Tutorial: Introduction to MPLS

Joseph M. Soricelli (jms@juniper.net)

NANOG 28, Salt Lake City, Utah

Caveats and Assumptions

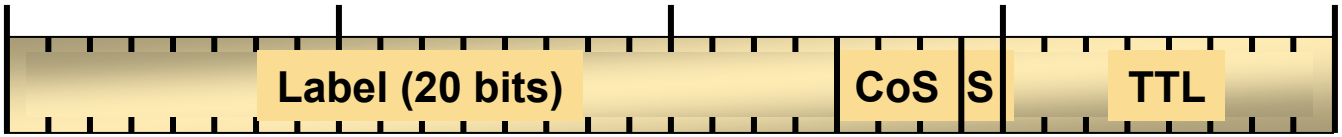
- **The views presented here are those of the author and they do not necessarily represent the views of Juniper Networks**
- **You will ask a question when you don't understand!**

What is MPLS?

- Forwarding of user data traffic using fixed sized headers which contain a label value
- Virtual Circuit for IP
  - Unidirectional path through the network
  - Tunnel through the network
- Traffic Engineering
  - Using paths other than the IGP shortest-path
- Mapping IP prefixes to LSPs
- Forwarding Equivalence Class (FEC)

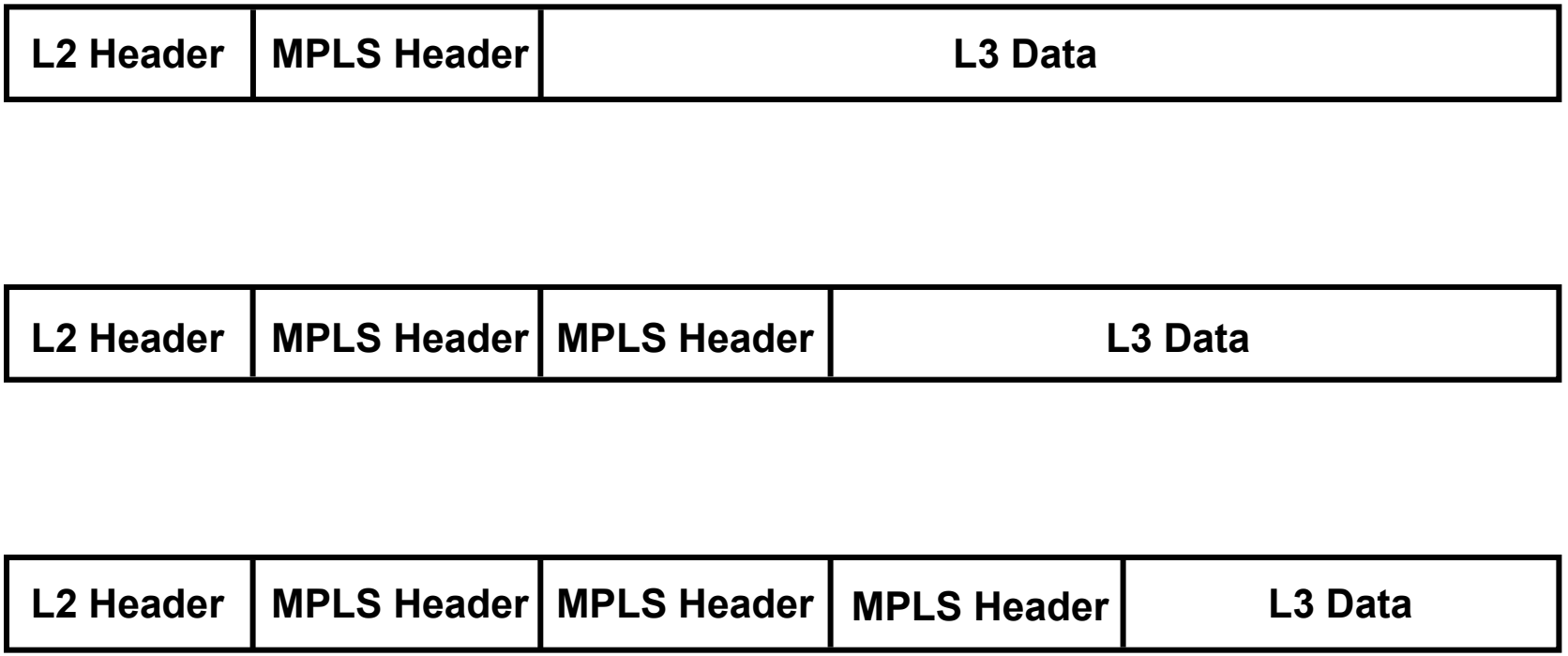
MPLS Labels

- Fixed Length
- Local Significance
- Labels usually change on each network segment
- Assigned upstream by signaling protocols
- Four defined fields
  - Label
  - Experimental
  - Stack Bit (0=additional labels 1=end of stack)
  - Time-to-Live



MPLS Shim Header

- Labels are placed between L2 header and L3 data
- Multiple labels may be stacked together



Label Space

- 20 bits of label allows for values between 0 and 1,048,575
- Labels 0 through 15 are reserved by IETF
  - Label 0 – IPv4 Explicit NULL
  - Label 1 – Router Alert
  - Label 2 – IPv6 Explicit NULL
  - Label 3 – IPv4 Implicit NULL
- All other labels may be allocated at random
  - Some vendors allocate dynamic labels from certain ranges
  - The JUNOS software begins at 100,000

MPLS Labels

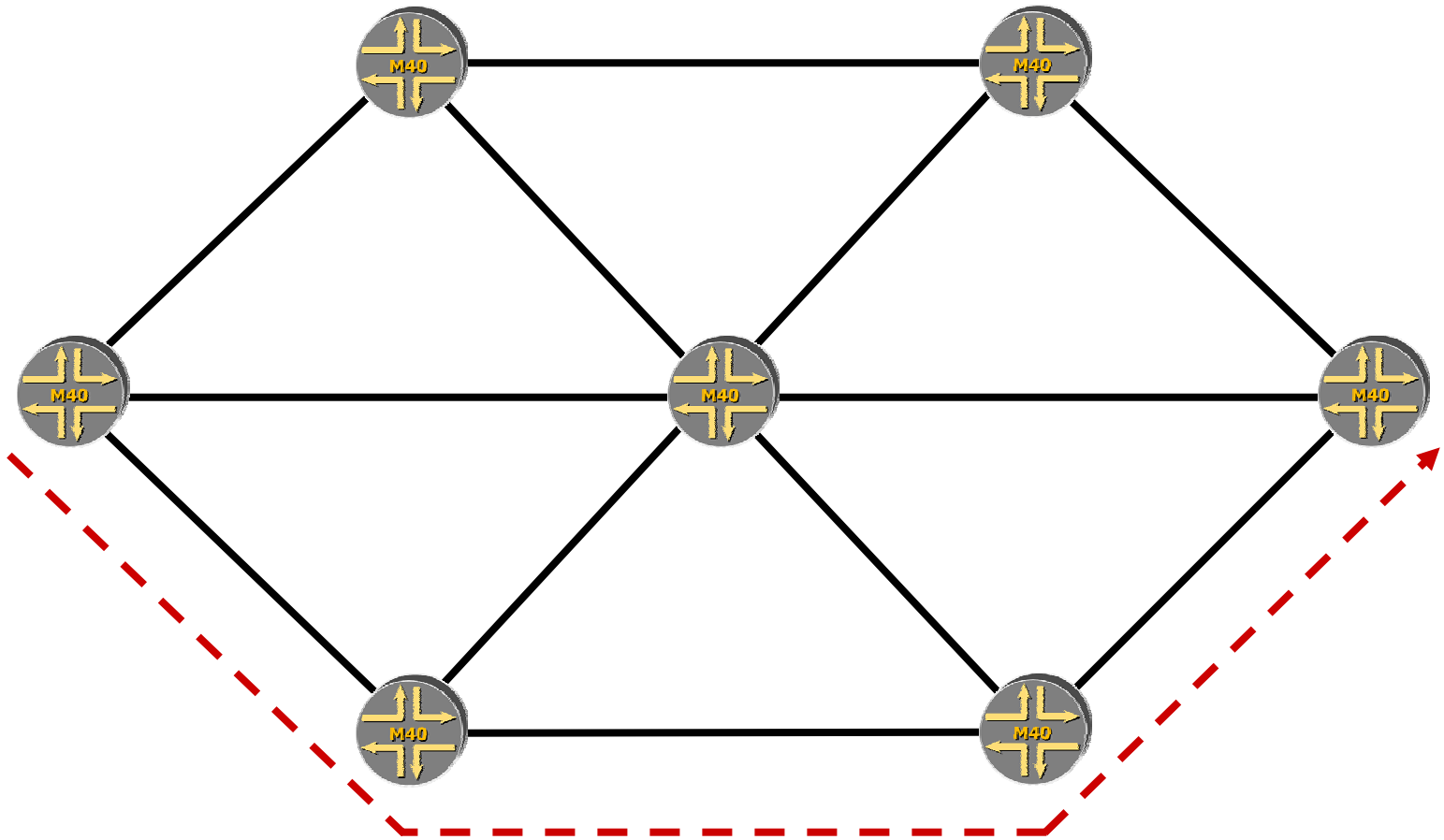
- **Label Distribution**
  - **Downstream-on-Demand**
    - **"Ask and you shall receive"**
  - **Unsolicited Downstream**
    - **Sent without a request**
    - **"Here's a label to use for this prefix"**
- **Label Retention**
  - **Liberal (Keep all received labels)**
  - **Conservative (Keep only labels you use)**
- **Label Control**
  - **Ordered**
    - **Allocate a label after receiving a label or if you are egress**
  - **Independent**
    - **Allocate a label at any time**

Link-Layer Support for MPLS

- PPP protocol ID value of 0x0281
- PPP NCP ID value of 0x8281
- All other Layer 2 encapsulations use 0x8847
  - Ethernet
  - HDLC
  - GRE Tunnel
  - Frame Relay
  - ATM AAL5 SNAP

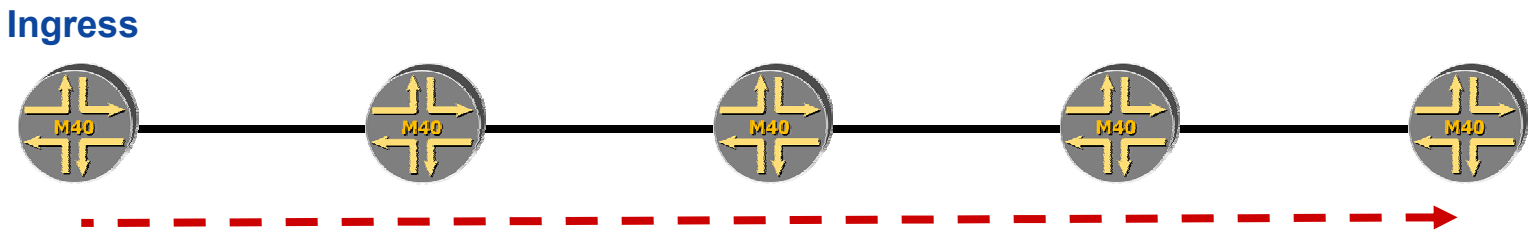
Label Switched Path (LSP)

- Unidirectional path through the network



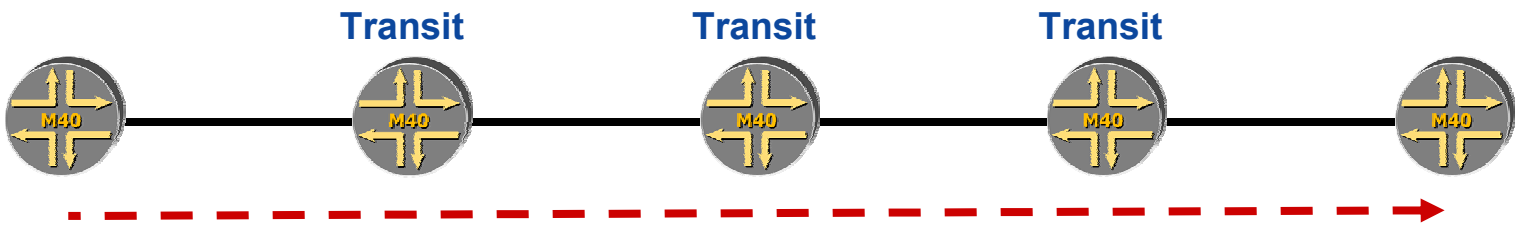
Router Types - Ingress

- **Ingress Router**
  - **Packets enter the LSP**
  - **Head-end router**
  - **Upstream from other routers in the LSP**
  - **"Gozinta" router**
  - **Performs a label push operation**



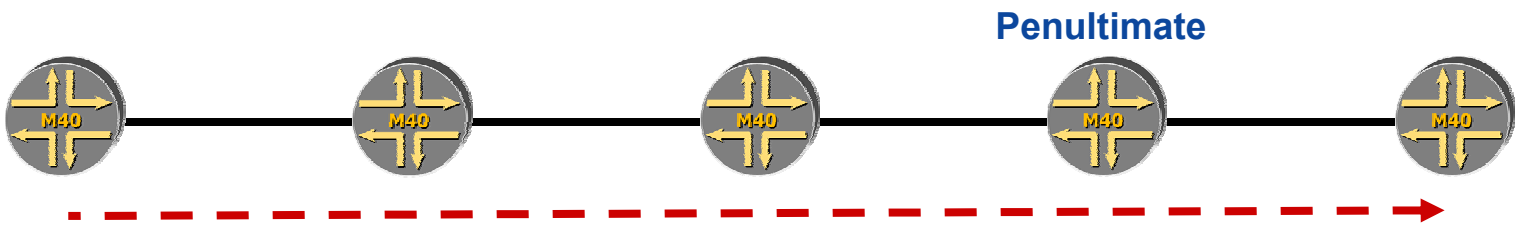
Router Types - Transit

- **Transit Router**
  - **Zero or more transit routers in an LSP**
    - **Maximum of 253**
  - **Sends traffic to the downstream physical next-hop of the LSP**
  - **Performs a label swap operation**



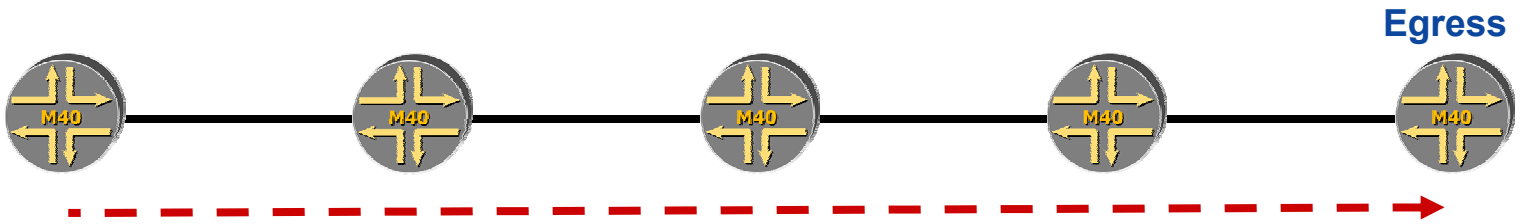
Router Types - Penultimate

- **Penultimate Router**
  - **Immediate upstream router from the egress router**
  - **Often performs a label pop operation**
    - **Penultimate Hop Popping (PHP)**
    - **Remaining packet contents sent to the egress router**
  - **Can perform a label swap operation**



Router Types - Egress

- **Egress Router**
  - **Packets exit the LSP**
  - **Tail-end router**
  - **Downstream from other routers in the LSP**
  - **"Gozoutta" router**
  - **Can perform a label pop operation**



Ultimate Hop Popping (UHP)

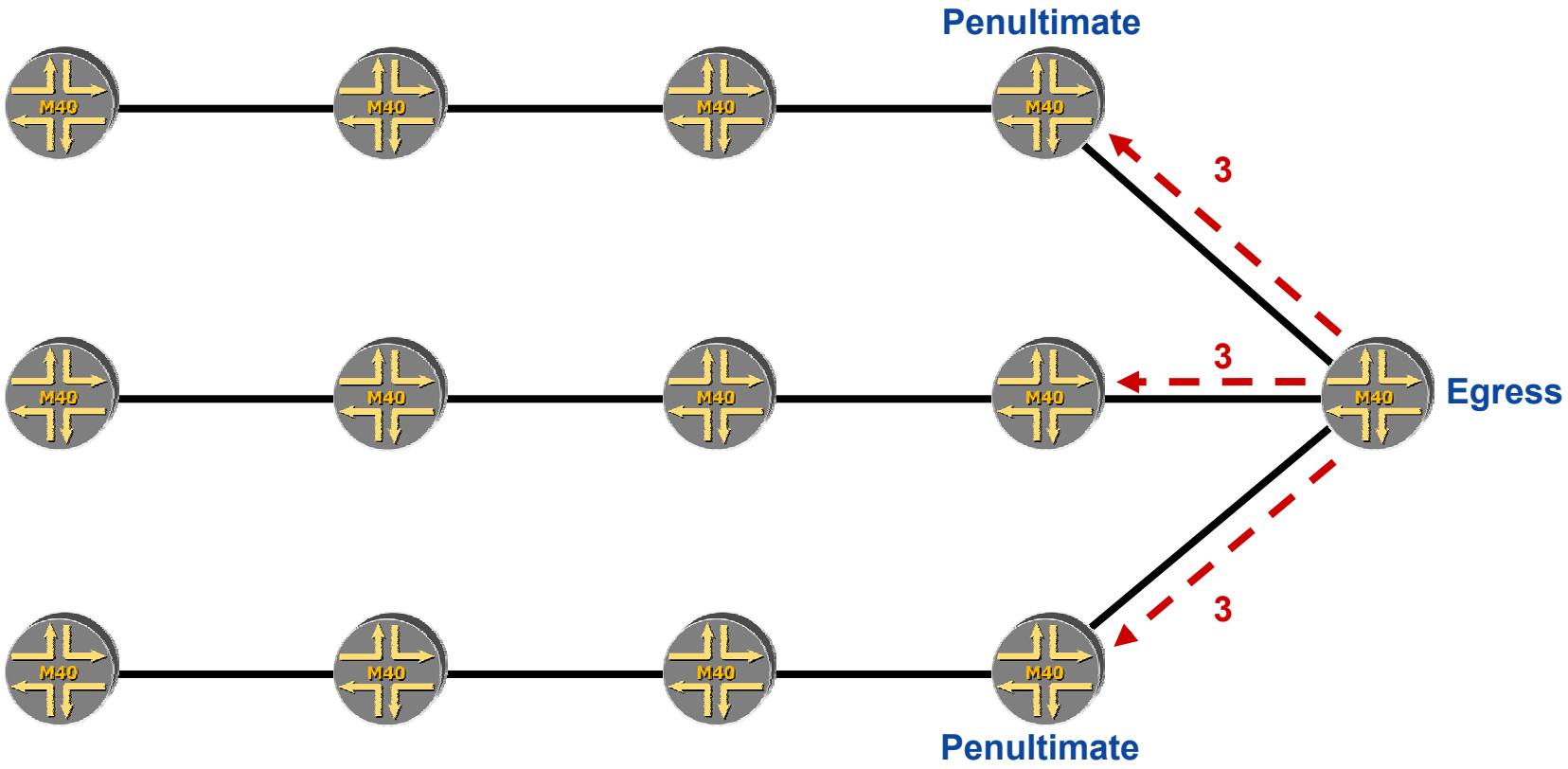
- The egress router signals a label value of 0 to the penultimate router
- The packet sent to the egress contains an MPLS header with a label value of 0
- Egress router pops the label and performs an IPv4 route lookup before forwarding the packet

Penultimate Hop Popping (PHP)

- The egress router signals a label value of 3 to the penultimate router
- The penultimate router pops the top label from the packet the forwards the remaining data to the egress router
  - Native IPv4 packet
  - MPLS header when label stacking is used
- Egress router performs an appropriate lookup
  - Route lookup for IPv4 packets
  - MPLS switching table lookup for labeled packets

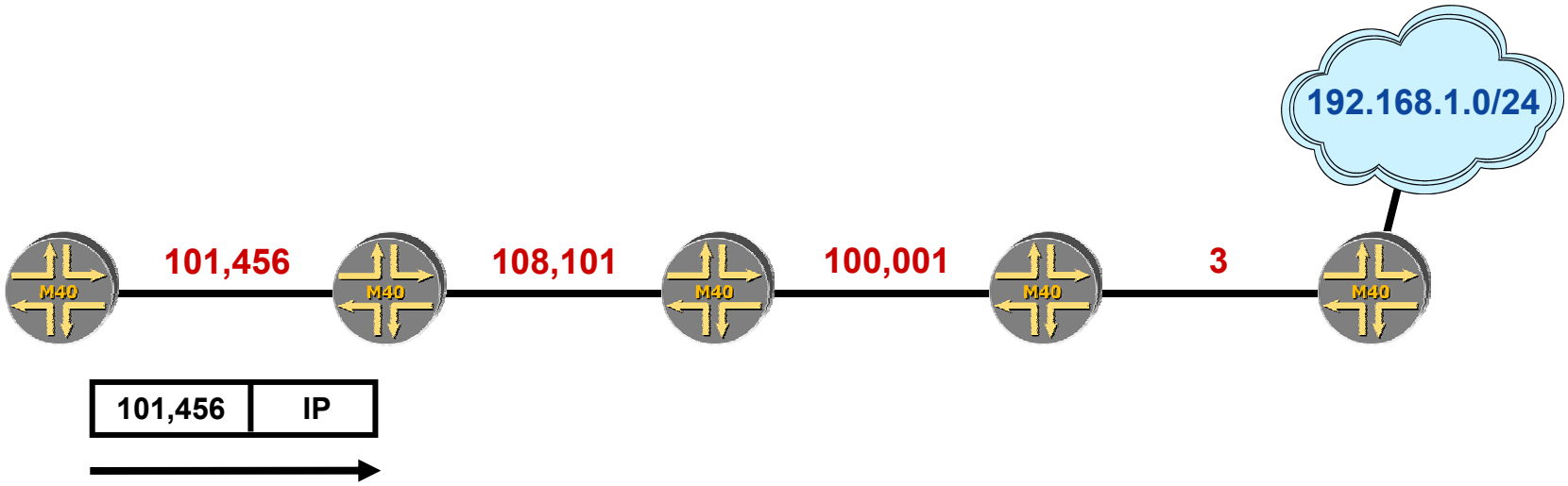
Penultimate Hop Popping (PHP)

- **Helps the egress router offload processing**
  - **Very beneficial for non-ASIC devices**



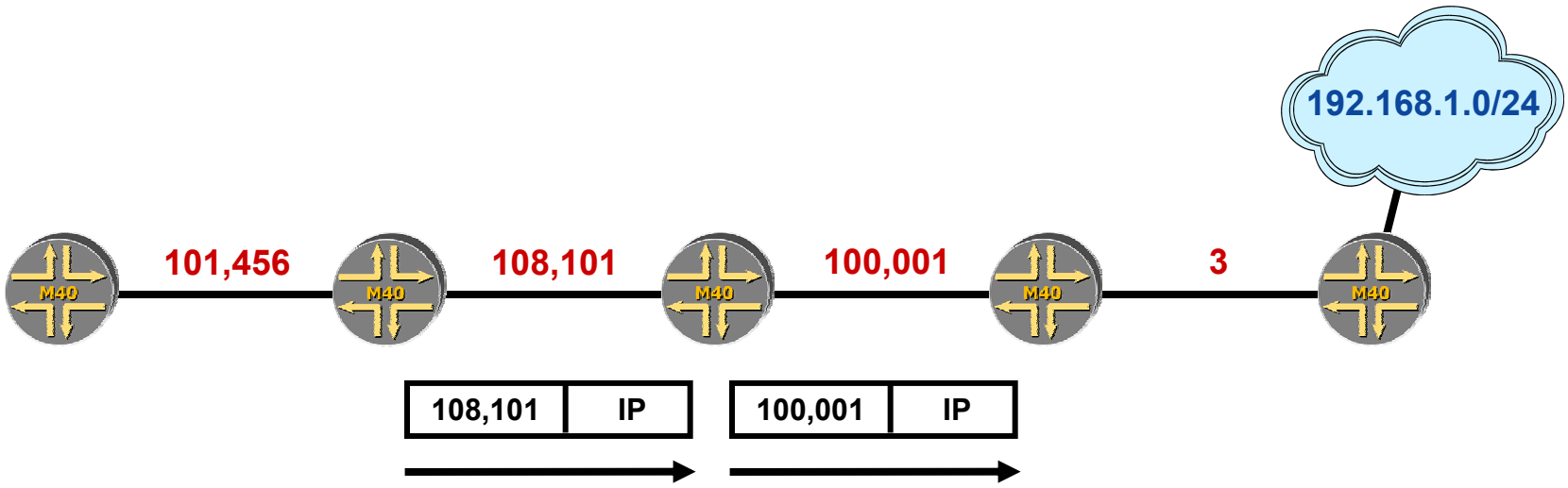
LSP Forwarding - Ingress

- An IPv4 packet arrives at the ingress router with a destination address of 192.168.1.1
- Ingress router has a route for 192.168.1.0/24 with a next-hop of the LSP
- MPLS header with a label of 101,456 is appended to the packet and forwarded downstream



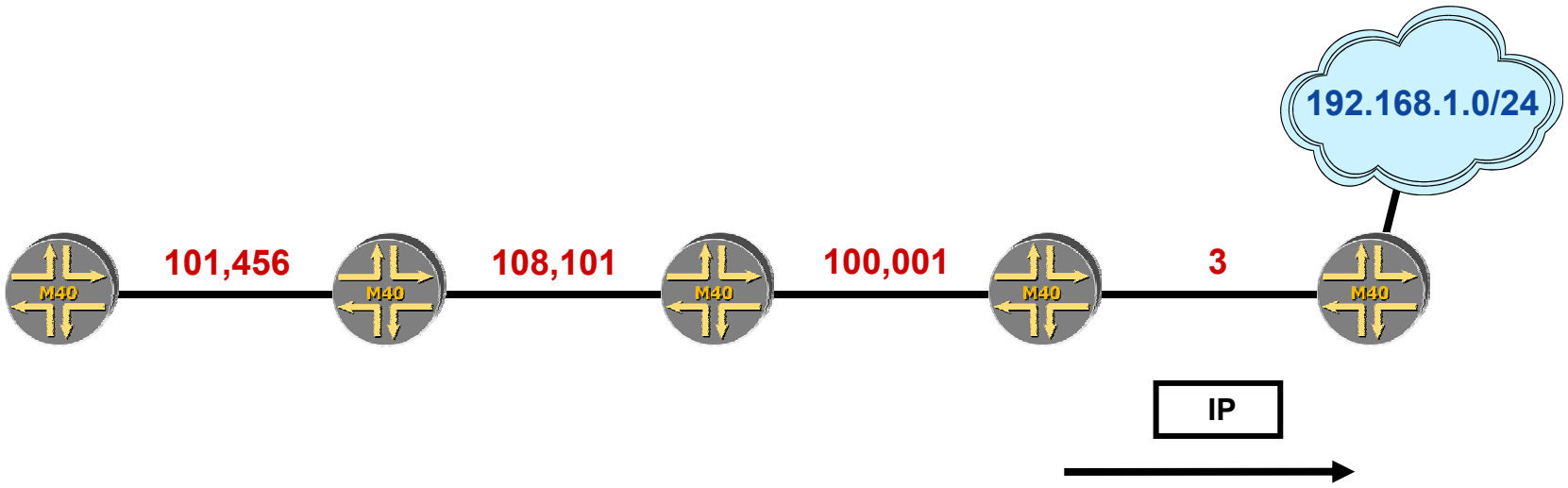
LSP Forwarding - Transit

- Each transit router receives a labeled packet and performs a switching table lookup
- Each transit router performs a label swap operation
  - 101,456 swapped for 108,101
  - 108,101 swapped for 100,001



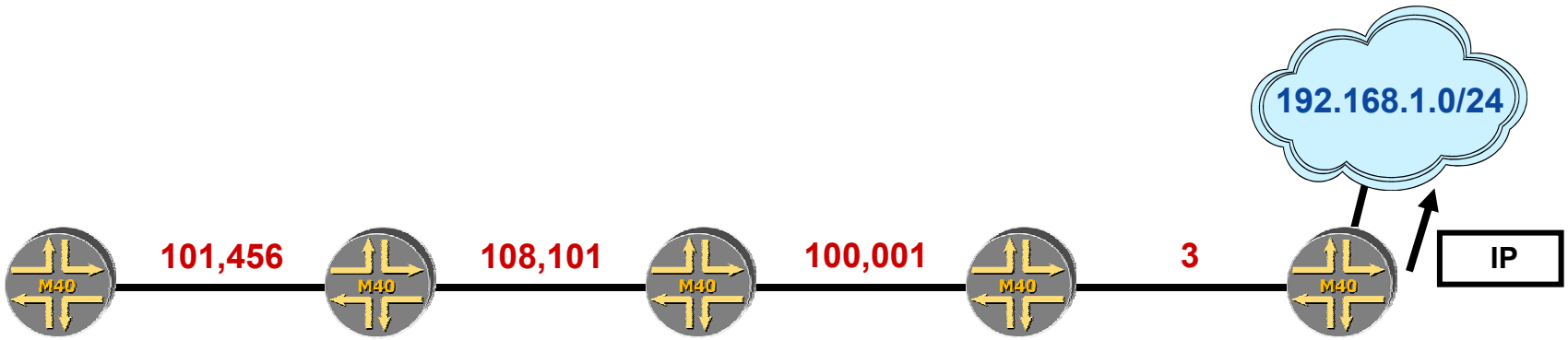
LSP Forwarding - Penultimate

- The penultimate router receives a labeled packet and performs a switching table lookup
- Since the egress router signaled a label value of 3, the penultimate router pops the top label and forwards the remaining data



LSP Forwarding - Egress

- The egress router receives a native IPv4 packet
  - Route lookup performed
  - Packet forwarded to the appropriate next-hop router



LSP Signaling Protocols

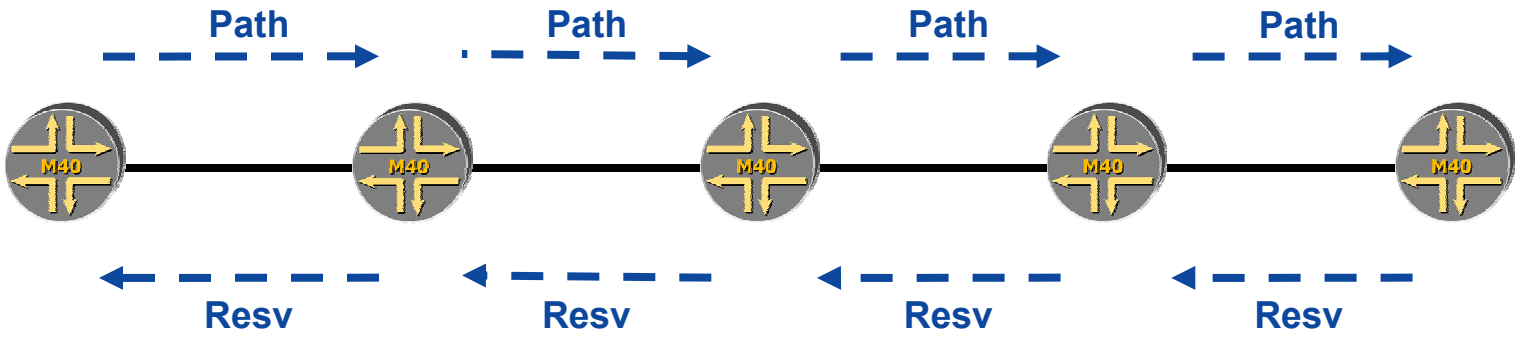
- **Resource Reservation Protocol (RSVP)**
  - Well-known signaling protocol
  - Extended to support traffic engineering
  - Supports explicit paths and bandwidth reservations
  - Labels allocated only along the defined LSP path
- **Label Distribution Protocol (LDP)**
  - Uses the same shortest-path as IGP for forwarding
  - Labels allocated and exchanged between neighbors
- **Constrained Routing LDP (CR-LDP)**
  - Adds traffic engineering capabilities to LDP
  - Limited support from vendors

RSVP Session

- Uniquely defines and identifies the LSP throughout the network
  - Destination address of LSP
  - Tunnel ID value
  - Protocol number (often set to 0)
- An individual session may have multiple defined senders
  - LSP ID defines a sender
  - Ingress router creating an additional path for the LSP
    - Secondary path
    - Fast Reroute Detour
- Routers become RSVP neighbors after session establishment

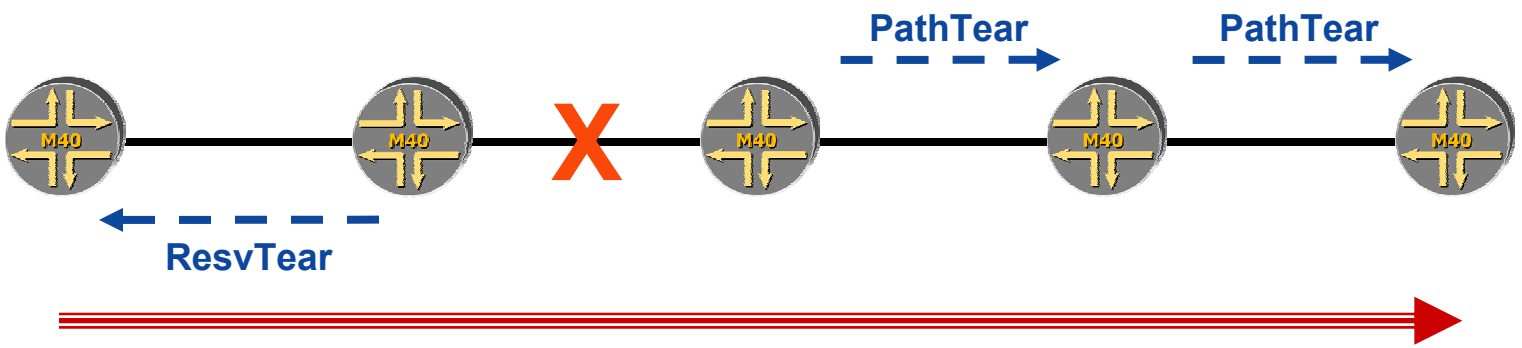
RSVP Path and Resv Messages

- **Path messages sent downstream**
  - **Addressed to the egress router**
  - **Contains the router alert option**
  - **Establishes protocol state along the way**
- **Resv messages sent upstream**
  - **Addressed to the next upstream node**
  - **Finalizes protocol state**
  - **Assigns and allocates resources**



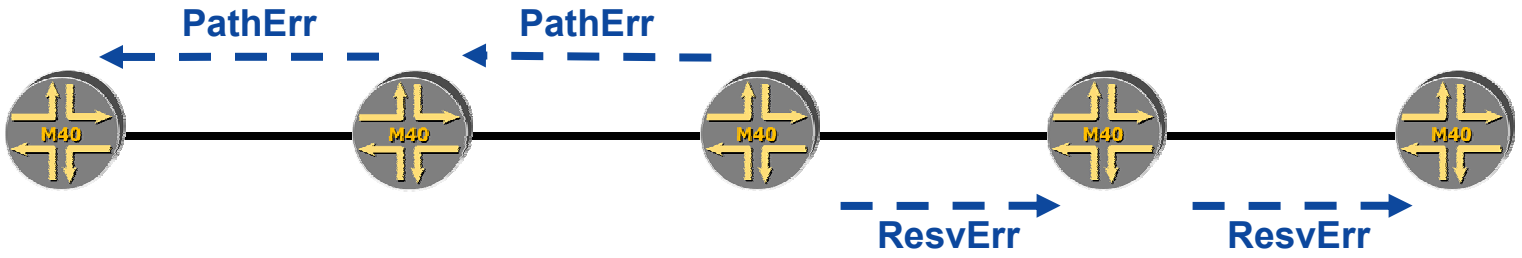
RSVP PathTear and ResvTear Messages

- LSP already established and operational
  - Link failure causes protocol state to be removed
- PathTear messages sent downstream
  - Addressed to the egress router
  - Contains the router alert option
  - Removes protocol state along the way
- ResvTear messages sent upstream
  - Addressed to the next upstream node
  - Removes protocol state



RSVP PathErr and ResvErr Messages

- Error messages signal problems to the ingress or egress routers
  - No protocol state removed by error messages
  - Ingress or egress routers may initiate a teardown of the LSP due to receipt of an error message
- PathErr messages sent upstream
- ResvErr messages sent downstream



RSVP Path Message Objects

- **Objects used to define the LSP and request resources**
  - **Session**
    - **Defines the address of the egress router**
    - **Contains the Tunnel ID value associated with the LSP**
  - **RSVP-Hop**
    - **Interface address of the previous hop of the LSP**
  - **Sender-Template**
    - **Defines the address of the ingress router**
    - **Contains a unique LSP ID**
  - **Sender-Tspec**
    - **Displays any request bandwidth reservations**
  - **Session Attribute**
    - **Contains LSP priority values as well as the ASCII name of the LSP**

RSVP Path Message Objects

- **Objects used to define the LSP and request resources**
  - **Label Request**
  - **Explicit Route**
    - **Defines the requested path of the LSP through the network**
    - **Can be manually created**
    - **Can be the output of Constrained SPF calculation**
  - **Record Route**
    - **Contains the actual path of the LSP through the network**
    - **Used for loop detection and prevention**

RSVP Resv Message Objects

- **Objects used to allocate resources and establish the LSP**
  - **Session**
    - **Contains the egress address and Tunnel ID**
  - **RSVP-Hop**
    - **Interface address of the downstream hop of the LSP**
  - **Style**
    - **Type of resource allocation performed**
    - **Fixed Filter (FF)**
    - **Shared Explicit (SE)**
  - **FlowSpec**
    - **Displays the bandwidth reserved by the LSP**
    - **Matches the information in the Sender-Tspec object**

RSVP Resv Message Objects

- **Objects used to define the LSP and request resources**
  - **Filter-Spec**
    - **Contains the ingress address and LSP ID**
    - **Matches the information in the Sender-Template object**
  - **Label**
    - **Contains the 20-bit label value to be used for traffic forwarding**
  - **Record Route**
    - **Contains the actual path of the LSP through the network**
    - **Used for loop detection and prevention**

RSVP Bandwidth

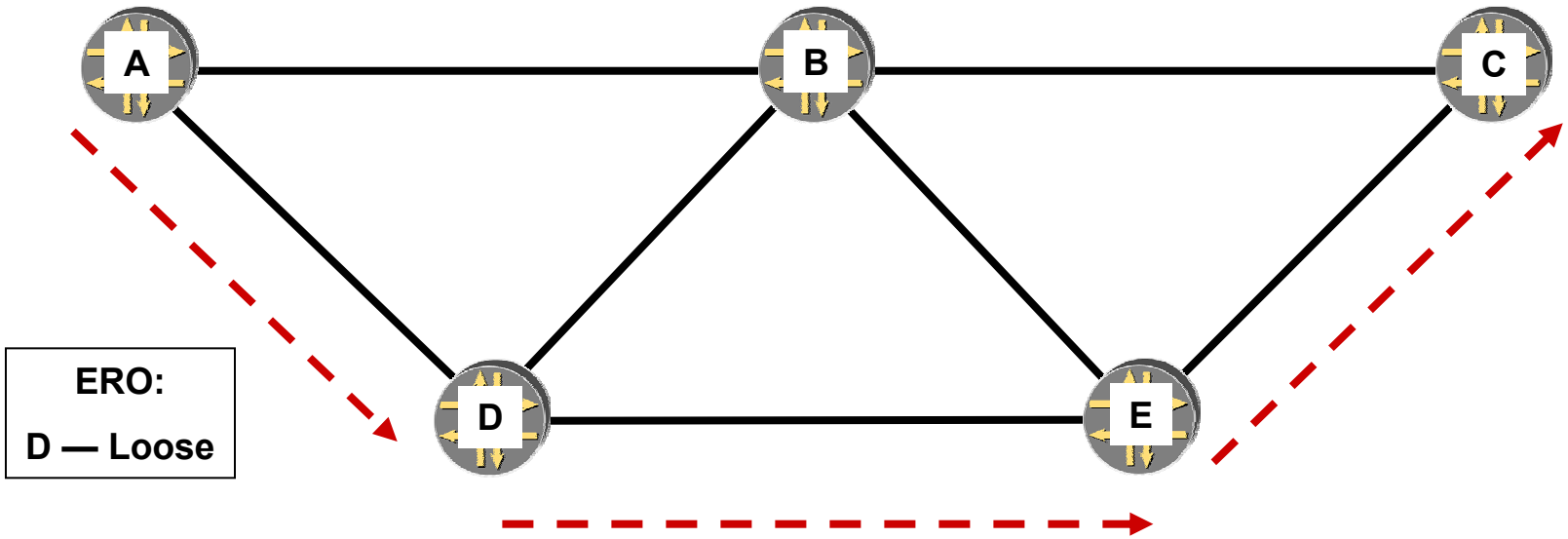
- By default, each interface uses 100% of its capacity as reservable bandwidth
  - You may change this percentage
- An LSP may request a bandwidth reservation during its establishment in the network
  - Only determines if the LSP is setup of not
  - The BW reservation is *NOT* used to police traffic

Explicit Route Object (ERO)

- Puts the "engineering" in TE
  - Allows the ingress router to define the path of the LSP through the network
- May contain loose hop information
  - Loose hop defines a node the LSP must pass through at some point
  - IPv4 shortest-path routing used for forwarding Path messages:
    - From ingress to first loose hop
    - Between loose hops
    - From last loose hop to egress router
- May contain strict hop information
  - Strict hop must be the next downstream router
  - Must be directly connected to the local router

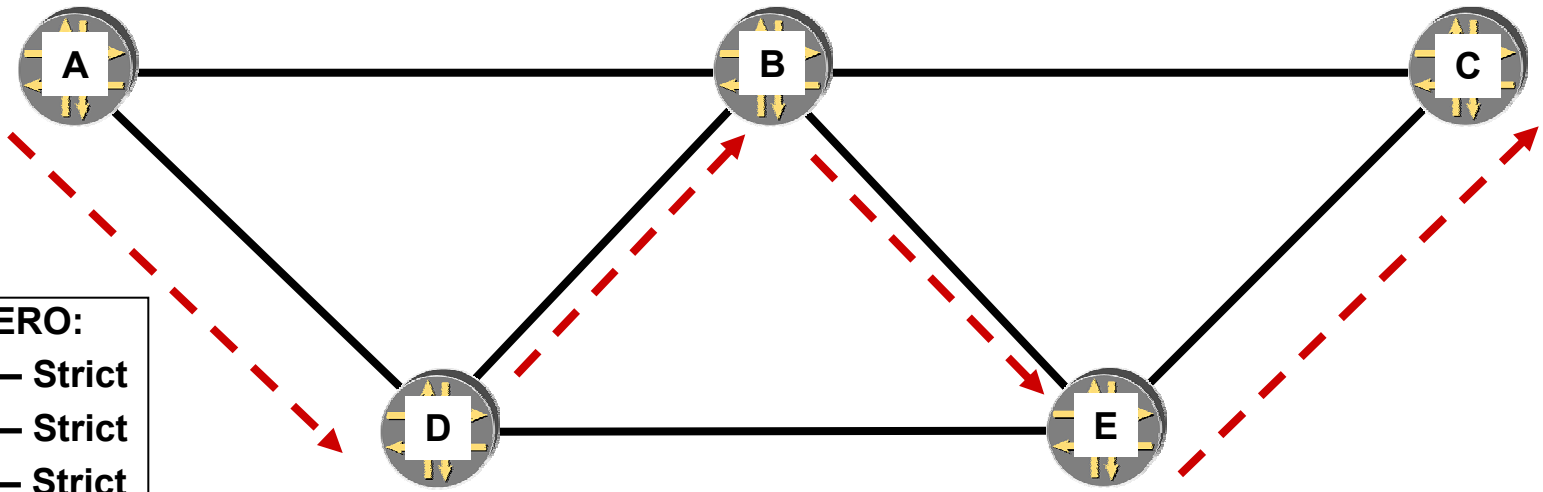
Loose Hop ERO

- Ingress uses routing table to forward the Path message towards RTR-D
- RTR-D uses routing table to forward message to egress router



Strict Hop ERO

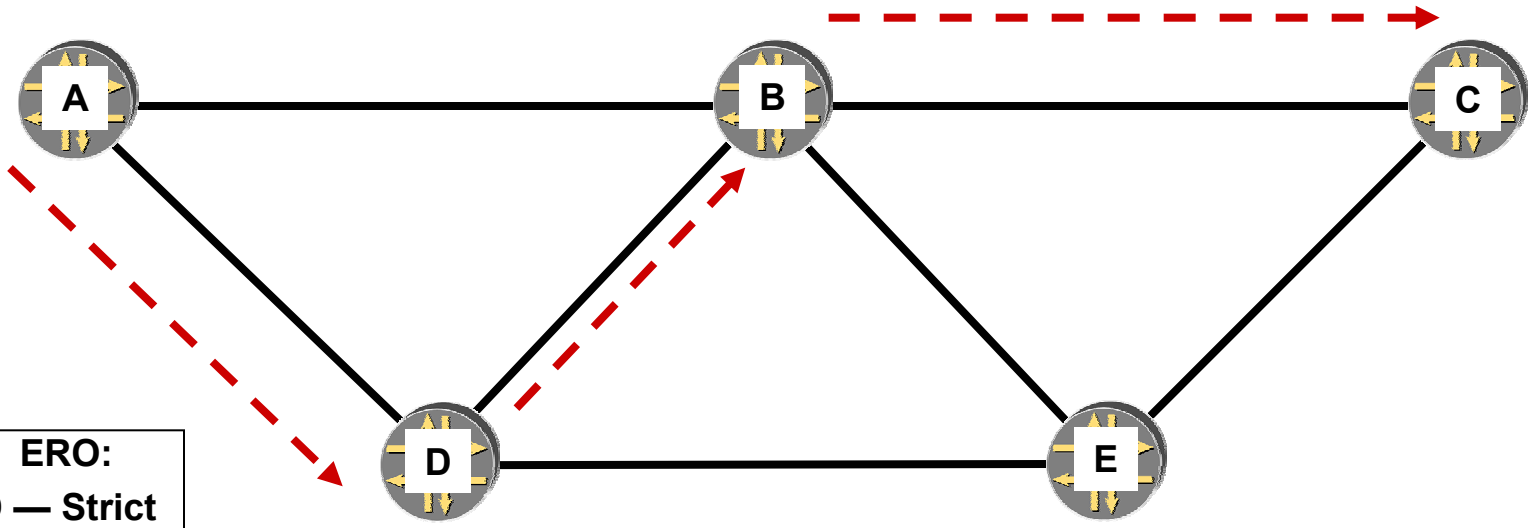
- Ingress consults ERO to locate the first strict hop
  - Forwards Path message out interface associated with that hop
- Each transit router ensure it is next strict hop in the path
  - If so, message forwarded to next strict hop
  - If not, PathErr message sent back to ingress router



ERO:
D — Strict
B — Strict
E — Strict
C — Strict

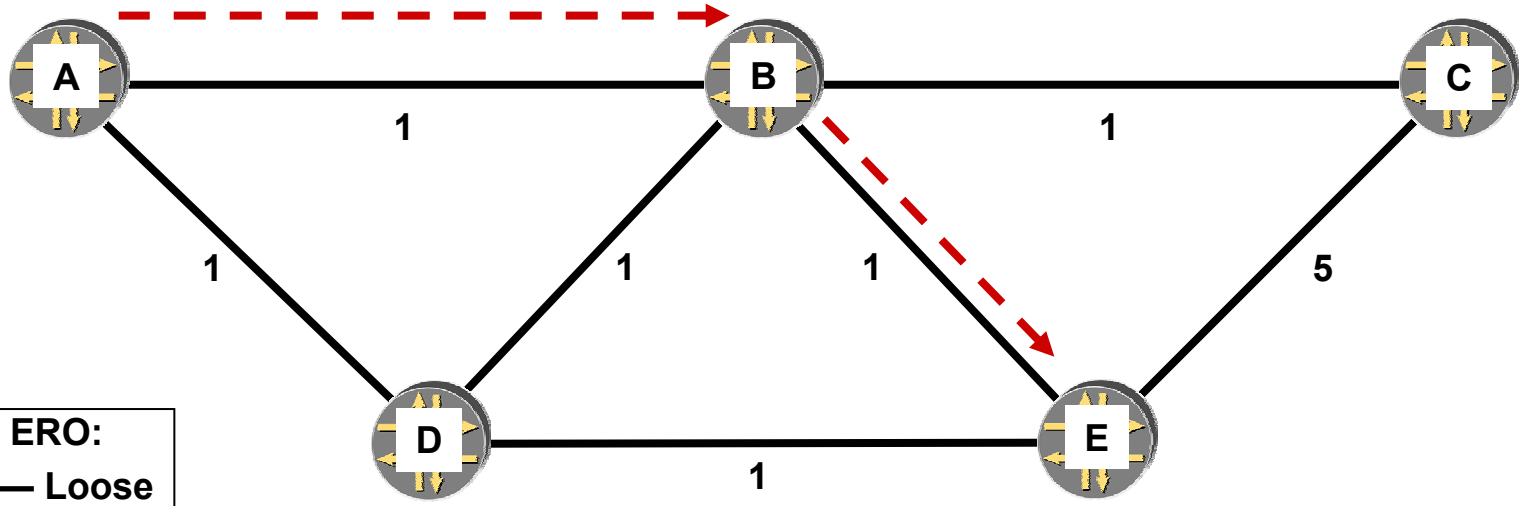
Mixing Strict and Loose Hops

- An ERO may contain both strict and loose hops
  - Loose hops are routed using routing table
  - Strict hops receive messages when they are directly connected



Manual ERO Concerns

- Manual use of EROs requires knowledge of the network topology
- Loop detection might cause LSP setup to fail
- IGP metrics shown below
  - RTR-E forwards Path message back to RTR-B
  - Loop detection in RRO by RTR-B prompts creation of PathErr message



Automatic EROs

- The ingress router can automatically create an ERO for the LSP
  - Contains all strict hops for the complete path
  - Formed from information contained in the Traffic Engineering Database (TED)
- TED is populated by information advertised by the Interior Gateway Protocols
  - Both OSPF and IS-IS have been extended to support traffic engineering
  - OSPF Opaque LSA 10 (Area-Scope Flooding)
  - IS-IS TLV 22 (Extended IS Reachability)
  - IS-IS TLV 135 (Extended IP Reachability)

IGP Extensions

- **Both link-state IGPs may advertise information which is stored in the TED**
  - **Interface and neighbor interface addresses**
  - **Maximum reservable bandwidth per network link**
  - **Current reservable bandwidth**
  - **Traffic Engineering metric**
  - **Administrative group information**
    - **Affinity classes**
    - **Colors**
- **The ingress router uses a modified form of the SPF algorithm within the TED to generate the ERO**
  - **Constrained Shortest-Path First (CSPF)**
  - **Takes user-defined constraints into account**

CSPF Algorithm

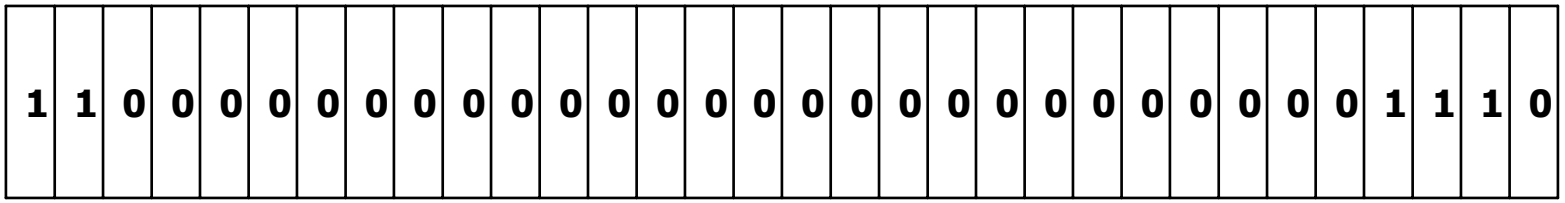
- When the ingress router invokes the CSPF algorithm, it creates a subset of the TED information based on the constraints provided
  1. Prune all links which don't have enough reservable BW
  2. Prune all links which don't contain an included administrative group color
  3. Prune all links which do contain an excluded administrative group color
  4. Calculate a shortest path from the ingress to egress using the subset of information
     - Manual ERO definitions taken into account
     - Run CSPF from ingress to first ERO node
     - Run second CSPF from ERO node to egress

CSPF Algorithm (Contd.)

- When the subset of information used by CSPF returns multiple equal-cost paths:
  5. Prefer the path where the last-hop address equals the egress address
  6. Should equal-cost paths still exist, select the one with the fewest physical hops
  7. Should equal-cost paths still exist, pick one based on the load-balancing configuration of the LSP
     - Random
     - Most-fill
     - Least-fill
- The result of CSPF (strict-hop ERO) is passed to RSVP for LSP signaling

Administrative Groups

- Each interface may be assigned to one or multiple administrative groups
  - Colors are often used to describe these groups (Gold, Silver, Bronze)
  - User-friendly names can be used as well (Voice, Management, Best-Effort)
  - Names are locally significant to the router
- Group information is propagated by the IGP as a 32-bit vector
  - Bits 0 through 31
  - Stored in the TED

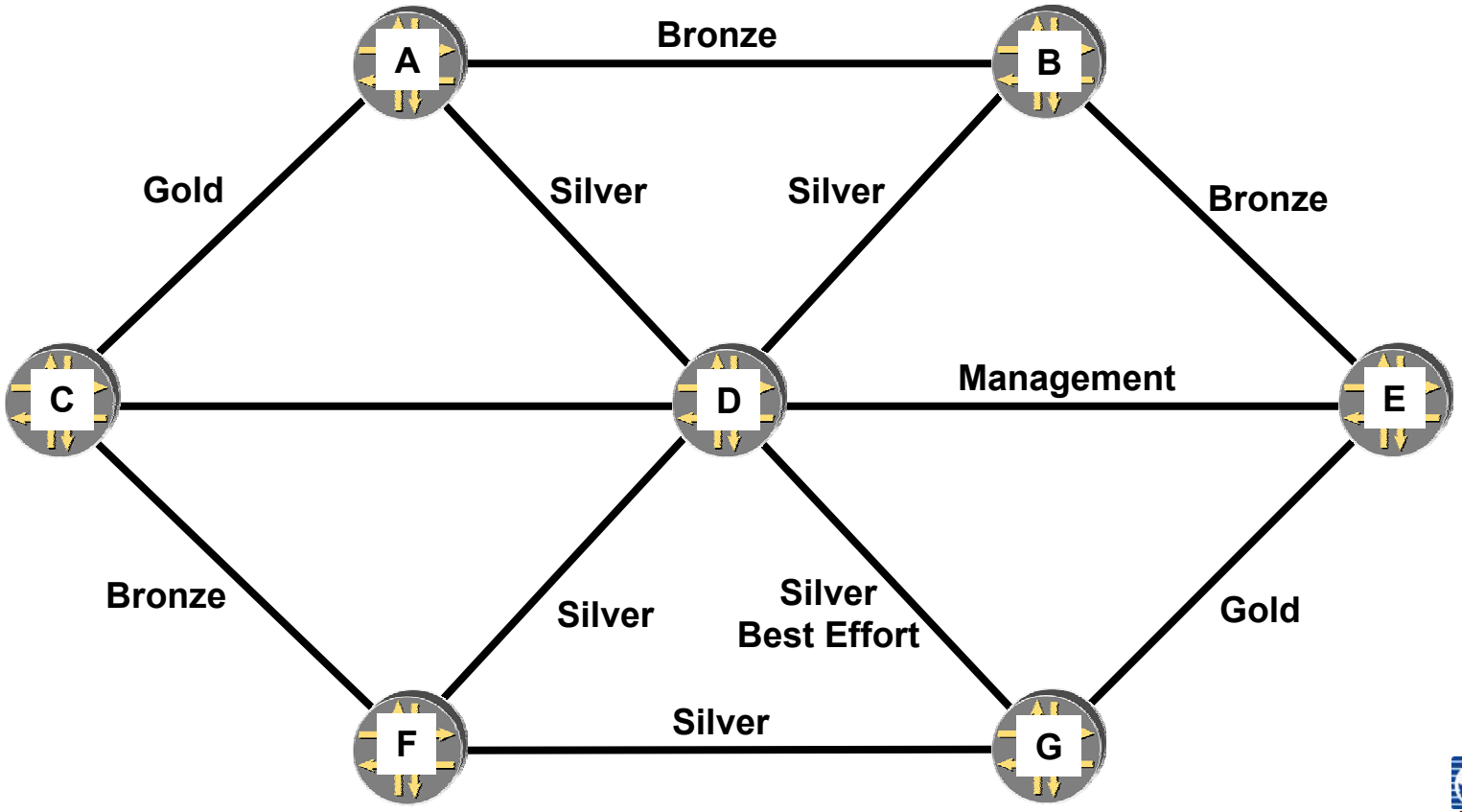


Administrative Groups

- The LSP can be configured to include or exclude certain group values
  - Include requires each link to contain the specified group value
    - Multiple values are combined as a logical OR
  - Exclude requires each link to not contain the value
    - Multiple values are combined as a logical OR
- LSP performs a logical AND on the include and exclude groups

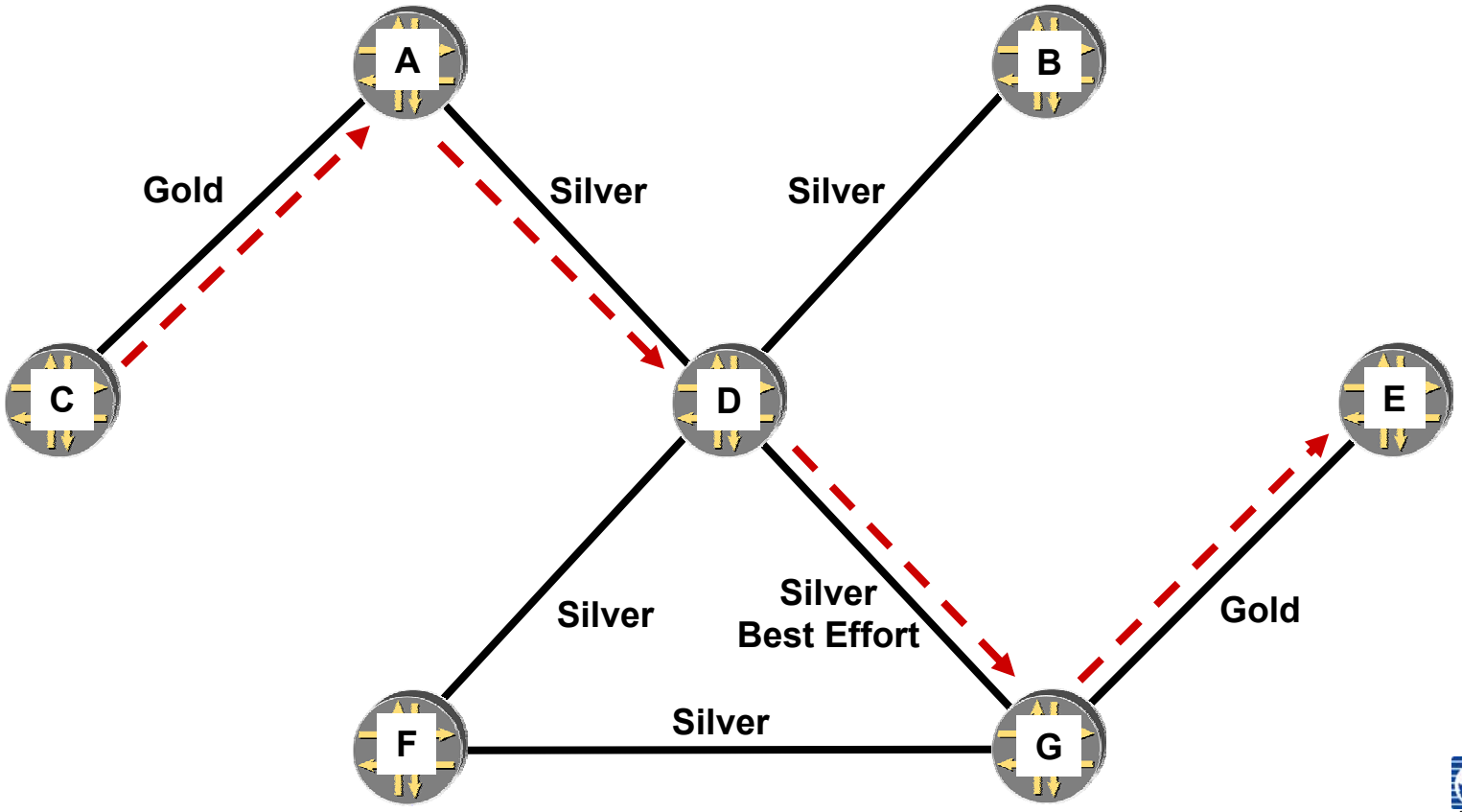
Administrative Groups Example-Include

- The LSP can be configured to include certain group values
  - LSP from C to E should include Gold or Silver
  - All IGP link metrics are set to 1



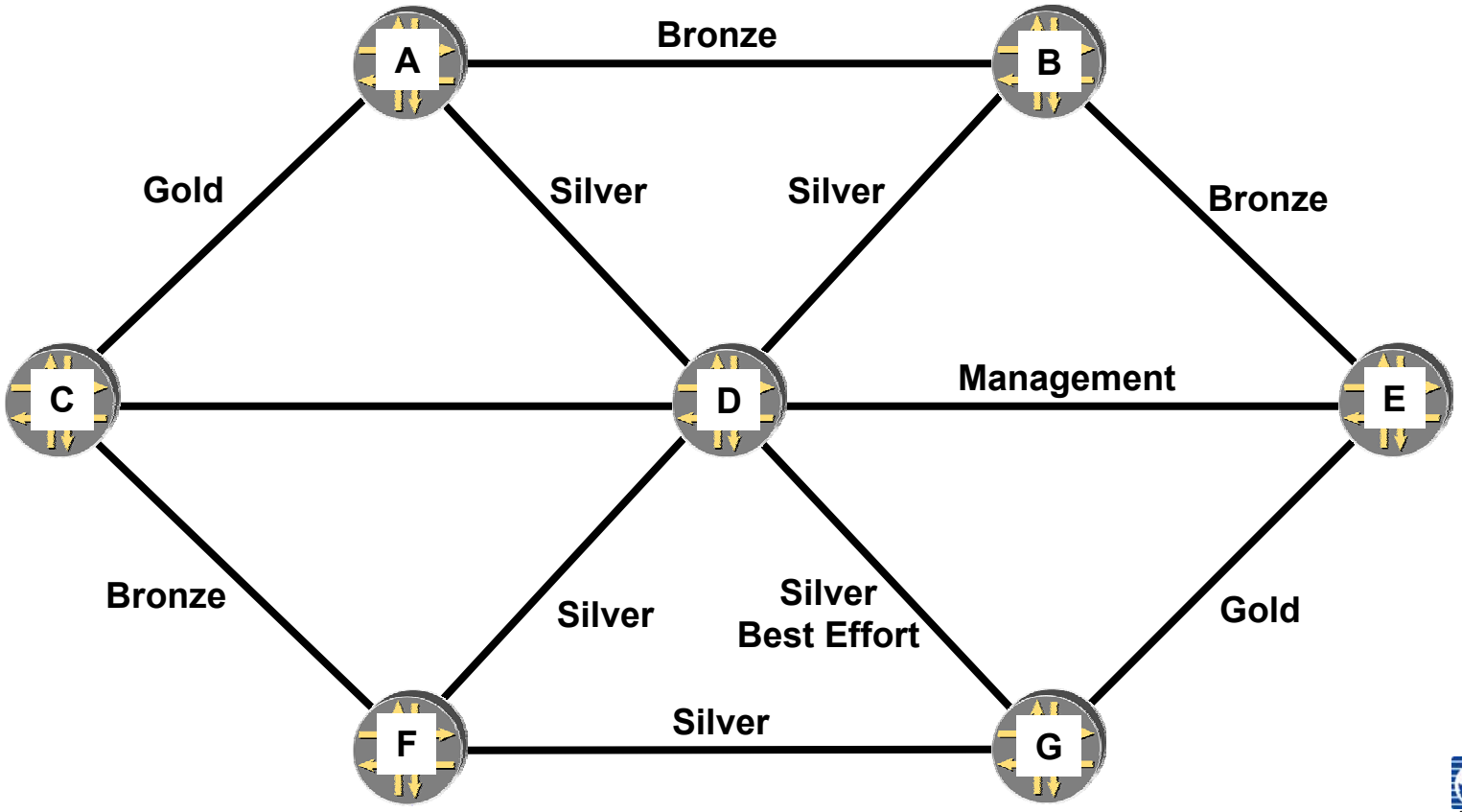
Administrative Groups Example-Include

- After pruning all links which do not contain either Gold or Silver, SPF is run and the ERO is formed



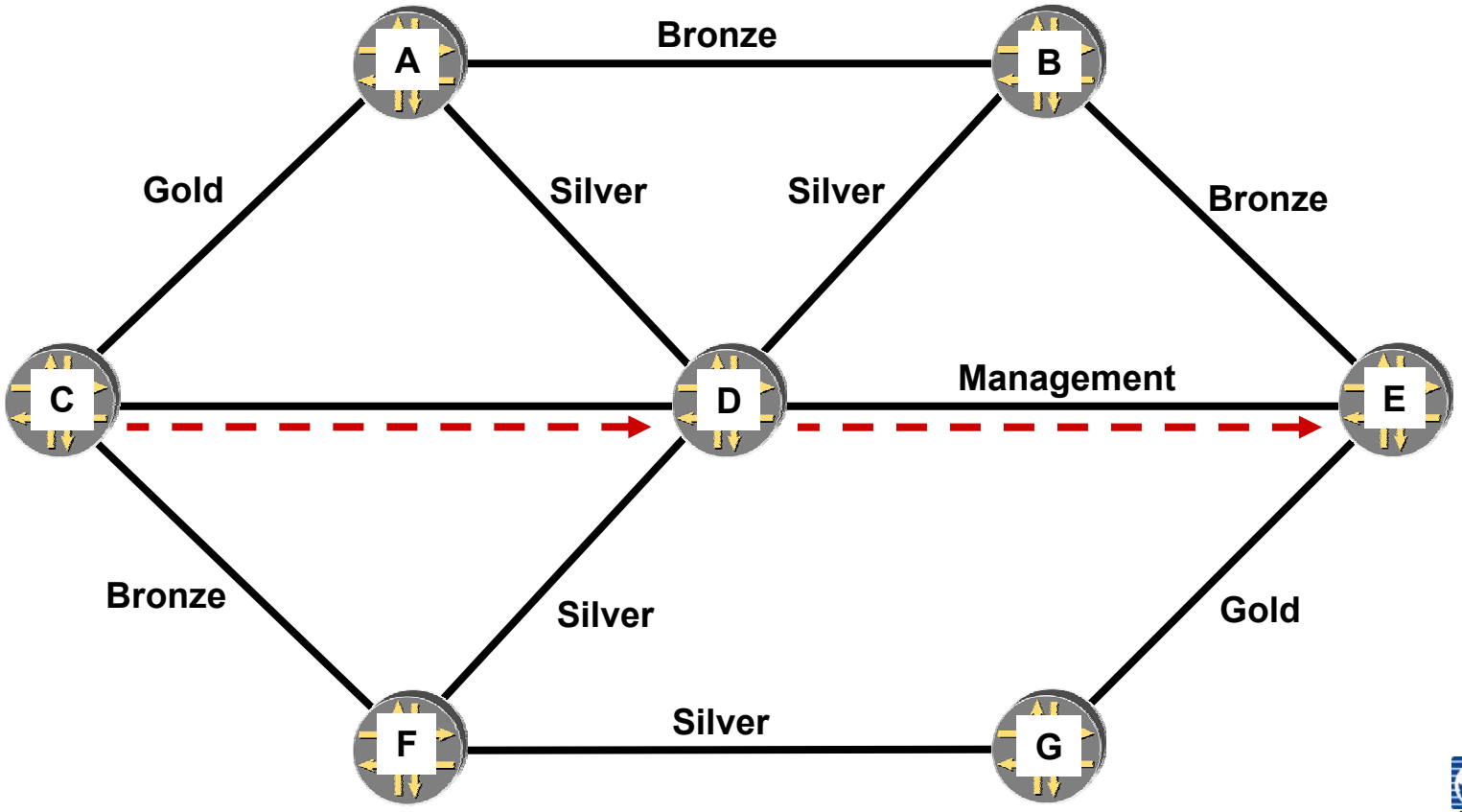
Administrative Groups Example-Exclude

- The LSP can be configured to exclude certain group values
  - LSP from C to E should exclude Best Effort
  - All IGP link metrics are set to 1



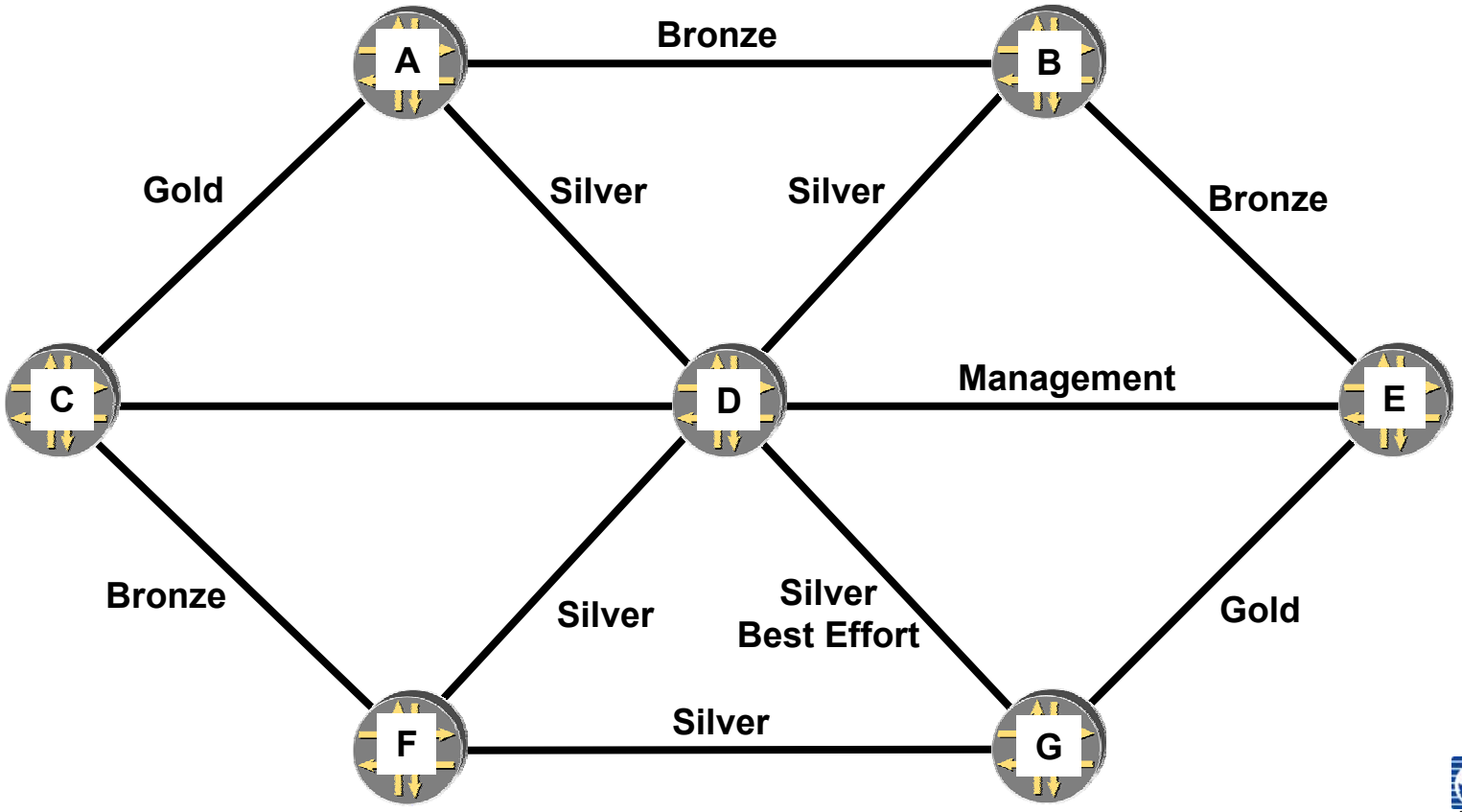
Administrative Groups Example-Exclude

- After pruning all links which contain Best Effort, SPF is run and the ERO is formed



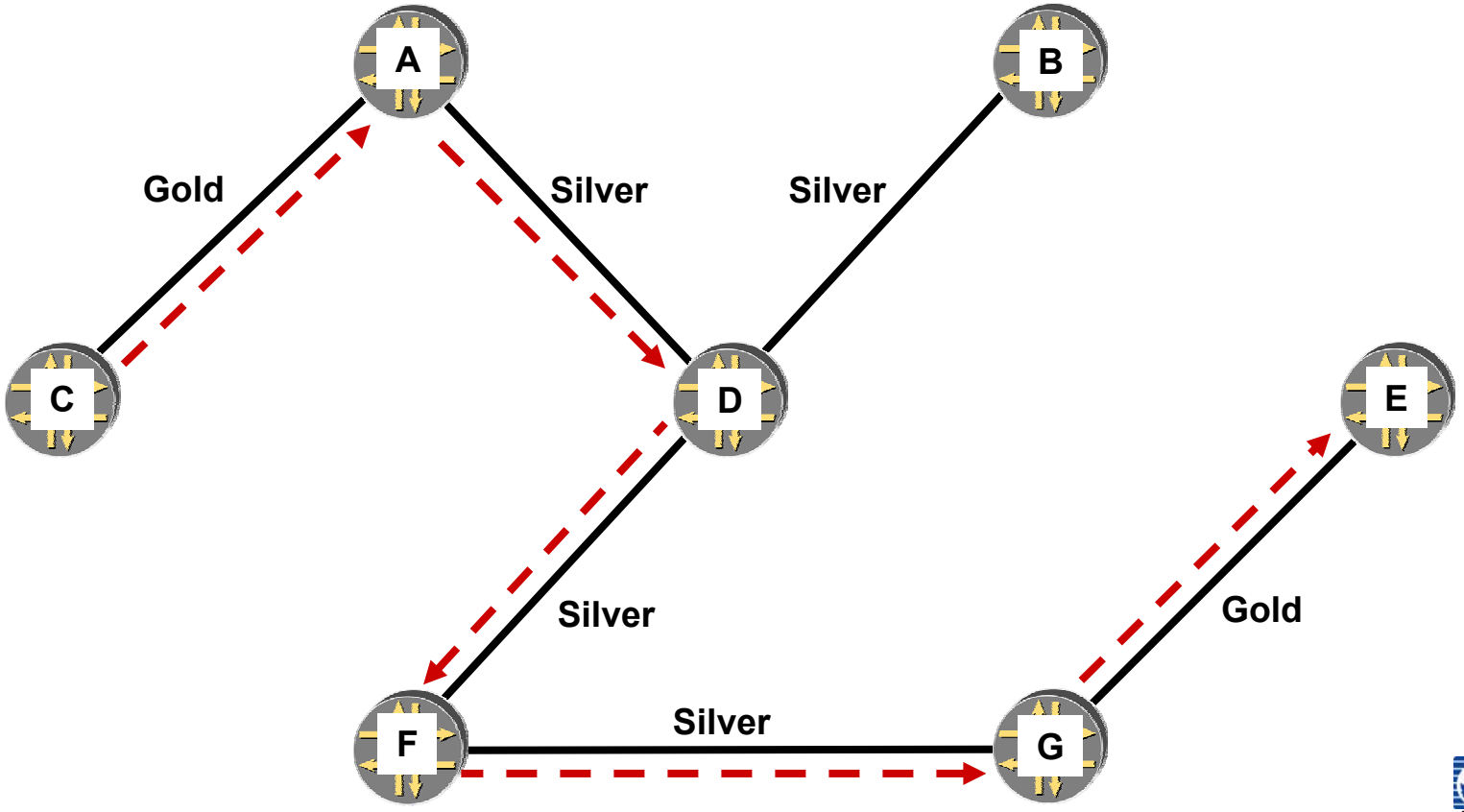
Administrative Groups Example-Both

- **Both include and exclude can be used together**
  - **LSP from C to E should (include Gold or Silver) and (exclude Best Effort)**
  - **All IGP link metrics are set to 1**



Administrative Groups Example-Both

- First, all links which do not contain either Gold or Silver are pruned
- Second, all links containing Best Effort are pruned
- Third, SPF is run and the ERO is formed



RSVP Configuration

- **Routers must be enabled to support MPLS and RSVP**

```
[edit]
user@host# show interfaces
ge-0/2/0 {
    unit 0 {
        family inet {
            address 10.222.29.1/24;
        }
        family mpls;
    }
}
[edit]
user@host# show protocols
rsvp {
    interface all;
}
mpls {
    interface all;
}
```

```
mpls traffic-eng tunnels
!
interface POS0/0/0
 ip address 10.222.28.1 255.255.255.0
 no ip directed-broadcast
 mpls traffic-eng tunnels
```

RSVP Configuration

- **Explicit networks paths can be established**

```
[edit]
user@host# show protocols mpls
path user-defined-ERO {                ip explicit-path name user-defined-ERO enable
    192.168.20.1 loose;                 next-address loose 192.168.20.1
    192.168.24.1 loose;                 next-address loose 192.168.24.1
    192.168.36.1 loose;                 next-address loose 192.168.36.1
    192.168.40.1 loose;                 next-address loose 192.168.40.1
}
```

RSVP Configuration

- **Configure the ingress router to support the LSP**

```
[edit]
user@host# show protocols mpls
label-switched-path ingress-to-egress {
    to 192.168.32.1;
    primary user-defined-ERO;
}
path user-defined-ERO {
    192.168.20.1 loose;
    192.168.24.1 loose;
    192.168.36.1 loose;
    192.168.40.1 loose;
}
```

```
interface Tunnel10
 ip unnumbered Loopback 0
 no ip directed-broadcast
 tunnel destination 192.168.32.1
 tunnel mode mpls traffic-eng
 tunnel mpls traffic-eng priority 1 1
 tunnel mpls traffic-eng path-option 1 explicit name user-defined-ERO
```

LDP Neighbors

- Two connected LDP routers form a neighbor relationship
  - Hello messages multicast to 224.0.0.2
  - UDP port 646
  - Included transport address or source IP address of message is used as session identifier
- Two remote LDP routers can also become neighbors
  - Targeted hello messages
  - Unicast to the remote neighbor
  - UDP port 646
  - Transport address included in message
- Label space advertised as part of LDP ID

LDP ID

- Each router creates an LDP ID
  - 6-byte value separated by a colon
    - 192.168.1.1:0
  - First 4 bytes are the router ID of the node
  - Last two bytes are used to define the type of labels allocated
    - A value of 0, the default, means labels are handed out on a per-node basis

LDP Sessions

- After becoming neighbors, each LDP router decides which router should be the active node
  - Highest router ID is the active node
- Active node creates a TCP connection between the routers
  - TCP port 646
  - Once connection between peers even when multiple physical links (multiple neighbors) exist
- Active node then creates an LDP session by sending initialization messages to the passive peer

LDP Session Initialization

- Initialization message includes:
  - Local LDP ID
  - Remote LDP ID
  - Protocol version
  - Negotiable hold time value
- The passive node examines the LDP ID values to verify that an active neighbor relationship is in place for this new session
  - If acceptable, a keepalive message is returned
- Passive node generates its own initialization message and sends it to the active node
  - If accepted, a keepalive message is returned
- The peers now have an operational session between themselves

Advertising Information in LDP

- Two new LDP peers exchange interface addresses in Address messages
  - All directly connected addresses are advertised
  - Allows each peer to associate a label advertisement from the session to a physical next-hop interface
- The peers then advertise FEC and label information
  - Local FEC includes prefixes reachable from the router as an egress router (loopback address)
  - FEC may also include information received from other LDP sessions
  - Labels are allocated for all reachable FEC information

LDP Database

- All LDP routers maintain an LDP database
  - FEC/Labels received across a particular session
  - FEC/Labels advertised across a particular session
- By default, FEC/Label information is flooded to all LDP peers for all possible prefixes
  - Full-mesh of information for the entire LDP network
  - Possible routing loop issues
- Full mesh of LSPs results from complete FEC prefix knowledge
  - Each router is an ingress router for every FEC
  - Each router is also an egress router for the FEC it advertised

LDP Database Example

```
user@host> show ldp database
Input label database, 192.168.16.1:0--192.168.20.1:0
  Label      Prefix
 100352      192.168.16.1/32
      3      192.168.20.1/32
 100368      192.168.24.1/32
 100304      192.168.32.1/32
 100320      192.168.36.1/32
 100336      192.168.40.1/32

Output label database, 192.168.16.1:0--192.168.20.1:0
  Label      Prefix
      3      192.168.16.1/32
 100112      192.168.20.1/32
 100096      192.168.24.1/32
 100128      192.168.32.1/32
 100144      192.168.40.1/32
```

LDP Loop Avoidance

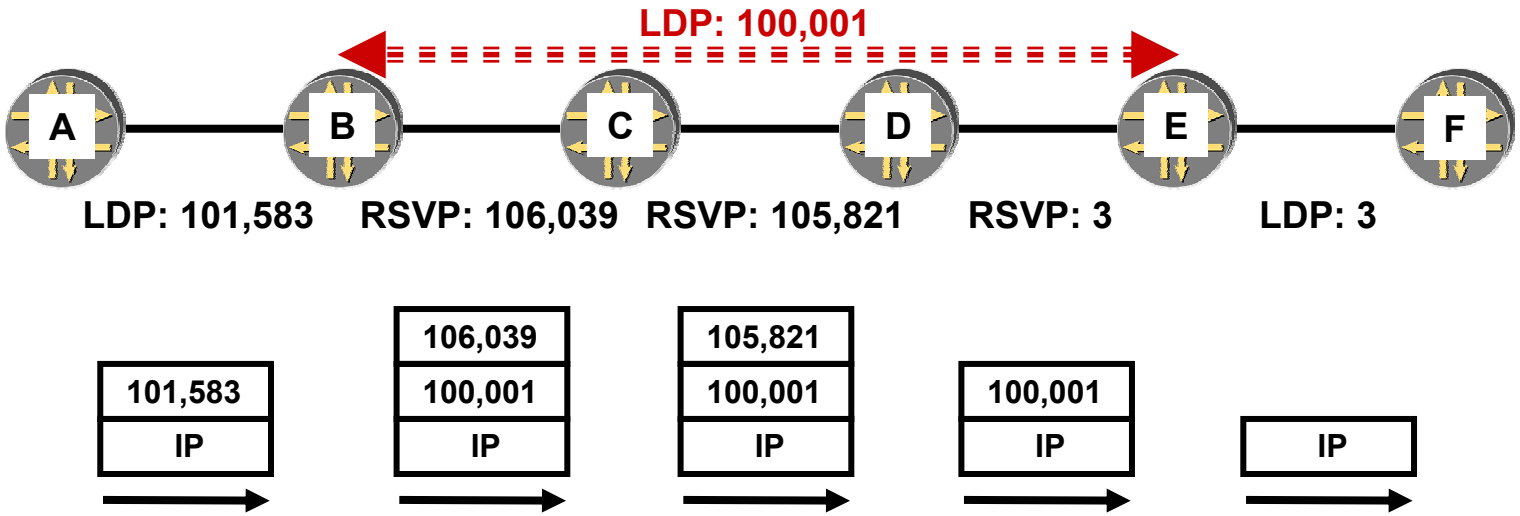
- **LDP consults the IGP shortest-path information to avoid loops**
  - **Results in the forwarding path for LDP and the IGP being identical**
- **For example:**
  - **Suppose that 192.168.1.1 is reachable via IS-IS over the `fe-0/0/0.0` interface**
  - **The LDP database reports that a label has been received from two LDP peers for 192.168.1.1**
    - **One is reachable over the `fe-0/0/0.0` interface**
    - **The other is reachable over the `fe-0/1/1.0` interface**
  - **The LDP process installs the label information associated with the `fe-0/0/0.0` interface**

Using RSVP for LDP Traffic Engineering

- Established RSVP tunnels can be used to engineer LDP forwarding paths through the network
- The RSVP ingress and egress routers form an LDP session using targeted hello messages
  - Once the session is formed, address and FEC information is advertised
  - Advertised and received FEC/Label information is stored in the LDP database
  - FEC/Label information is re-advertised to other LDP peers
- Label stacking is performed across the RSVP LSP

LDP Tunneling and Label Stacking

- LDP neighbor relationships between:
  - A and B via the physical interface
  - B and E via bi-directional RSVP LSPs
  - E and F via the physical interface
- RTR-B performs a swap and push operation
  - Swap label 101,583 for label 100,001 (advertised by E)
  - Push label 106,039 (advertised by C) to reach RTR-E



Questions and Comments

- We've just barely scratched the surface here
  - LSP protection
    - Primary and secondary paths
    - Fast Reroute
  - Policy control and LSP selection
  - LSP Preemption
  - Class of Service
  - Using MPLS to support services (L3VPN, L2VPN, VPLS)
- Feedback on this presentation is highly encouraged
  - jms@juniper.net
- Questions?

Juniper® NETWORKS

Thank you!

http://www.juniper.net