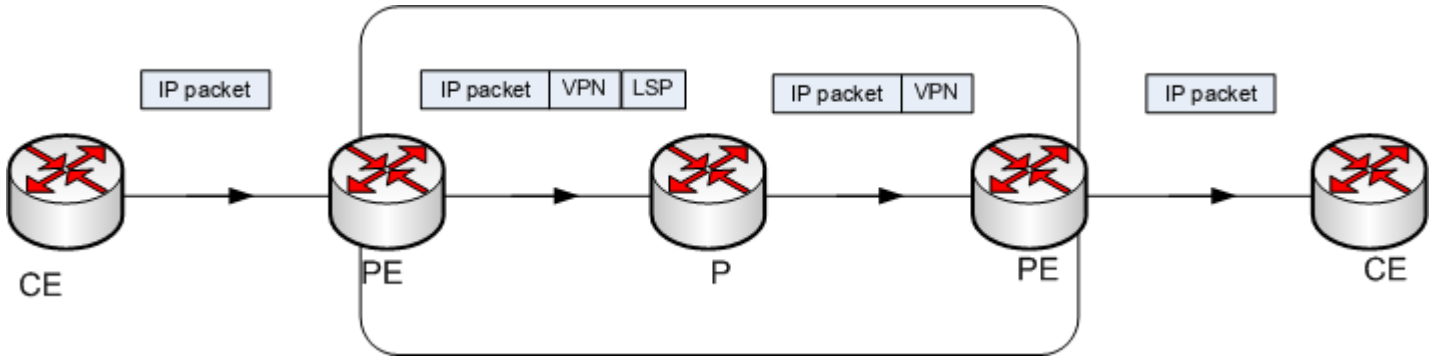
- BGP extended community attached to every VPN route
- Export RT – added when a VRF route is exported to MP-BGP
- Import RT – a PE imports a VPN route into every VRF whose import RT matches
- RD makes routes unique, RT decides where they go
- Topologies by RT policy: any-to-any, hub-and-spoke, extranet / shared services
- Use the provider's ASN: route-target both 65300:1

MPLS VPN – labels

- VPN traffic uses two labels
- Outer (top) label – transport LSP to the egress PE (LDP or SR)
- Inner (bottom, S=1) label – VPN label, allocated by the egress PE and advertised in MP-BGP
- P routers look only at the transport label
- VPN label allocation: per-prefix (Cisco default), per-CE or per-VRF

MPLS VPN – packet walk

1. CE sends IP packet
2. PE pushes VPN label (S bit = 1) and transport label
3. Penultimate P router pops the transport label (PHP)
4. Egress PE uses the VPN label to select the VRF / next hop



L3VPN control plane – step by step

- CE advertises 192.168.1.0/24 to PE1 (eBGP, OSPF or static)
- PE1 installs it in VRF RED and allocates a VPN label (e.g. 24)
- PE1 sends a VPNv4 update to PE2:
  - prefix 65300:1:192.168.1.0/24, RT 65300:1, label 24, next hop = PE1 loopback
- PE2 imports it into every VRF whose import RT matches
- PE2 resolves next hop PE1 via the IGP → transport label (LDP / SR)
- Verify: show bgp vpnv4 unicast all [labels]

Scaling – VPNv4 route reflectors

- iBGP between PEs needs a full mesh: n(n-1)/2 sessions
- Route reflectors for address-family vpnv4 / vpnv6
  - Each PE peers only with two RRs
  - RR doesn’t have to be in the forwarding path and needs no VRFs
- RT constraint (RFC 4684): RR sends a PE only routes it will import
- Unique RD per PE: the RR reflects both paths of a multihomed site

IPv6 over MPLS – 6PE and 6VPE

- The core stays IPv4-only (IGP + LDP)
- 6PE (RFC 4798) – global IPv6 over MP-BGP labelled unicast
- 6VPE (RFC 4659) – IPv6 in VRFs, address-family vpnv6
- BGP next hop = IPv4-mapped IPv6 address of the PE loopback
- IPv4 LSP carries the traffic; the BGP label hides IPv6 from P routers

MPLS L2VPN

L2VPN – pseudowires (VPWS)

- L2VPNs are built with Pseudowire (PW) technology
- Any Transport Over MPLS (AToM) is Cisco's implementation of VPWS for IP/MPLS networks
- Inner label – PW label, signalled by targeted LDP (RFC 8077)
- P routers look only at the transport label
- Optional control word after the labels (sequencing, avoids ECMP mis-hashing)

MPLS L2VPN example

- https://cshark.fit.vutbr.cz/captures/5b325952aa76

```
▷ Frame 23: 140 bytes on wire (1120 bits), 140 bytes captured (1120 bits) on interface 0
▷ Ethernet II, Src: aa:bb:cc:00:04:10 (aa:bb:cc:00:04:10), Dst: aa:bb:cc:00:05:10 (aa:bb:cc:00:05:10)
▷ MultiProtocol Label Switching Header, Label: 18, Exp: 0, S: 0, TTL: 254
▷ MultiProtocol Label Switching Header, Label: 20, Exp: 0, S: 1, TTL: 255
▷ PW Ethernet Control Word
▷ Ethernet II, Src: aa:bb:cc:00:01:00 (aa:bb:cc:00:01:00), Dst: aa:bb:cc:00:07:00 (aa:bb:cc:00:07:00)
▷ Internet Protocol Version 4, Src: 192.168.1.1, Dst: 192.168.1.2
▷ Internet Control Message Protocol
```

MPLS VPLS

VPLS

- L2 connectivity achieved across L3 network
- Whole L2 frame can be encapsulated using MPLS
- L2 interconnectivity between geographical different buildings

VPLS ①

- CE routers see the connection as L2 switch
- VPLS needs to achieve the same properties
  - Ethernet forwarding
  - Broadcasting ethernet frames with unknown ethernet MAC address
  - Broadcast/multicast replication
  - Loop prevention (split horizon, no STP across the core)
  - Dynamic MAC learning

VPLS ②

- Connection between PE routers – **pseudowire**
- CE router sends Ethernet broadcast frame to PE router
  - PE floods the frame to all local VPLS ports and all pseudowires (except the one it came from)
- PE routers must form full-meshed network
  - Split horizon: a frame from a PW is never sent to another PW → hence the full mesh

VPLS ③

- Two labels are used for packet forwarding
- Outer – transport (LSP) label
- Inner – pseudowire label

- Full mesh with LDP signalling (RFC 4762) = configuration on every PE
  - PW signalling uses targeted LDP (unicast), not multicast discovery
- BGP auto-discovery / signalling (RFC 4761, 6074) removes the manual mesh
- H-VPLS reduces the mesh in large networks

EVPN

- EVPN (RFC 7432) – current best practice for L2 services, replaces VPLS
- MAC addresses learnt in the control plane, advertised in MP-BGP (route type 2)
- All-active multihoming of CEs (Ethernet Segment), per-flow load balancing
- ARP/ND suppression, less flooding, fast MAC mobility
- Runs over MPLS (or VXLAN in data centres); point-to-point variant: EVPN-VPWS (RFC 8214)

Segment Routing and TE

Segment Routing (SR-MPLS)

- Source routing: ingress puts the path into the packet as a label stack
- Labels (SIDs) are advertised by IS-IS / OSPF – no LDP, no RSVP-TE
  - RFC 8402, 8665 (OSPF), 8667 (IS-IS)
- Prefix SID – global index per node loopback; label = SRGB base + index (16000 + 2 = 16002)
- Adjacency SID – local label for one link
- Same MPLS data plane: push / swap / pop, PHP
- Default design for new SP backbones; SRv6 = the same idea over IPv6

Traffic engineering

- Goal: send traffic off the IGP shortest path (latency, bandwidth, disjoint paths)
- RSVP-TE – path signalled hop by hop, state on every router, bandwidth reservation
- SR-TE – path = label stack at the ingress, no per-LSP state in the core
  - SR policies, computed locally or by a PCE controller
- Steering traffic into TE paths: by destination, by BGP colour, by class

Fast reroute

- Goal: restore traffic in < 50 ms after a failure, before the IGP converges
- RSVP-TE FRR – pre-signalled backup tunnels (facility backup)
- LFA / remote LFA – loop-free alternate next hop for IP and LDP
- TI-LFA (Segment Routing) – 100 % coverage, backup = post-convergence path
- Combine with BFD for fast failure detection

Operations

MTU and TTL

- Each label adds 4 bytes
  - L3VPN: +8 B, L2VPN with control word: +12 B (+ Ethernet header), more with TE/FRR
  - Core MTU must be larger than the customer MTU (e.g. 9000+ in the core)
- TTL propagation: traceroute shows P routers by default
  - no mpls ip propagate-ttl hides them
- RFC 4950: ICMP time exceeded carries the label stack → labels visible in traceroute
- P routers send ICMP along the LSP to the egress (they don’t know customer routes)

Operational best practices

- LSP ping / traceroute (RFC 8029): ping mpls ipv4 x.x.x.x/32
- LDP–IGP synchronization: IGP avoids links where LDP is not up yet
- LDP session protection: keep bindings during a link flap
- BFD for fast failure detection
- /32 loopbacks, never summarized; LDP router-ID pinned to the loopback
- Verify: show mpls forwarding-table, show mpls ldp neighbor, show bgp vpnv4 unicast all labels