Appendix 3 - MPLS (v6.1)

MPLS

Multi-Protocol Label Switching

Agenda

- Review ATM
- IP over WAN Problems (Traditional Approach)
- MPLS Principles
- Label Distribution Methods
- MPLS Details (Cisco)
- RFCs

ATM Principles

- **ATM**
  - Asynchronous Transfer Mode
  - Based on asynchronous TDM
    - Hence buffering and address information is necessary
    - Variable delay (!)
- **Cell switching technology**
  - Based on store-and-forward of cells
  - Connection-oriented type of service with PVC and SVC
  - But no error recovery (!)
- **ATM cell**
  - Small packet with constant length
  - 53 bytes long (5 bytes header + 48 bytes data)

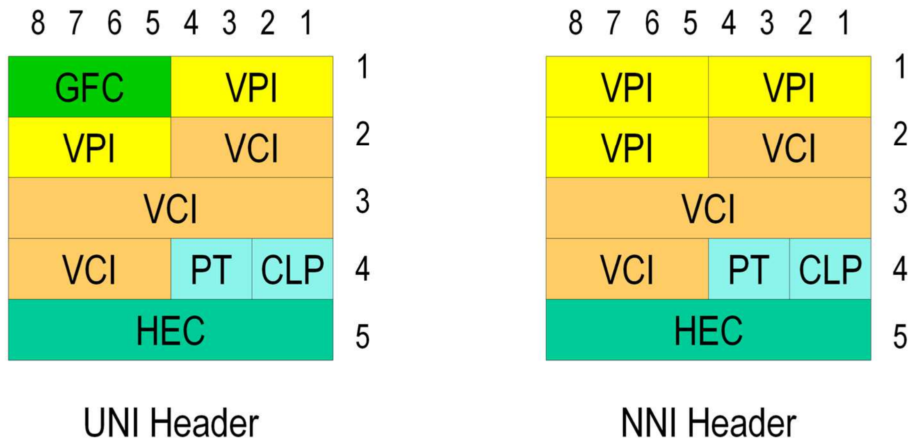
 MPLS v6.1 3

ATM is packet switching in the "Virtual Call" service mode and offers high-speed virtual circuits. Although connection-oriented no error-recovery or flow control is performed in the network itself. It is like in frame relay or IP up to the end-system to take appropriate actions in case reliable transport is necessary.

We call it cell switching because of constant frame length. The reason for cells will be explained soon. ATM is based on statistical multiplexing hence transport of frames will experience a variable delay.

A service provider can offer WAN (Wide Area Network) service (PVC and SVC) although ATM originally was planned to be B-(Broadband)-ISDN. Hence it should be the universal interface for all types of traffic (voice, video, data) and all types of networks (LAN (Local Area Network), MAN (Metropolitan area network) and WAN. In LAN and MAN environment ATM disappeared because of the success of the Ethernet family, allowing nowadays speed up to 10 Gbit/s reaching distances up to hundreds of kilometers. We will learn about that later in the Ethernet chapter.

Cell Format



- **Two slightly different formats**
  - UNI … 8 bits for VPI
  - NNI … 12 bits for VPI

 MPLS v6.1 4

Cell Size 53 byte: 5 byte header and 48 byte payload.

VPI - Virtual Path Identifier / VCI - Virtual Channel Identifier -> local connection identfier.

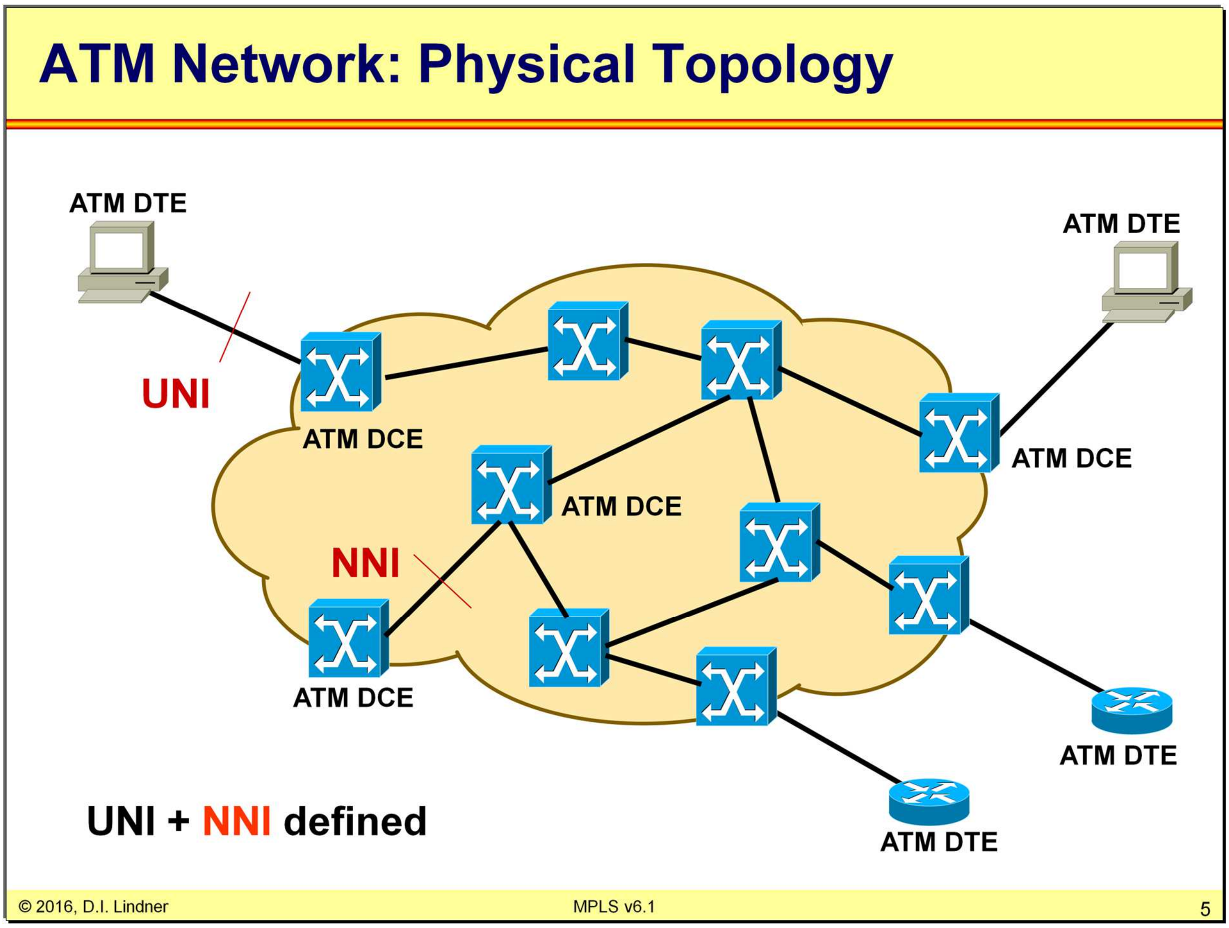
VPI/VCI identifies the virtual connection, similar function as the X.25 logical channel identifier or the Frame Relay DLCI.

The Virtual Path Identifier (VPI) is four bits longer inside the network (on NNIs) in order to support better traffic aggregation (Virtual Path Switching).

Reserved values used fo signaling, operation and maintenance, resource management

The Generic Flow Control (GFC) field is only used on the UNI but not transported into the network. The GFC is not used today as there are better methods available (special flow-control cells).

Appendix 3 - MPLS (v6.1)



In contrast to X.25 and Frame Relay the operation within the ATM cloud can be standard-based. In X.25 and Frame Relay the operation within the corresponding cloud is vendor specific.

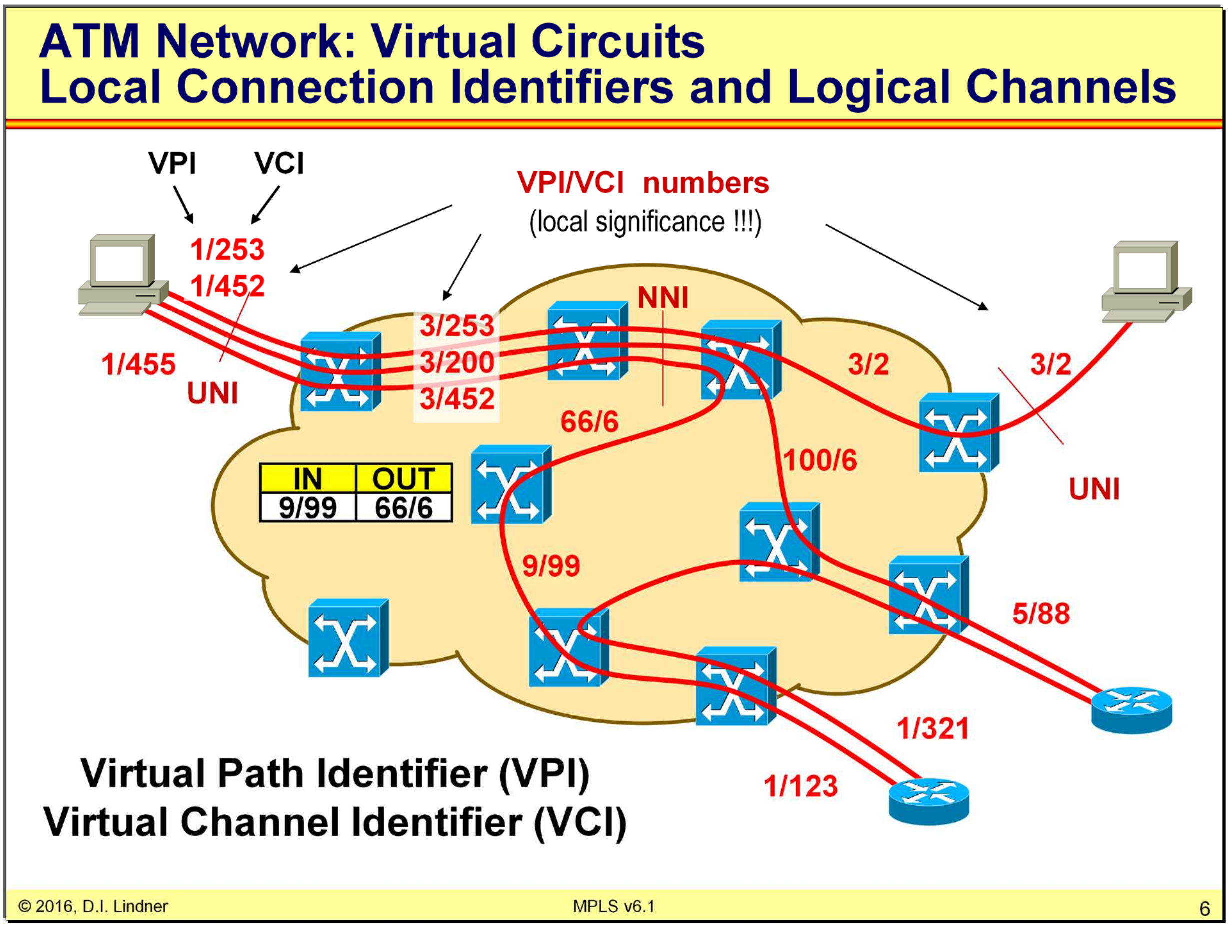
Typically, end device or a router is an ATM DTE the ATM switch is DCE.

The ATM cell header can be in two formats, UNI and NNI

UNI – User Network Interface, for public and private ATM network access

NNI – Network Network Interface, defines communication between ATM switches.

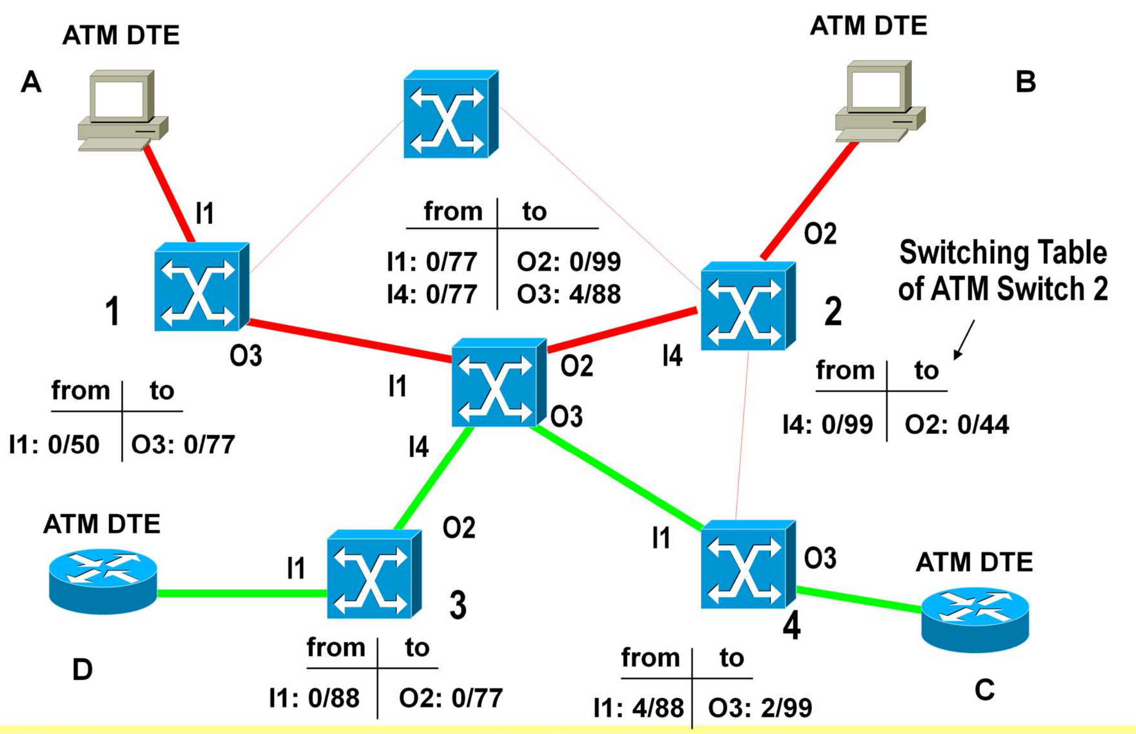
Appendix 3 - MPLS (v6.1)



Virtual Circuits in ATM could be Switched (SVC) or Permanent (PVC).

There are two types of connections: Virtual Channel (VC) and Virtual Path (VP). These two types were defined for better management. A transmission path (physical connection) consists of a bundle of VPs. A VP consists of a bundle of VCs. Virtual Path Identifier (VPI) is the number of VP in bundle. Virtual Channel Identifier (VCI) is the number of VC bundle. ATM switch uses VPI/VCI values for forwarding of ATM cells.

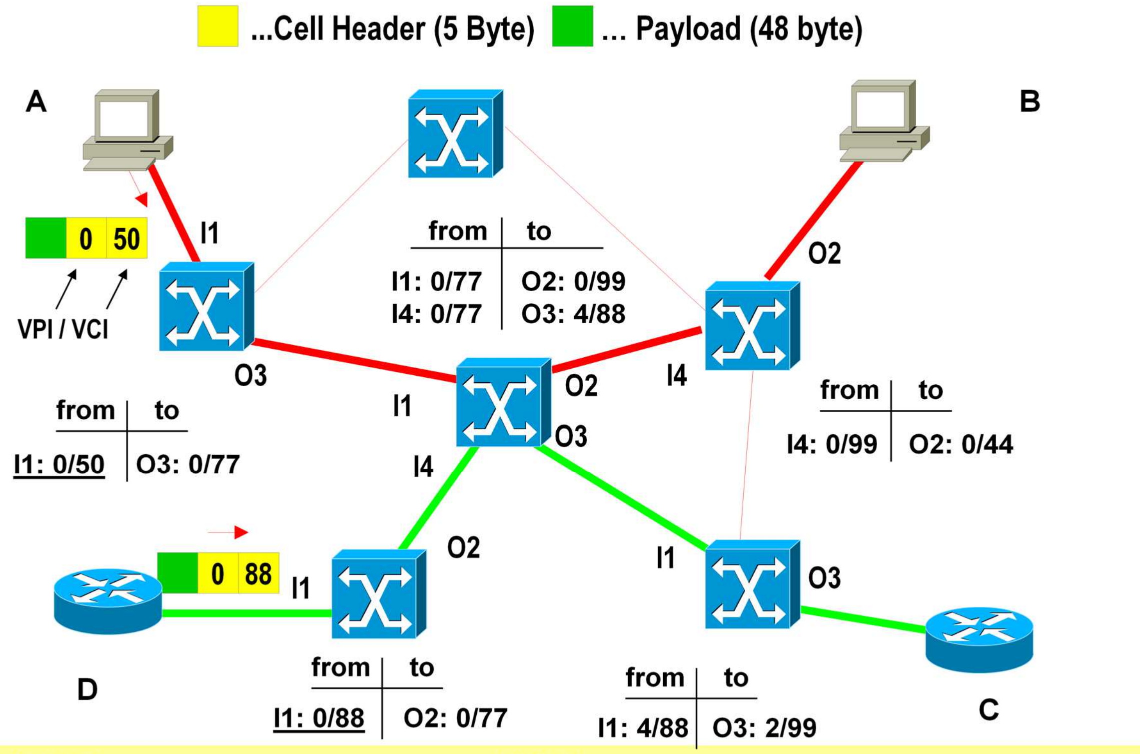
ATM Switching Tables



© 2016, D.I. Lindner MPLS v6.1 7

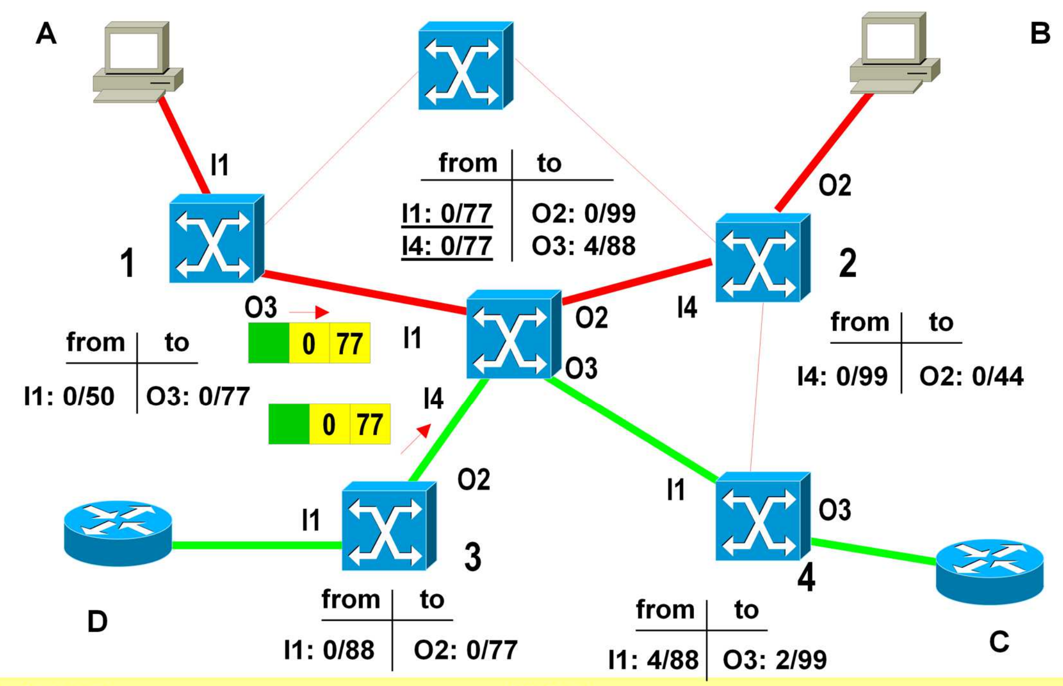
Appendix 3 - MPLS (v6.1)

Cell Forwarding / Label Swapping 1



Appendix 3 - MPLS (v6.1)

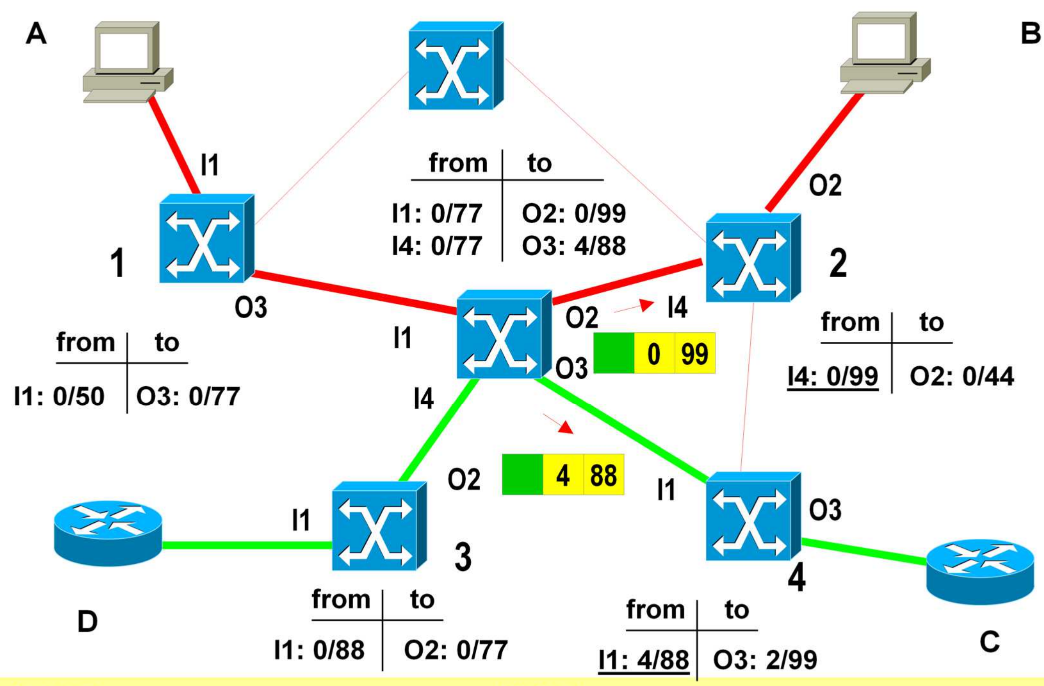
Cell Forwarding / Label Swapping 2



© 2016, D.I. Lindner MPLS v6.1 9

Appendix 3 - MPLS (v6.1)

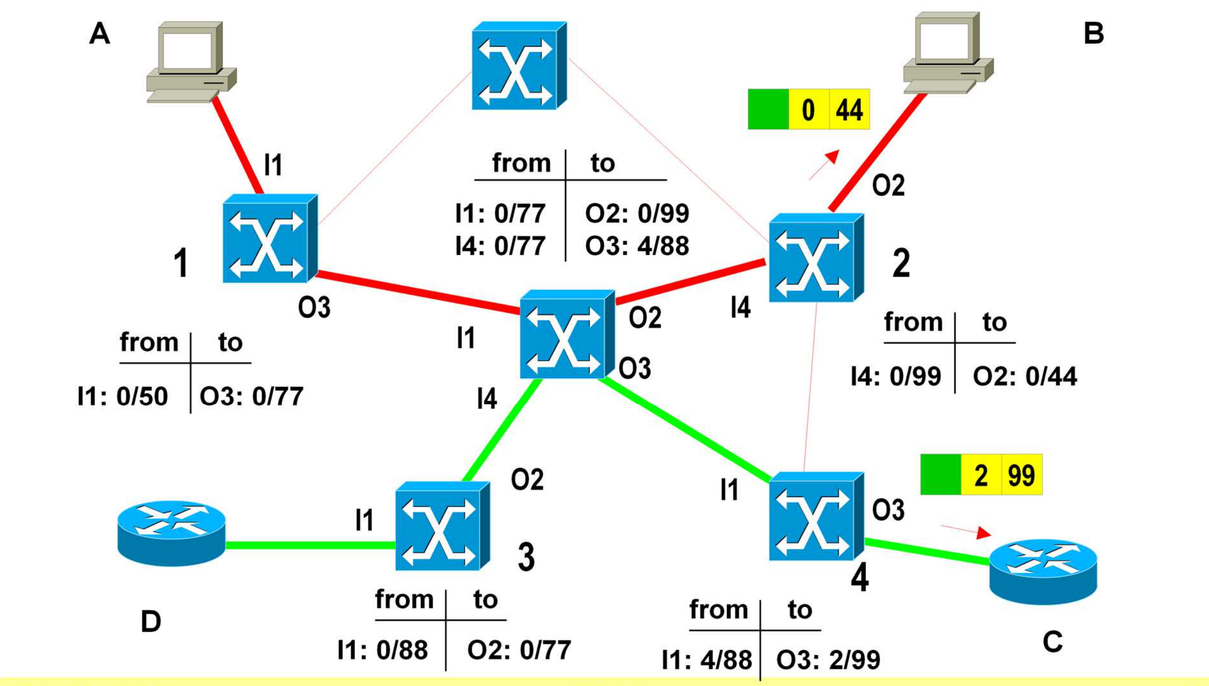
Cell Forwarding / Label Swapping 3



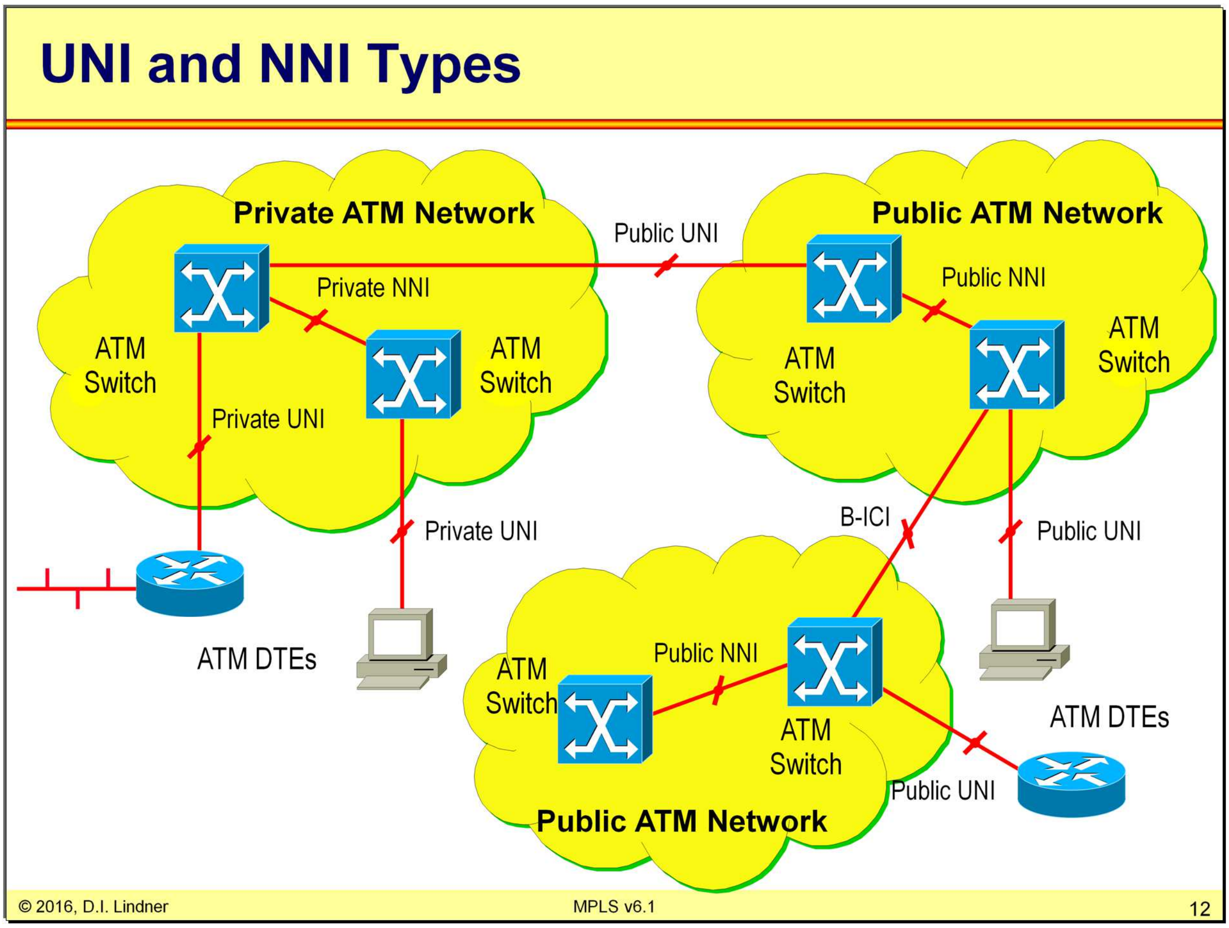
 MPLS v6.1 10

Appendix 3 - MPLS (v6.1)

Cell Forwarding / Label Swapping 4



Appendix 3 - MPLS (v6.1)



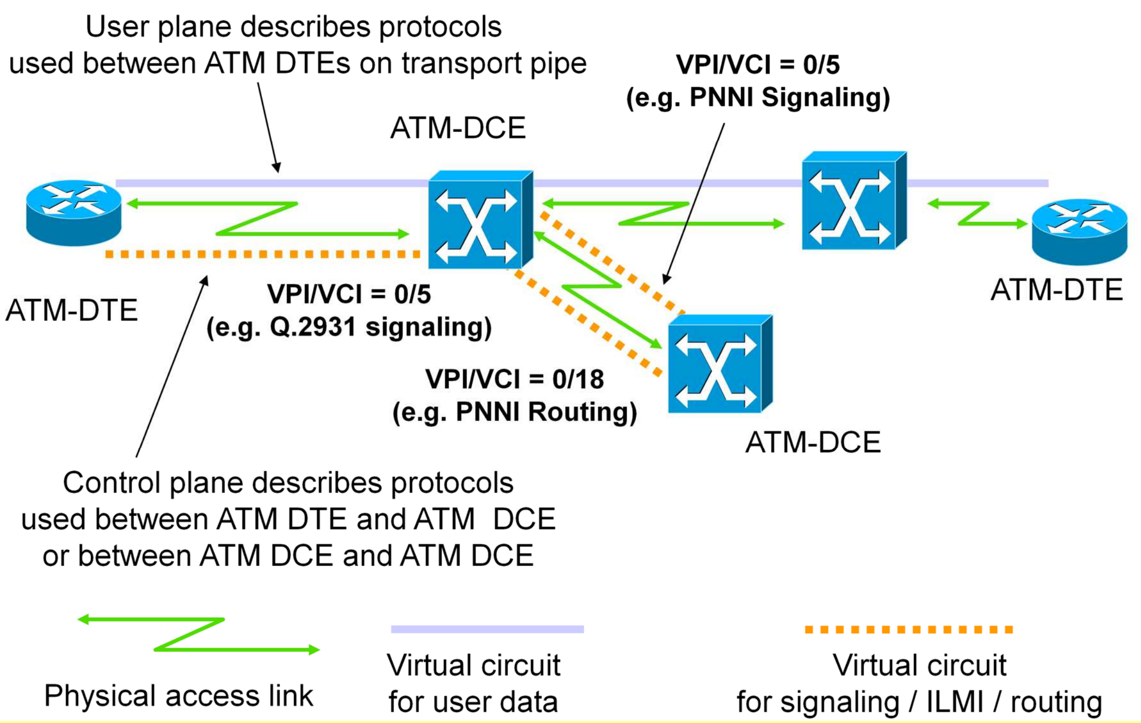
NNI-ISSI (Public NNI)
ISSI = Inter Switch System Interface
Used to connect two switches of one public service provider.

NNI-ICI (B - ICI)
ICI - Inter Carrier Interface
Used to connect two ATM networks of two different service providers.

Private NNI
Used to connect two switches of different vendors in private ATM networks.

Appendix 3 - MPLS (v6.1)

Control Plane <-> User Plane



 MPLS v6.1 13

Appendix 3 - MPLS (v6.1)

Service Classes

Guaranteed Service "Bandwidth on Demand"	CBR	Constant Bit Rate Circuit Emulation, Voice
	VBR	Variable Bit Rate Full Traffic Characterization Real-Time VBR and Non Real-Time VBR
"Best Effort" Service	UBR	Unspecified Bit Rate No Guarantees, "Send and Pray"
	ABR	Available Bit Rate No Quantitative Guarantees, but Congestion Control Feedback assures low cell loss

Traffic Contract per Service Class

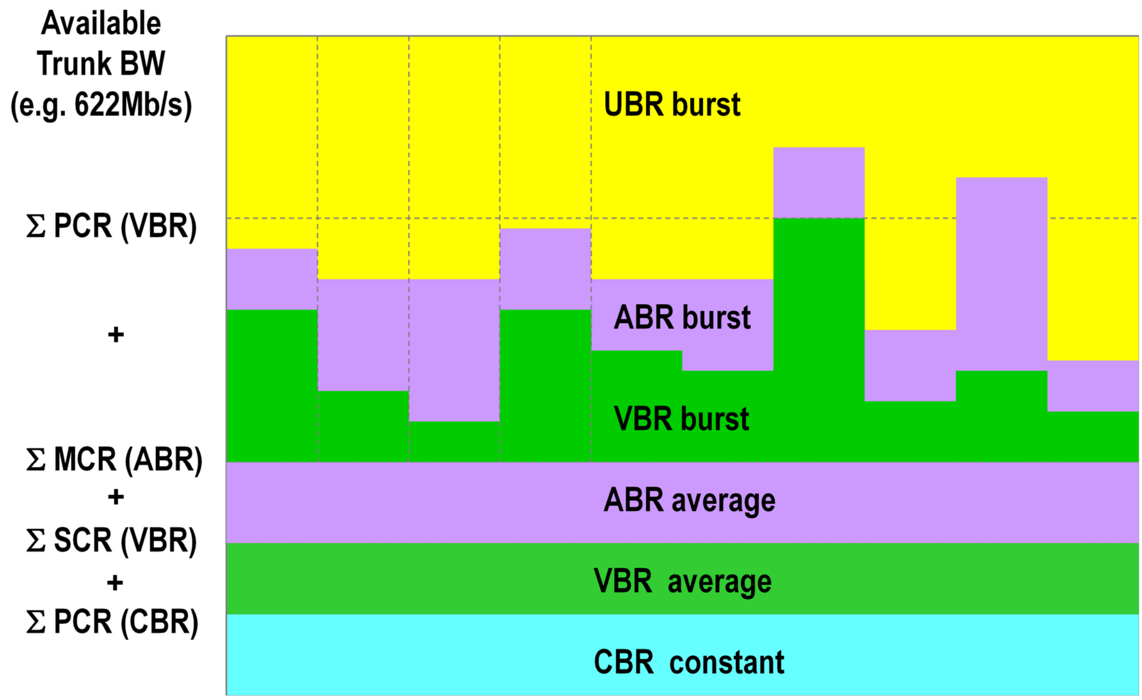
- Specified for each service class

ATTRIBUTE	CBR	rt-VBR	nrt-VBR	ABR	UBR
PCR & CDVT	Specified			Specified	
SCR, MBS, CDVT	n/a	Specified		n/a	
MCR	n/a			Specified	n/a
max CTD & ptp CDV	Specified		Unspecified	Unspecified	
CLR	Specified			Optional	Unspecified

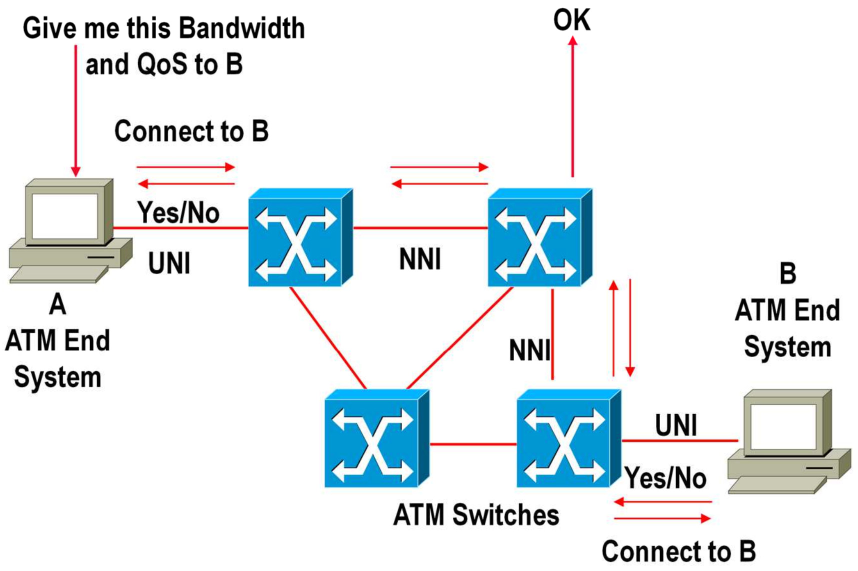
CLR = Cell Loss Ratio
CTD = Cell Transfer Delay
CDV = Cell Delay Variation
MBS = Maximum Burst Size

PCR = Peak Cell Rate
CDVT = CDV Tolerance
SCR = Sustainable CR
MCR = Minimum CR

ATM as an Intelligent Bandwidth Management System



ATM Goal: Bandwidth on Demand with QoS Guarantees

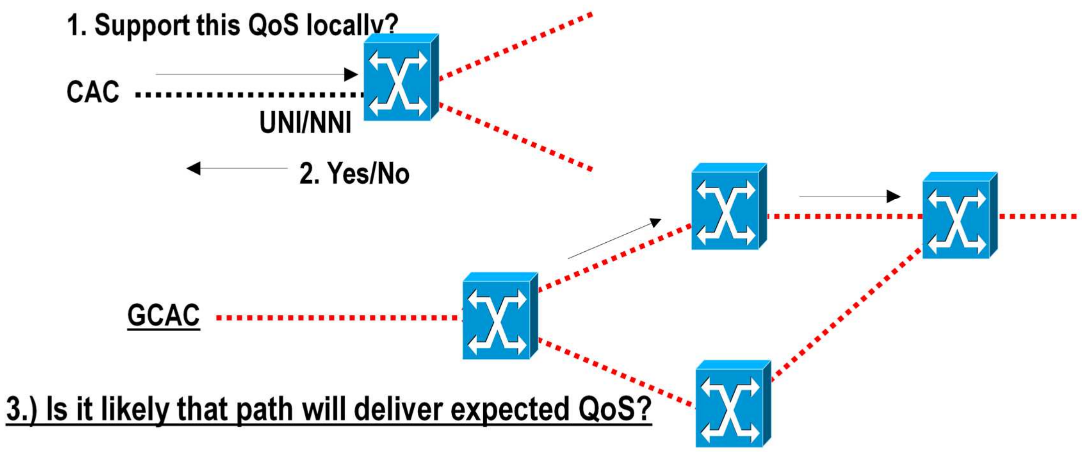


ATM Routing in Private ATM Networks

- **PNNI is based on Link-State technique**
  - like OSPF
- **Topology database**
  - Every switch maintains a database representing the states of the links and the switches
  - Extension to link state routing !!!
  - Announce status of node (!) as well as status of links
    - Contains dynamic parameters like delay, available cell rate, etc. versus static-only parameters of OSPF (link up/down, node up/down, nominal bandwidth of link)
- **Path determination based on metrics**
  - Much more complex than with standard routing protocols because of ATM-inherent QoS support

PNNI Routing

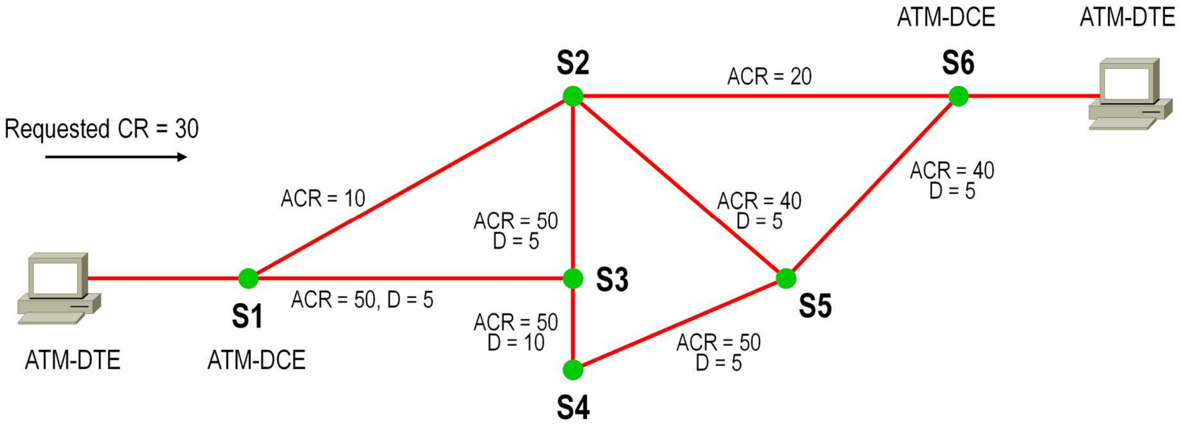
- **Generic Connection Admission Control (GCAC)**
  - Used by the source switch to select a path through the network
  - Calculates the expected CAC (Connection Admission Control) behavior of another node



Appendix 3 - MPLS (v6.1)

PNNI Routing (Simple QoS -> ACR only)

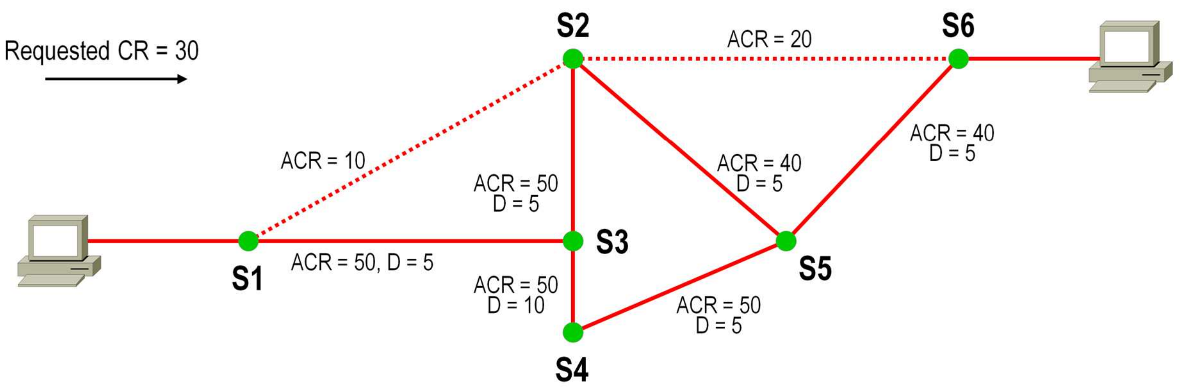
- Operation of the GCAC
  - CR … Cell Rate
  - ACR … Available Cell Rate
  - D … Distance like OSPF costs



© 2016, D.I. Lindner MPLS v6.1 20

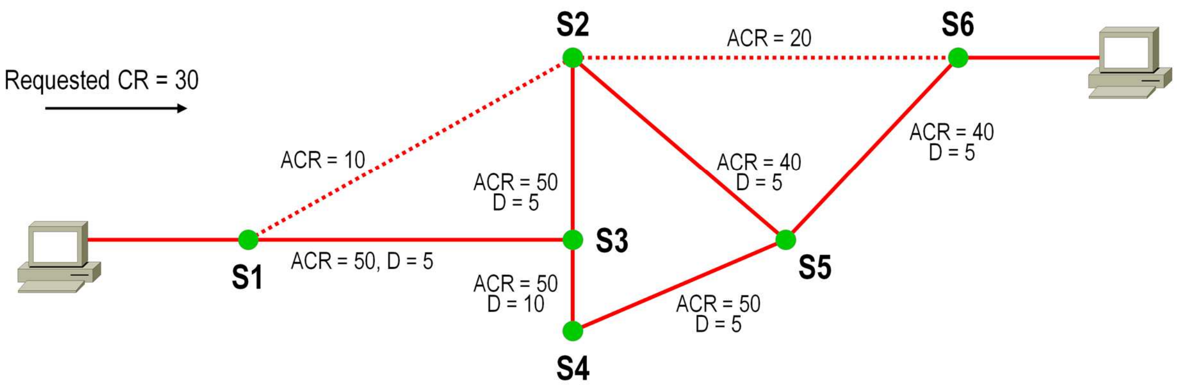
PNNI Routing

- **Operation of the GCAC**
  - 1) Links not supporting requested CR are eliminated -> ....
    - Metric component -> ACR value used



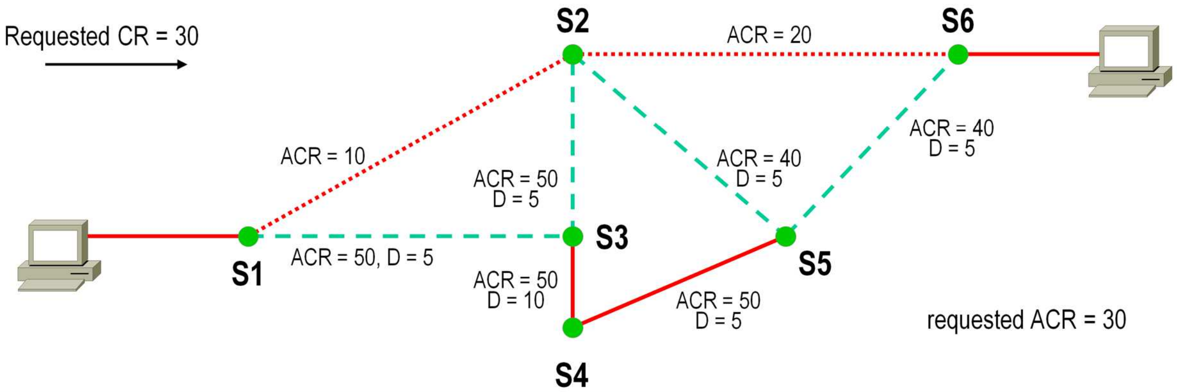
PNNI Routing

- **Operation of the GCAC**
  - 2) Next, shortest path(s) to the destination is (are) calculated
    - Metric component -> Distance value used



PNNI Routing

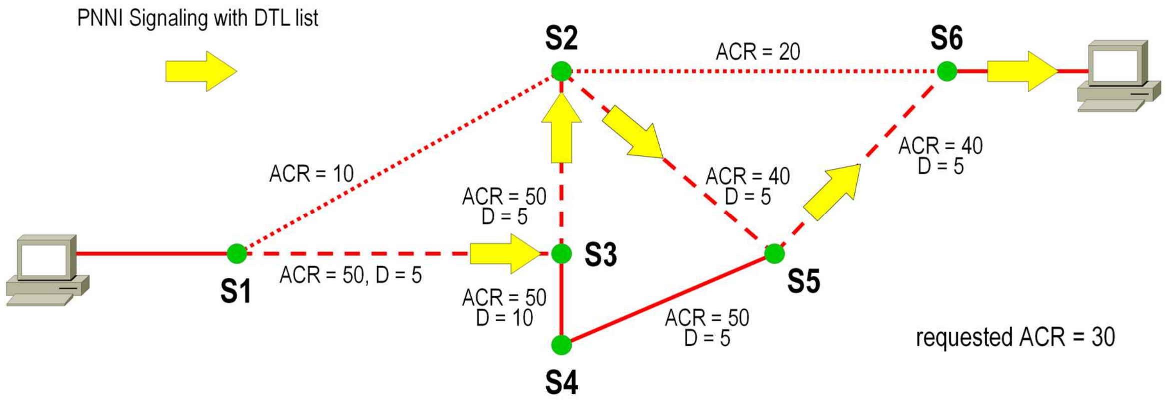
- **Operation of the GCAC**
  - 3) One path is chosen and source node S1 constructs a Designated Transit List (DTL) -> source routing -->
    - Describes the complete route to the destination



 MPLS v6.1 23

PNNI Routing - Source Routing

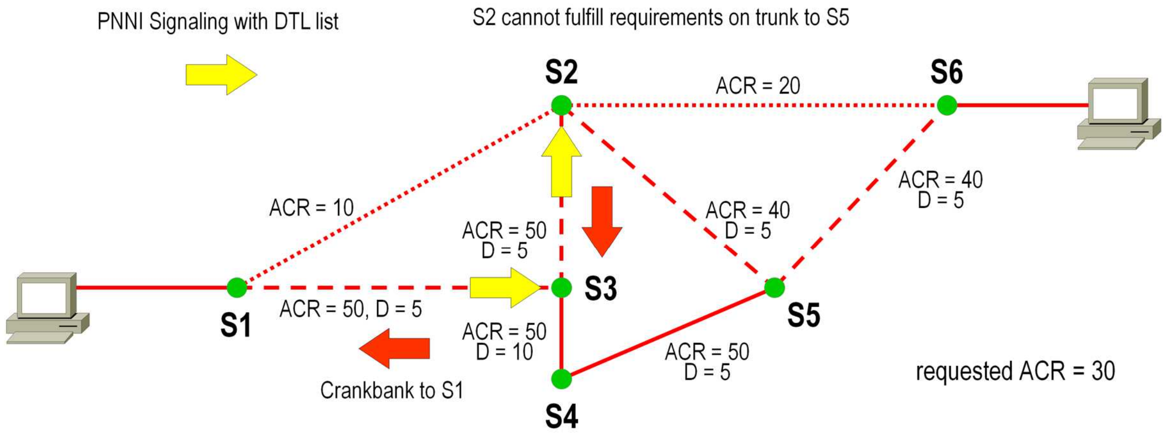
- **Operation of the GCAC**
  - 4) DTL is inserted into signaling request and moved on to next switch
  - 5) After receipt next switch perform local CAC
    - 5a) if ok -> pass PNNI signaling message on to next switch of DTL
  - 6a) finally signaling request will reach destination ATM-DTE -> VC ok



Appendix 3 - MPLS (v6.1)

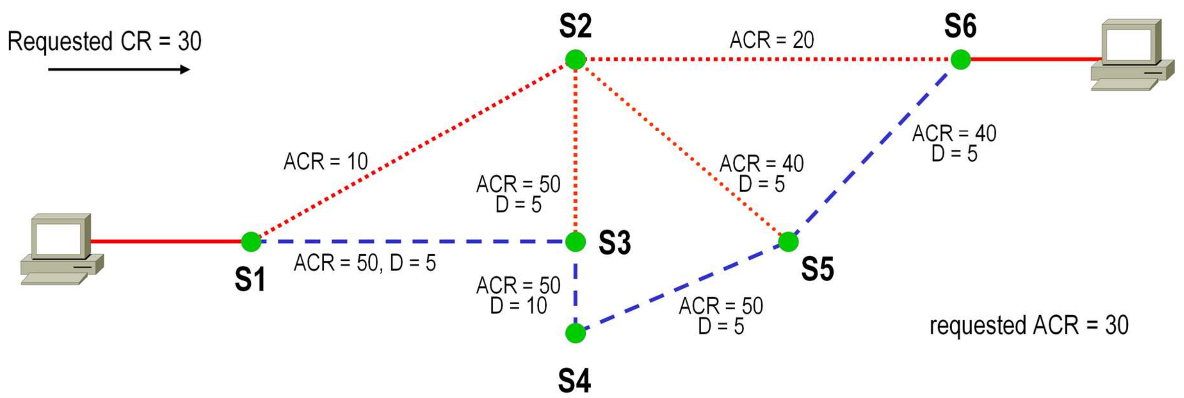
PNNI Routing - Crankbank

- **Operation of the GCAC**
  - 5) After receipt next switch (S2) perform local CAC
    - 5b) if nok -> return PNNI signaling message to originator of DTL
  - 6b) S1 will construct alternate source route



PNNI Routing - New Trial

- **Operation after Crankbank**
  - 7b) The other possible path is chosen - source node constructs again a new Designated Transit List (DTL)

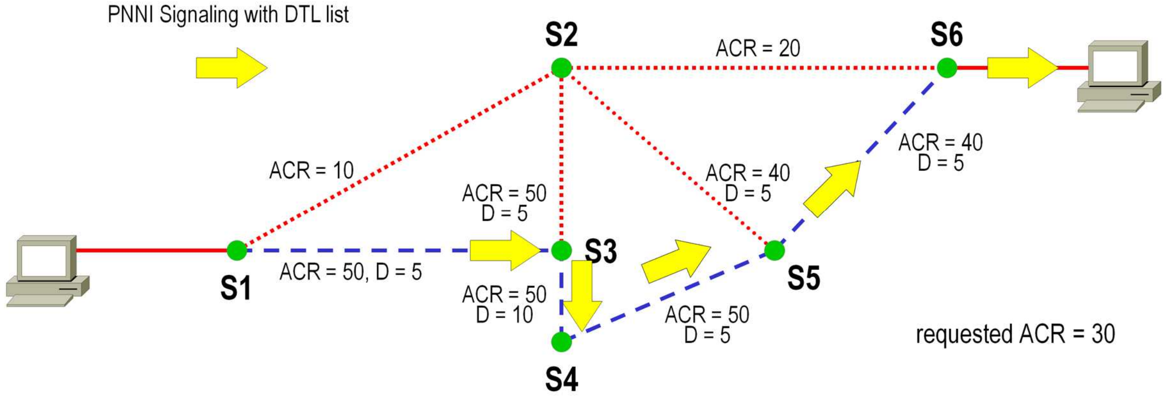


 MPLS v6.1 26

Appendix 3 - MPLS (v6.1)

PNNI Routing - Source Routing

- **Operation of the GCAC**
  - 8b) DTL is inserted into signaling request
  - 9b) After receipt next switch perform local CAC
    - if ok -> pass PNNI signaling message on to next switch of DTL
  - 10b) finally signaling request will reach destination ATM-DTE -> VC ok



Agenda

- **Review ATM**
- **IP over WAN Problems (Traditional Approach)**
  - Introduction, Base Problem 1
    - Non-NBMA-View
    - NMBA-View
  - Base Problem 2, Solution
- **MPLS Principles**
- **Label Distribution Methods**
- **MPLS Details (Cisco)**
- **RFCs**

IP Overlay Model - Scalability

- **Base problem Nr.1**
  - IP routing separated from ATM routing because of the normal IP overlay model
  - no exchange of routing information between IP and ATM world
  - leads to scalability and performance problems
    - many peers, configuration overhead, duplicate broadcasts
  - note:
    - IP system requests virtual circuits from the ATM network
    - ATM virtual circuits are established according to PNNI routing
    - virtual circuits are treated by IP as normal point-to-point links
    - IP routing messages are transported via this point-to-point links to discover IP neighbors and IP network topology

IP Performance

- **Base problem Nr.2**
  - IP forwarding is slow compared to ATM cell forwarding
    - IP routing paradigm
    - hop-by-hop routing with (recursive) IP routing table lookup, IP TTL decrement and IP checksum computing
    - destination based routing (large tables in the core of the Internet)
  - Load balancing
    - in a stable network all IP datagram's will follow the same path (least cost routing versus ATM´s QoS routing)
  - QoS (Quality of Service)
    - IP is connectionless packet switching (best-effort delivery versus ATM´s guarantees)
  - VPN (Virtual Private Networks)
    - ATM VC´s have a natural closed user group (=VPN) behavior

 MPLS v6.1 30

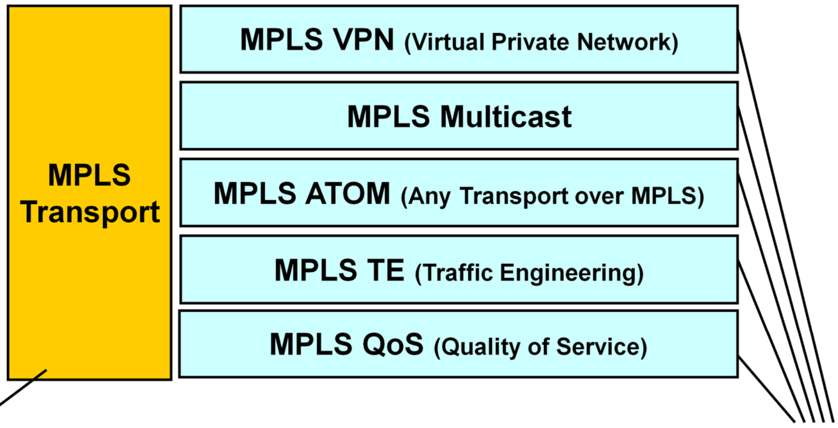
Basic Ideas to Solve the Problems

- **Make ATM topology visible to IP routing**
  - to solve the scalability problems
  - a classical ATM switch gets IP router functionality
- **Divide IP routing from IP forwarding**
  - to solve the performance problems
  - IP forwarding based on ATM´s label swapping paradigm (connection-oriented packet switching)
  - IP routing based on classical IP routing protocols
- **Combine best of both**
  - forwarding based on ATM label swapping paradigm
  - routing done by traditional IP routing protocols

MPLS

- **Several similar technologies were invented in the mid-1990s**
  - IP Switching (Ipsilon)
  - Cell Switching Router (CSR, Toshiba)
  - Tag Switching (Cisco)
  - Aggregated Route-Based IP Switching (ARIS, IBM)
- **IETF merges these technologies**
  - MPLS (Multi Protocol Label Switching)
    - note: multiprotocol means that IP is just one possible protocol to be transported by a MPLS switched network
  - RFC 3031

MPLS Building Blocks



You always need this! MPLS Transport solves most of the mentioned problems (scalability / performance)

If you need "Advanced Features like VPN or Multicast support you optionally may choose from these building blocks riding on top of a MPLS Transport network

 MPLS v6.1 33

The MPLS technology supports different types of so called MPLS Applications like the one shown in the graphic above.

MPLS Transport is the base MPLS Application which needs to be configured if you want to use other MPLS Applications like MPLS VPN, MPLS TE etc. MPLS Transport can be used to replace pure layer 3 IP forwarding with Label switching.

MPLS VPN can be used to built closed user groups on top of the MPLS Transport system.

MPLS Multicast is needed if Multicast transport through an MPLS cloud is desired.

MPLS Atom allows you to tunnel Ethernet, Frame-relay and ATM traffic through an MPLS domain.

MPLS TE can be used to overcome load-balancing limitations of IP routing protocols by the use of traffic engineering tunnels.

MPLS QoS is used if you want to support different traffic classes inside your MPLS network.

Several reasons lead to a label stack. For example, with MPLS VPNs, the top label identifies the egress router while a second label identifes the VPN itself. Thus the egress router can (as soon as the packet arrived) pop the outermost label and forward the packet to the right interface according to the inner label. Another example is MPLS Traffic Engineering (TE), where the outer label points to the TE tunnel endpoint and the inner label to the final destination itself.

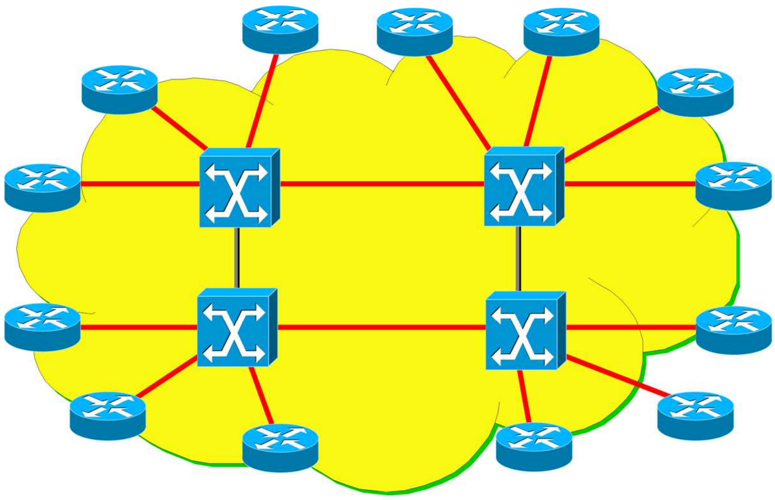
Agenda

- Review ATM
- IP over WAN Problems (Traditional Approach)
  - Introduction, Base Problem 1
    - Non-NBMA-View
    - NMBA-View
  - Base Problem 2, Solution
- MPLS Principles
- Label Distribution Methods
- MPLS Details (Cisco)
- RFCs

 MPLS v6.1 34

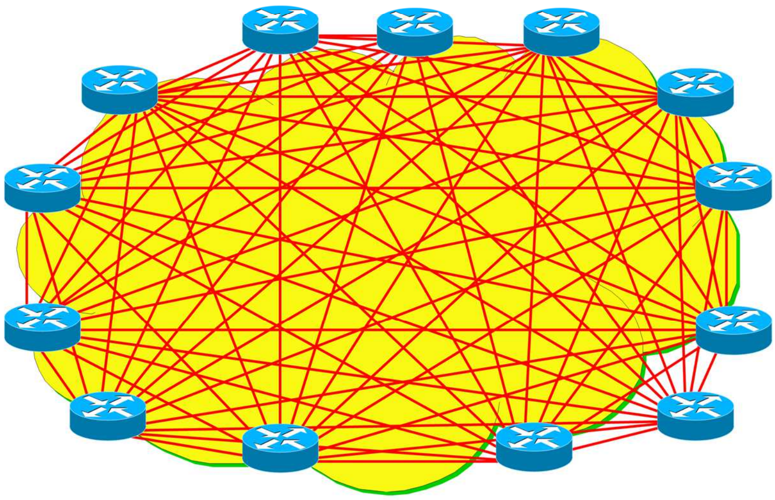
A Simple Physical Network ...

Physical wiring

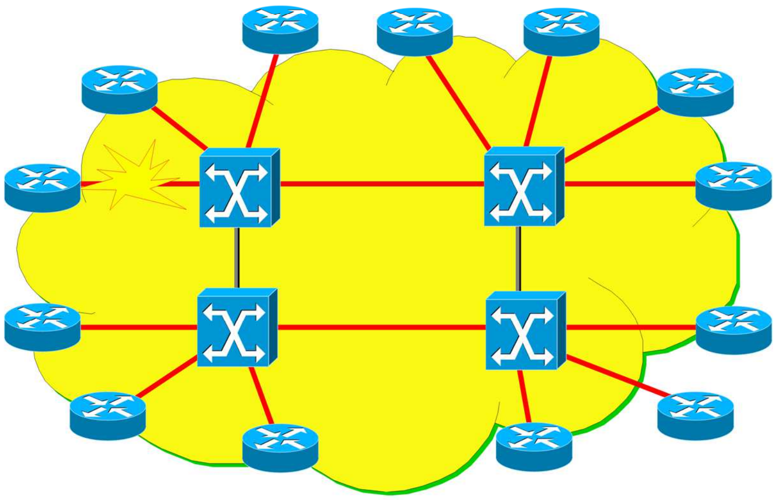


IP Data Link View (Non-NBMA)

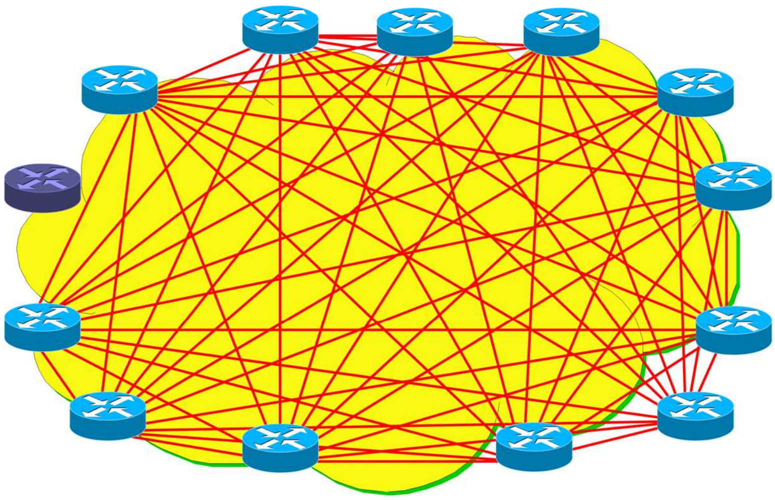
Every virtual circuit has its own IP Net-ID (subinterface technique)



A Single Network Failure ...

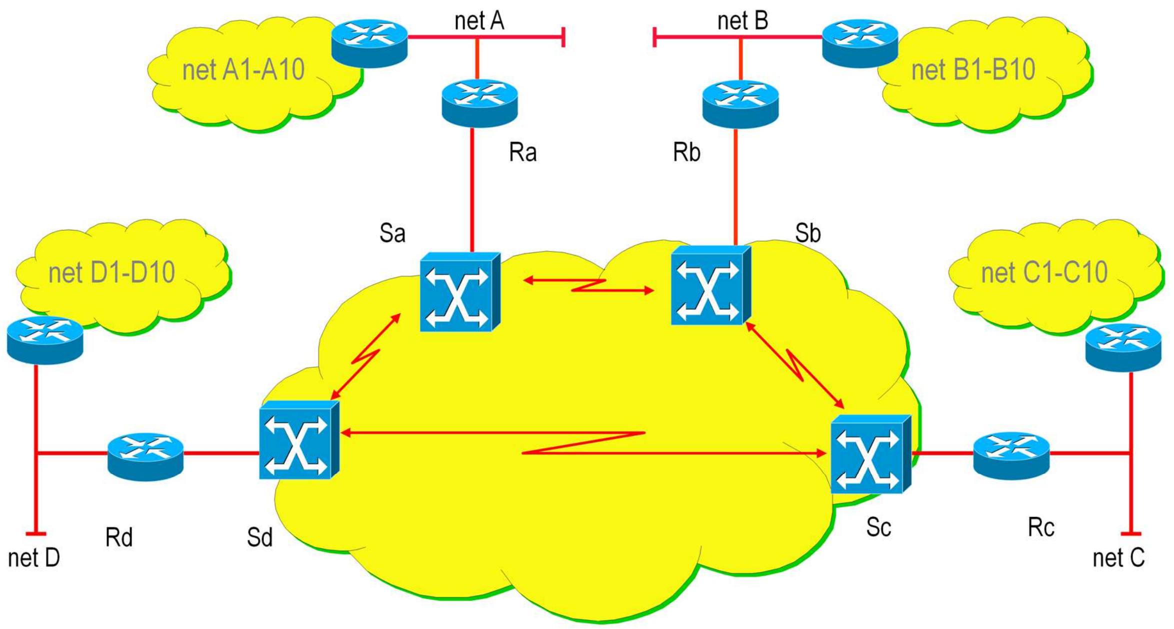


Causes Loss of Multiple IP Router Peers !!!



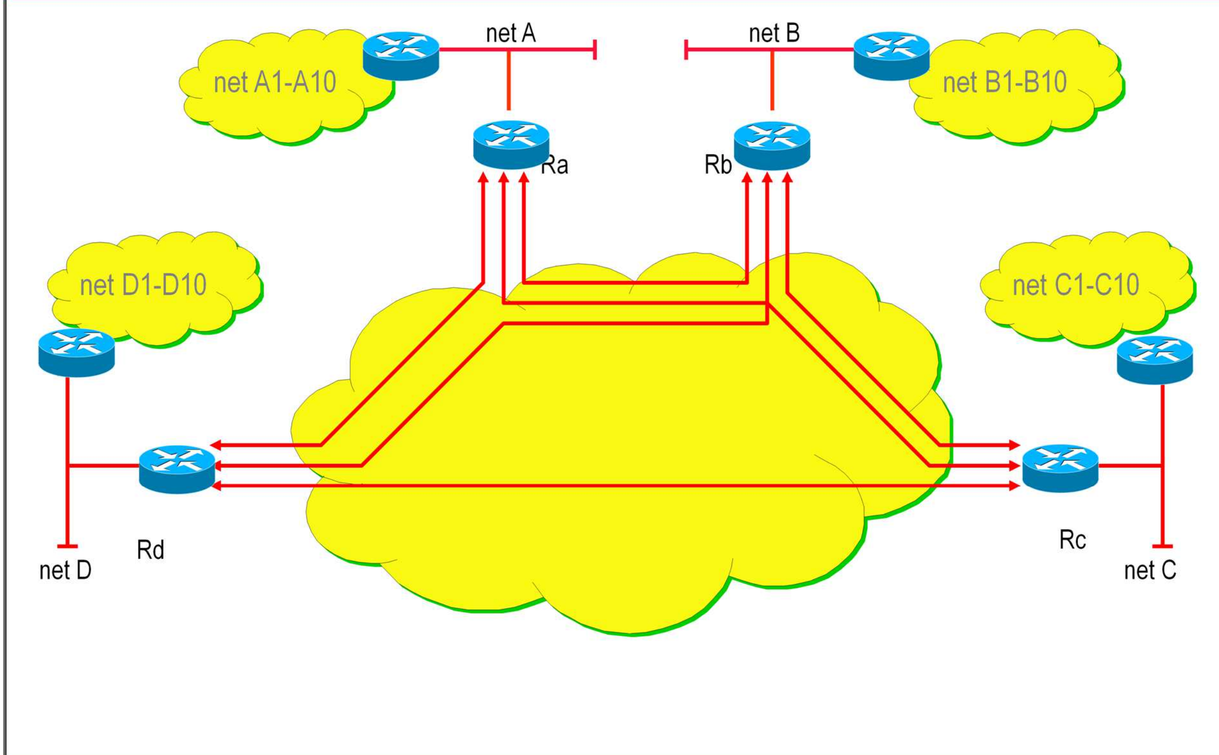
Appendix 3 - MPLS (v6.1)

Example - Physical Topology



Appendix 3 - MPLS (v6.1)

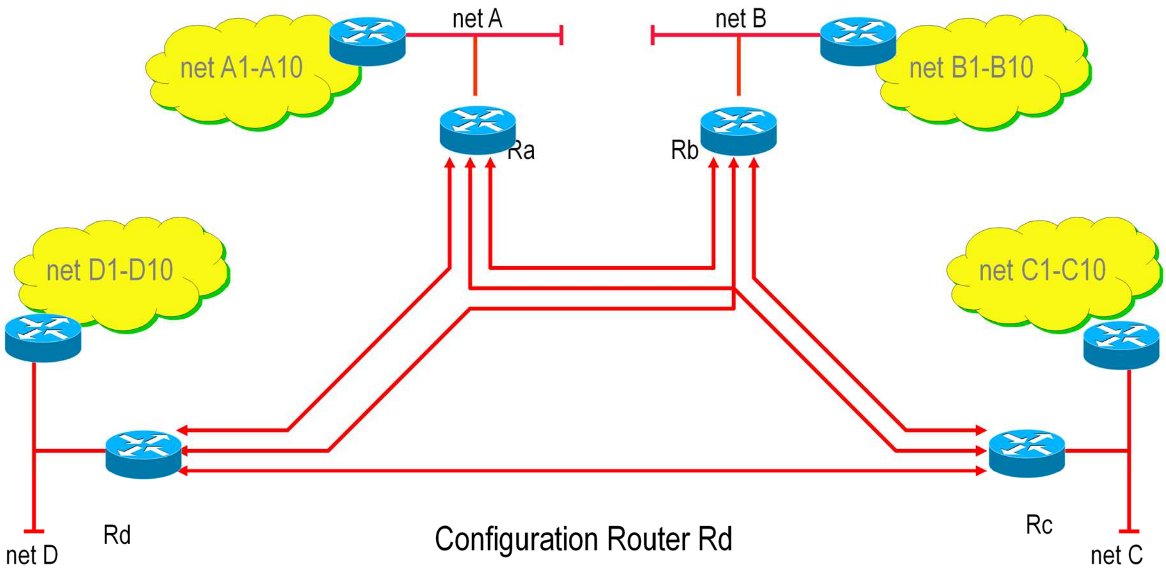
IP Connectivity through Full-mesh VC´s



 MPLS v6.1 40

Appendix 3 - MPLS (v6.1)

Static Routing/No Routing Broadcasts



Configuration Router Rd

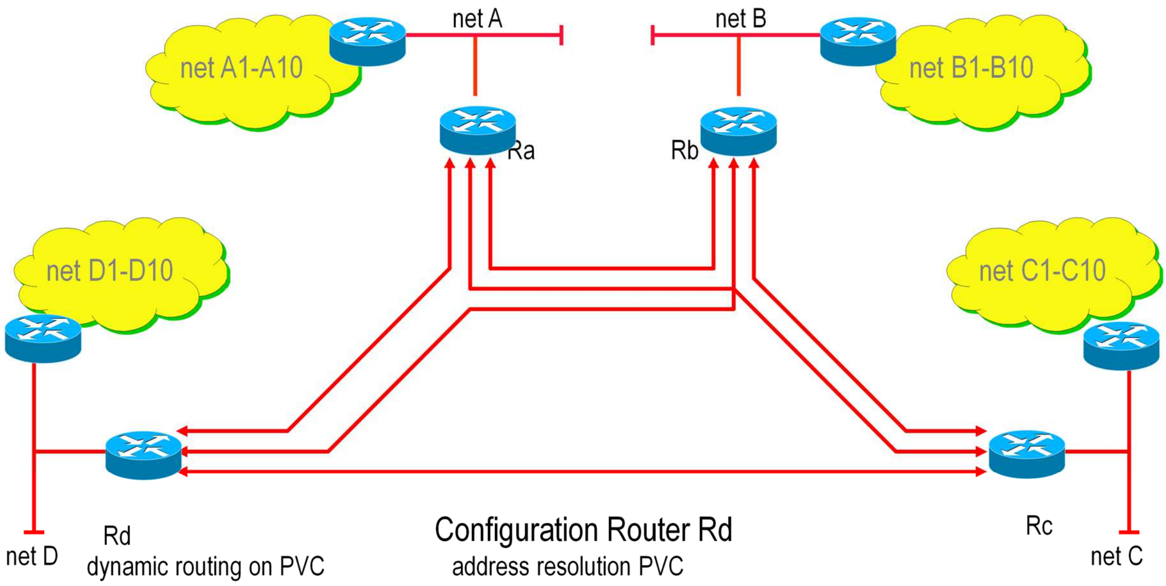
static routing
net A via next hopRa
net B via next hopRb
net C via next hopRc
every remote network listed here!

address resolution PVC
Ra map VPI/VCI Rd ⇨ Ra
Rb map VPI/VCI Rd ⇨ Rb
Rc map VPI/VCI Rd ⇨ Rc

address resolution SVC
Ra map ATM addr. Ra
Rb map ATM addr. Rb
Rc map ATM addr. Rc

Appendix 3 - MPLS (v6.1)

Dynamic Routing/Routing Broadcasts



Configuration Router Rd

dynamic routing on PVC	address resolution PVC
VPI/VCI Rd ⇨ Ra broadcast	Ra map VPI/VCI Rd ⇨ Ra
VPI/VCI Rd ⇨ Rb broadcast	Rb map VPI/VCI Rd ⇨ Rb
VPI/VCI Rd ⇨ Rc broadcast	Rc map VPI/VCI Rd ⇨ Rc

note: SVCs may be possible if Cisco neighbor command is specified for Cisco routing process because no automatic neighbor discovery is possible in this case

 MPLS v6.1 42

Observations

- **This clearly does not scale**
- **Switch/router interaction needed**
  - peering model
- **Without MPLS**
  - Only outside routers are layer 3 neighbors
  - one ATM link failure causes multiple peer failures
  - routing traffic does not scale (number of peers)
- **With MPLS**
  - Inside MPLS switch is the layer 3 routing peer of an outside router
  - one ATM link failure causes one peer failure
  - highly improved routing traffic scalability

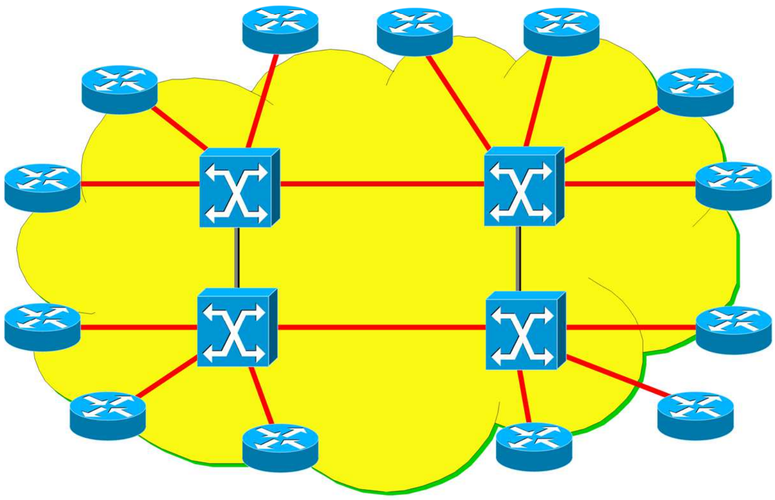
 MPLS v6.1 43

Agenda

- **Review ATM**
- **IP over WAN Problems (Traditional Approach)**
  - Introduction, Base Problem 1
    - Non-NBMA-View
    - NMBA-View
  - Base Problem 2, Solution
- **MPLS Principles**
- **Label Distribution Methods**
- **MPLS Details (Cisco)**
- **RFCs**

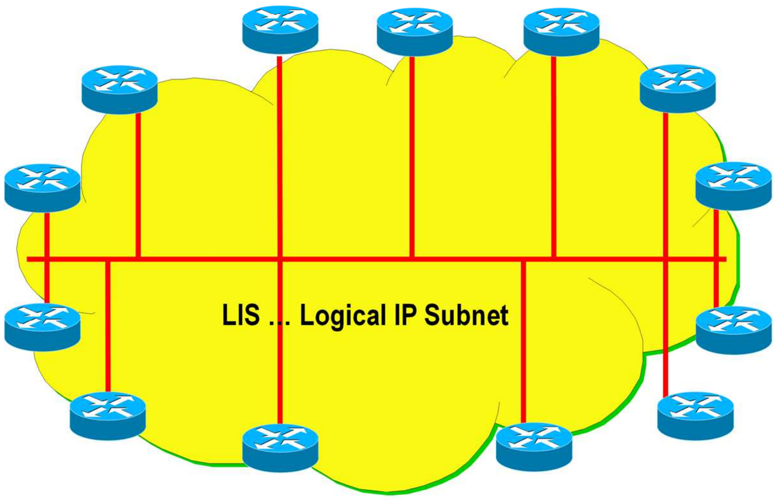
A Simple Physical Network ...

Physical wiring and NBMA behavior



IP Data Link View (NBMA)

Routers assume a LAN behavior because all interfaces have the same IP Net-ID but LAN broadcasting to reach all others is not possible



Some Solutions for the NBMA Problem

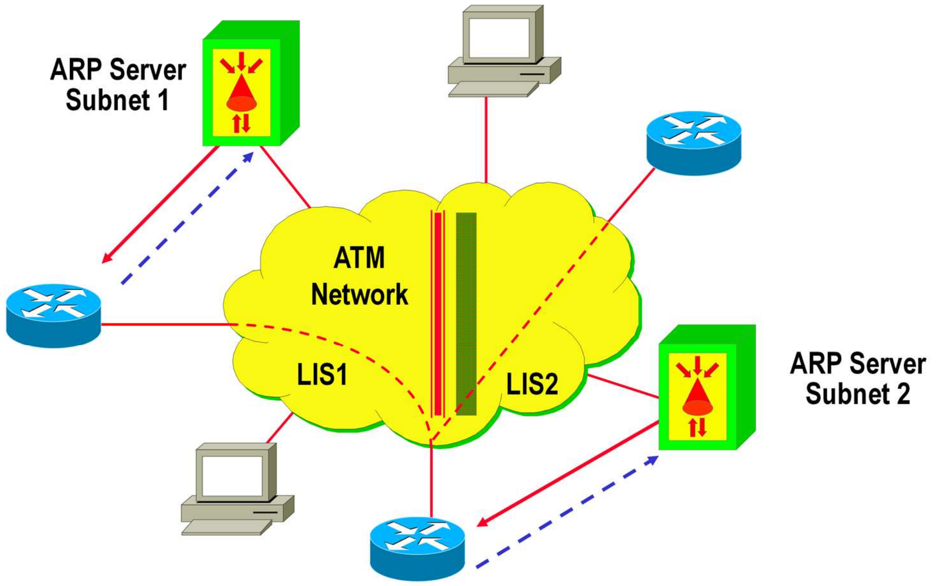
- ARP (Address Resolution Protocol) Server
  - keeps configuration overhead for address resolution small
  - but does not solve the routing issue (neighbor discovery and duplicate routing broadcasts on a single wire)
- MARS/MCS (Multicast Address Resolution Server / Multicast Server)
  - additional keeps configuration overhead for routing small
  - and does solve broadcast/multicast problem with either full mesh of point-to-multipoint circuits or by usage of MCS server
- LANE (LAN Emulation = ATM VLAN´s)
  - simulates LAN behavior where address resolution and routing broadcasts are not a problem
- All of them
  - require a lot of control virtual circuits (p-t-p and p-t-m) and SVC support of the underlying ATM network

 MPLS v6.1 47

Appendix 3 - MPLS (v6.1)

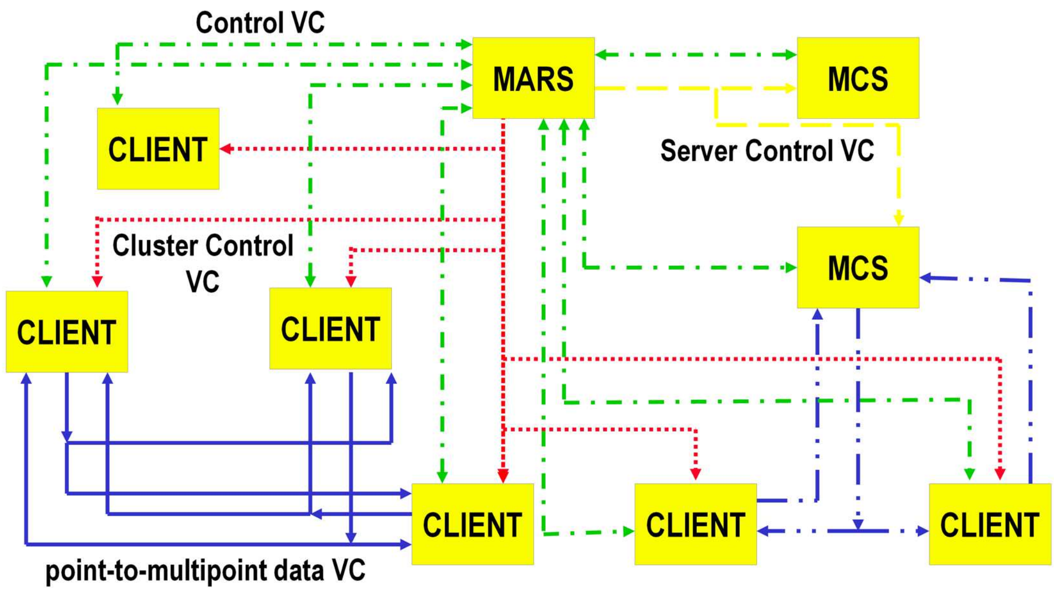
RFC 2225 Operation (Classical IP over ATM)

- **ARP server for every LIS**
  - multiple hops for communication between Logical IP Subnets



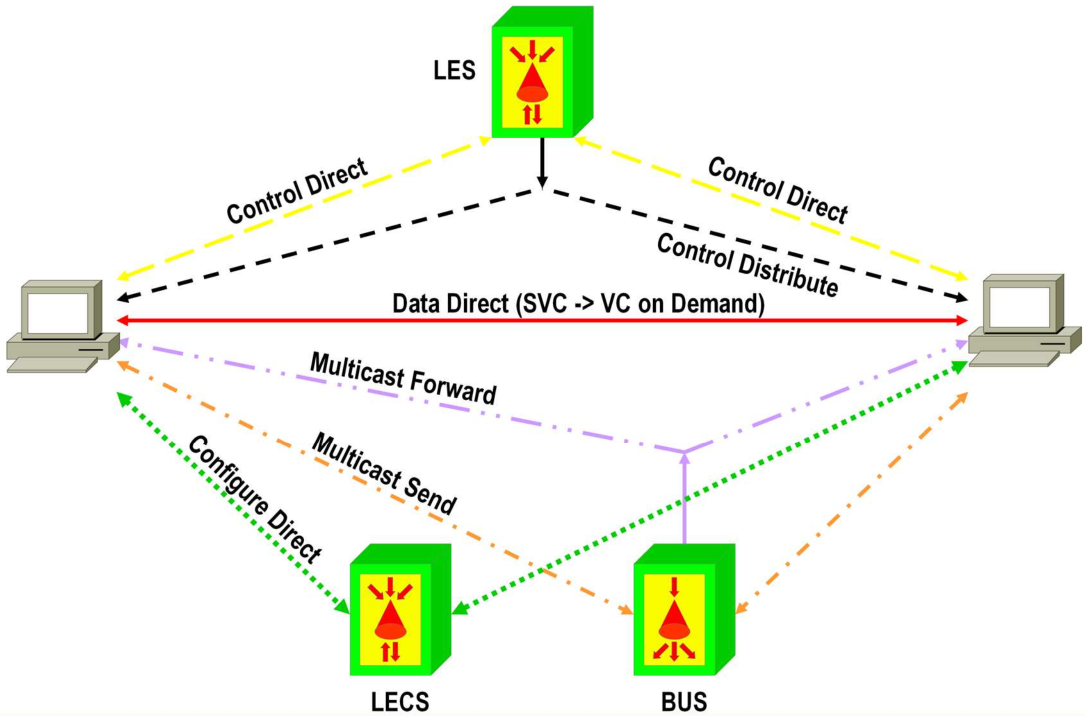
Appendix 3 - MPLS (v6.1)

MARS/MCS Architecture



Appendix 3 - MPLS (v6.1)

LANE Connections

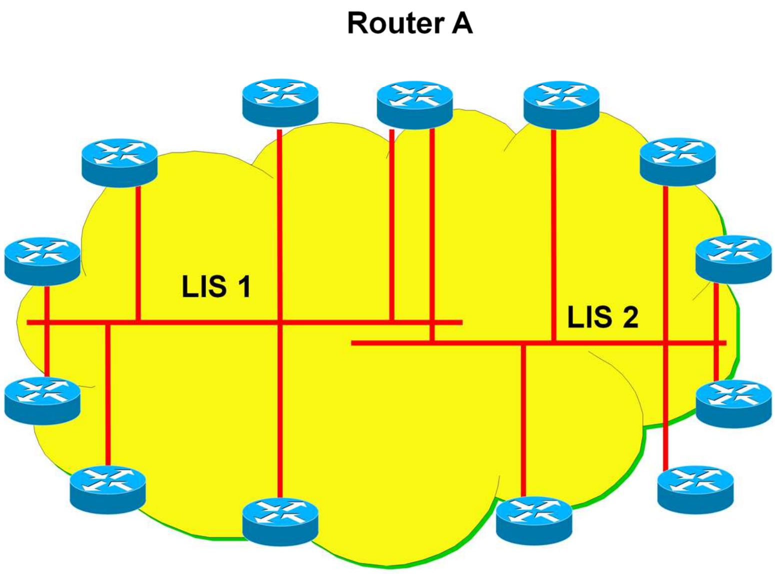


Scalability Aspects

- **Number of IP peers determines**
  - number of data virtual circuits
  - number of control virtual circuits
  - number of duplicate broadcasts on a single wire
- **Method to solve the broadcast domain problem**
  - split the network in several LIS (logical IP subnets)
  - connect LIS's by normal IP router (ATM-DCE) which is of course outside the ATM network
- **But then another problem arise**
  - traffic between to two systems which both are attached to the ATM network but belong to different LIS´s must leave the ATM network and enter it again at the connecting IP router (-> SAR delay)

IP Multiple LIS´s in case of ROLC (Routing Over Large Clouds)

IP router A connects LIS1 and LIS2

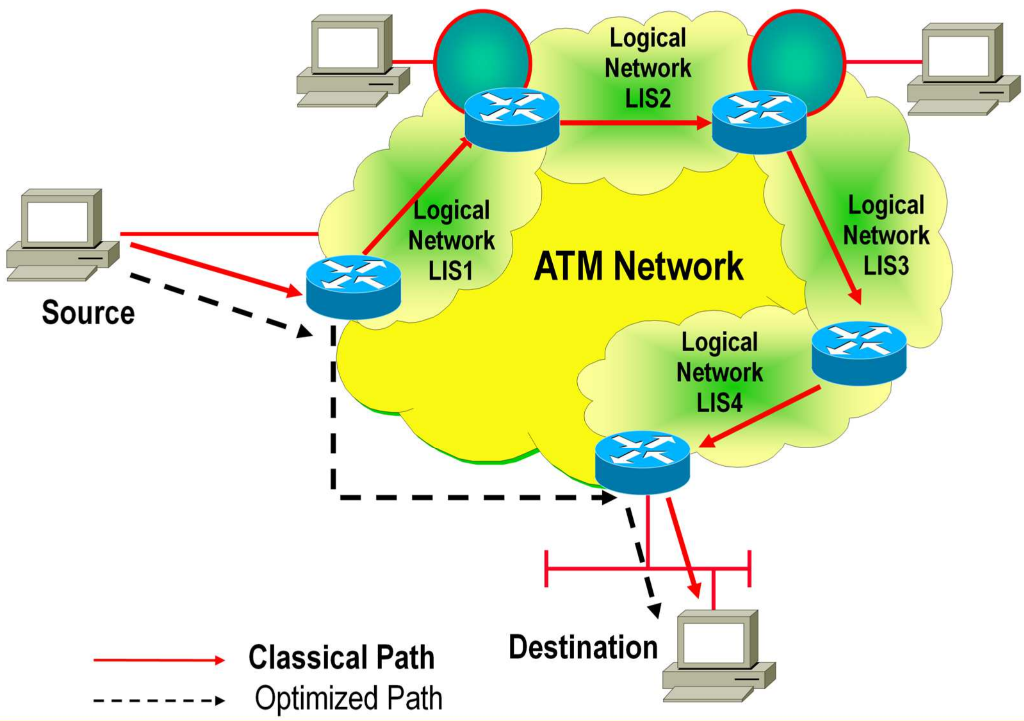


Some Solutions for the ROLC Problem

- **NHRP (Next Hop Resolution Protocol)**
  - creates an ATM shortcut between two systems of different LIS´s
- **MPOA (Multi Protocol Over ATM)**
  - LANE + NHRP combined
  - creates an ATM shortcut between two systems of different LIS´s
- **In both methods**
  - the ATM shortcut is created if traffic between the two systems exceeds a certain threshold -> data-flow driven
  - a lot of control virtual circuits (p-t-p and p-t-m) is required

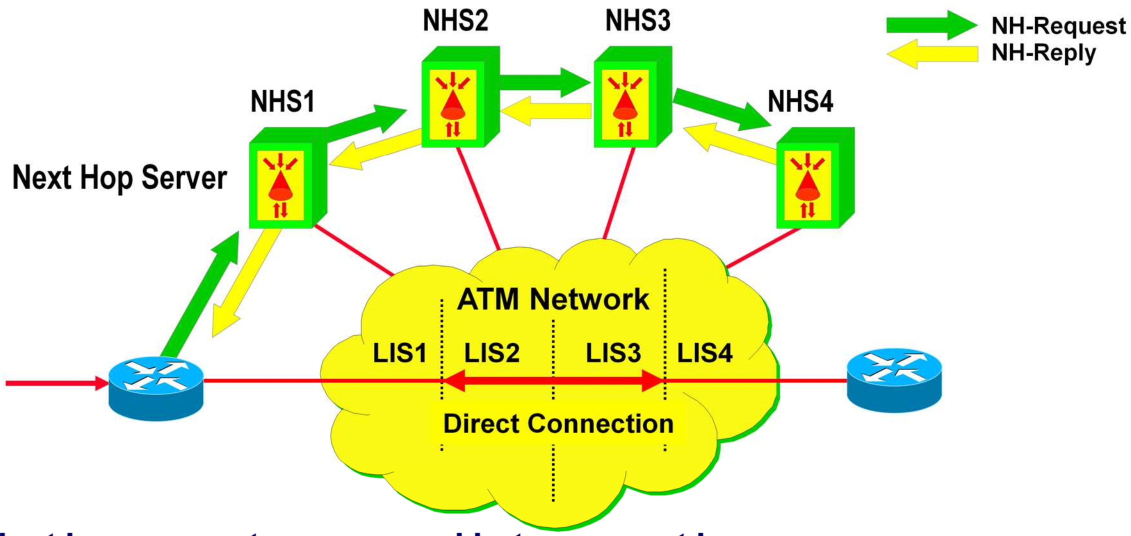
Appendix 3 - MPLS (v6.1)

Wish for Optimized Connectivity



© 2016, D.I. Lindner MPLS v6.1 54

Next Hop Resolution Protocol (RFC 2332)



- **Next hop requests are passed between next hop servers**
  - Next hop servers do not forward data
- **NHS that knows about the destination sends back a NH-reply**
  - Allows direct connection between logical IP subnets across the ATM cloud
  - Separates data forwarding path from reachability information

Agenda

- Review ATM
- IP over WAN Problems (Traditional Approach)
  - Introduction, Base Problem 1
    - Non-NBMA-View
    - NMBA-View
  - Base Problem 2, Solution
- MPLS Principles
- Label Distribution Methods
- MPLS Details (Cisco)
- RFCs

 MPLS v6.1 56

IP Performance

- **Base problem Nr.2**
  - IP forwarding is slow compared to ATM cell forwarding
    - IP routing paradigm
    - hop-by-hop routing with (recursive) IP routing table lookup, IP TTL decrement and IP checksum computing
    - destination based routing (large tables in the core of the Internet)
  - Load balancing
    - in a stable network all IP datagram's will follow the same path (least cost routing versus ATM´s QoS routing)
  - QoS (Quality of Service)
    - IP is connectionless packet switching (best-effort delivery versus ATM´s guarantees)
  - VPN (Virtual Private Networks)
    - ATM VC´s have a natural closed user group (=VPN) behavior

Basic Ideas to Solve the Problems

- **Make ATM topology visible to IP routing**
  - to solve the scalability problems
  - a classical ATM switch gets IP router functionality
- **Divide IP routing from IP forwarding**
  - to solve the performance problems
  - IP forwarding based on ATM´s label swapping paradigm (connection-oriented packet switching)
  - IP routing based on classical IP routing protocols
- **Combine best of both**
  - forwarding based on ATM label swapping paradigm
  - routing done by traditional IP routing protocols

MPLS

- **Several similar technologies were invented in the mid-1990s**
  - IP Switching (Ipsilon)
  - Cell Switching Router (CSR, Toshiba)
  - Tag Switching (Cisco)
  - Aggregated Route-Based IP Switching (ARIS, IBM)
- **IETF merges these technologies**
  - MPLS (Multi Protocol Label Switching)
    - note: multiprotocol means that IP is just one possible protocol to be transported by a MPLS switched network
  - RFC 3031

Agenda

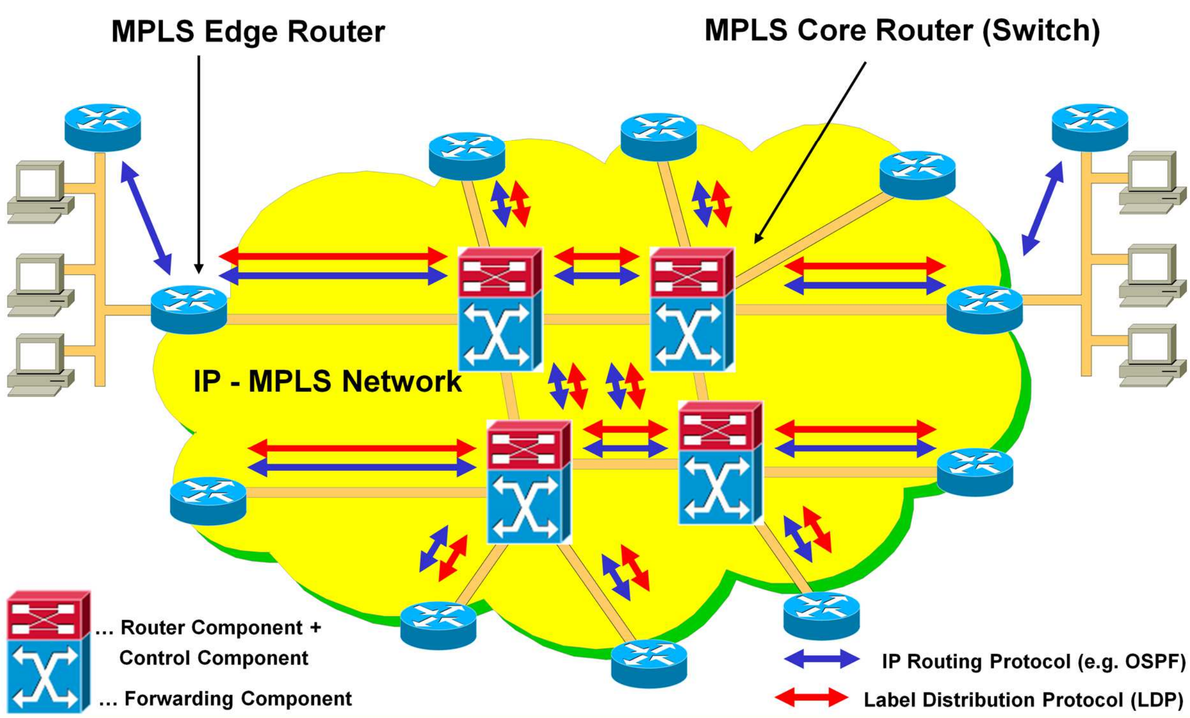
- Review ATM
- IP over WAN Problems (Traditional Approach)
- MPLS Principles
- Label Distribution Methods
- MPLS Details (Cisco)
- RFCs

MPLS Approach

- **Traditional IP uses the same information for**
  - path determination (routing)
  - packet forwarding (switching)
- **MPLS separates the tasks**
  - L3 addresses used for path determination
  - labels used for switching
- **MPLS Network consists of**
  - MPLS Edge Routers and MPLS Switches
- **MPLS Edge Routers and MPLS Switches**
  - exchange routing information about L3 IP networks
  - exchange forwarding information about the actual usage of labels

 MPLS v6.1 61

MPLS Network

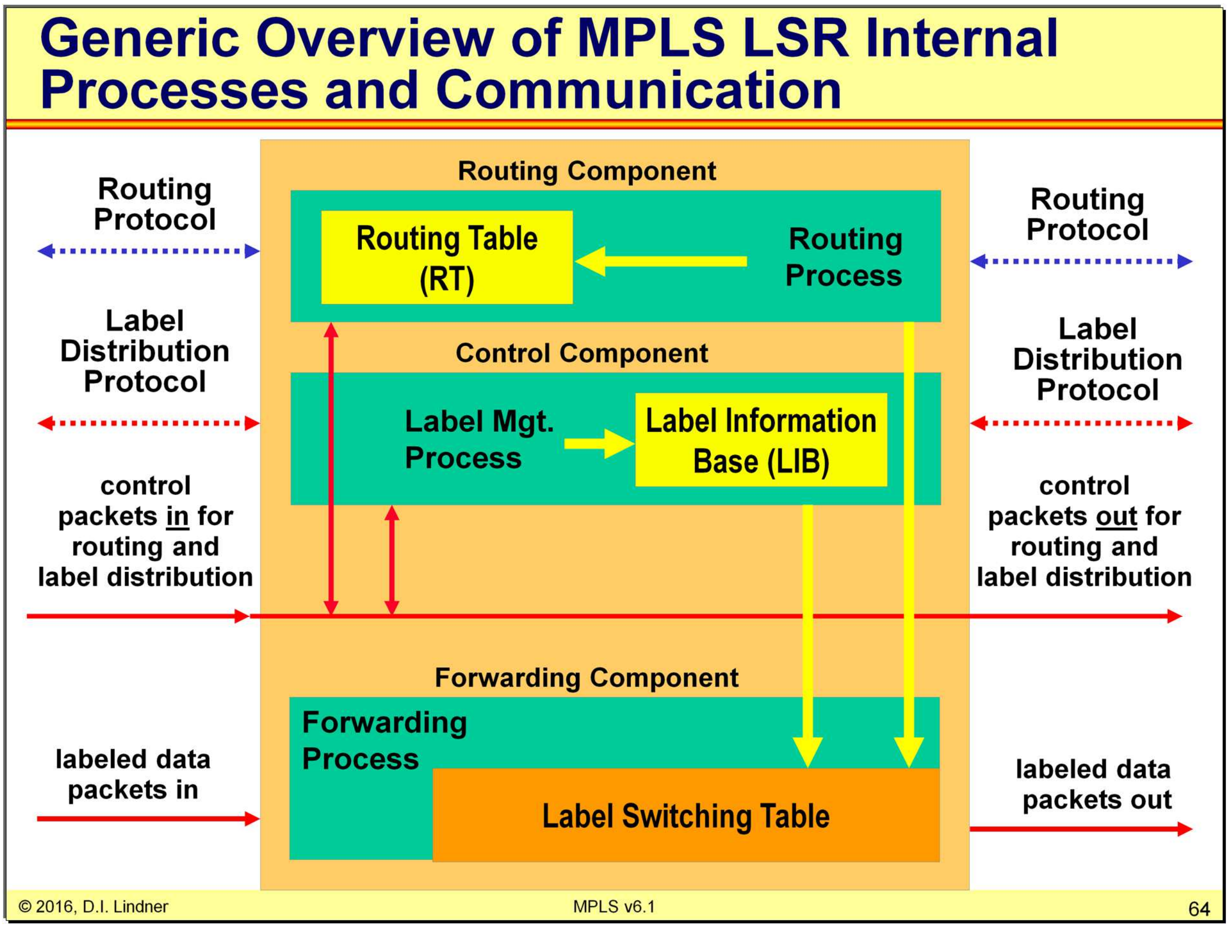


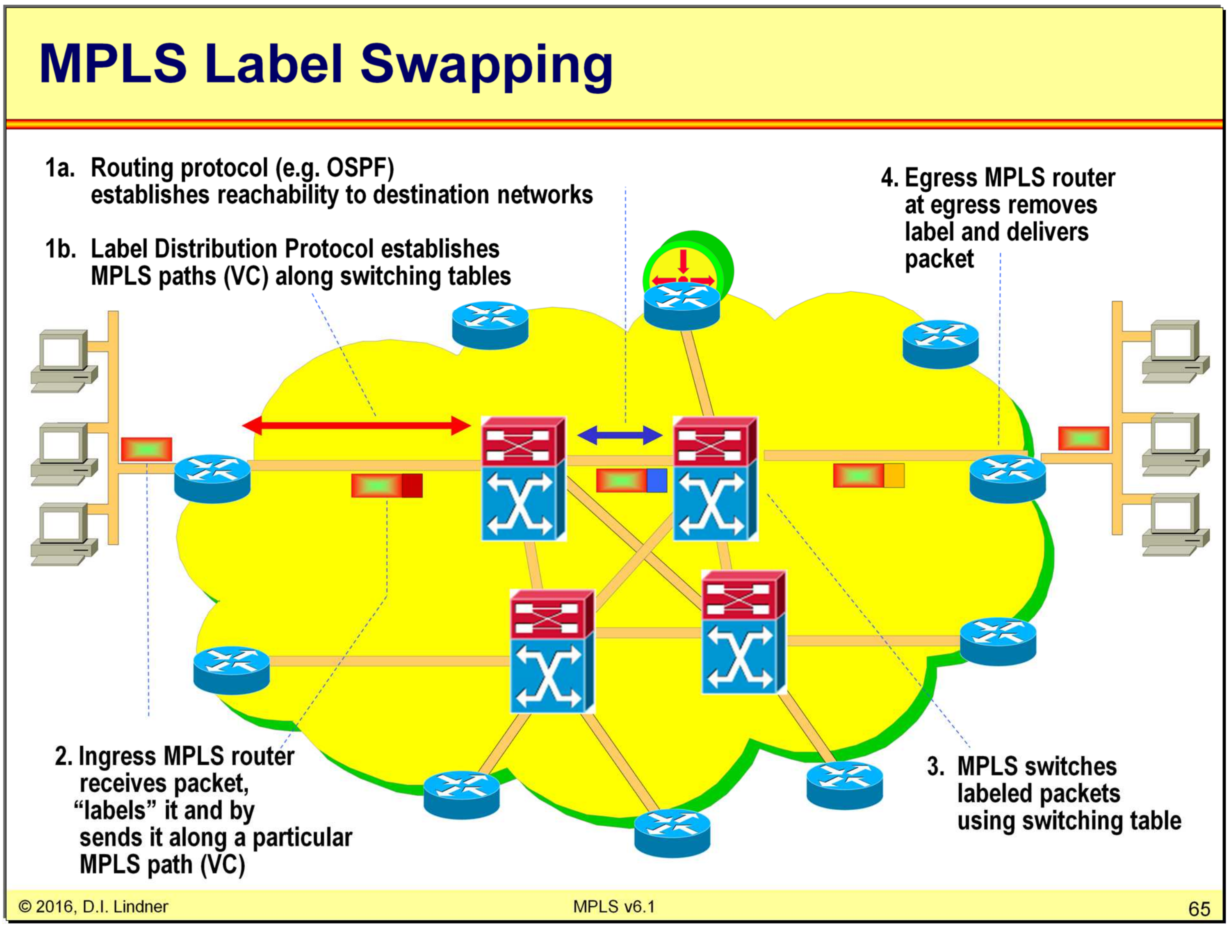
© 2016, D.I. Lindner MPLS v6.1 62

MPLS LSR Internal Components

- **Routing Component**
  - still accomplished by using standard IP routing protocols creating routing table
- **Control Component**
  - maintains correct label distribution among a group of label switches
  - Label Distribution Protocol for communication
    - between MPLS Switches
    - between MPLS Switch and MPLS Edge Router
- **Forwarding Component**
  - uses labels carried by packets plus label information maintained by a label switch (classical VC switching table) to perform packet forwarding -> "label swapping"

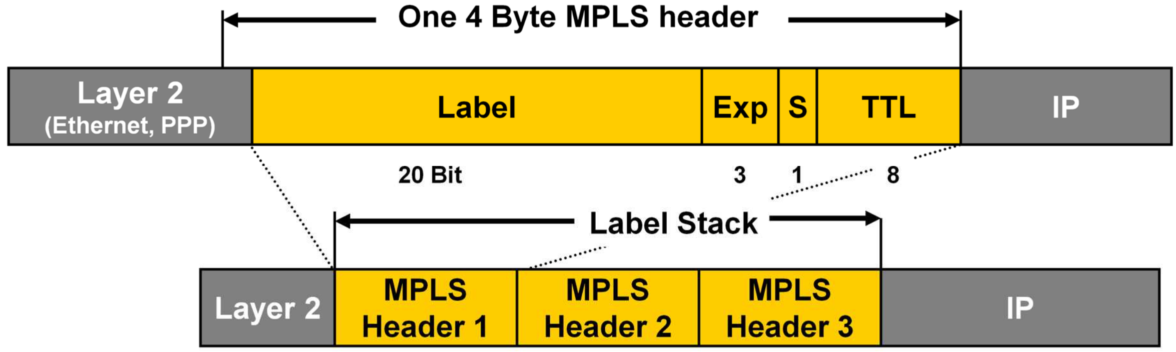
Appendix 3 - MPLS (v6.1)




MPLS Label Swapping
1a. Routing protocol (e.g. OSPF) establishes reachability to destination networks
1b. Label Distribution Protocol establishes MPLS paths (VC) along switching tables
4. Egress MPLS router at egress removes label and delivers packet
2. Ingress MPLS router receives packet, “labels” it and by sends it along a particular MPLS path (VC)
3. MPLS switches labeled packets using switching table
© 2016, D.I. Lindner
MPLS v6.1
65

Appendix 3 - MPLS (v6.1)

MPLS Header: Frame Mode



- **"Layer 2.5" can be used over Ethernet, 802.3 or PPP links**
  - note: 2.5 means 32 bit
  - 20-bit MPLS label (Label)
  - 3-bit experimental field (Exp)
    - could be copy of IP Precedence -> MPLS QoS like IP QoS with DiffServ Model based on DSCP
  - 1-bit bottom-of-stack indicator (S)
    - Labels could be stacked (Push & Pop)
    - MPLS switching performed always on the first label of the stack
  - 8-bit time-to-live field (TTL)

 MPLS v6.1 66

The MPLS Header is made up of four bytes and is located between the layer two header and the layer three header. The existence of an MPLS header is indicated by the layer two type field entry 0x8848.

The MPLS header is made up of a:

20 bit label field used for forwarding,

3 Experimental bits typically used to carry IP Precedence settings,

1 bit bottom of stack (0 indicates last label in the stack, 1 indicates there are some more labels on top of the bottom label)

TTL field in which by default the IP TTL value is copied to when a Label is inserted.

If MPLS is used on top of ATM, the VPI/VCI field of the standard ATM cell header is used to carry the label information. There is no additional MPLS header involved because this would require hardware changes if you want to migrate existing ATM devices to support MPLS.

Note: The labels 0 to 15 are reserved. Therefore the lowest usable label number is 16 and the highest possible label is 1,048,575 (which is actually 2^20-1). Only four out of the 16 reserved labels had been defined (RFC 3032), which are: 0 "IPv4 Explicit Null Label", 1 "Router Alert Label", 2 "IPv6 Explicit Null Label", 3 "Implicit Null Label".

Several reasons lead to a label stack. For example, with MPLS VPNs, the top label identifies the egress router while a second label identifies the VPN itself. Thus the egress router can (as

Appendix 3 - MPLS (v6.1)

soon as the packet arrived) pop the outermost label and forward the packet to the right interface according to the inner label. Another example is MPLS Traffic Engineering (TE), where the outer label points to the TE tunnel endpoint and the inner label to the final destination itself.

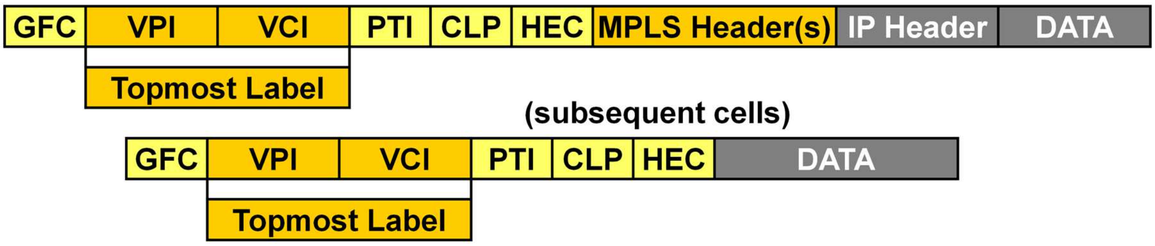
MPLS Header: Cell Mode

ATM Convergence Sublayer (CS):

- **ATM Switches can only switch VPI/VCI—no MPLS labels!**
  - Only the topmost label is inserted in the VPI/VCI field

ATM Segmentation and Reassembling Sublayer (SAR):

(first cell)



Labels and FEC

- **A label is used to identify a certain subset of packets**
  - which take the same MPLS path or which get the same forwarding treatment in the MPLS label switched network
  - The path is so called Label Switched Path (LSP)
    - "The MPLS Virtual Circuit"
- **Thus a label represents**
  - a so called Forwarding Equivalence Class (FEC)
- **The assignment of a packet to FEC**
  - is done just once by the MPLS Edge Router, as the packet enters the network
  - most commonly this is based on the IP network layer destination address

 MPLS v6.1 68

Label Binding

- **Two neighboring LSRs R1 and R2**
  - may agree that when R1 transmits a packet to R2, R1 will label with packet with label value L if and only if the packet is a member of a particular FEC F
- **They agree**
  - on a so called "binding" between label L and FEC F for packets moving from R1 to R2
- **As a result**
  - L becomes R1´s "outgoing label" or “remote label” representing FEC F
  - L becomes R2´s "incoming label" or “local label” representing FEC F

Creating and Destroying Label Binding 1

- **Control Driven (favored by IETF-WG)**
  - creation or deconstruction of labels is triggered by control information such as
    - OSPF routing, IS-IS routing
    - PIM Join/Prune messages in case of IP multicast routing
    - IntSrv RSVP messages in case of IP QoS IntSrv Model
    - DiffSrv Traffic Engineering in Case of IP QoS DiffSrv Model
  - hence we have a pre-assignment of labels based on reachability information
    - and optionally based on QoS needs
  - also called Topology Driven

Creating and Destroying Label Binding 2

- **Data Driven**
  - creation or deconstruction of labels is triggered by data packets
    - but only if a critical threshold number of packets for a specific communication relationship is reached
    - may have a big performance impact
  - hence we have dynamic assignment of labels based on data flow detection
  - also called Traffic Driven

Some FEC Examples for Topology Driven

- **FECs could be for example**
  - a set of unicast packets whose network layer destination address matches a particular IP address prefix
    - MPLS application: Destination Based (Unicast) Routing
  - a set of multicast packets with the same source and destination network layer address
    - MPLS application: Multicast Routing
  - a set of unicast packets whose network layer destination address matches a particular IP address prefix and whose Type of Service (ToS) or DSCP bits are the same
    - MPLS application: Quality of Service
    - MPLS application: Traffic Engineering or Constraint Based Routing

Label Distribution

- **MPLS architecture allows an LSR to distribute bindings to LSRs that have not explicitly requested them**
  - "Unsolicited Downstream" label distribution
  - usually used by Frame-Mode MPLS
- **MPLS architecture allows an LSR to explicitly request, from its next hop for a particular FEC, a label binding for that FEC**
  - "Downstream-On-Demand" label distribution
  - must be used by Cell-Mode MPLS

 MPLS v6.1 73

Label Binding

- **The decision to bind a particular label L to a particular FEC F**
  - is made by the LSR which is DOWNSTREAM with respect to that binding
  - the downstream LSR then informs the upstream LSR of the binding
  - thus labels are "downstream-assigned“
  - thus label bindings are distributed in the "downstream to upstream“ direction
- **Discussion were about if**
  - labels should also be “upstream-assigned“
  - not any longer part of current MPLS-RFC

Label Retention Mode 1

- **A LSR may receive a label binding**
  - for a particular FEC from another LSR, which is not next hop based on the routing table for that FEC
- **This LSR then has the choice**
  - of whether to keep track of such bindings, or whether to discard such bindings
- **A LSR supports "Liberal Label Retention Mode"**
  - if it maintains the bindings between a label and a FEC which are received from LSR´s which are not its next hop for that FEC

Label Retention Mode 2

- **A LSR supports "Conservative Label Retention mode "**
  - If it discards the bindings between a label and a FEC which are received from LSR´s which are not its next hop for that FEC
- **Liberal Label Retention mode**
  - allows for quicker adaptation to routing changes
  - LSR can switch over to next best LSP
- **Conservative Label Retention mode**
  - requires an LSR to maintain fewer labels
  - LSR has to wait for new label bindings in case of topology changes

 MPLS v6.1 76

Independent versus Ordered Control

- **Independent Control:**
  - each LSR may make an independent decision to assign a a label to a FEC and to advertise the assignment to its neighbors
  - typically used in Frame-Mode MPLS for destination based routing
  - loop prevention must be done by other means (-> MPLS TTL) but there is faster convergence
- **Ordered Control:**
  - label assignment proceeds in an orderly fashion from one end of a LSP to the other
  - under ordered control, LSP setup may be initiated by the ingress (header) or egress (tail) MPLS Edge Router

Ordered Control - Egress

- in case of egress method the only LSR which can initiate the process of label assignment is the egress LSR
- a LSR knows that it is the egress for a given FEC if its next hop for this FEC is not an LSR
- this LSR will sent a label advertisement to all neighboring LSRs
- a neighboring LSR receiving such a label advertisement from a interface which is the next hop to a given FEC will assign its own label and advertise it to all other neighboring LSRs
- inherent loop prevention
- slower convergence

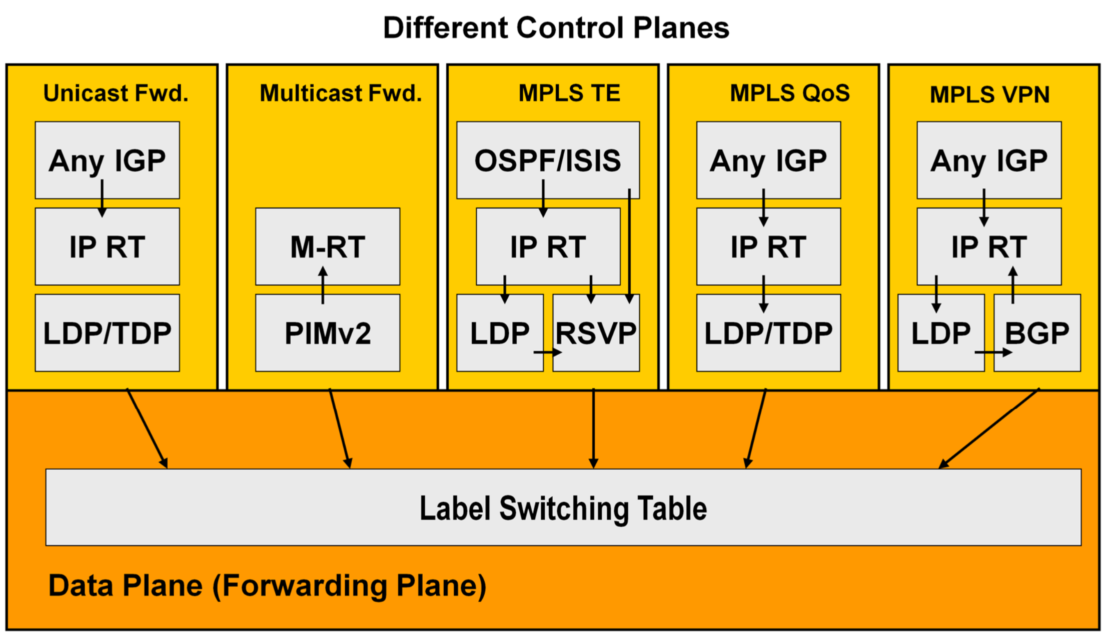
Ordered Control - Ingress

- in case of ingress method the LSR which initiates the process of label assignment is the ingress LSR
- the ingress LSR constructs a source route and pass on requests for label bindings to the next LSR
- this is done until LSR which is the end of the source route is reached
- from this LSR label bindings will flow upstream to the ingress LSR
- used for MPLS Traffic Engineering (TE)

 MPLS v6.1 79

Appendix 3 - MPLS (v6.1)

MPLS Applications and MPLS Control Plane



© 2016, D.I. Lindner MPLS v6.1 80

The diagram above illustrates how different MPLS applications use a different control plane. It is in fact the control plane which determines the FECs—in other words, what label-based forwarding is good for.

But all applications use the same (primitive) data plane.

Note that there are different types of MPLS-based Multicast. MPLS Multicast is discussed in another chapter, soon...

Agenda

- **Review ATM**
- **IP over WAN Problems (Traditional Approach)**
- **MPLS Principles**
- **Label Distribution Methods**
  - Unsolicited Downstream
  - Downstream On Demand
  - MPLS and ATM, VC Merge Problem
- **MPLS Details (Cisco)**
- **RFCs**

 MPLS v6.1 81

Appendix 3 - MPLS (v6.1)

Routing Table Created by Routing Protocol

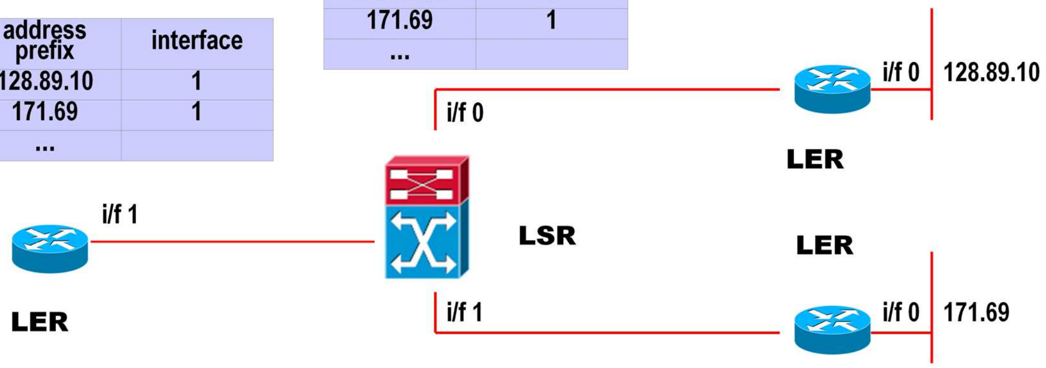
FEC Label Binding:
Control Driven
Destination Based Routing

Routing Table

address prefix	interface
128.89.10	0
...	

address prefix	interface
128.89.10	0
171.69	1
...	

address prefix	interface
128.89.10	1
171.69	1
...	

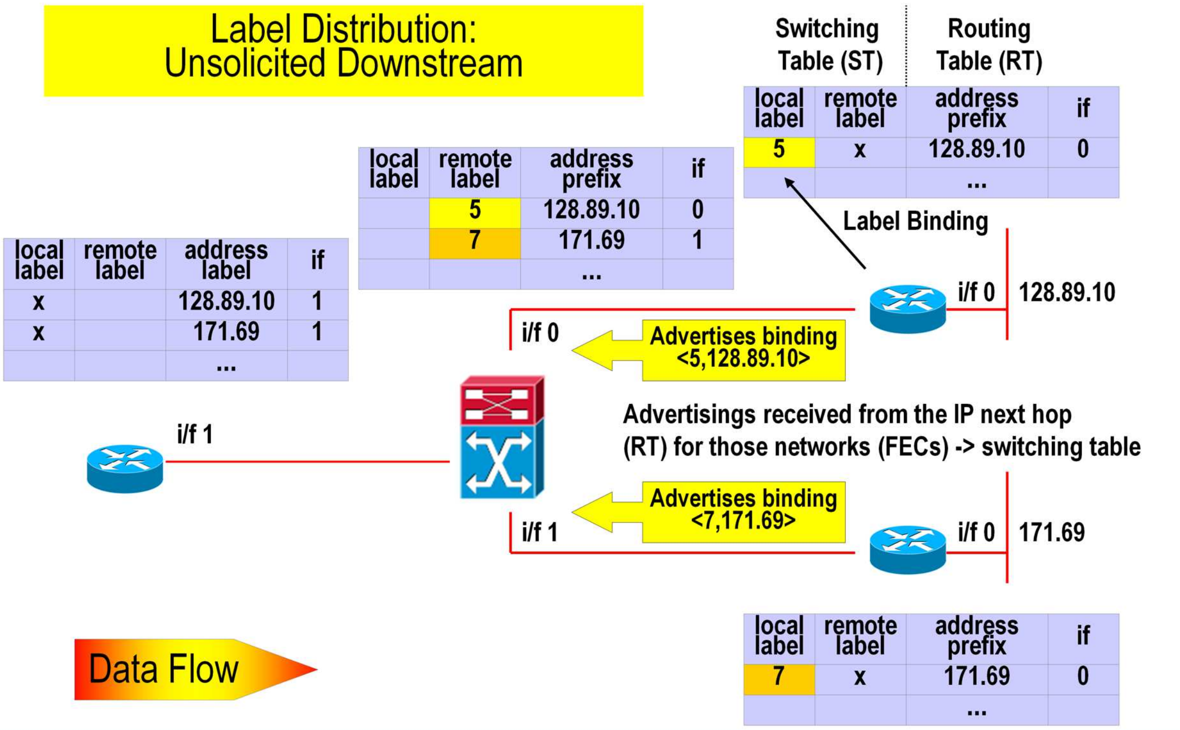


Data Flow

address prefix	interface
171.69	0
...	

Appendix 3 - MPLS (v6.1)

Labels Sent by LDP



local label	remote label	address label	if
x		128.89.10	1
x		171.69	1
		...	

local label	remote label	address prefix	if
	5	128.89.10	0
	7	171.69	1
		...	

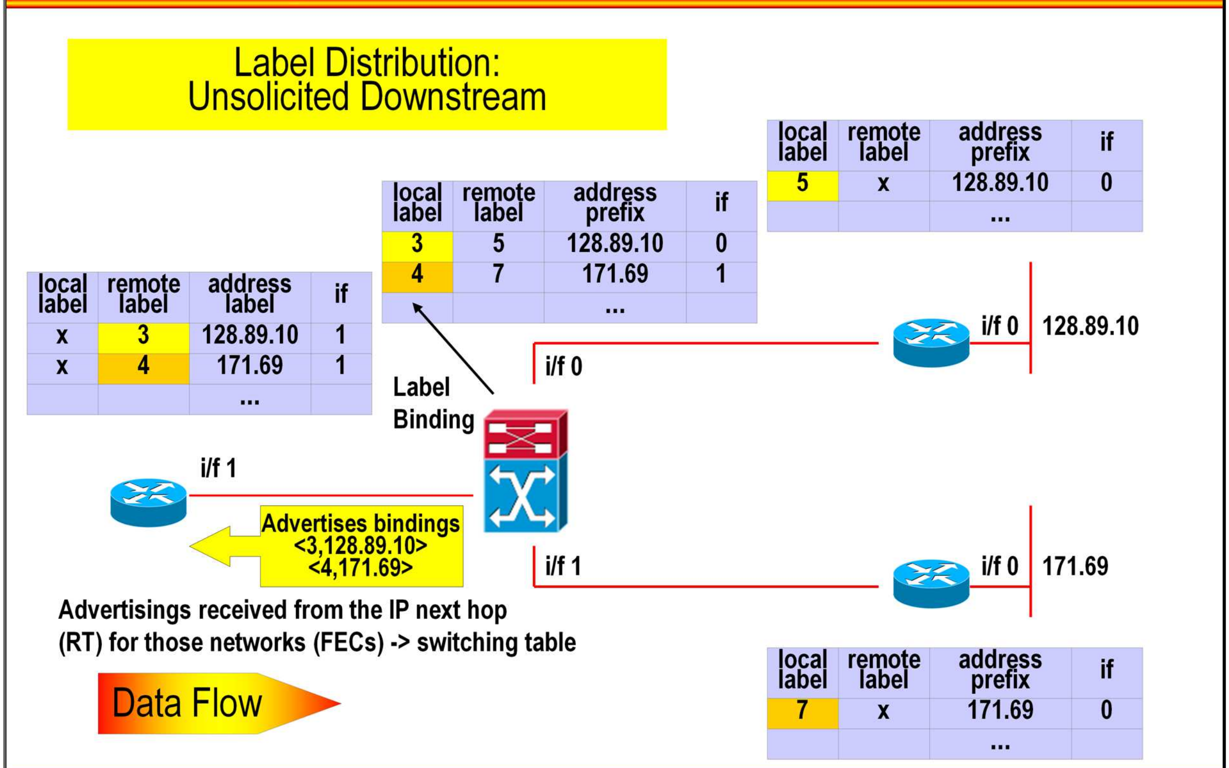
local label	remote label	address prefix	if
5	x	128.89.10	0
		...	

local label	remote label	address prefix	if
7	x	171.69	0
		...	

 MPLS v6.1 83

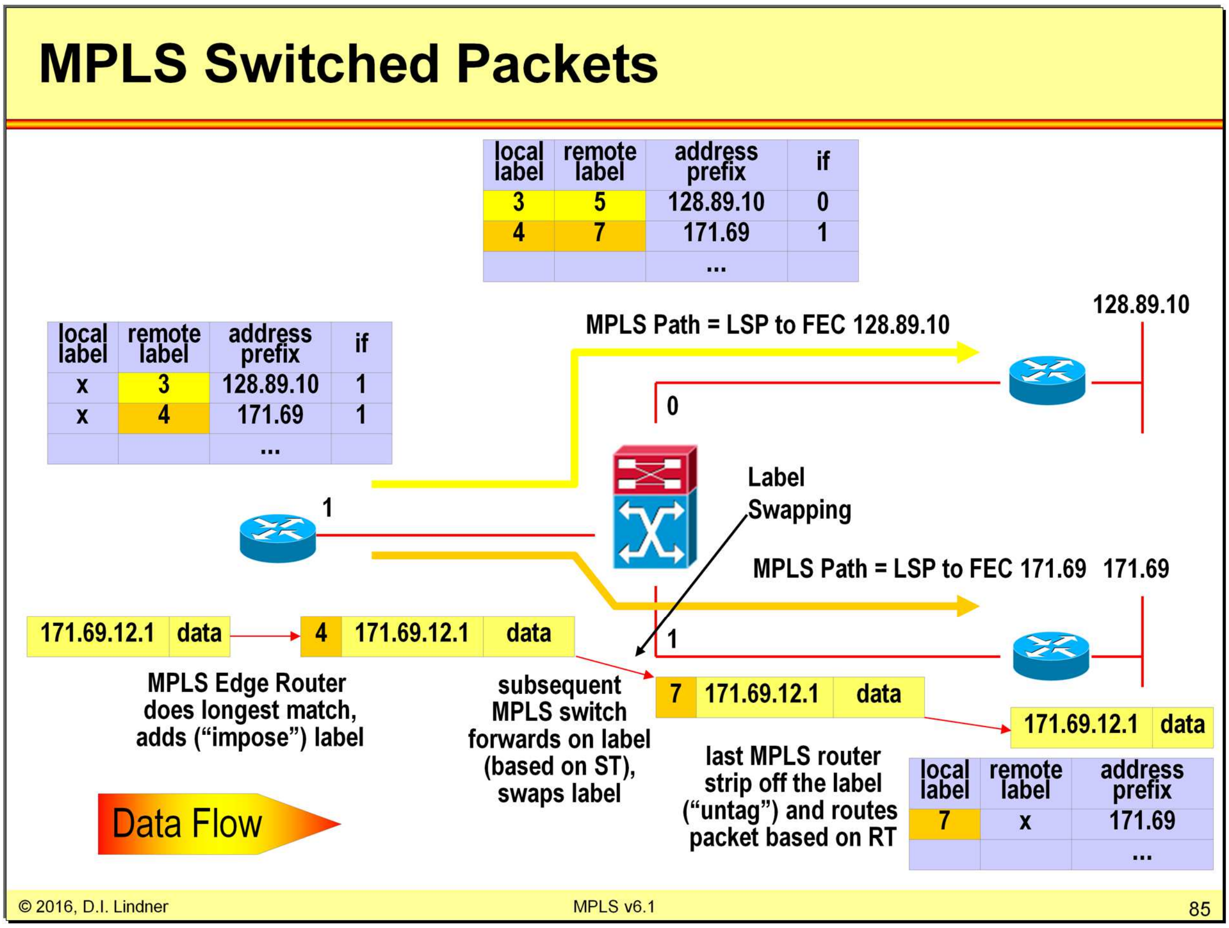
Appendix 3 - MPLS (v6.1)

Labels Sent and Switching Table Entry Created by MPLS Switch



© 2016, D.I. Lindner MPLS v6.1 84

Appendix 3 - MPLS (v6.1)



Appendix 3 - MPLS (v6.1)

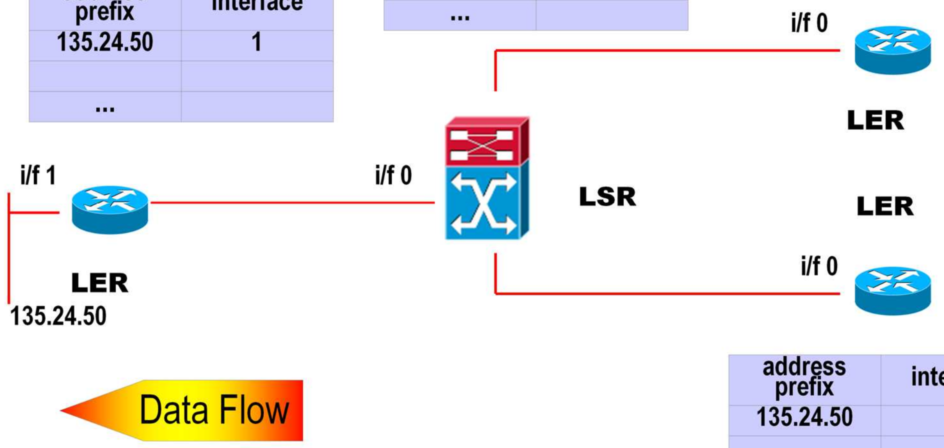
Routing Table Created by Routing Protocol

FEC Label Binding:
Control Driven
Destination Based Routing

address prefix	interface
135.24.50	0
...	

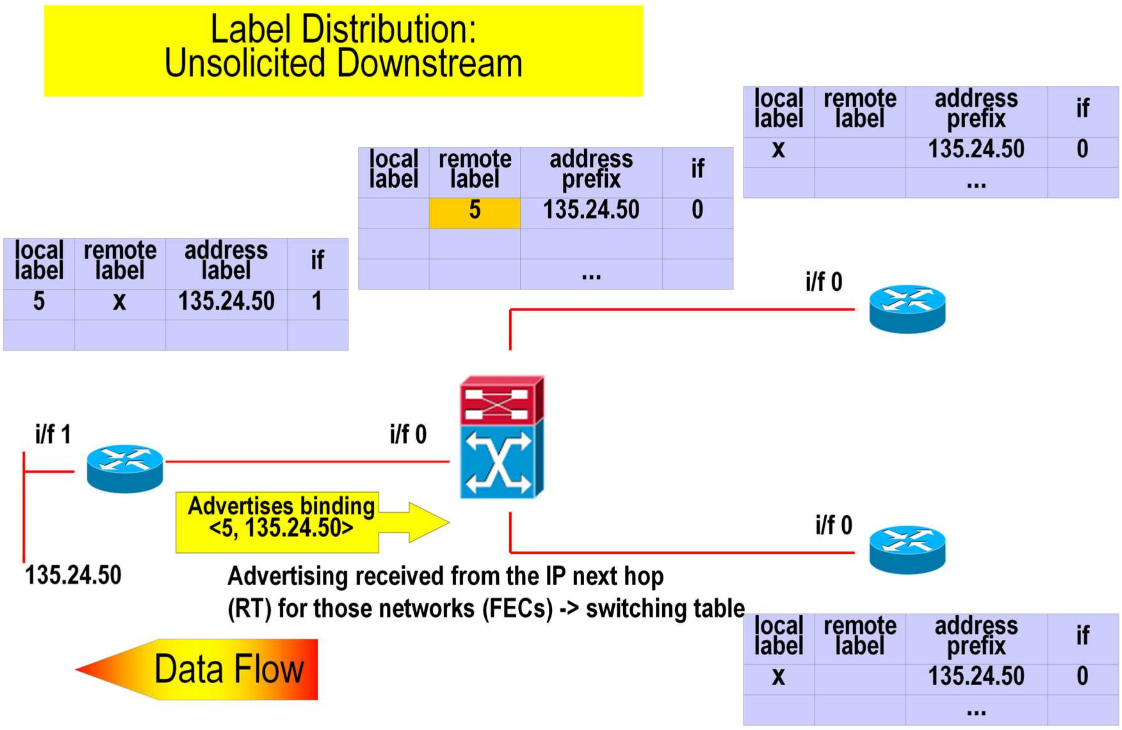
address prefix	interface
135.24.50	0
...	

address prefix	interface
135.24.50	1
...	



address prefix	interface
135.24.50	0
...	

Labels Sent by LDP



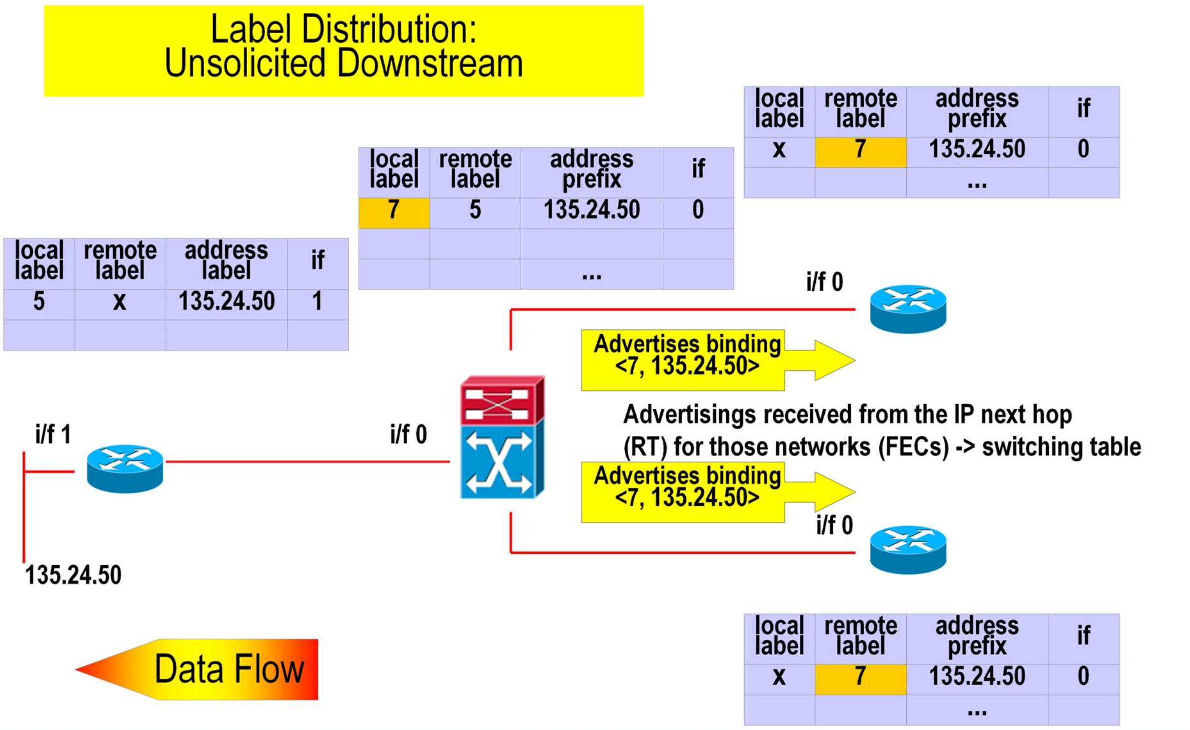
local label	remote label	address label	if
5	x	135.24.50	1

local label	remote label	address prefix	if
	5	135.24.50	0
		...	

local label	remote label	address prefix	if
x		135.24.50	0
		...	

local label	remote label	address prefix	if
x		135.24.50	0
		...	

Labels Sent and Switching Table Entry Created by MPLS Switch



local label	remote label	address label	if
5	x	135.24.50	1

local label	remote label	address prefix	if
7	5	135.24.50	0
		...	

local label	remote label	address prefix	if
x	7	135.24.50	0
		...	

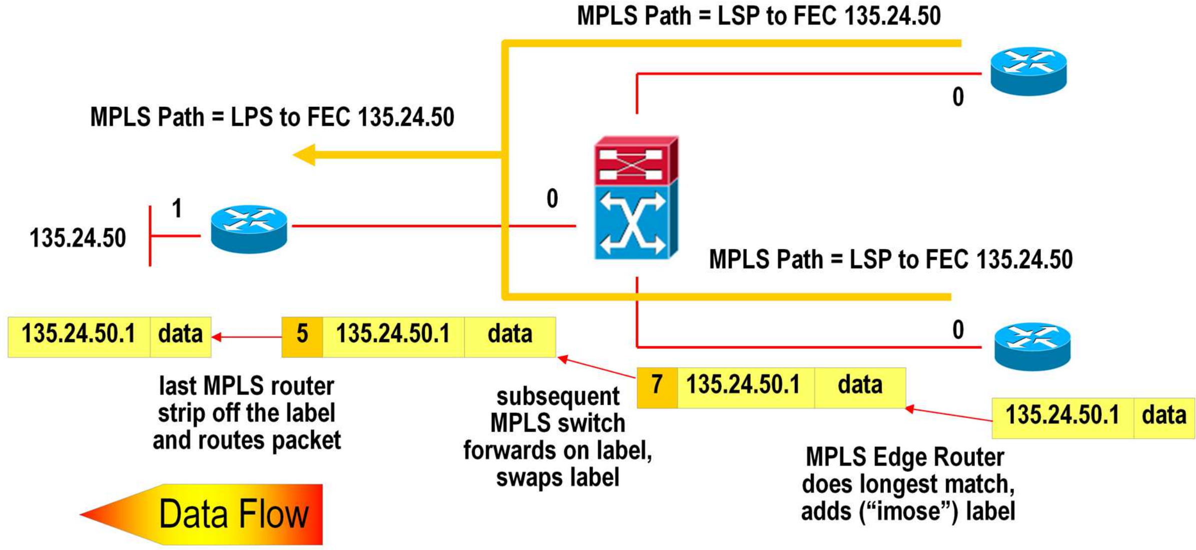
local label	remote label	address prefix	if
x	7	135.24.50	0
		...	

Label Merging - LSP Merging

local label	remote label	address label	if
5	x	135.24.50	1

local label	remote label	address prefix	if
7	5	135.24.50	0
		...	

local label	remote label	address prefix	if
x	7	135.24.50	0
		...	



 MPLS v6.1 89

Agenda

- **Review ATM**
- **IP over WAN Problems (Traditional Approach)**
- **MPLS Principles**
- **Label Distribution Methods**
  - Unsolicited Downstream
  - Downstream On Demand
  - MPLS and ATM, VC Merge Problem
- **MPLS Details (Cisco)**
- **RFCs**

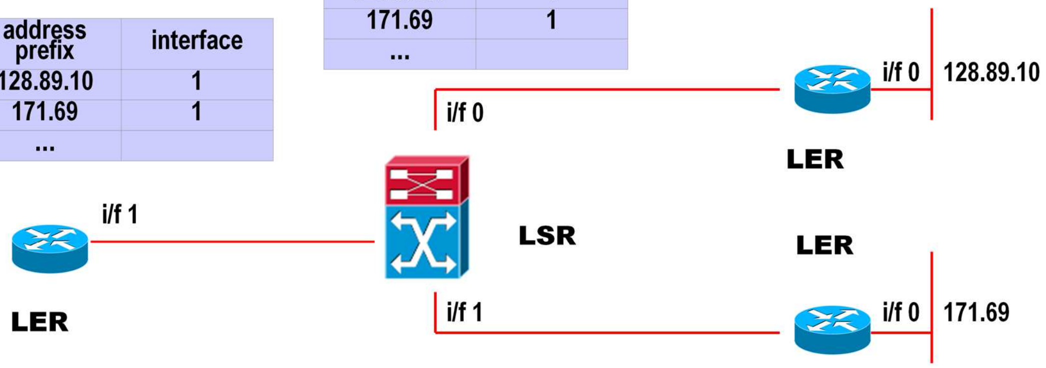
Routing Table Created by Routing Protocol

FEC Label Binding:
Control Driven
Destination Based Routing

address prefix	interface
128.89.10	0
...	

address prefix	interface
128.89.10	0
171.69	1
...	

address prefix	interface
128.89.10	1
171.69	1
...	



address prefix	interface
171.69	0
...	

Appendix 3 - MPLS (v6.1)

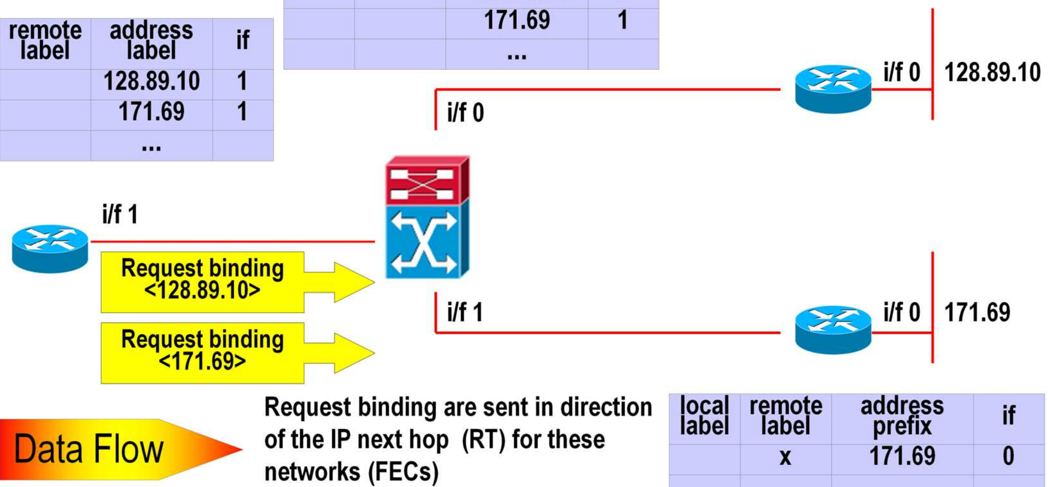
Labels Requested by MPLS Edge Routers

Label Distribution: Downstream-On-Demand

local label	remote label	address prefix	if
	x	128.89.10	0
		...	

local label	remote label	address prefix	if
		128.89.10	0
		171.69	1
		...	

local label	remote label	address label	if
x		128.89.10	1
x		171.69	1
		...	



Request binding are sent in direction of the IP next hop (RT) for these networks (FECs)

local label	remote label	address prefix	if
	x	171.69	0
		...	

Appendix 3 - MPLS (v6.1)

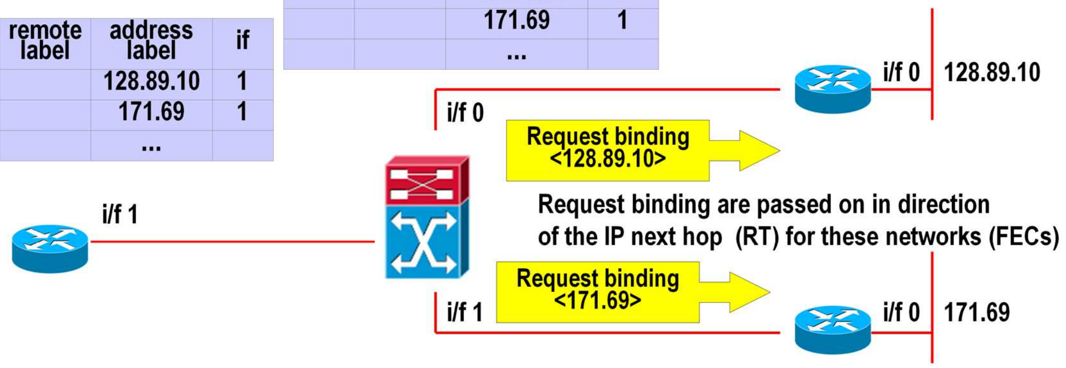
Labels Requested by MPLS Switch

Label Distribution:
Downstream-On-Demand

local label	remote label	address prefix	if
	x	128.89.10	0
		...	

local label	remote label	address prefix	if
		128.89.10	0
		171.69	1
		...	

local label	remote label	address label	if
x		128.89.10	1
x		171.69	1
		...	



local label	remote label	address prefix	if
	x	171.69	0
		...	

Data Flow

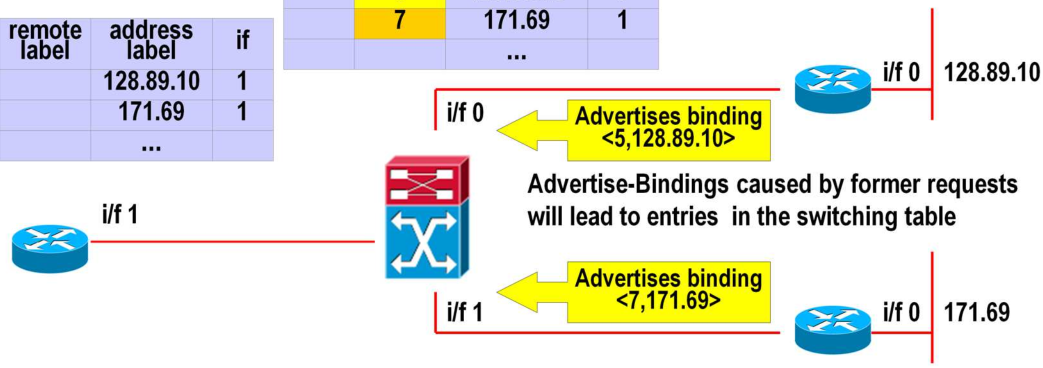
Labels Allocated by MPLS Edge Router

Label Distribution: Downstream-On-Demand

local label	remote label	address prefix	if
5	x	128.89.10	0
		...	

local label	remote label	address prefix	if
	5	128.89.10	0
	7	171.69	1
		...	

local label	remote label	address label	if
x		128.89.10	1
x		171.69	1
		...	



local label	remote label	address prefix	if
7	x	171.69	0
		...	

Data Flow

Appendix 3 - MPLS (v6.1)

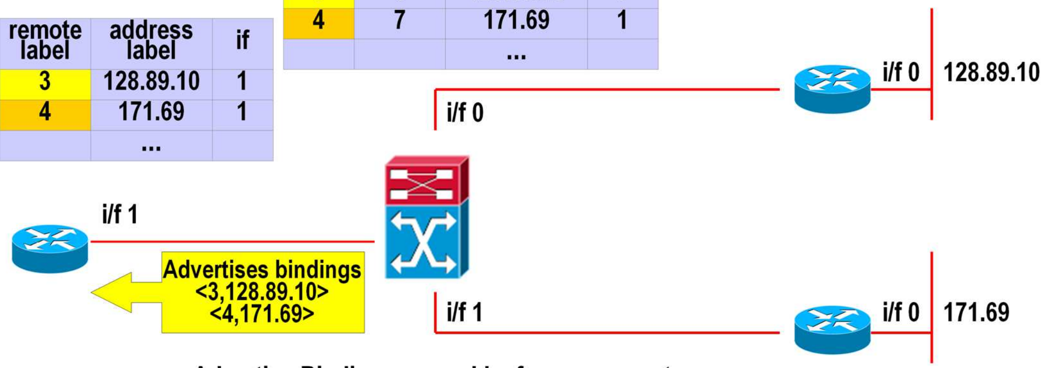
Labels Allocated and Switching Table Built by MPLS Switch

Label Distribution:
Downstream-On-Demand

local label	remote label	address prefix	if
5	x	128.89.10	0
		...	

local label	remote label	address prefix	if
3	5	128.89.10	0
4	7	171.69	1
		...	

local label	remote label	address label	if
x	3	128.89.10	1
x	4	171.69	1
		...	

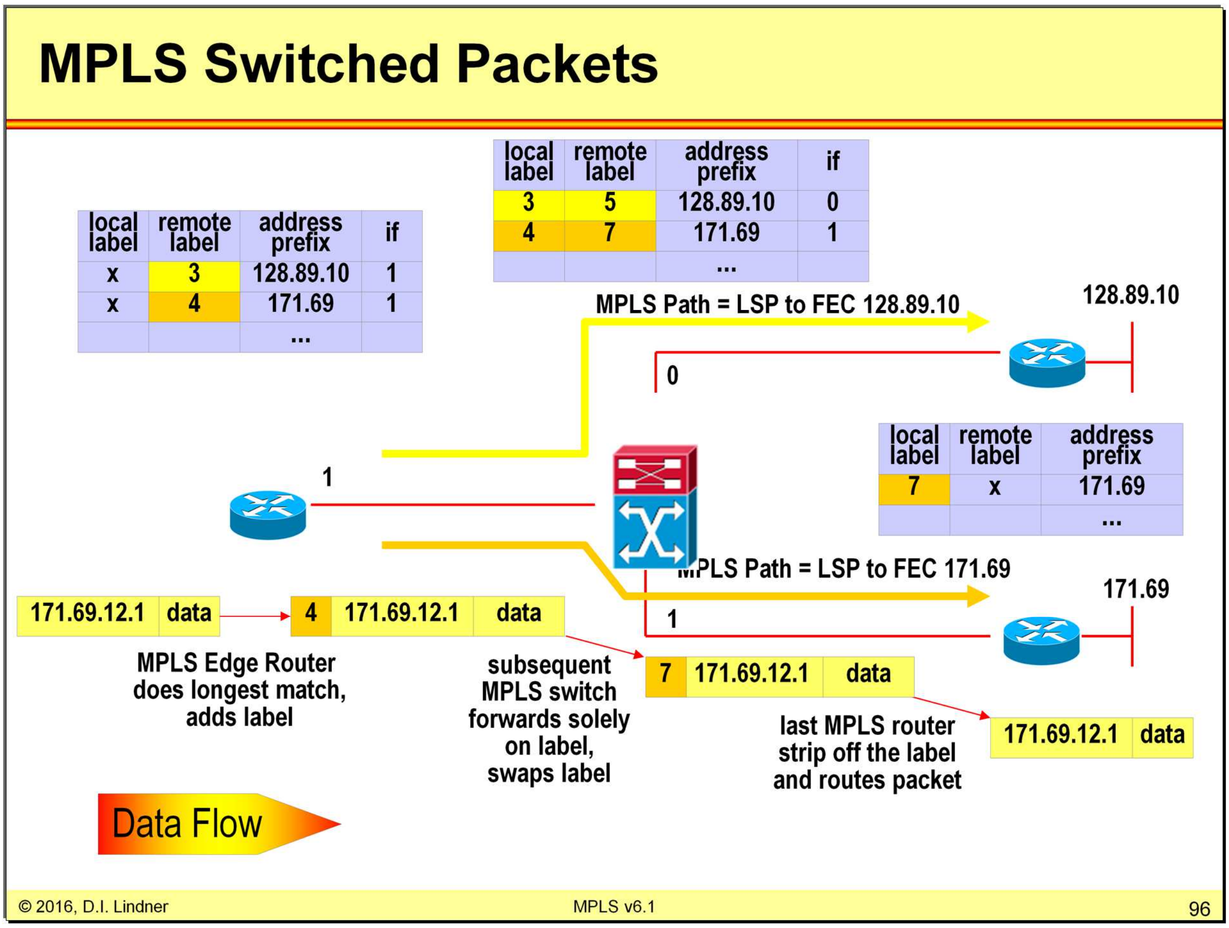


Advertise-Bindings caused by former requests will lead to entries in the switching table

local label	remote label	address prefix	if
7	x	171.69	0
		...	

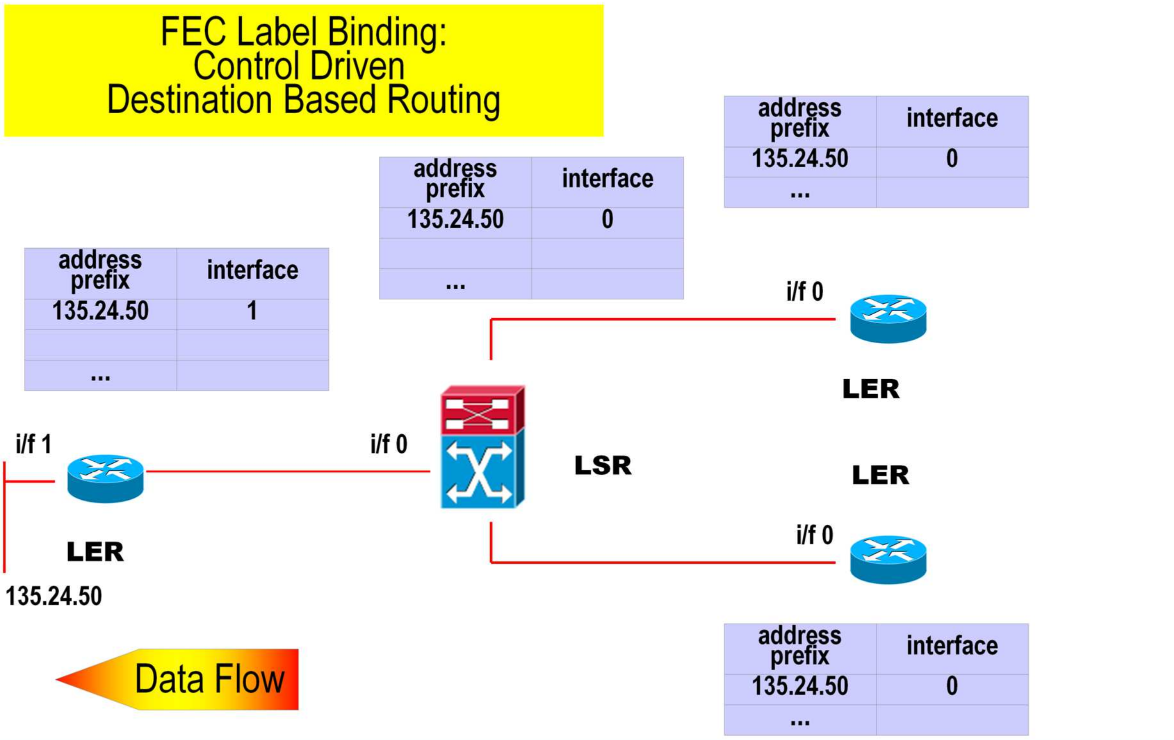
Data Flow

Appendix 3 - MPLS (v6.1)



Appendix 3 - MPLS (v6.1)

Routing Table Created by Routing Protocol



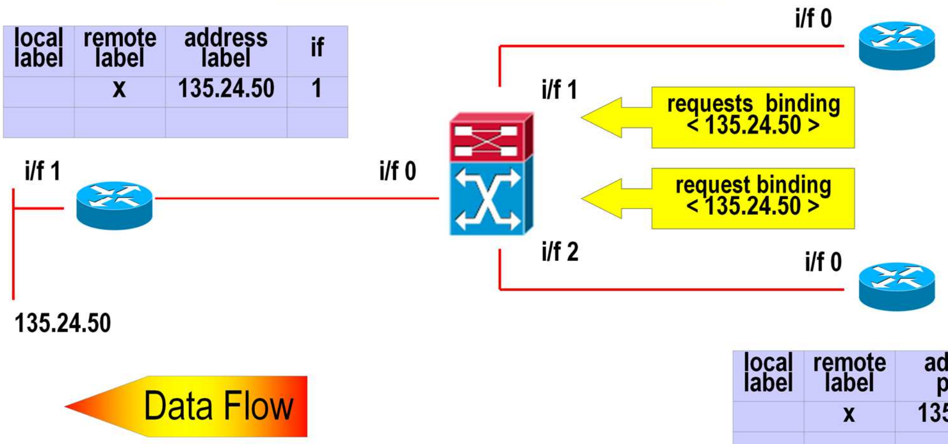
Labels Requested by MPLS Edge Routers

Label Distribution:
Downstream-On-Demand

local label	remote label	address prefix	if
	x	135.24.50	0
		...	

in-if	local label	remote label	address prefix	out-if
1			135.24.50	0
2				
			...	

local label	remote label	address label	if
	x	135.24.50	1



local label	remote label	address prefix	if
	x	135.24.50	0
		...	

Appendix 3 - MPLS (v6.1)

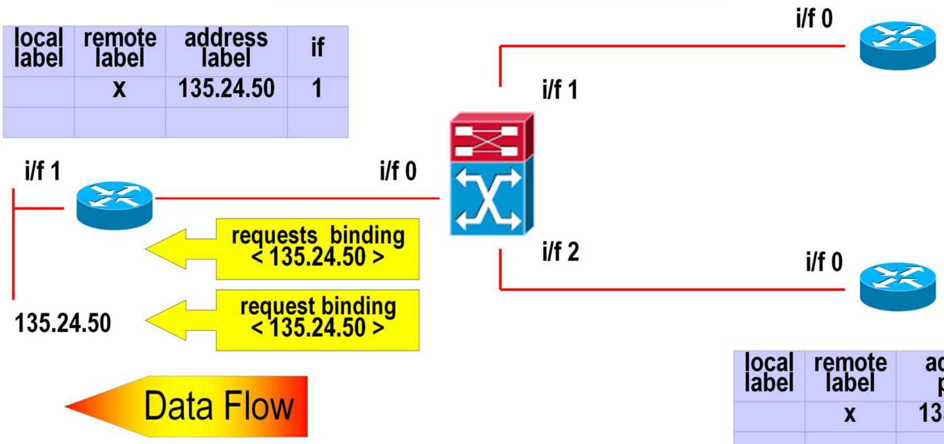
Labels Requested by MPLS Switch

Label Distribution: Downstream-On-Demand

local label	remote label	address prefix	if
	x	135.24.50	0
		...	

in-if	local label	remote label	address prefix	out-if
1			135.24.50	0
2				
			...	

local label	remote label	address label	if
	x	135.24.50	1



local label	remote label	address prefix	if
	x	135.24.50	0
		...	

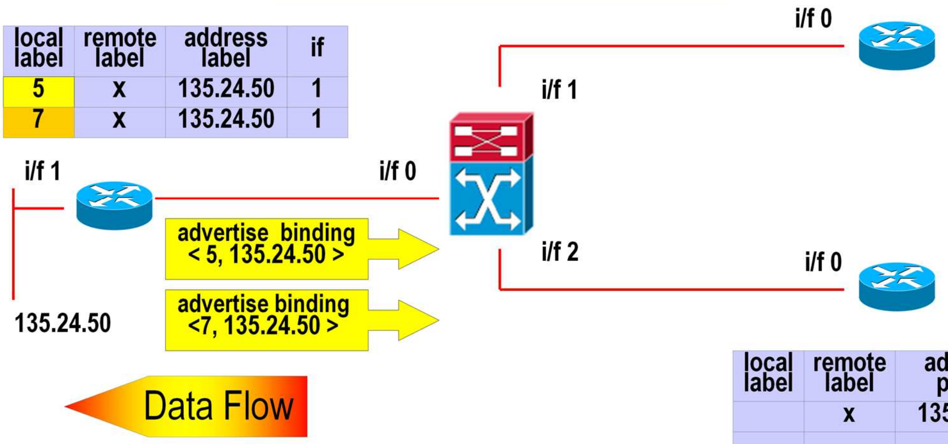
Labels Allocated by MPLS Edge Router

Label Distribution:
Downstream-On-Demand

local label	remote label	address prefix	if
	x	135.24.50	0
		...	

in-if	local label	remote label	address prefix	out-if
1		5	135.24.50	0
2		7	135.24.50	0
			...	

local label	remote label	address label	if
5	x	135.24.50	1
7	x	135.24.50	1



local label	remote label	address prefix	if
	x	135.24.50	0
		...	

Appendix 3 - MPLS (v6.1)

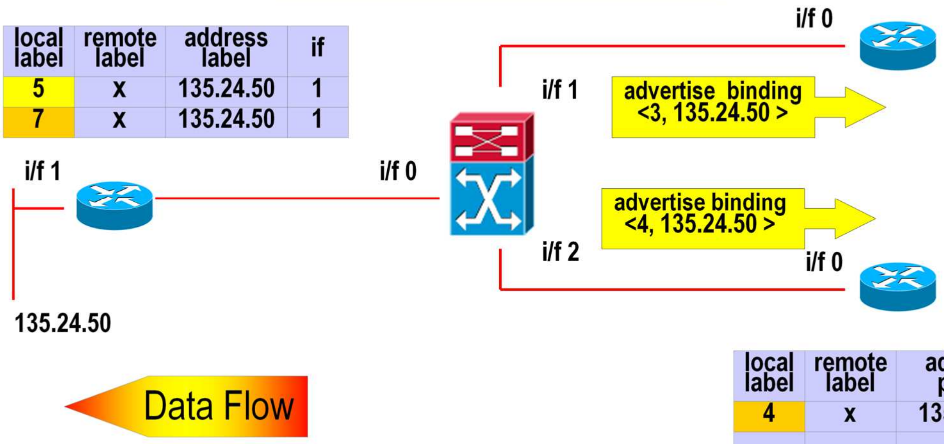
Labels Allocated and Switching Table Built by MPLS Switch

Label Distribution:
Downstream-On-Demand

local label	remote label	address prefix	if
3	x	135.24.50	0
		...	

in-if	local label	remote label	address prefix	out-if
1	3	5	135.24.50	0
2	4	7	135.24.50	0
			...	

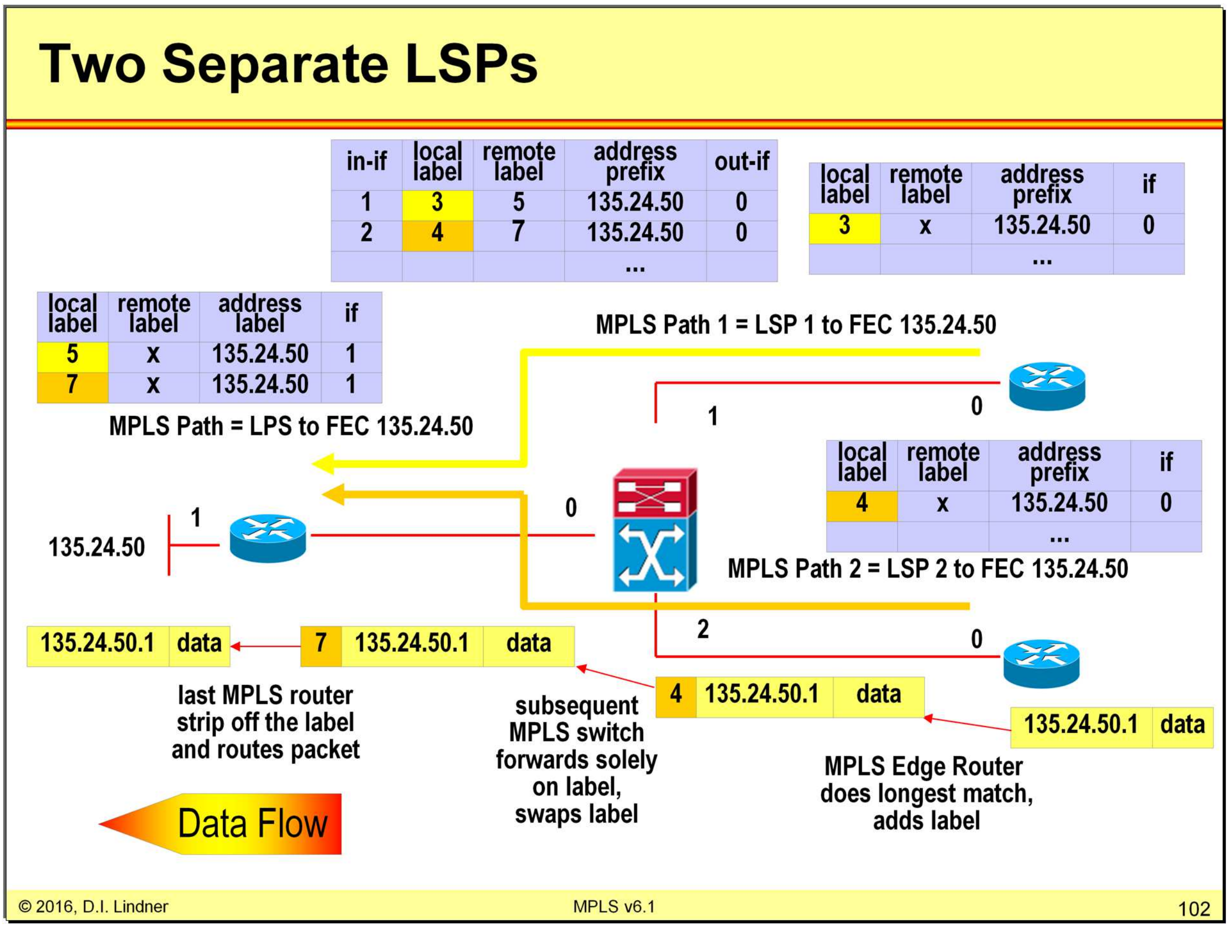
local label	remote label	address label	if
5	x	135.24.50	1
7	x	135.24.50	1



local label	remote label	address prefix	if
4	x	135.24.50	0
		...	

 MPLS v6.1 101

Appendix 3 - MPLS (v6.1)



Agenda

- Review ATM
- IP over WAN Problems (Traditional Approach)
- MPLS Principles
- Label Distribution Methods
  - Unsolicited Downstream
  - Downstream On Demand
  - MPLS and ATM, VC Merge Problem
- MPLS Details (Cisco)
- RFCs

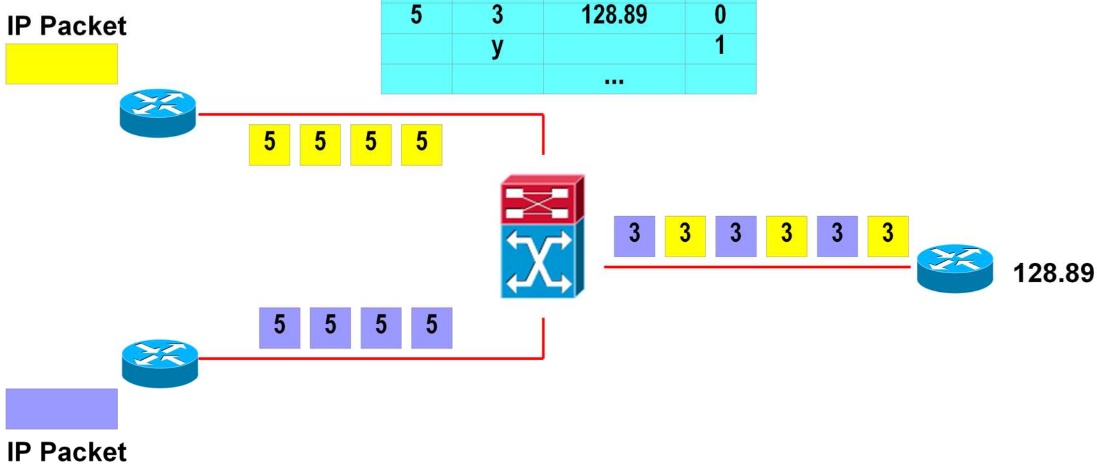
Label Switching and ATM

- **Can be easily deployed with ATM because ATM uses label swapping**
  - VPI/VCI is used as a label
- **ATM switches needs to implement control component of label switching**
  - ATM attached router peers with ATM switch (label switch)
    - exchange label binding information
- **Differences**
  - how labels are set up
    - label distribution -> downstream on demand allocation
  - label merging
    - in order to scale, merging of multiple streams (labels) into one stream (label) is required

Appendix 3 - MPLS (v6.1)

Label Switching and ATM

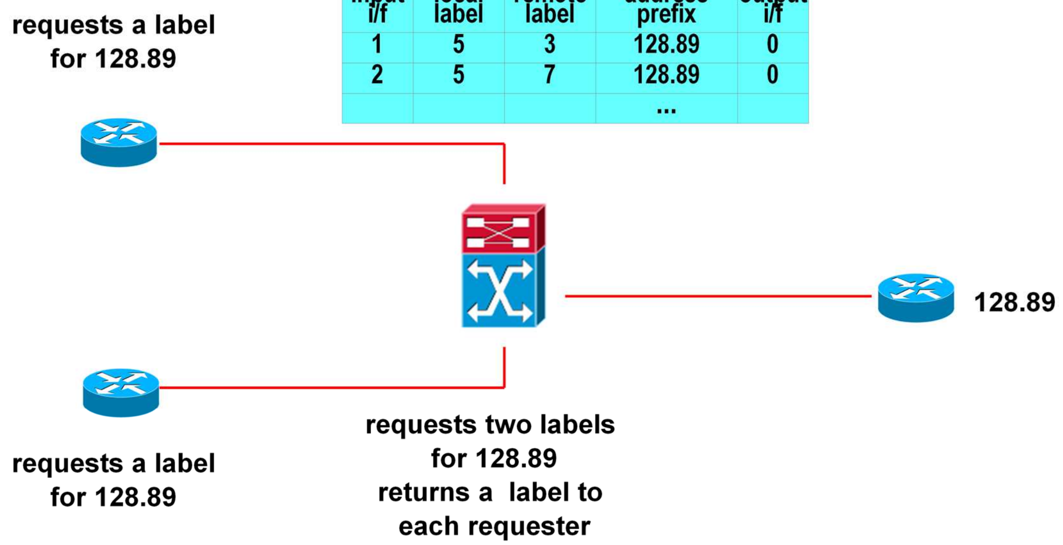
local label	remote label	address prefix	if
5	3	128.89	0
	y		1
		...	



ATM switch interleaves cells of different packets onto same label.
That is a problem in case of AAL5 encapsulation.
No problem in case of AAL3/AAL4 encapsulation because of AAL3/AAL4´s inherent multiplexing capability.

Label Distribution Solution for ATM

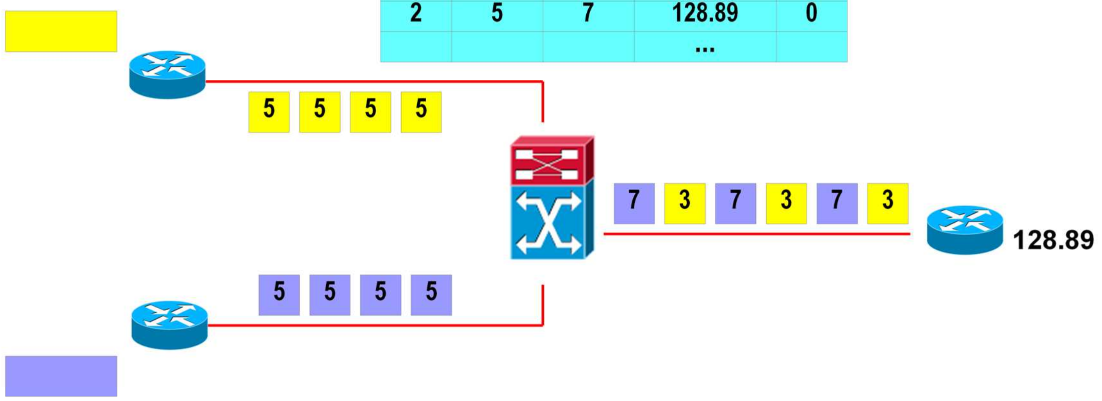
input i/f	local label	remote label	address prefix	output i/f
1	5	3	128.89	0
2	5	7	128.89	0
			...	



- "Downstream On Demand" Label Distribution

Label Distribution Solution for ATM

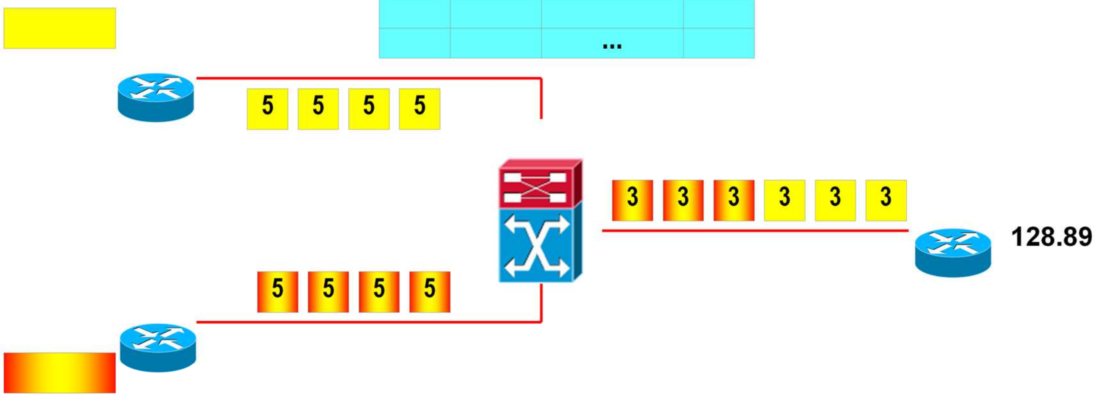
input i/f	local label	remote label	address prefix	output i/f
1	5	3	128.89	0
2	5	7	128.89	0
			...	



- **Downstream On Demand label distribution is necessary**
  - multiple labels per FEC may be assigned
  - one label per (ingress, egress) router pair
- **Label space can be reduced with VC-merge technique**

VC Merge Technique

local label	remote label	address prefix	if
5	3	128.89	0
		...	

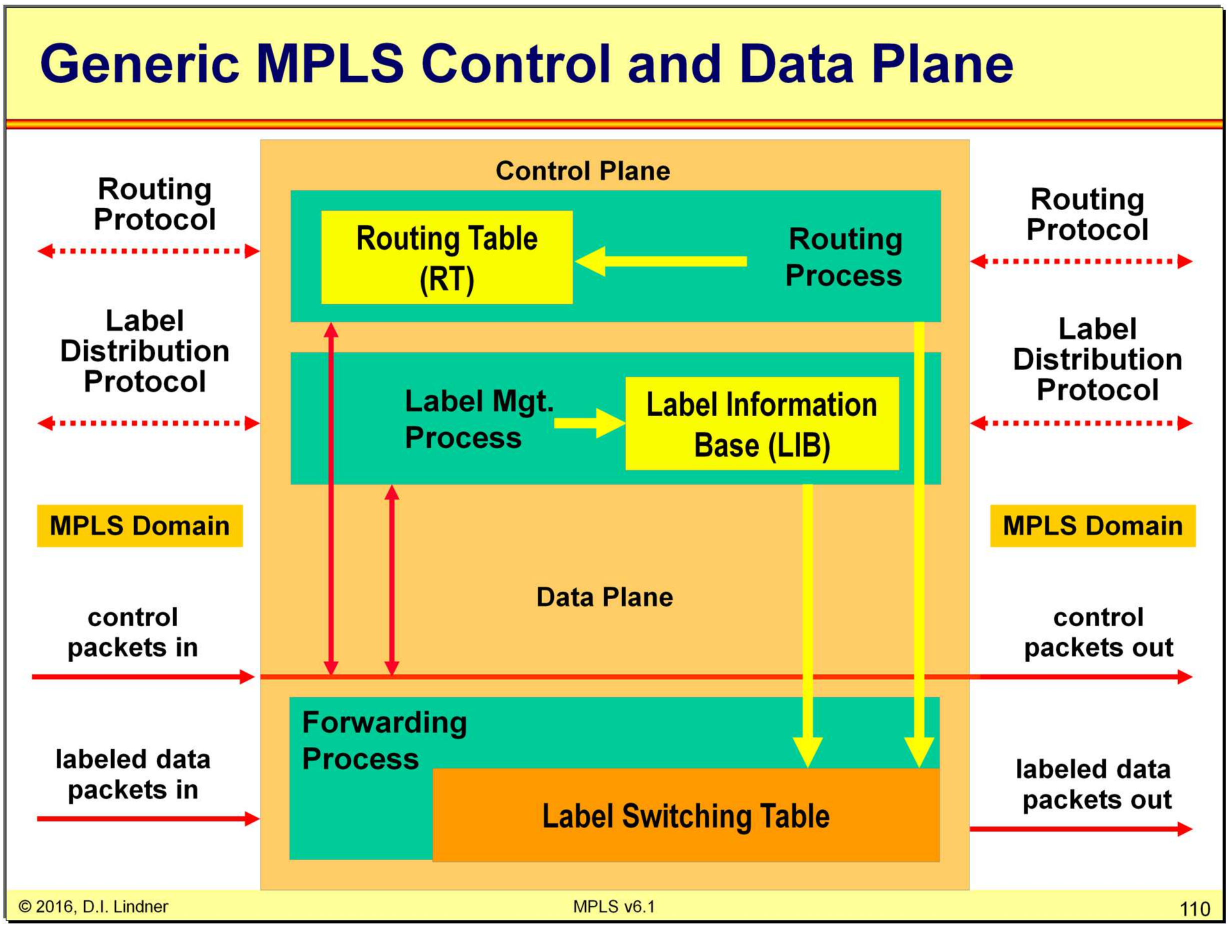


- **ATM switch avoids interleaving of frames**
  - VC Merge technique
  - looking for AAL5 trailers and storing corresponding cells of a frame until AAL5 trailer is seen

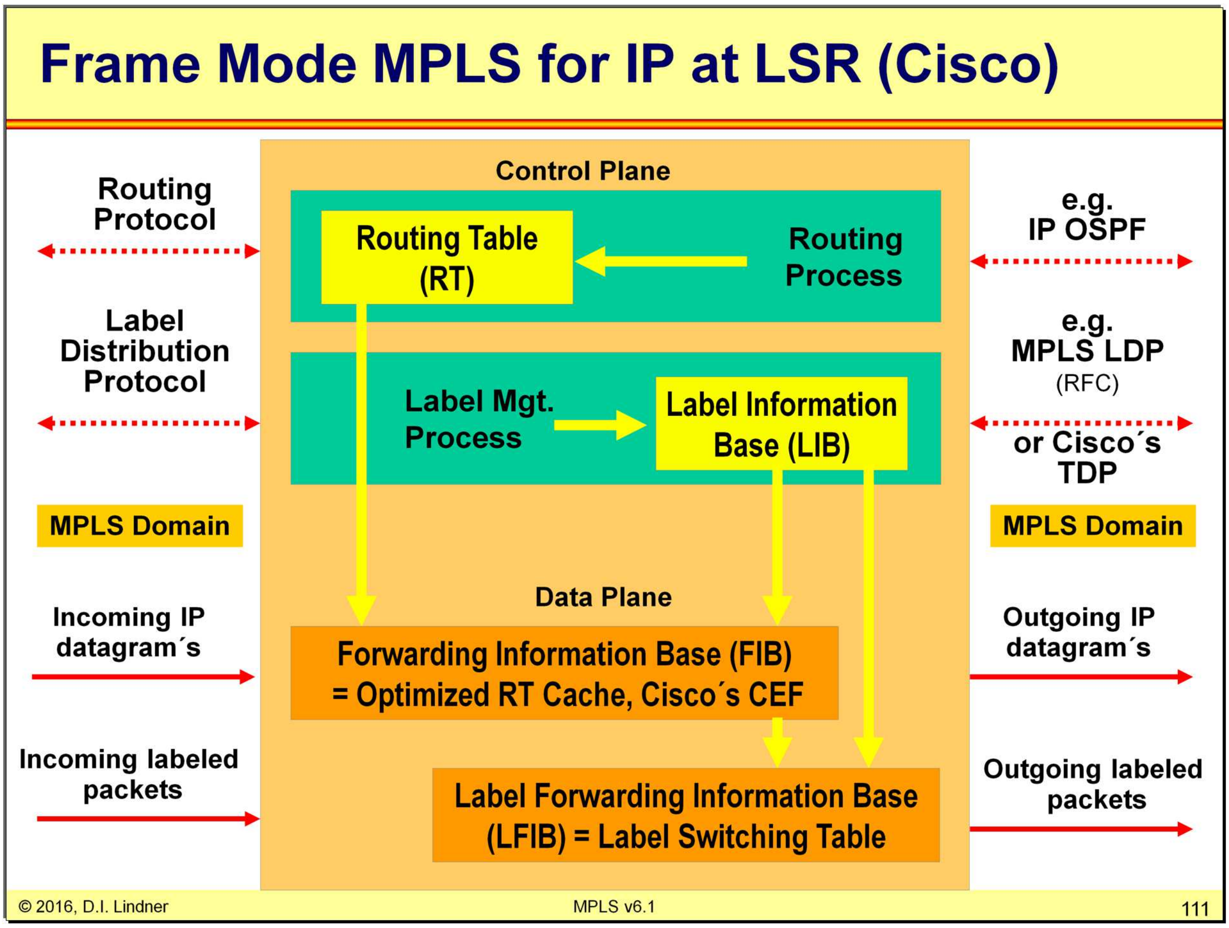
Agenda

- **Review ATM**
- **IP over WAN Problems (Traditional Approach)**
- **MPLS Principles**
- **Label Distribution Methods**
- **MPLS Details (Cisco)**
  - Internal Components
  - MPLS in Action
  - TDP, LDP
  - TTL
  - Traffic Engineering
  - MPLS and BGP
- **RFCs**

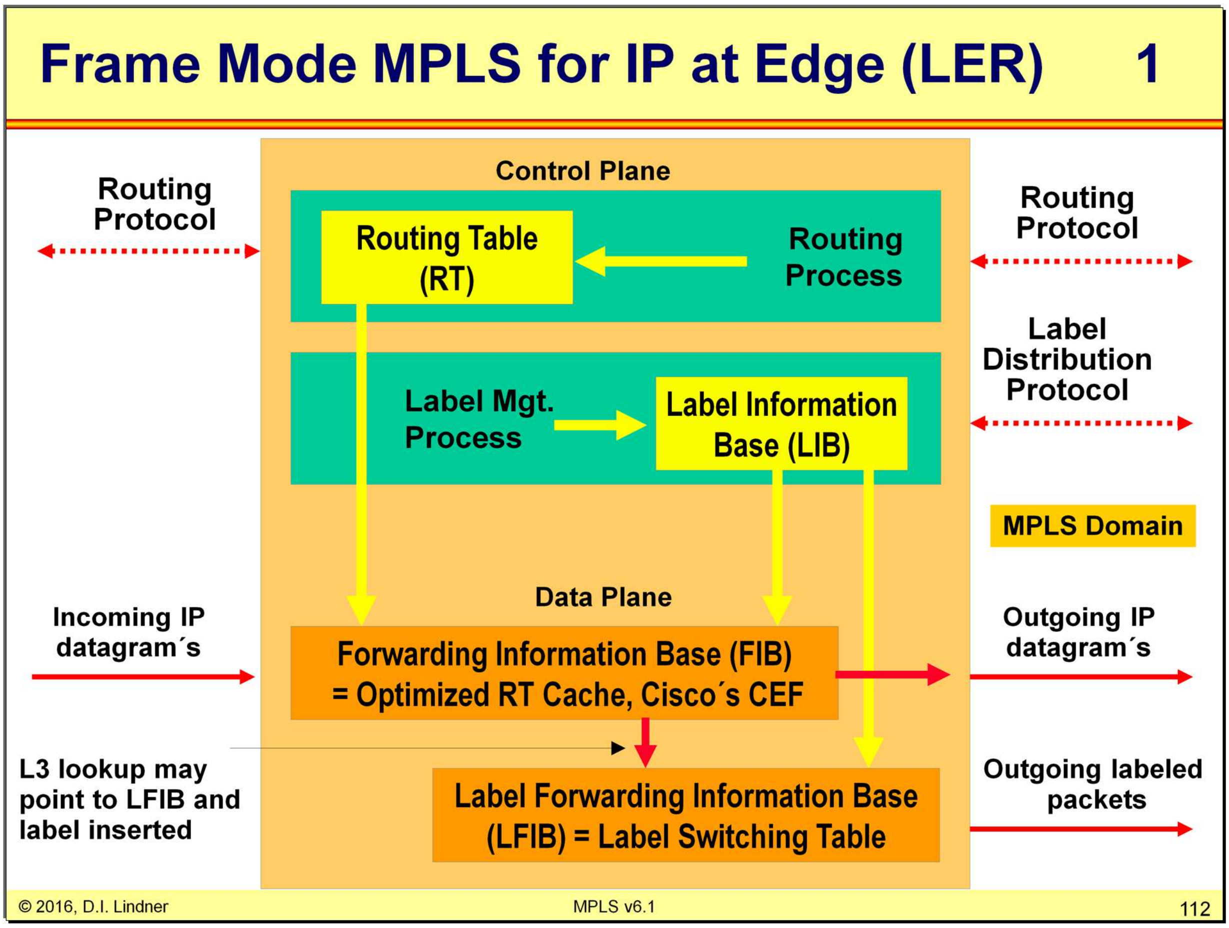
Appendix 3 - MPLS (v6.1)



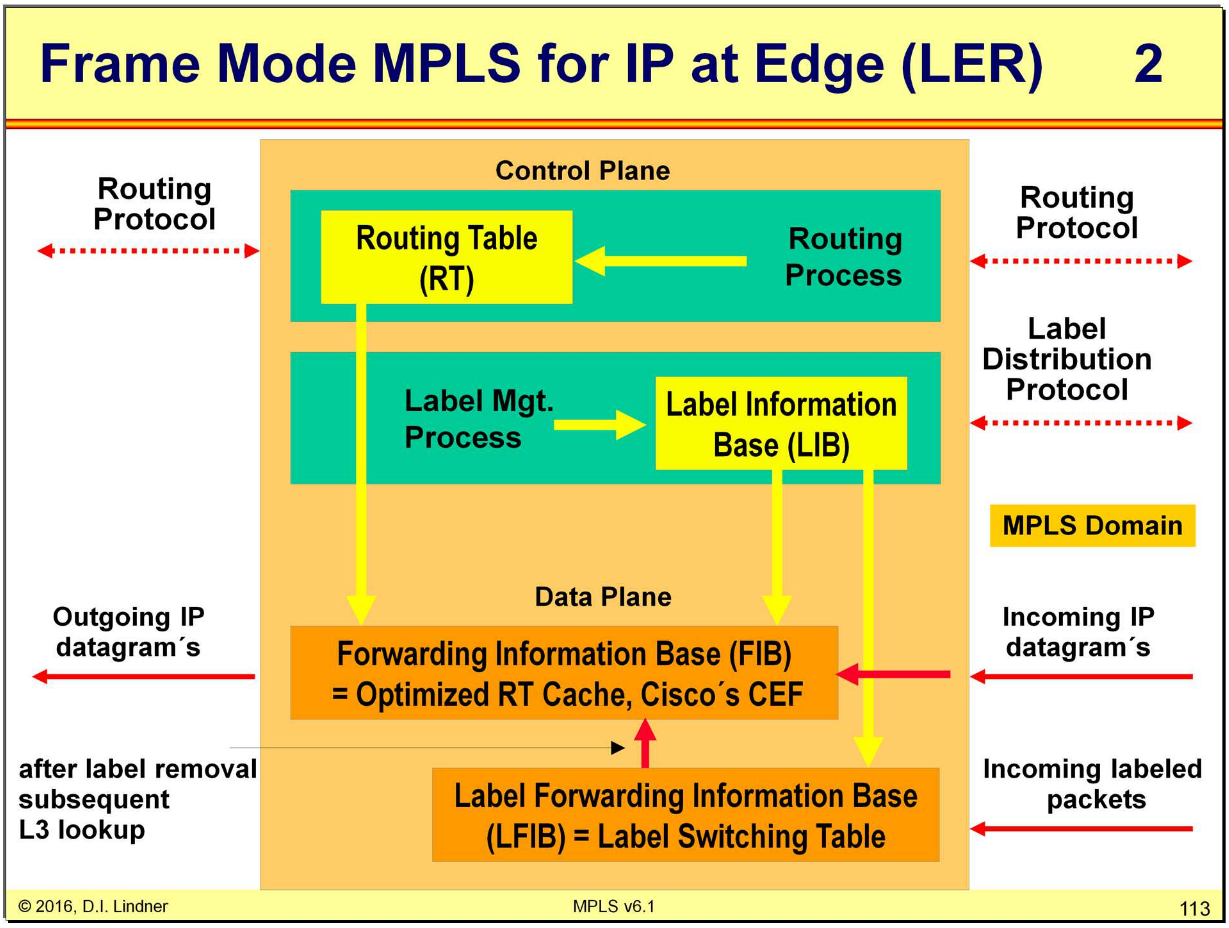
Appendix 3 - MPLS (v6.1)



Appendix 3 - MPLS (v6.1)



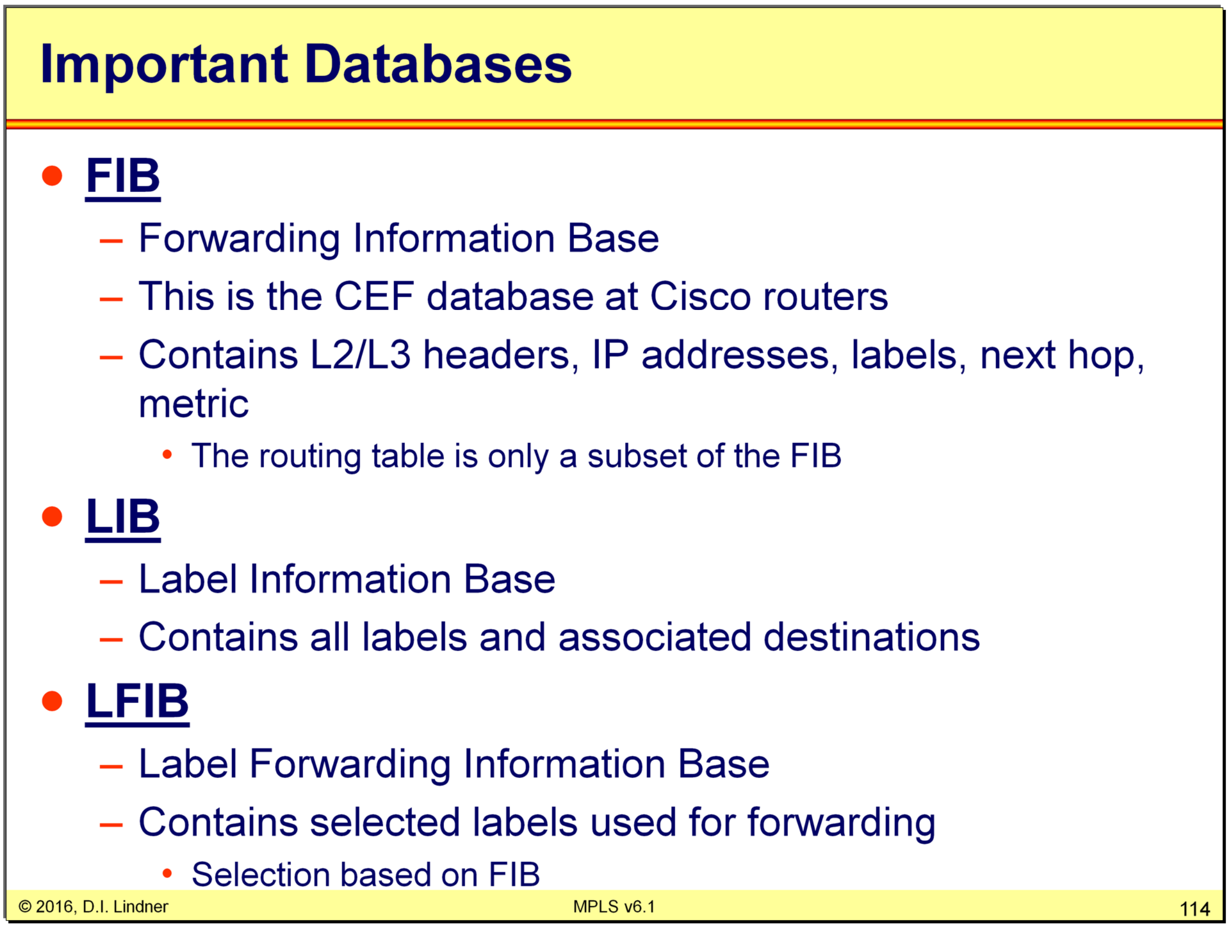
Appendix 3 - MPLS (v6.1)



MPLS is basically a software solution. With Cisco IOS version 12.0, routers are able to perform CEF switching (explained soon in detail), which is the basis for MPLS. That is, nearly any Cisco router (except the smallest home office devices) are able to do MPLS.

MPLS routers are also called "Label Switch Routers" (LSRs) and must be able to perform the following basic operations: Insert (or "impose") a label (this is essential for edge routers), remove (or "pop") a label (this is essential for last hop routers), and swap labels (this is always done during packet forwarding).

Several reasons lead to a label stack. For example, with MPLS VPNs, the top label identifies the egress router while a second label identifies the VPN itself. Thus the egress router can (as soon as the packet arrived) pop the outermost label and forward the packet to the right interface according to the inner label. Another example is MPLS Traffic Engineering (TE), where the outer label points to the TE tunnel endpoint and the inner label to the final destination itself.



This slide summarized the three important databases which had been introduced with MPLS.

MPLS needs different types of tables which are interacting to provide MPLS forwarding functionality.

The IP routing table is a common routing table which is built by the IGP in use.

The FIB table is processed from the information held in the routing table plus all necessary layer 2 information and label Information needed for packet forwarding. All incoming IP packets are forwarded related to the information kept in the FIB table.

The LIB table holds all the corresponding Label – IP Destination relationships. The LIB is built using either LDP or TDP updates. Both protocols distribute Label to IP prefix bindings. The LIB can be seen like a Label Topology database.

The LFIB only holds the best Labels out of the LIB and is actually used to forward MPLS packets. Whats the best label in the LIB is determined by the Next Hop information supplied by the local IGP.

Cisco Express Forwarding (CEF)

- **Requirement for MPLS**
  - Forwarding information (L2-headers, addresses, labels) are maintained in FIB for each destination
  - Newest and fastest IOS switching method
  - Critical in environments with frequent route changes and large RT's: The Internet backbone!
- **Invented to overcome Fast Switching problems:**
  - Originally Hash table, since 10.2 2-way radix-tree
  - No overlapping cache entries
  - Any change of RT or ARP cache invalidates route cache
  - First packet is always process-switched to build route cache entry
  - Inefficient load balancing when "many hosts to one server"

 MPLS v6.1 115

Many route changes occur in the Internet backbone, causing cache entries to be invalidated frequently. Therefore, a significant percentage of Internet traffic is process switched. First tests with IOS "ISP Geek images" under extreme conditions. Now CEF is the default switching mode in Cisco IOS Release 12.0 and the only switching mode on Cisco 12000 routers and Catalyst 8500.

Cisco IOS 12.0 knows several switching methods: Process Switching, Fast Switching, Autonomous Switching, Silicon Switching Engine (SSE) Switching, Optimum Switching, Distributed Fast Switching, CEF, Distributed CED (dCEF).

Process Switching was the first switching method implemented in IOS. It is simple (brute-force), slow, CPU demanding, non-optimized but at least platform independent.

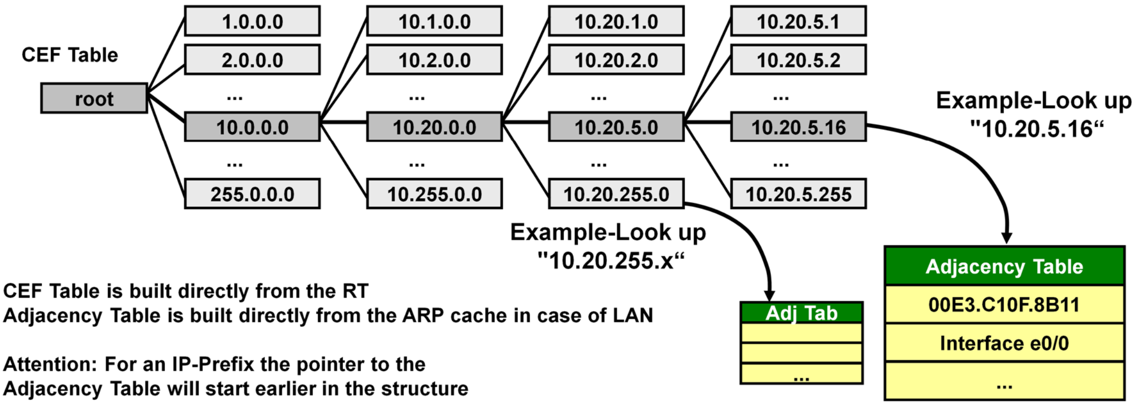
Fast Switching: Cached subset of the routing table and MAC address tables. During Process Switching (which is still done for the first packet), the information learned is stored in a fast cache. This information contains route (next hop), interface and MAC header combinations. In order to avoid collisions in the fast cache, beginning with IOS 12.0, radix trees instead of hash tables are used.

Compared to process switching and fast switching technologies, CEF supports packet manipulation on the fly. This means the FIB table lookup also provides some additional information (e.g. precedence settings, Label information etc.) which are implemented in the outgoing data packet.

Appendix 3 - MPLS (v6.1)

How CEF Works

- CEF "Fast Cache" consists of
  - CEF table: Stripped-down version of the RT (256-way mtrie data structure)
  - Adjacency table: Actual forwarding information (MAC, interfaces, ...)
- CEF cache is pre-built before any packets are switched
  - No packet needs to be process switched
- CEF entries never age out
  - Any RT or ARP changes are immediately mapped into CEF cache



 MPLS v6.1 116

The CEF (FIB) table holds all the necessary information needed to rewrite the layer 2 and 3 header of an forwarded data packet. Changes in the routing table has to be reflected in the CEF table immediately.

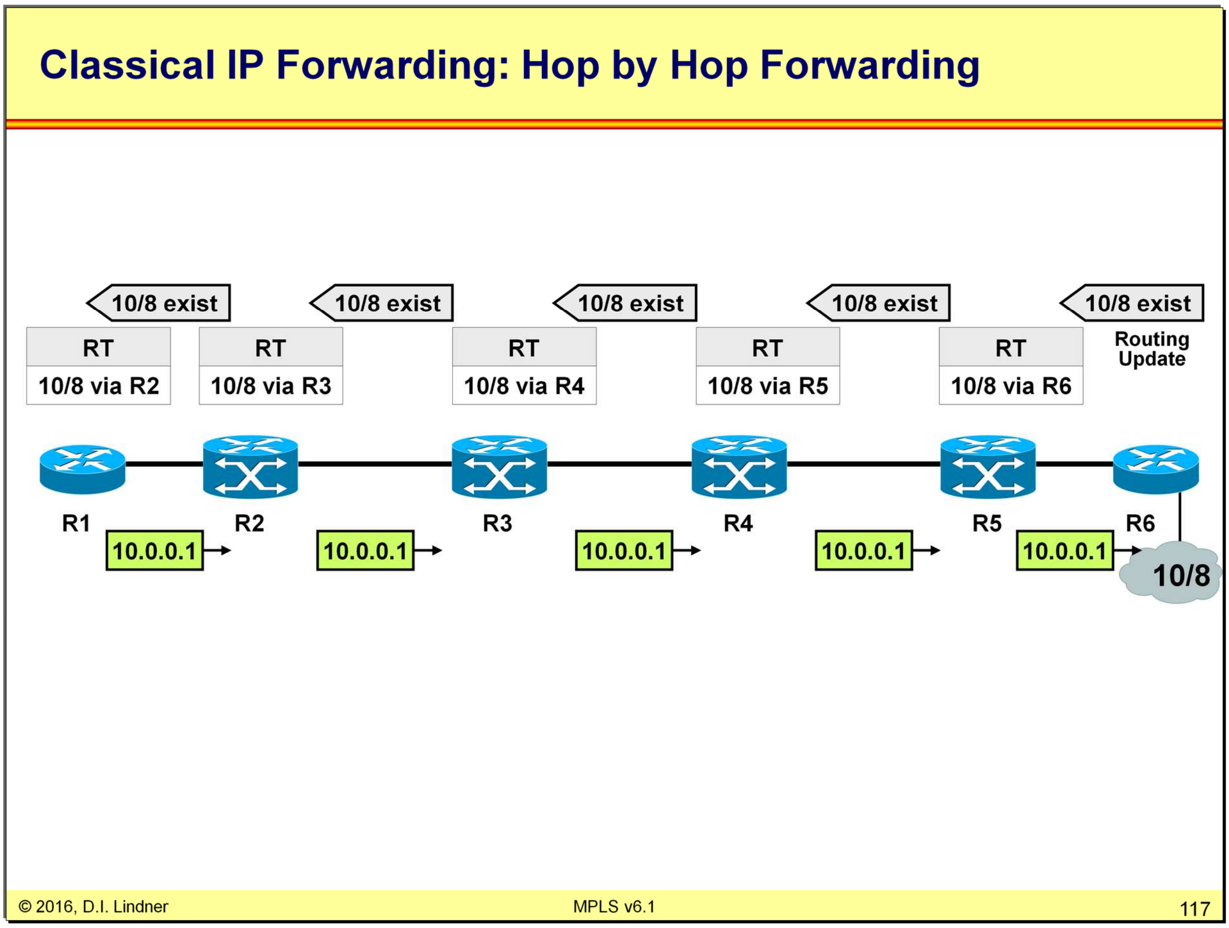
mtrie: tree of pointers; data is stored elsewhere.

Display CEF table information using show ip cef summary.

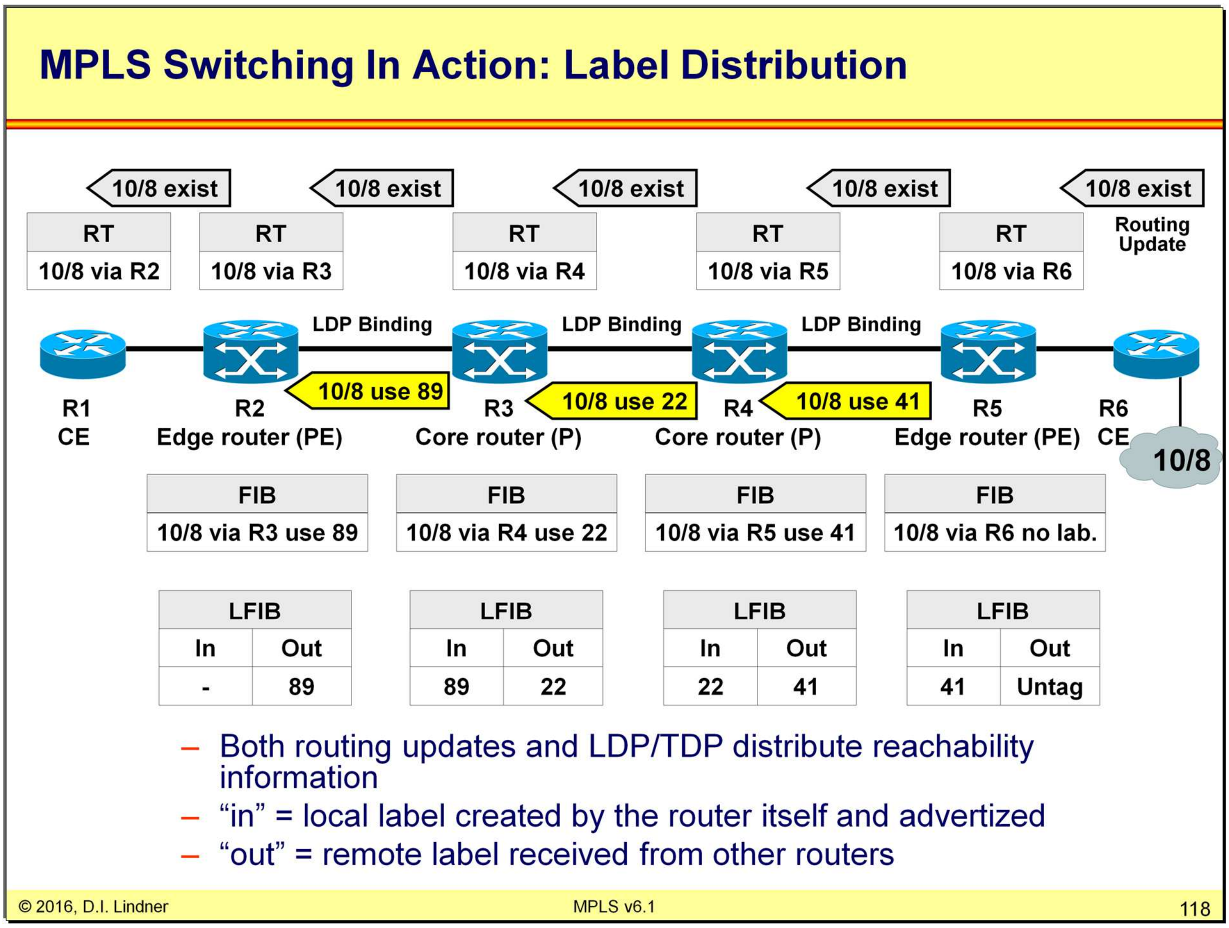
Display Adjacency table information: show adjacency.

dCEF: Very high performance boost. Each interface holds its own CEF table and is able to forward packets autonomously. Available on GSR, Cisco 7500 router

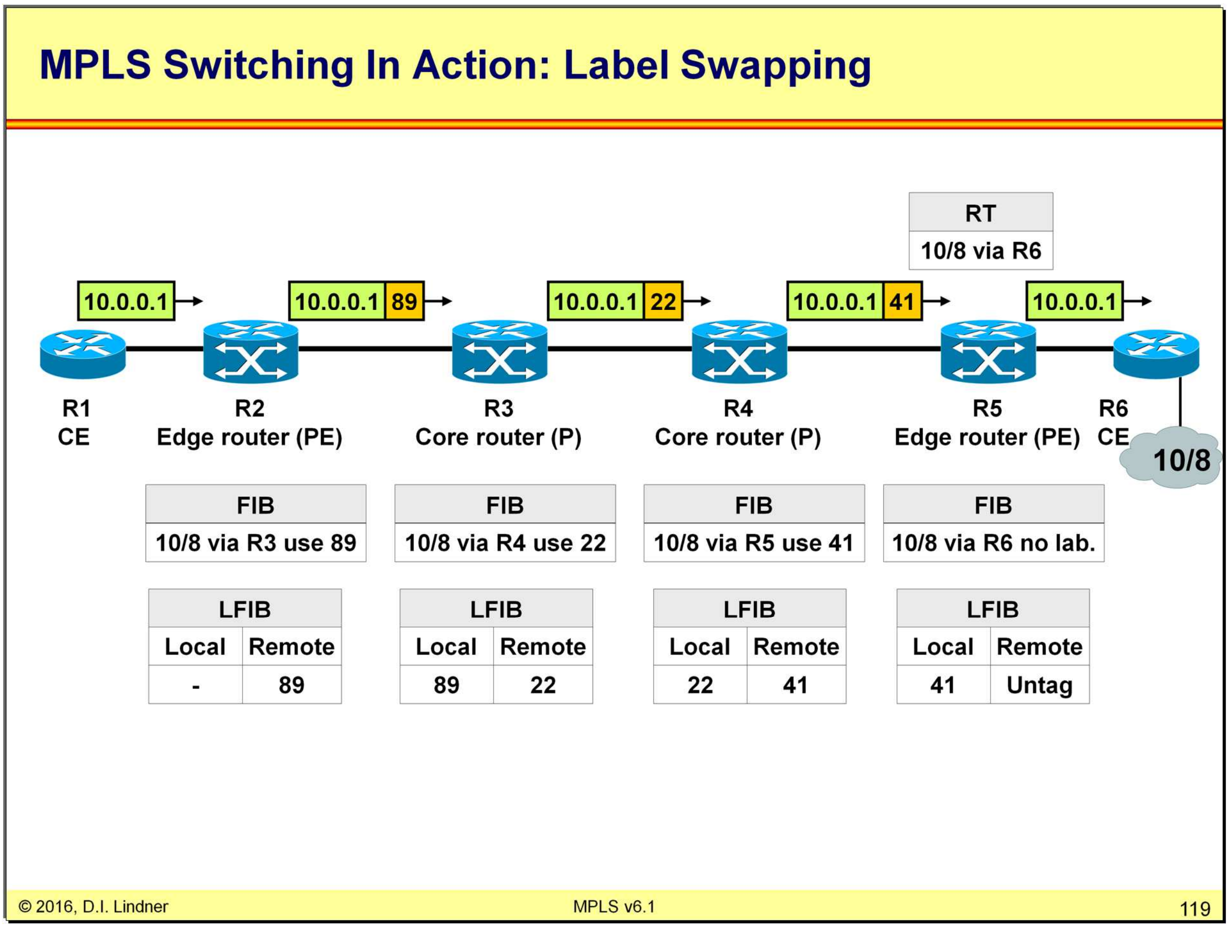
mtree: data is stored in the tree (optimum switching)



The picture above shows classical IP hop-by-hop routing using signposts established by routing protocols and stored in the corresponding routing table.

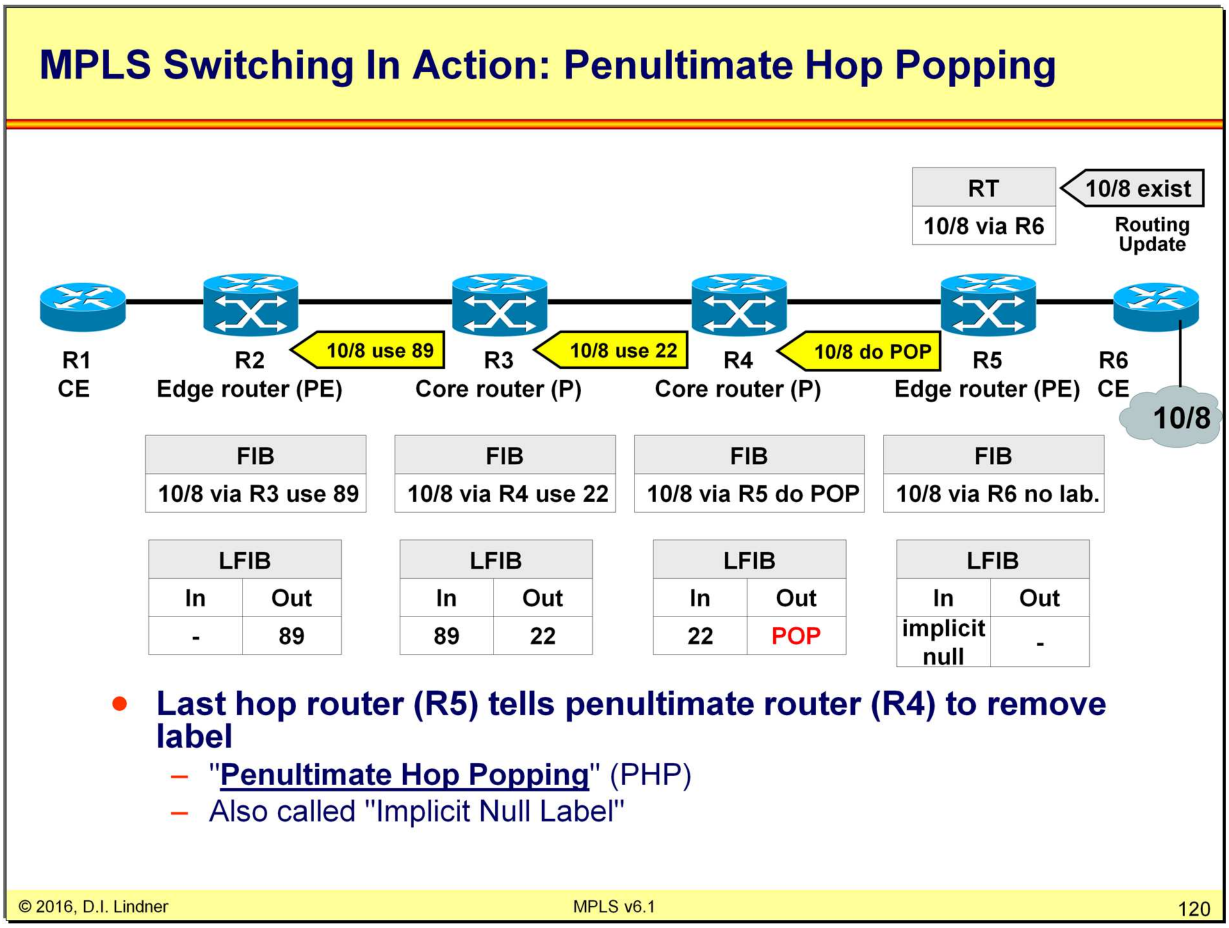


The picture above shows how a label-switched path is established from left to the right. Both routing updates as well as a label distribution protocol (LDP or TDP) distribute reachability information for this destination network.



The picture above shows how packets can now be sent using a MPLS header. Label switching is performed on each hop (LSR) inside the provider domain (R2, R3, R4, R5). The LFIB tables are used to perform a fast lookup.

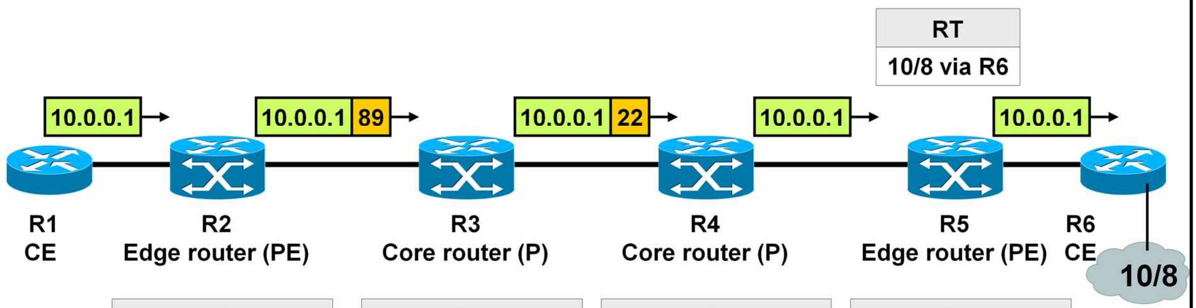
But R5 cannot find any outgoing label in its LFIB. After this unsuccessful lookup, R5 looks into the FIB and determines the next hop. Note that this double lookup would be done for every packet! Therefore it would be reasonable to remove the label even one hop earlier (the penultimate hop, R4) in order to leave R5's LFIB empty.



In this scenario "Penultimate Hop Popping" (PHP) is illustrated. Now R5 does not allocate an incoming label for this destination but rather announces to R4 to use an "implicit null" label. It is also said, that R4 should perform the "POP" operation. The label number "3" had been reserved to represent the "do POP" command.

Implicit Null Label and hence POP upstream sent out only for directly connected networks or aggregates of advertising router

MPLS Switching In Action: Penultimate Hop Popping



FIB (R2)	FIB (R3)	FIB (R4)	FIB (R5)
10/8 via R3 use 89	10/8 via R4 use 22	10/8 via R5 do POP	10/8 via R6 no lab.

LFIB (R2)

In	Out
-	89

LFIB (R3)

In	Out
89	22

LFIB (R4)

In	Out
22	POP

LFIB (R5)

In	Out
implicit null	-

- R5 only performs single lookup in FIB

Appendix 3 - MPLS (v6.1)

Cisco IOS Standard Behavior 1

- Routers with packet interfaces (Frame-Mode MPLS)
  - Per-platform Label Space !!!
    - a label assigned by an LSR to a given FEC is used on all interfaces in advertisements of this LSR
  - Unsolicited Downstream Label Distribution
    - label distribution is done unsolicited
  - Liberal Label Retention Mode
    - received labels which are not used by a given LSR are still stored in the LIB
    - allows faster convergence of LSP after rerouting
  - Independent Control
    - labels are assigned by LSR independently from each other

 MPLS v6.1 122

This slide summarized the main differences.

Note that routers performs a per-platform label allocation. That is, the LFIB does not contain any incoming interface, so the label must be unqiue on the entire router for a given destination. In other words, the same label can be used for a packet on any interface and will be forwarded to the same destination—this is the positive version.

Which label distribution and retention behavior is used depends on the interface type in use.

Unsolicited label distribution means that labels are advertised automatically without being asked...

Liberal label retention: All advertised labels are accepted, even from LSRs which are not next hop to the destination.

Conservative label retention: Advertised labels are only accepted from LSRs which are next hop LSRs for a given destination.

Cisco IOS Standard Behavior 2

- Routers with ATM interfaces (Cell-Mode MPLS)
  - Per-interface Label Space
    - a different label for the same FEC is used on each single interface in advertisements of this LSR
  - Downstream On Demand Label Distribution
    - label distribution is done on request
  - Conservative or Liberal Label Retention Mode
    - received labels which are not used by a given LSR are not stored in the LIB in case of conservative mode
  - Independent Control

Cisco IOS Standard Behavior 3

– ATM switches (Cell-Mode MPLS)
  - Per-interface Label Space
  - Downstream On Demand Label Distribution
  - Conservative Label Retention Mode
  - Ordered control
    – labels are assigned by LSR in a controlled fashion from egress to ingress

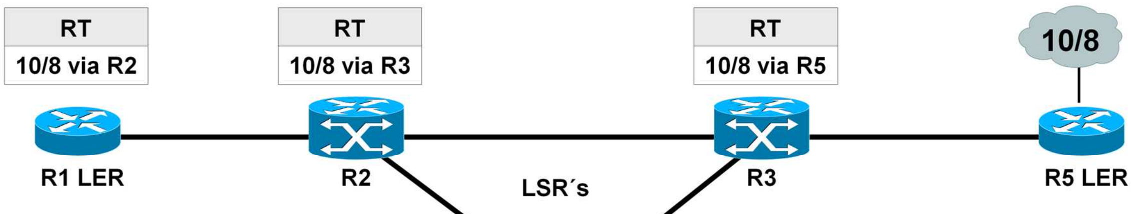
Agenda

- **Review ATM**
- **IP over WAN Problems (Traditional Approach)**
- **MPLS Principles**
- **Label Distribution Methods**
- **MPLS Details** **(Cisco)**
  - Internal Components
  - MPLS in Action
  - TDP, LDP
  - TTL
  - Traffic Engineering
  - MPLS and BGP
- **RFCs**

 MPLS v6.1 125

Appendix 3 - MPLS (v6.1)

Building Routing Tables



FIB R2		
Net	Next Hop	Label
10/8	R3	none

LIB R2		
Net	Label	Type
-	-	-
-	-	-
-	-	-

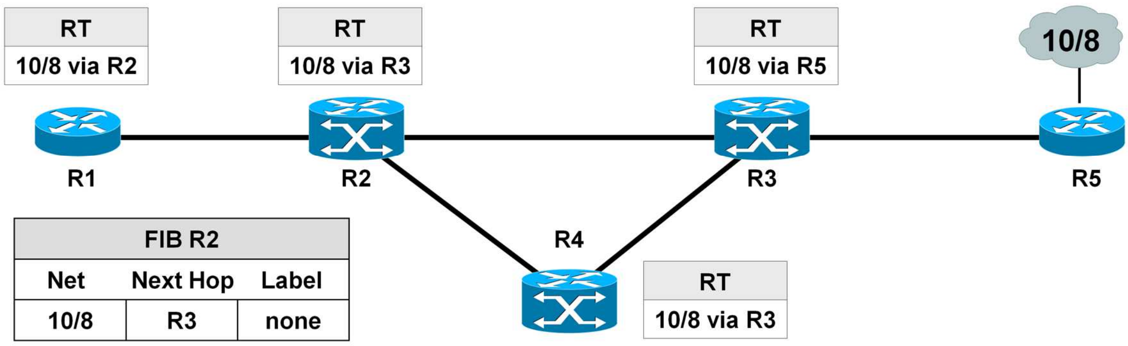
LFIB R2		
Label	Action	Next Hop
-	-	-

- **Routing Protocol**
  - establish routing tables RT in all routers
  - best path based on metric is stored in RT
- **RT**
  - contains next hop information (outgoing interface)
- **FIB**
  - additionally contains outgoing label information (which label can be used towards next hop)

 MPLS v6.1 126

Appendix 3 - MPLS (v6.1)

Allocating Labels



FIB R2		
Net	Next Hop	Label
10/8	R3	none

LIB R2		
Net	Label	Type
10/8	49	local
-	-	-
-	-	-

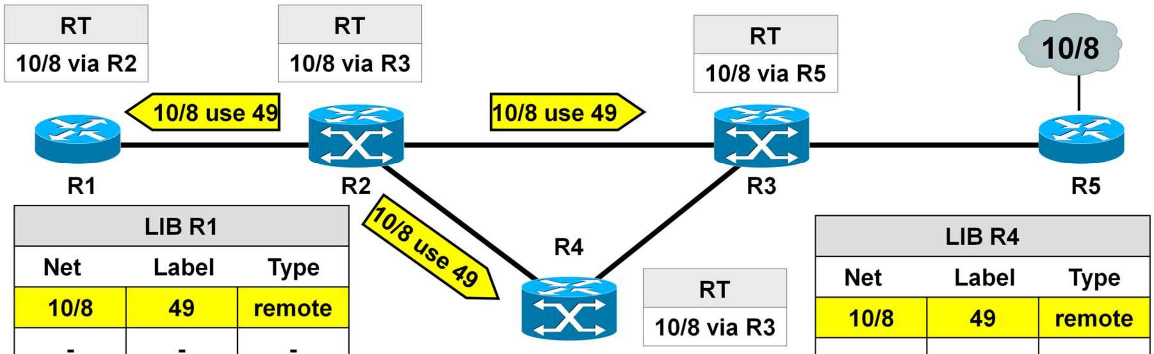
LFIB R2		
Label	Action	Next Hop
49	untag	-

- **R2**
  - allocates label 49 to FEC 10/8
  - stored in LIB with type local
  - stores action untag in LFIB because no other router has advertised a label for that FEC
- **Every MPLS router**
  - allocates labels for all IP destinations found in the routing table
  - this is done independently from each other
  - a label has only local significance

 MPLS v6.1 127

Appendix 3 - MPLS (v6.1)

Advertising and Receiving Labels via LDP



LIB R1		
Net	Label	Type
10/8	49	remote
-	-	-

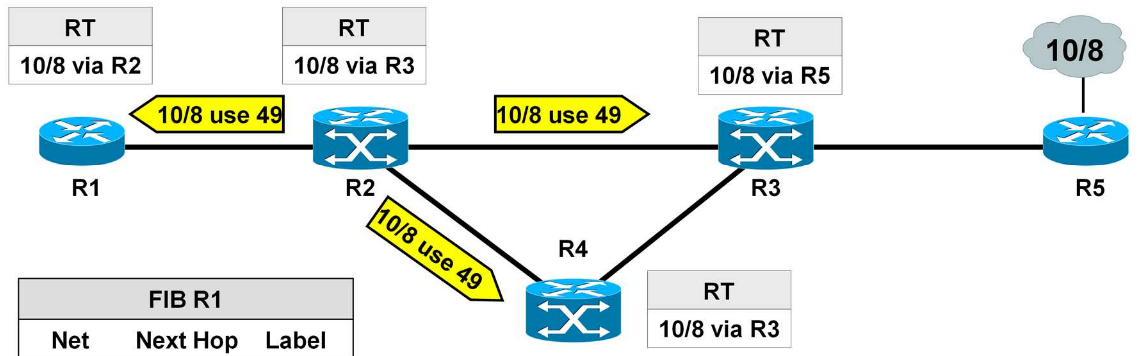
LIB R4		
Net	Label	Type
10/8	49	remote
-	-	-

LIB R3		
Net	Label	Type
10/8	49	remote
-	-	-

- **R2**
  - advertises label 49 for FEC 10/8 to all neighbor routers
- **Per platform label allocation**
  - same label on all interfaces
  - LFIB may not contain an incoming interface (next HOP) field at that moment
- **Every neighbor MPLS router**
  - stores received label for IP destination 10/8 in the corresponding LIB

 MPLS v6.1 128

Actions on Receiving Labels on R1



FIB R1		
Net	Next Hop	Label
10/8	R2	49

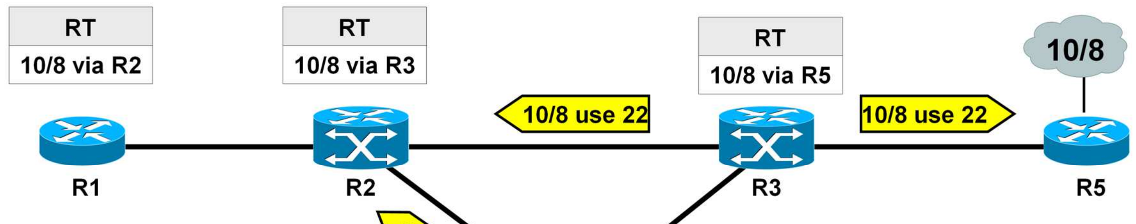
LIB R1		
Net	Label	Type
10/8	49	remote
-	-	-

LFIB R1		
Label	Action	Next Hop
-	49	R2

- **R1**
  - receives label 49 for FEC 10/8
  - label is advertised by router which is the next hop in the routing table -> therefore populates the FIB
  - LFIB is adapted to use label 49 for FEC 10/8 towards R2
  - action in LFIB has the meaning of outgoing label or remote label
  - label in LFIB has the meaning of incoming label or local label

Appendix 3 - MPLS (v6.1)

Actions on Receiving of Labels from R3 and R4 on Router R2



FIB R2		
Net	Next Hop	Label
10/8	R3	22

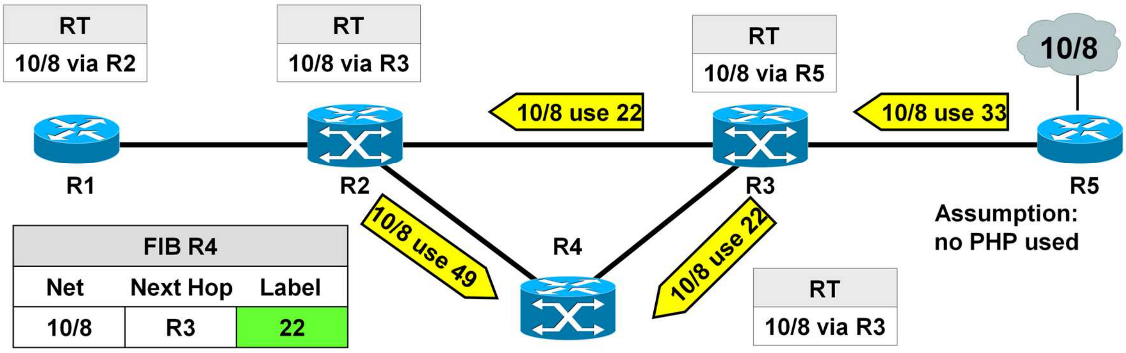
LIB R2		
Net	Label	Type
10/8	49	local
10/8	22	remote
10/8	55	remote

LFIB R2		
Label	Action	Next Hop
49	22	R3

- **R2**
  - receives label 22 for FEC 10/8
    - this label is advertised by router which is the next hop in the routing table -> therefore populates the FIB
    - LFIB is adapted to use (swap) label 22 for FEC 10/8 towards R3
  - receives label 55 for FEC 10/8
    - this label is advertised by router which is not the next hop in the routing table but will be still stored in the LIB -> liberal retention mode

 MPLS v6.1 130

Receiving of Labels from R2 and R3 on Router R4



FIB R4		
Net	Next Hop	Label
10/8	R3	22

LIB R4		
Net	Label	Type
10/8	55	local
10/8	22	remote
10/8	49	remote

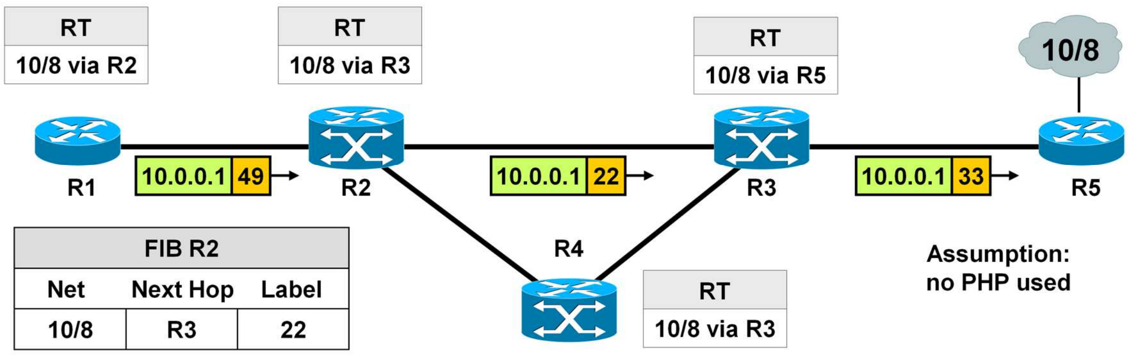
LFIB R4		
Label	Action	Next Hop
55	22	R3

- **R4**
  receives label 22 for FEC 10/8
  - this label is advertised by router which is the next hop in the routing table-> therefore populates the FIB
  - LFIB is adapted to use label 22 for FEC 10/8 towards R3

  already received label 49 for FEC 10/8
  - this label is advertised by router which is not the next hop in the routing table but will be still stored in the LIB

Appendix 3 - MPLS (v6.1)

Label Switching



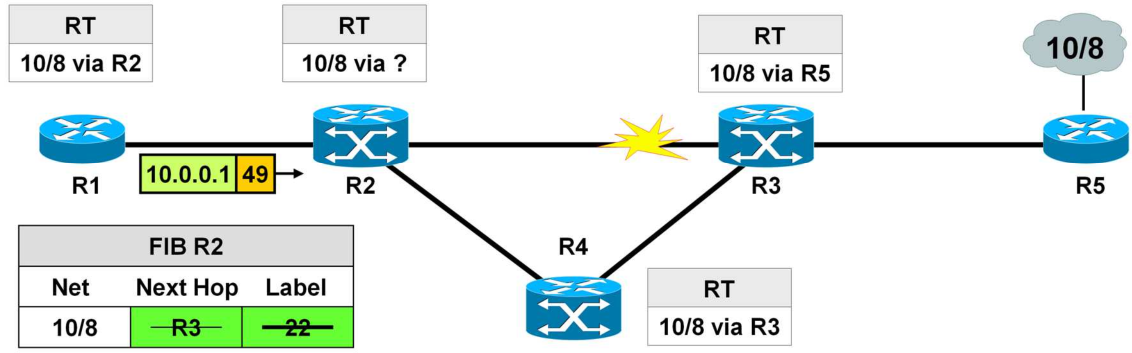
FIB R2		
Net	Next Hop	Label
10/8	R3	22

LIB R2		
Net	Label	Type
10/8	49	local
10/8	22	remote
10/8	55	remote

LFIB R2		
Label	Action	Next Hop
49	22	R3

- **Packets of FEC 10/8 will follow the corresponding Label Switched Path**

Link Failure R2 <-> R3



FIB R2		
Net	Next Hop	Label
10/8	~~R3~~	~~22~~

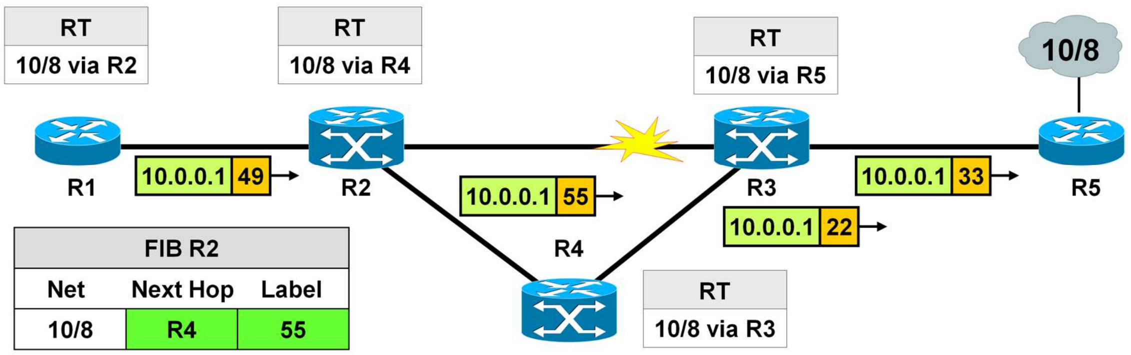
LIB R2		
Net	Label	Type
10/8	49	local
10/8	~~22~~	~~remote~~
10/8	55	remote

LFIB R2		
Label	Action	Next Hop
49	~~22~~	~~R3~~

- **Routing protocol neighbors and LDP neighbors are lost after failure**
  - corresponding entries in FIB, LIB and LFIB are removed
- **Traffic**
  - to FEC 10/8 will not be forwarded until routing table converges

Appendix 3 - MPLS (v6.1)

Routing Protocol Convergence



FIB R2		
Net	Next Hop	Label
10/8	R4	55

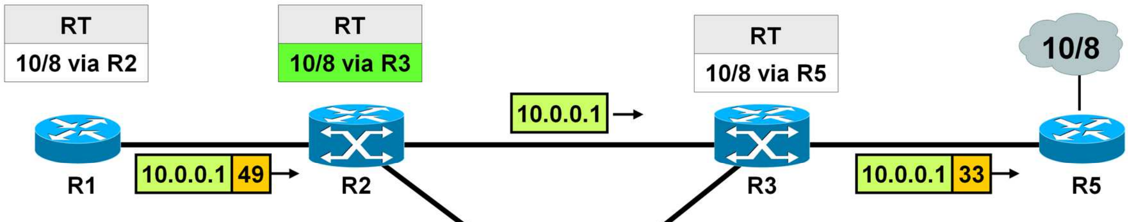
LIB R2		
Net	Label	Type
10/8	49	local
10/8		
10/8	55	remote

LFIB R2		
Label	Action	Next Hop
49	55	R4

- **After routing protocol convergence**
  - R2 can switch over immediately to other LSP if alternative label advertisements were stored in LIB and labeled packets will flow again
  - Otherwise R2 must wait for new bindings and can forward packets only based on IP address in the meantime (action untag in LFIB)
- **Packets of FEC 10/8 will follow the new Label Switched Path via R4**

Appendix 3 - MPLS (v6.1)

Link Failure Repair 1



FIB R2		
Net	Next Hop	Label
10/8	R3	-

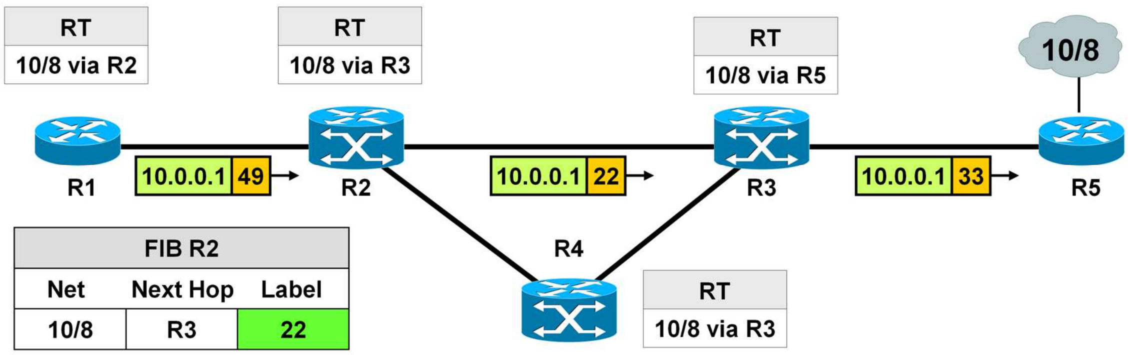
LIB R2		
Net	Label	Type
10/8	49	local
10/8		
10/8	55	remote

LFIB R2		
Label	Action	Next Hop
49	untag	-

- **After link repair**
  - Routing protocol neighbor detection and routing table adaptation
  - R2 must wait for new bindings and can forward packets only based on IP address in the meantime (action untag in LFIB)

 MPLS v6.1 135

Link Failure Repair 2



FIB R2		
Net	Next Hop	Label
10/8	R3	22

LIB R2		
Net	Label	Type
10/8	49	local
10/8	22	remote
10/8	55	remote

LFIB R2		
Label	Action	Next Hop
49	22	R3

- **After LDP session to R3 is up and binding for FEC 10/8 from R3 received**
  - Packets of FEC 10/8 will follow the corresponding Label Switched Path again

Agenda

- **Review ATM**
- **IP over WAN Problems (Traditional Approach)**
- **MPLS Principles**
- **Label Distribution Methods**
- **MPLS Details (Cisco)**
  - Internal Components
  - MPLS in Action
  - TDP, LDP
  - TTL
  - Traffic Engineering
  - MPLS and BGP
- **RFCs**

 MPLS v6.1 137

TDP Key Facts

- **Tag Distribution Protocol (TDP)**
  - invented by Cisco
  - for distributing <label, prefix> bindings
  - enabled by default
- **Session establishment: UDP/TCP port 711**
  - Hello messages via UDP
  - destination address -> 224.0.0.2
    - well-known multicast address for all subnet routers
  - TDP session via TCP, incremental updates
- **Not compatible with LDP**
  - but can co-exist as long as two peers use the same protocol

 MPLS v6.1 138

The TDP protocol was developed by Cisco and is used to distribute Lable-Prefix bindings between adjacent LSRs. Only in the case of MPLS TE TDP updates are also exchanged between not adjacent LSRs through so called Tunnel interfaces.

The TDP protocol is using both UDP and TCP at the transport layer. The TDP server process is addressed by the port number 711 and the updates are sent using the well known all routers Multicast address 224.0.0.2.

UDP is used in combination with a Hello procedure to detect neighboring LSRs.

The TCP protocol is used to reliable transport label binding information.

TDP is incompatible with LDP so neighboring LSRs need to use the same Protocol to allow a TDP/LDP session to come up.

LDP Key Facts

- Label Distribution Protocol
- IETF standard RFC 3036
  - descendent of Cisco's proprietary TDP
- Same concept but port 646
- LDP-Identifier
  - Router ID (4 bytes)
  - Label Space ID (2 bytes)
    - in case of per-platform label space this field is set to zero
    - note: in ATM you need a per-interface label space
- TCP session initiated from router with highest address

 MPLS v6.1 139

The LDP protocol is the standard protocol specified by the IETF. It works the same way like TDP does but they are incompatible as you can see just by the port numbers in use.

Reference: draft-ietf-mpls-ldp-07.txt

Combination of frame-mode and cell-mode (or multiple cell-mode) links result in multiple LDP sessions.

An LDP session is established by the router with the higher IP address.

Non-adjacent neighbors are discovered by unicast messages.

LDP Message Types

- **Four basic types:**
  - Discovery (UDP):
    - getting into contact with neighbor LSR´s
  - Adjacency (TCP):
    - Initialization, Keepalive and Shutdown of LDP sessions
  - Label Advertisement (TCP):
    - Label Binding - Advertisement, - Request, - Withdrawal, - Release
  - Notification (TCP):
    - Signal of Error Information, Advisory Information
- **TLV (Type/Length/Value)**
  - encoding is used for easy extension of the protocol

Discovery Message

- **Basic discovery of directly connected LSRs:**
  - Hello Message with targeted bit set to 0
    - UDP to port 646
    - IP multicast address "all routers on this subnet" (224.0.0.2)
- **Extended discovery of non-directly connected LSR's:**
  - Hello Message with targeted bit set to 1 (Targeted Hello)
    - UDP to port 646
    - IP unicast address of neighbor
  - used e.g. in case of MPLS Traffic Engineering
- **After discovery**
  - LDP session is created running on top of TCP
    - well known port 646

 MPLS v6.1 141

Adjacency Messages

- **Adjacency**
  - Initialization
    - negotiates
      - protocol version (current version = 1)
      - label advertisement discipline
        - » Unsolicited Downstream = 0
        - » Downstream-on-Demand = 1
      - keepalive time
  - Keepalive
    - maintains LDP session

 MPLS v6.1

Label Advertisement Messages

- **Label Advertisement**
  - Label Mapping
    - advertise a binding between a FEC and a label
  - Label Withdrawing
    - reverse the mapping process
    - e.g. if FEC is not longer valid because address prefix has been removed from the routing table
  - Label Release
    - issued by a LSR which has previously received a label mapping and no longer has a need for that mapping
  - Label Request / Label Request Abort
    - for Downstream-on-Demand method
    - abort is used to revoke a request before it has been satisfied

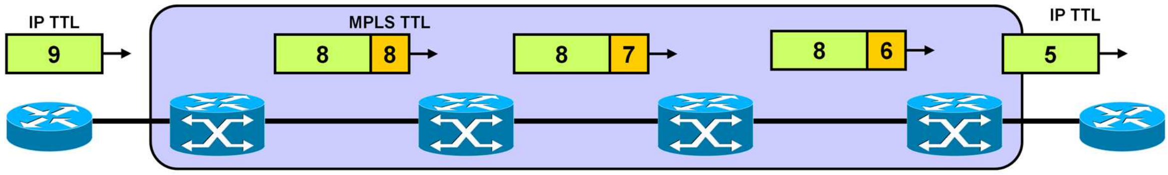
Agenda

- Review ATM
- IP over WAN Problems (Traditional Approach)
- MPLS Principles
- Label Distribution Methods
- MPLS Details (Cisco)
  - Internal Components
  - MPLS in Action
  - TDP, LDP
  - TTL
  - Traffic Engineering
  - MPLS and BGP
- RFCs

 MPLS v6.1 144

Normal TTL Usage

- **Loop detection**
  - LDP and TDP basically rely on IGP loop detection, therefore no additional tasks are necessary for MPLS control packets
  - Additionally a TTL field in the MPLS header prevents endless routing of MPLS data packets
- **TTL Propagation:**
  - IP TTL is copied into MPLS header
  - Done by Ingress LSR (LER)
  - MPLS TTL decremented by every LSR
  - Egress LSR copies MPLS-TTL back to IP TTL
  - Enabled by default on Cisco routers

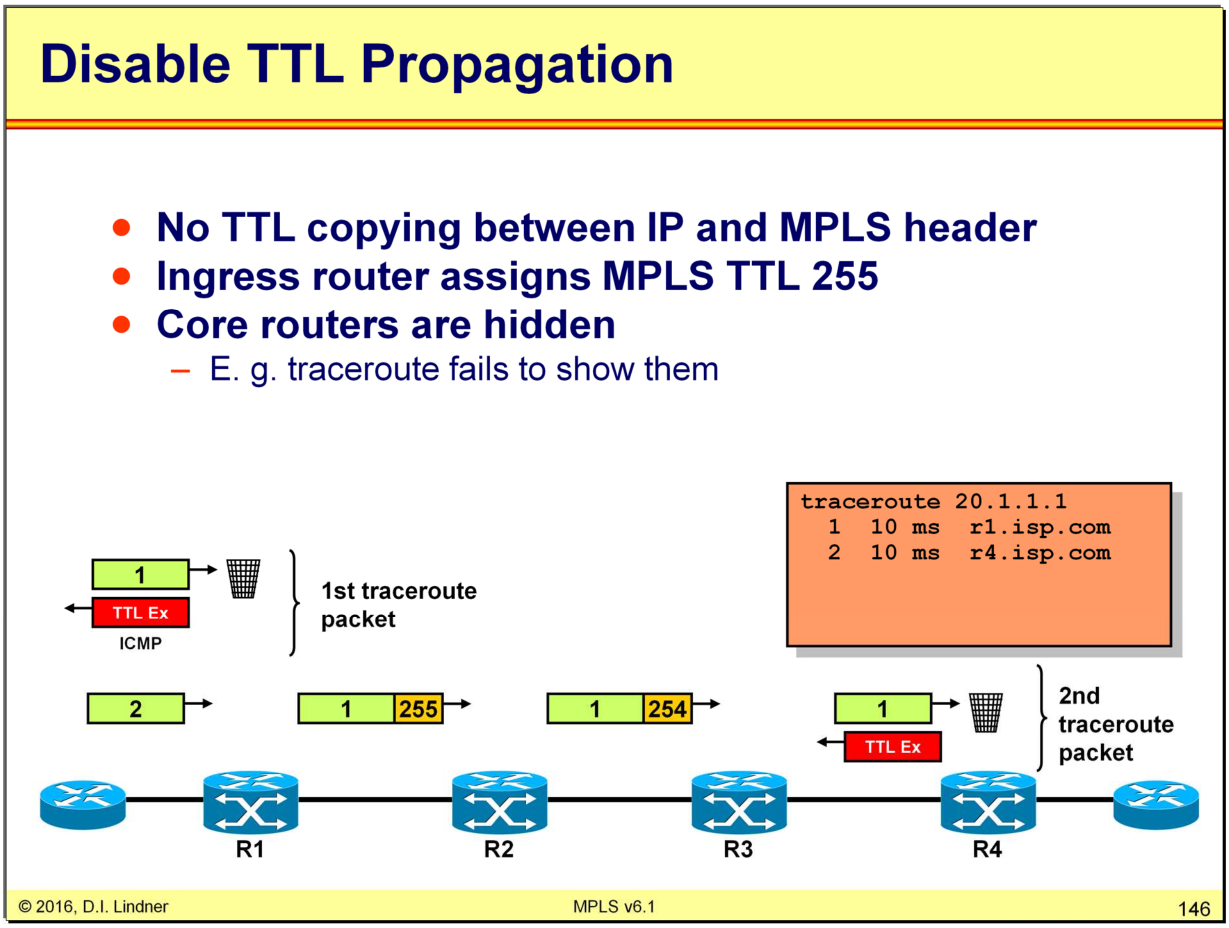


 MPLS v6.1 145

IGP protocols typically provide strong mechanisms to avoid routing loops. Nevertheless, the MPLS header carries a TTL field which provides additional protection against endless looping—for example caused by misconfigured static routes.

TTL Propagation: This mechanism is enabled by default (at least on Cisco routers) and ensures that the IP TTL value is also processed inside the MPLS domain. Actually, the IP TTL value is copied into the MPLS header. Within the MPLS domain only the MPLS TTL value is decremented.

Upon ingress, the IP TTL is copied to the MPLS header, upon egress the MPLS TTL is copied back to the IP header.



As the example above shows, only the ingress and the egress LSRs are seen by traceroute.

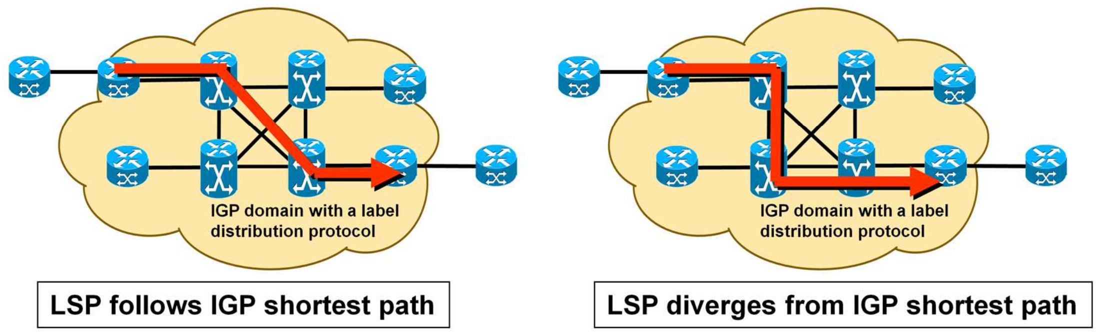
Note: If a traceroute would be started from any LSR (e. g. R1) every downstream router would be visible in the traceroute output. This is because TTL propagation can only be disabled for forwarded traffic. Traceroute from LSRs does not use the initial TTL value of 255.

Note: When TTL propagation should be disabled, it has to be disabled on all LSRs in the core! Frequently, ISPs forget to disable TTL propagation on some core routers. This typically lead to wrong traceroute results.

Agenda

- **Review ATM**
- **IP over WAN Problems (Traditional Approach)**
- **MPLS Principles**
- **Label Distribution Methods**
- **MPLS Details (Cisco)**
  - Internal Components
  - MPLS in Action
  - TDP, LTP
  - TTL
  - Traffic Engineering
  - MPLS and BGP
- **RFCs**

Label Switch Path (LSP)



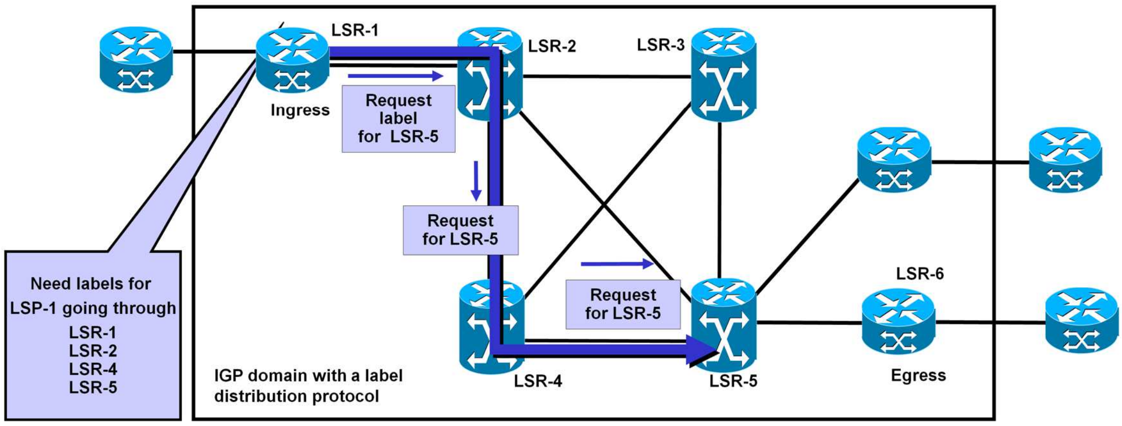
- **Normal MPLS Destination Based Routing**
  - FEC is determined in LSR-ingress
  - LSP's derive from IGP routing information
- **If LSPs should diverge from IGP shortest path**
  - LSP Explicit Routing (LSP Tunnel) is necessary
  - MPLS Traffic Engineering

Traffic Engineering via LSP - Tunnels

- **Explicit Routing:**
  - Source Routing
  - Constraint-Based Path Selection Algorithm
    - similar to ATM PNNI
  - OSPF / IS-IS extension for flooding of resources / policy information
    - traffic class, resource requirements and the available network resources (bandwidth)
  - RSVP as the mechanism for establishing LSP's
    - uses new RSVP objects in PATH and RESV messages
      - Explicit-Route (ERO) in Path, Label found in RSV
  - Usage of ER-LSPs in the forwarding table
    - label stack

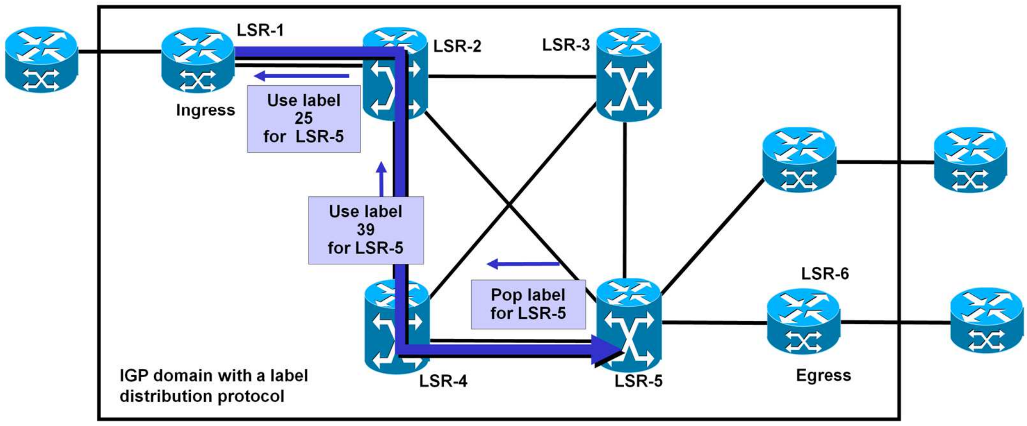
Appendix 3 - MPLS (v6.1)

Explicit Routing 1



- **LSR-1 request an explicit LSP to LSR-5:**
  - LSR-1, LSR-2, LSR-4, LSR-5
- **The request travels hop-by-hop**
  - using RSVP PATH messages

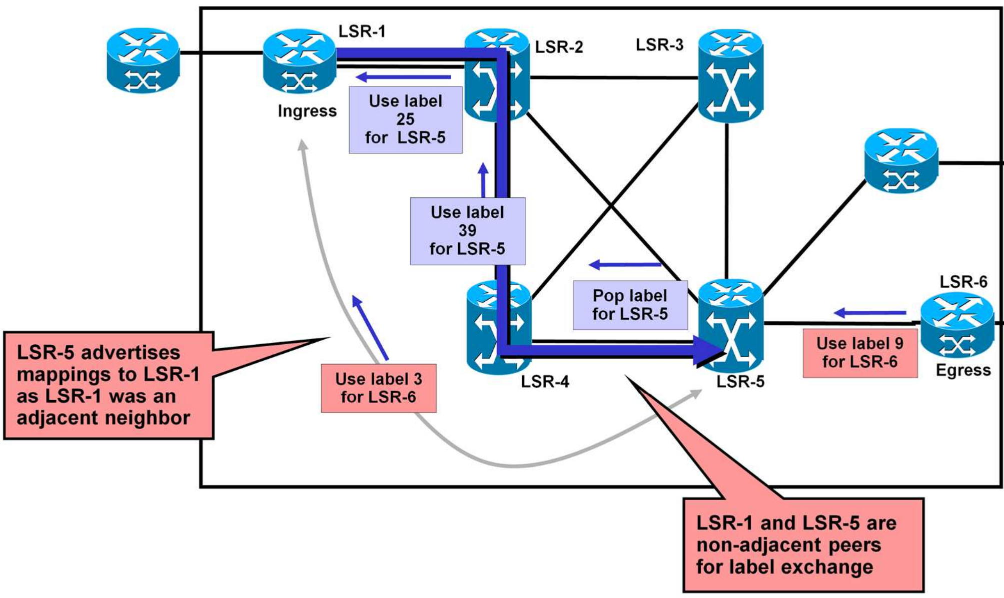
Explicit Routing 2



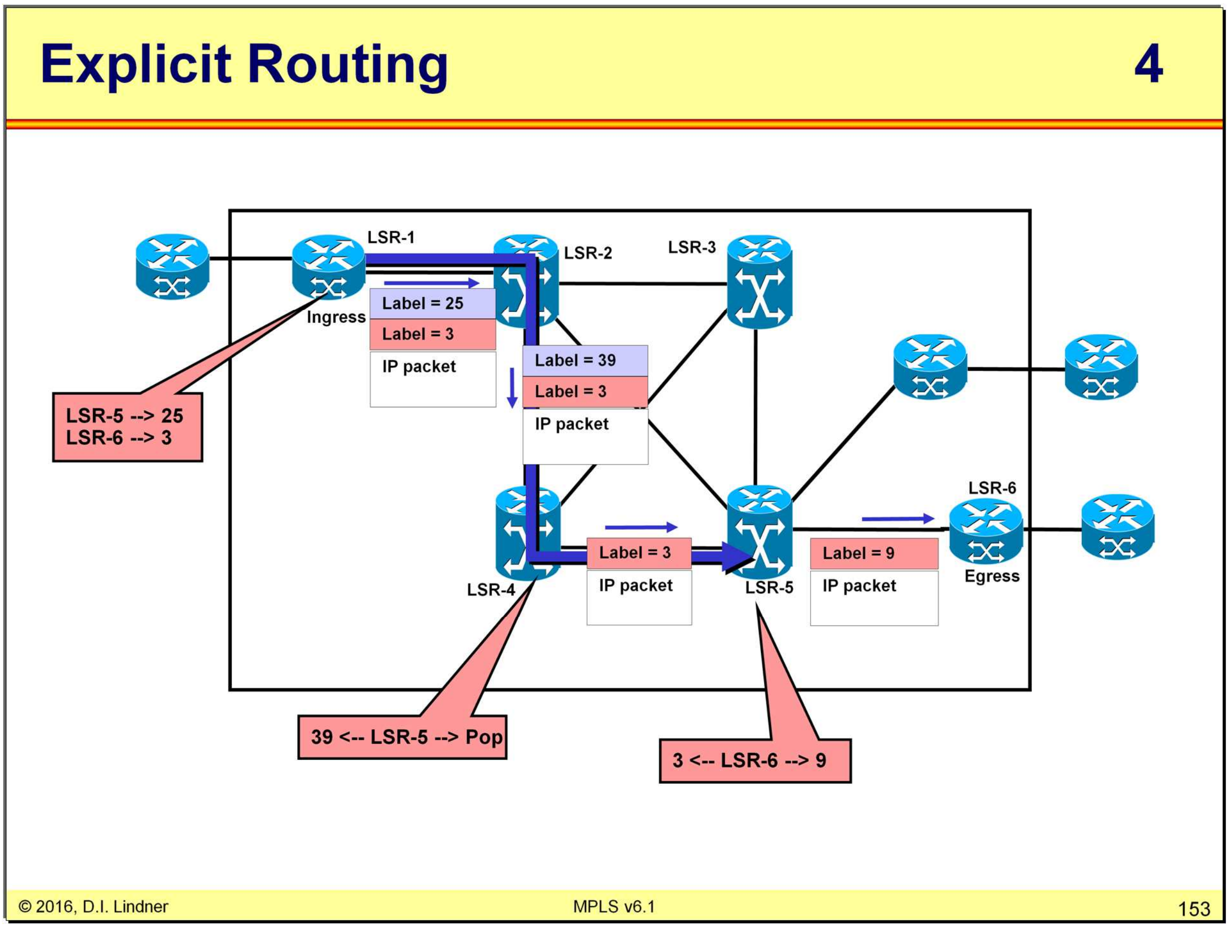
- **When the request reaches the egress point labels are advertised back to the ingress LSR**
  - via RSVP RESV messages

Appendix 3 - MPLS (v6.1)

Explicit Routing 3



Appendix 3 - MPLS (v6.1)



Several reasons lead to a label stack. For example, with MPLS VPNs, the top label identifies the egress router while a second label identifes the VPN itself. Thus the egress router can (as soon as the packet arrived) pop the outermost label and forward the packet to the right interface according to the inner label. Another example is MPLS Traffic Engineering (TE), where the outer label points to the TE tunnel endpoint and the inner label to the final destination itself.

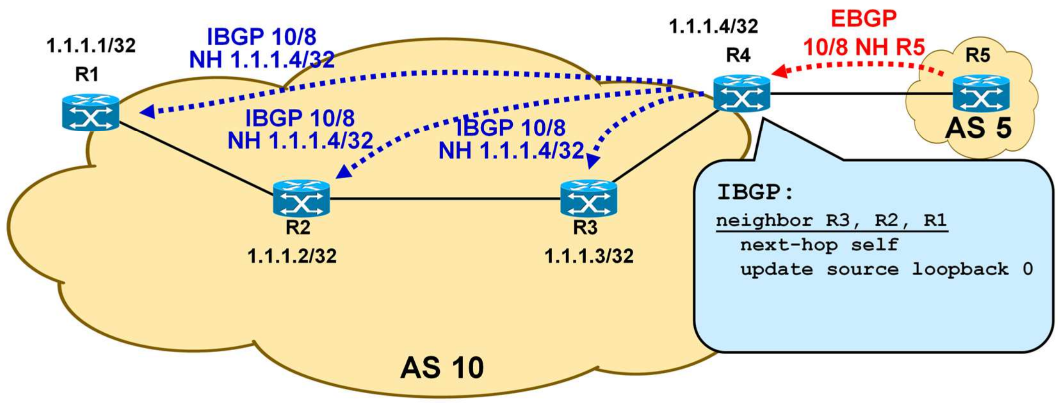
Agenda

- **Review ATM**
- **IP over WAN Problems (Traditional Approach)**
- **MPLS Principles**
- **Label Distribution Methods**
- **MPLS Details (Cisco)**
  - Internal Components
  - MPLS in Action
  - TDP, LTP
  - TTL
  - Traffic Engineering
  - MPLS and BGP
- **RFCs**

 MPLS v6.1 154

Appendix 3 - MPLS (v6.1)

BGP Standard Behavior



- **Good style: Use loopback addresses and next hop self**
  - BUT: Full mesh IBGP !!!
  - BUT: Each router has full routing table !!!
- **IGP is used to propagate loopback addresses**
  - 1.1.1.1/32, 1.1.1.2/32, 1.1.1.3/32, and 1.1.1.4/32
- **Note: BGP Synchronization Off**
  - Otherwise IBGP routes would never be copied into the routing table
  - IBGP updates would only be propagated by PE-router if this network is reachable via IGP

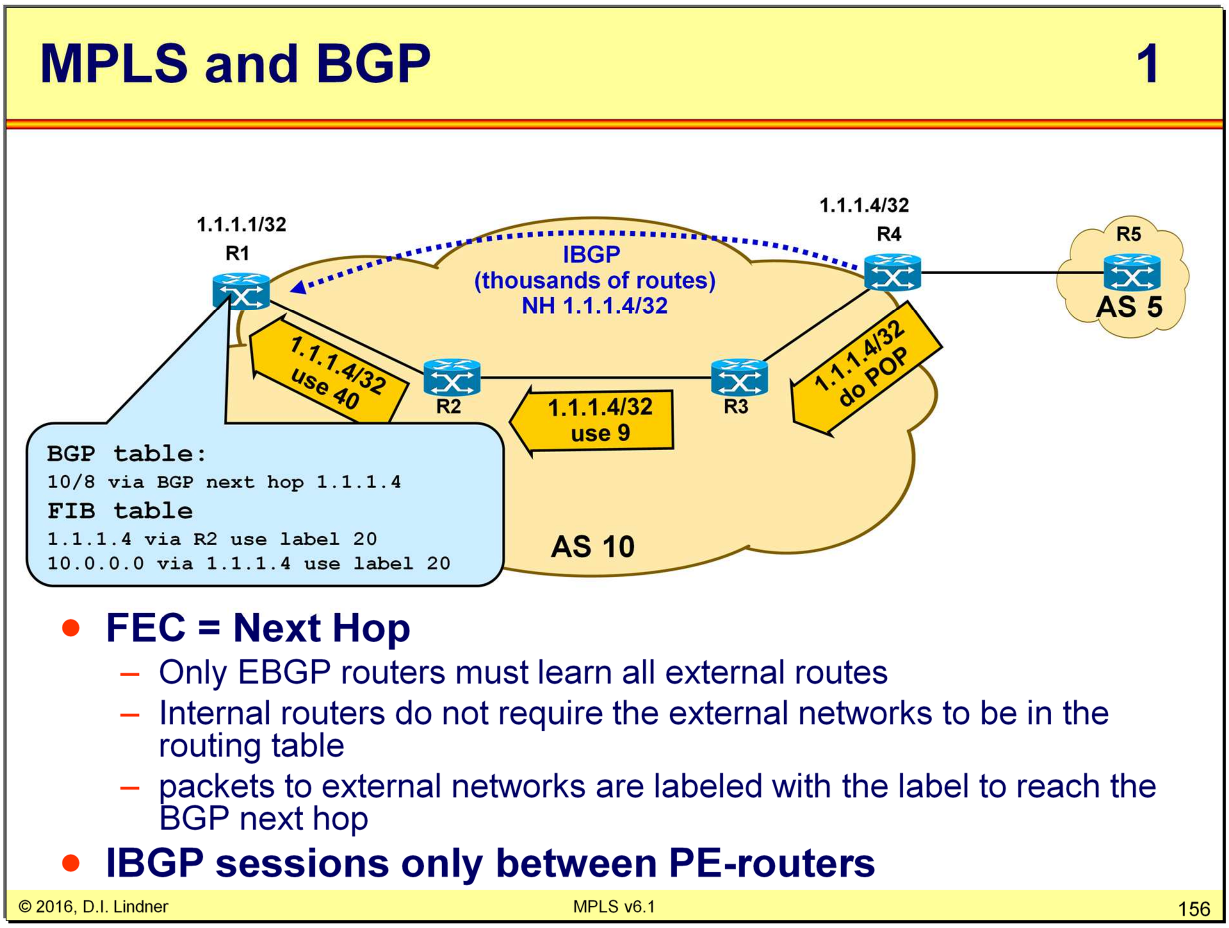
 MPLS v6.1 155

Note: Sync is on by default (Cisco). "Update source loopback" makes IBGP updates using the loopback address as source address of update messages.

Note: The loopback addresses are specified as neighbor addresses.

Note: Next-hop self is necessary for the PE-routers because BGP otherwise assumes R5 to be the next hop AND there is no label to R5 if the IGP was not started on the external link.

Do not summarize PE loopback addresses as it would break the label-switching path. Therefore it is a good practice to use host-route loopback addresses with subnet masks of 32 bits. Equivalently do not use next-hop-self on confederation boundaries as it would also break the label-switching path.



!!!!!!!! For IP Prefix learned by BGP no label is assigned. Instead the label of the BGP Next Hop address is used. !!!!!!!

For IGP derived routes a FEC represents an IP destination network.

For BGP derived routes a FEC represents the BGP Next Hop attribute.

This means that all routes which are imported by an EBGP Peer into an autonomous system are reachable via one and the same Label which points towards the EBGP Peers loopback address in the case NEXT HOP SELF is used on the EBGP Peer.

Therefore P routers don t need to run BGP because they are able to forward packets for external locations using the Label information derived from the EBGP Peers loopback address.

Advantages summary:

The BGP topology has been much simplified—only the AS edge routers need to run BGP with full Internet routing.

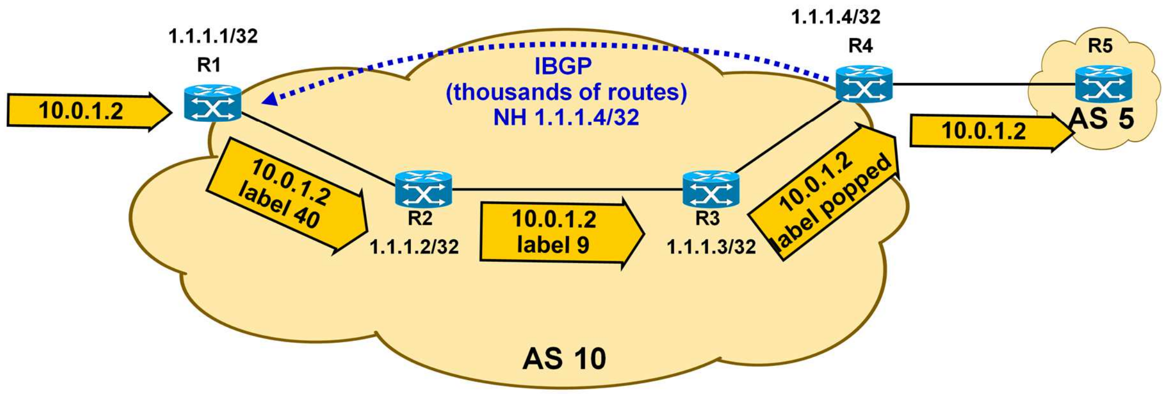
Core routers do not require much memory. The Internet routing table (by 2002) comprises about 100,000 routes which may require more than 50 MB of memory for the BGP table, IP routing table, and CEF's FIB table and distributed FIB tables).

Changes in the Internet do not impact core routers!

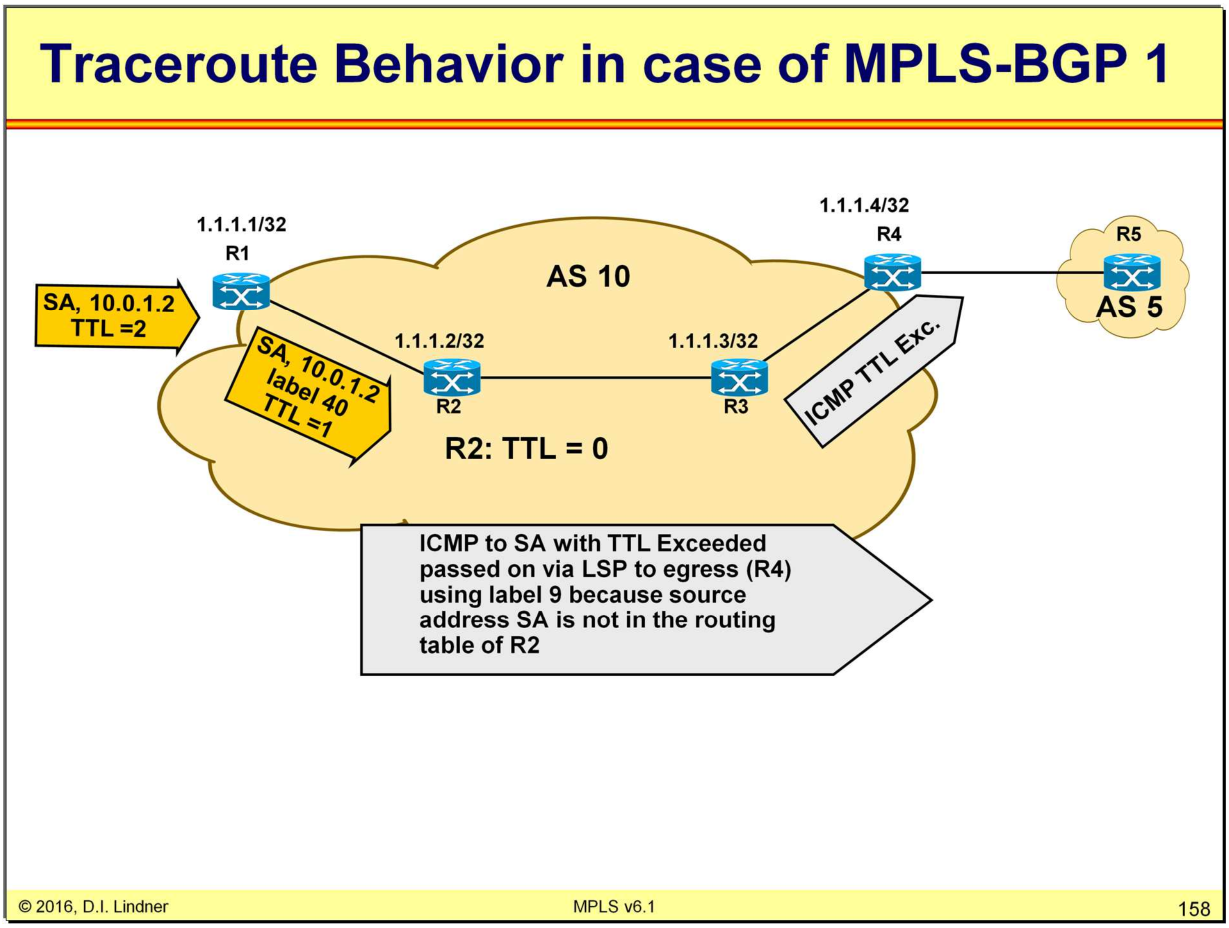
Private (RFC 1918) addresses can be used inside the core. Note that in this case the TTL propagation must be disabled—otherwise a traceroute would show private addresses.

Appendix 3 - MPLS (v6.1)

MPLS and BGP 2

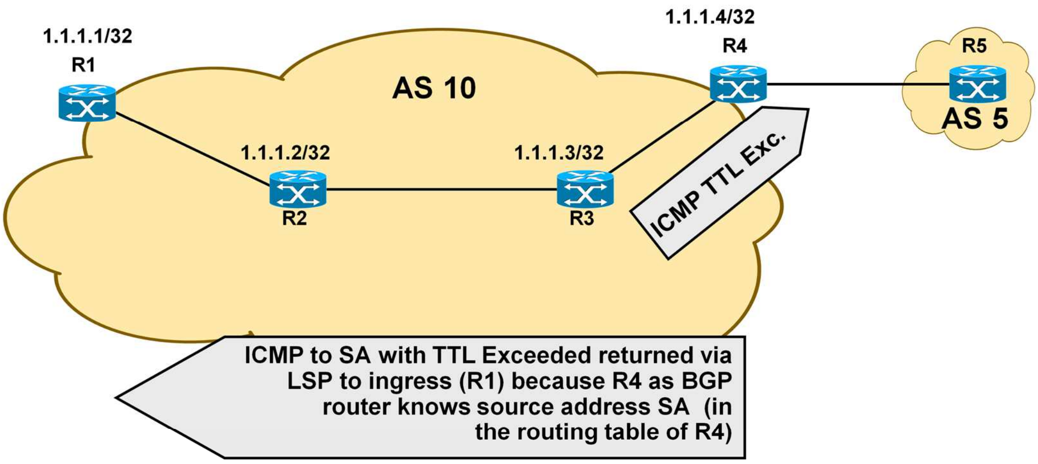


Appendix 3 - MPLS (v6.1)



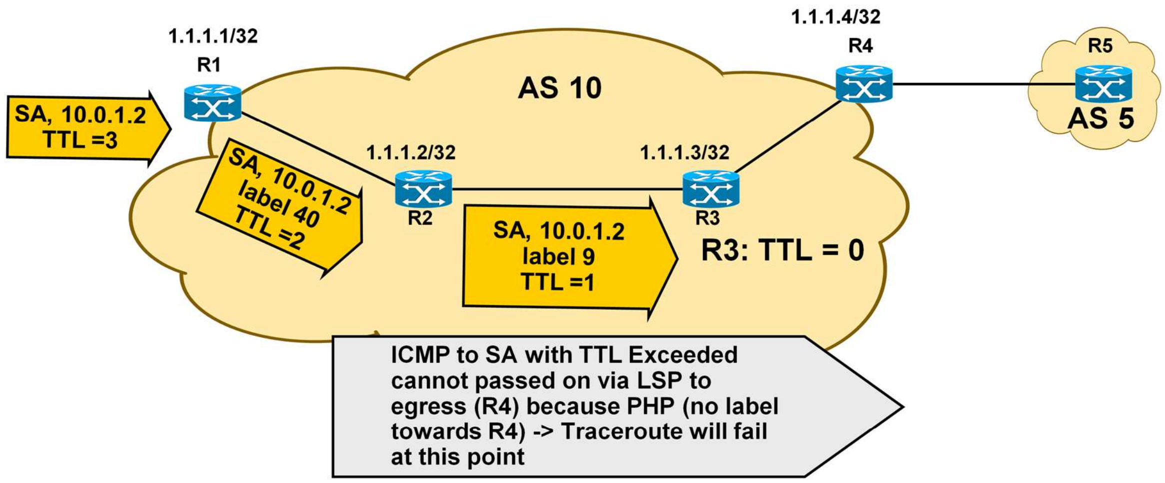
Appendix 3 - MPLS (v6.1)

Traceroute Behavior in case of MPLS-BGP 2



Appendix 3 - MPLS (v6.1)

Traceroute Behavior in case of MPLS-BGP 3



Agenda

- Review ATM
- IP over WAN Problems (Traditional Approach)
- MPLS Principles
- Label Distribution Methods
- MPLS Details (Cisco)
- RFCs

 MPLS v6.1 161

RFC References 1

- **RFC 3031**
  - Multiprotocol Label Switching Architecture
- **RFC 3032**
  - MPLS Label Stack Encoding
- **RFC 3036**
  - LDP Specification
- **RFC 3063**
  - MPLS Loop Prevention Mechanism
- **RFC 3270**
  - MPLS Support of Differentiated Services

RFC References 2

- **RFC 3443**
  - Time To Live (TTL) Processing in MPLS
- **RFC 3469**
  - Framework for Multi-Protocol Label Switching (MPLS)-based Recovery
- **RFC 3478**
  - Graceful Restart Mechanism for Label Distribution Protocol
- **RFC 3479**
  - Fault Tolerance for the Label Distribution Protocol (LDP)

