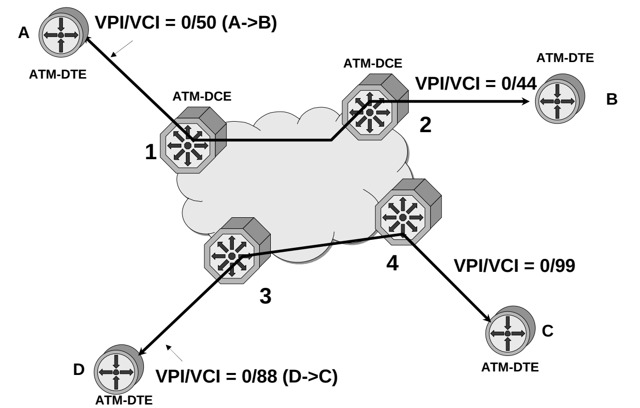
MPLS

Multi-Protocol Label Switching

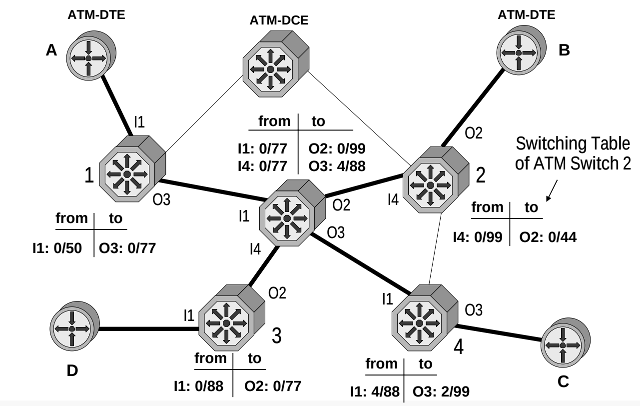
Agenda

- **Review ATM**
- **IP over WAN Problems (Traditional Approach)**
- **MPLS Principles**
- **Label Distribution Methods**
- **RFC's**

ATM Scenario -> VPI / VCI

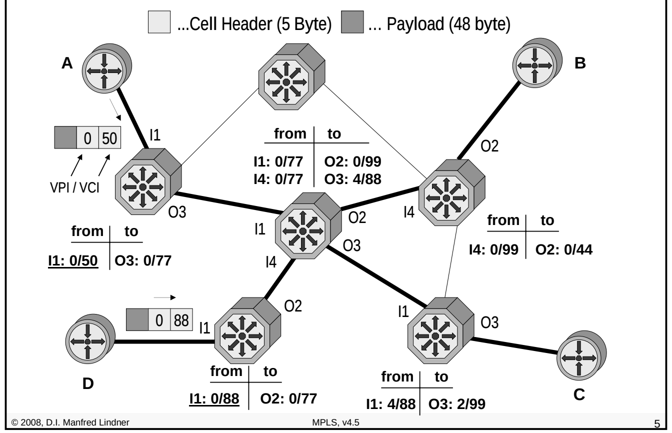


ATM Switching Tables

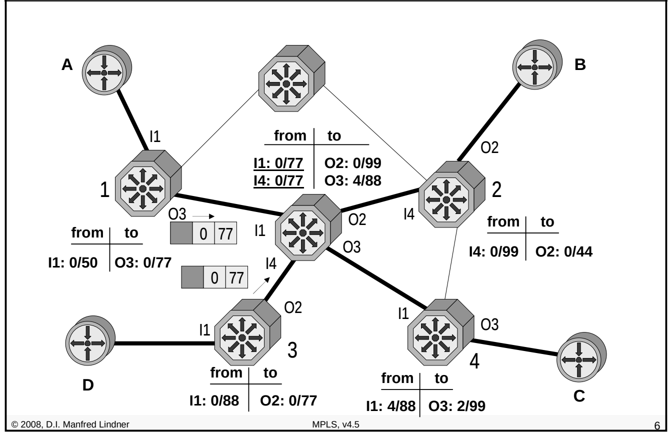


Appendix 3 - Multiprotocol Label Switching

Cell Forwarding / Label Swapping 1

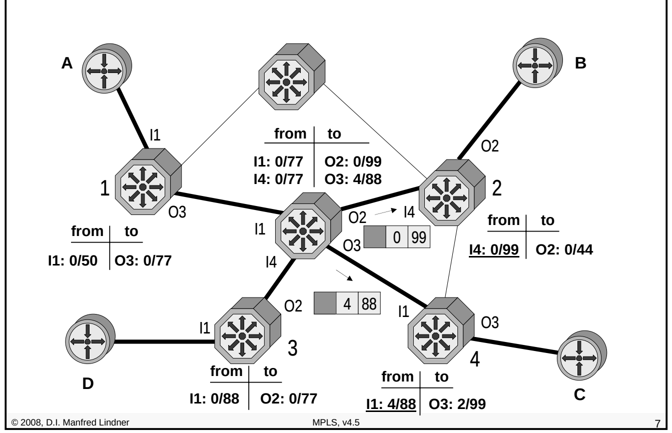


Cell Forwarding / Label Swapping 2

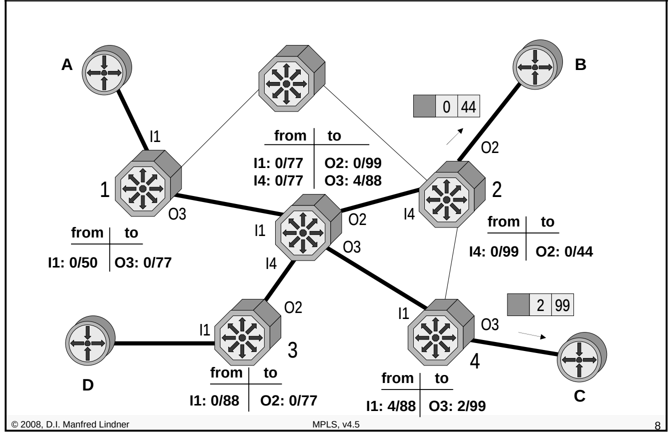


Appendix 3 - Multiprotocol Label Switching

Cell Forwarding / Label Swapping 3

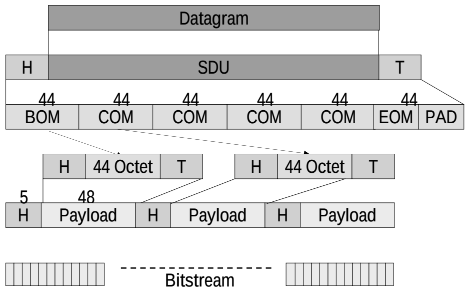


Cell Forwarding / Label Swapping 4



Segmentation Principle

- **Cells are much smaller than data packets**
  - Segmentation and Reassembly is necessary in ATM DTE's (!!!)
  - ATM DCE's (ATM switches) are not involved in that

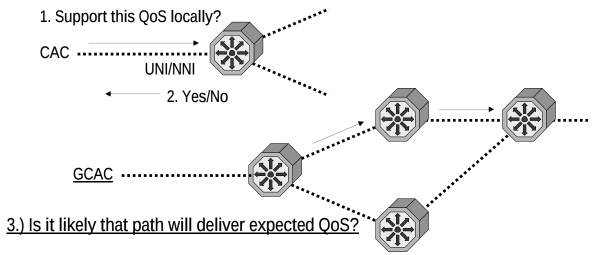


ATM Routing in Private ATM Networks

- **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

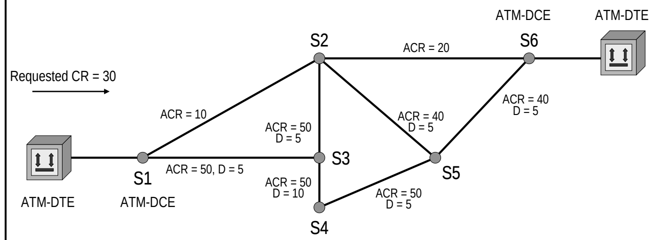
PNNI Routing

- **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



PNNI Routing (Simple QoS -> ACR only)

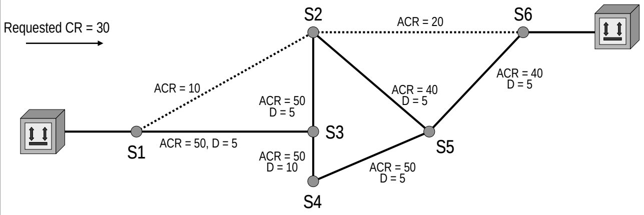
- **Operation of the GCAC**
  - CR … Cell Rate
  - ACR … Available Cell Rate
  - D … Distance like OSPF costs



Appendix 3 - Multiprotocol Label Switching

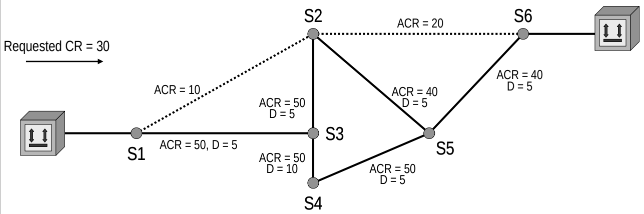
PNNI Routing

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



PNNI Routing

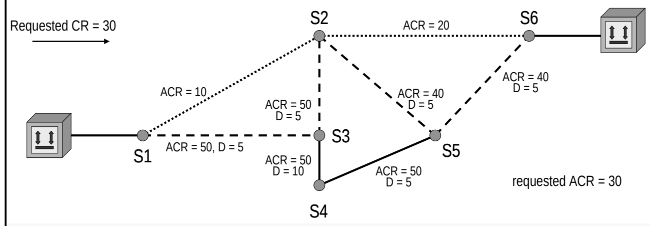
- **Operation of the GCAC**
  - 2) Next, shortest path(s) to the destination is (are) calculated
    - Metric component -> Distance value used



Appendix 3 - Multiprotocol Label Switching

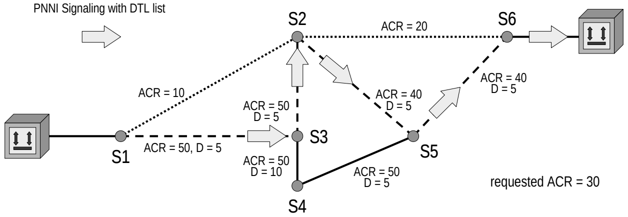
PNNI Routing

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



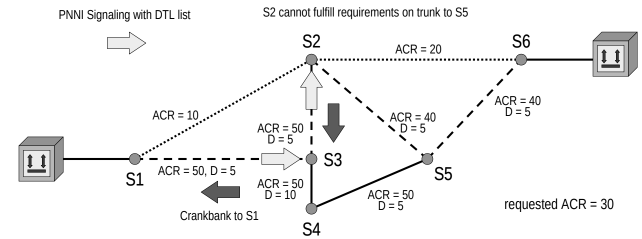
PNNI Routing - Source Routing

- **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



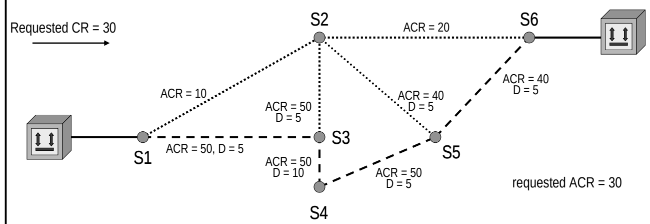
PNNI Routing - Crankbank

- **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



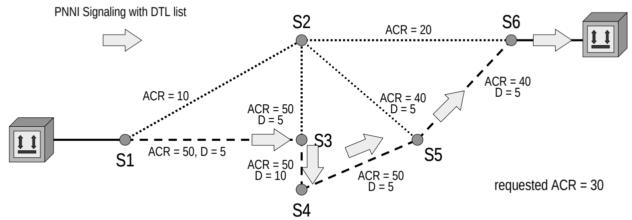
PNNI Routing - New Trial

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



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

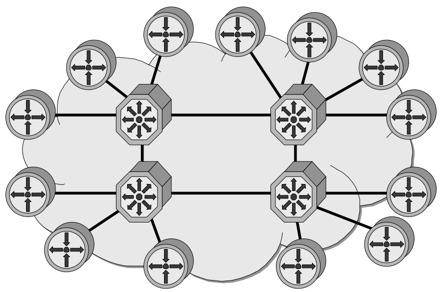
- **Review ATM**
- **<u>IP over WAN Problems (Traditional Approach)</u>**
- **MPLS Principles**
- **Label Distribution Methods**
- **RFC's**

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

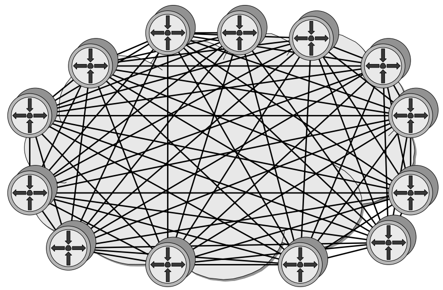
A Simple Physical Network ...

Physical wiring

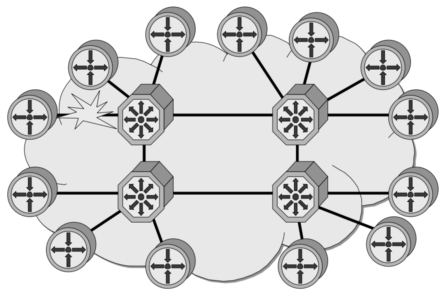


IP Data Link View (Non-NBMA)

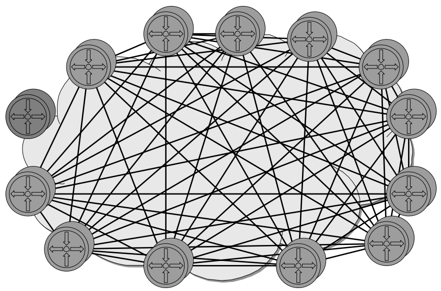
Every virtual circuit has its own IP Net-ID (subinterface technique)



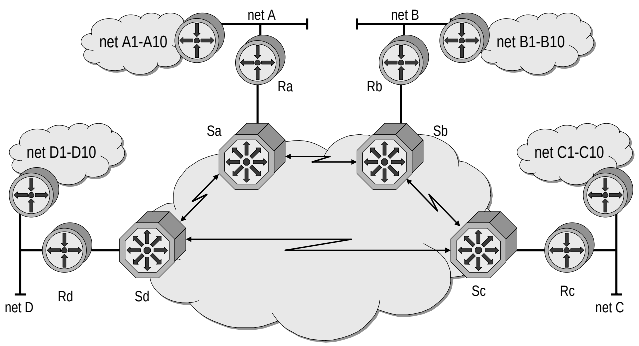
A Single Network Failure ...



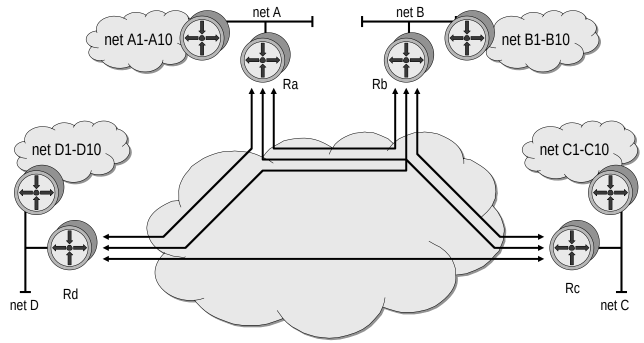
Causes Loss of Multiple IP Router Peers !!!



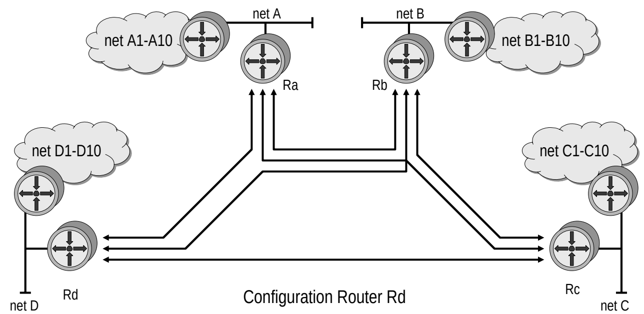
Example - Physical Topology



IP Connectivity through Full-mesh VC´s



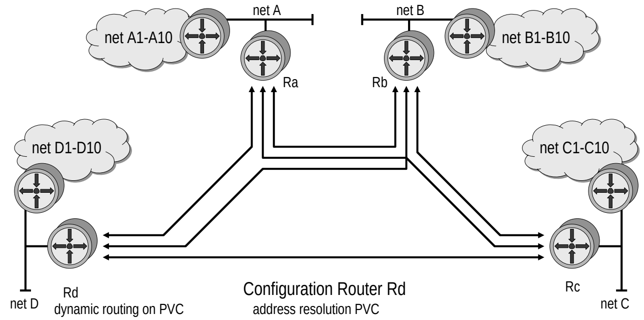
Static Routing/No Routing Broadcasts



Configuration Router Rd

static routing	address resolution PVC	address resolution SVC
net A via next hopRa	Ra map VPI/VCI Rd ⇨ Ra	Ra map ATM addr. Ra
net B via next hopRb	Rb map VPI/VCI Rd ⇨ Rb	Rb map ATM addr. Rb
net C via next hopRc	Rc map VPI/VCI Rd ⇨ Rc	Rc map ATM addr. Rc
every remote network listed here!		

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

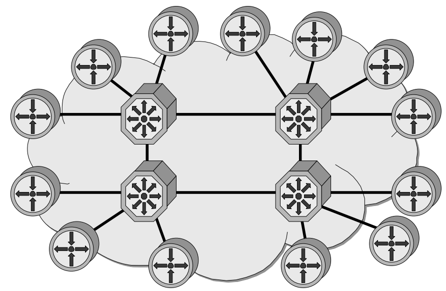
note: SVCs may be possible if Cisco neighbor command is specified for Cisco routing process because no automatic neighbor discovery is possible in this case

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

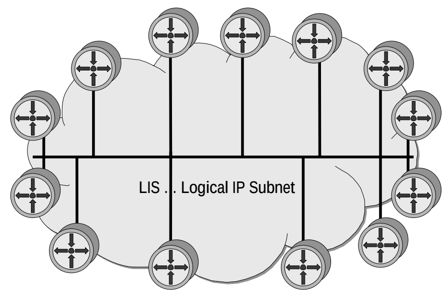
A Simple Physical Network ...

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

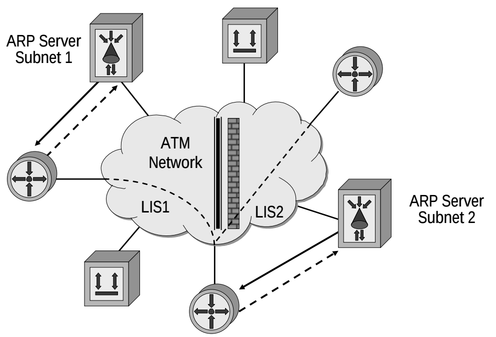


Some Solutions for the NBMA Problem

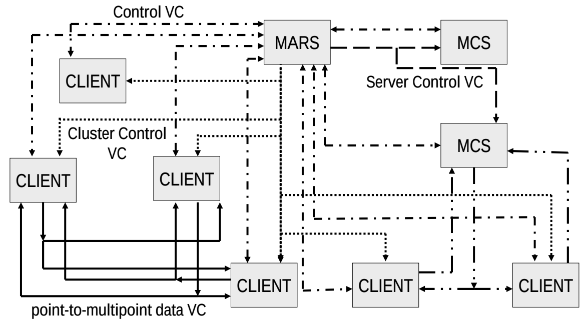
- ARP (Address Resolution Protocol) Server
  - keeps configuration overhead for address resolution small
  - but does not solve the routing issue (neighbor discovery and duplicate routing broadcasts on a single wire)
- MARS/MCS (Multicast Address Resolution Server / Multicast Server)
  - additional keeps configuration overhead for routing small
  - but does not solve the duplicate broadcast problem
- 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

RFC 2225 Operation (Classical IP over ATM)

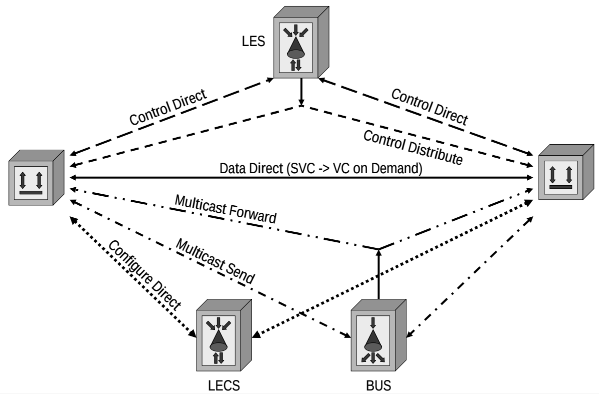
- **ARP server for every LIS**
  - multiple hops for communication between Logical IP Subnets



MARS/MCS Architecture



LANE Connections

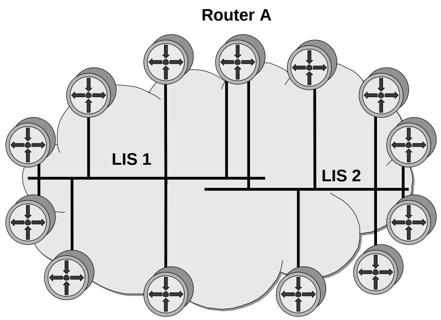


Scalability Aspects

- **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 duplicate broadcast 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)

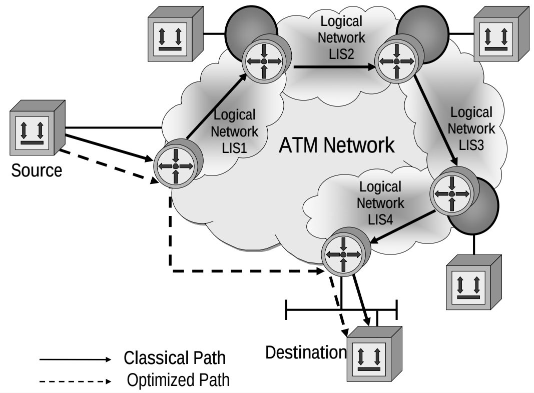
IP router A connects LIS1 and LIS2



Some Solutions for the ROLC Problem

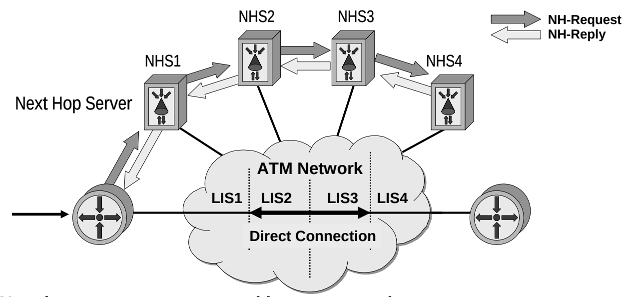
- **NHRP (Next Hop Resolution Protocol)**
  - creates an ATM shortcut between two systems of different LIS´s
- **MPOA (Multi Protocol Over ATM)**
  - LANE + NHRP combined
  - creates an ATM shortcut between two systems of different LIS´s
- **In both methods**
  - the ATM shortcut is created if traffic 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

Wish for Optimized Connectivity



© 2008, D.I. Manfred Lindner

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

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

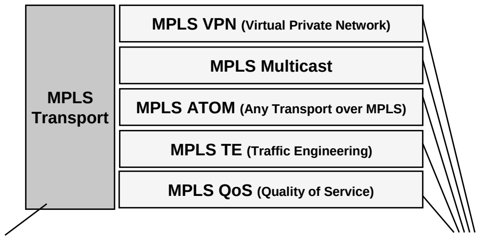
Basic Ideas to Solve the Problems

- **Make ATM topology visible to IP routing**
  - to solve the scalability problems
  - an 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)
- **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

Agenda

- **Review ATM**
- **IP over WAN Problems (Traditional Approach)**
- **<u>MPLS Principles</u>**
- **Label Distribution Methods**
- **RFC's**

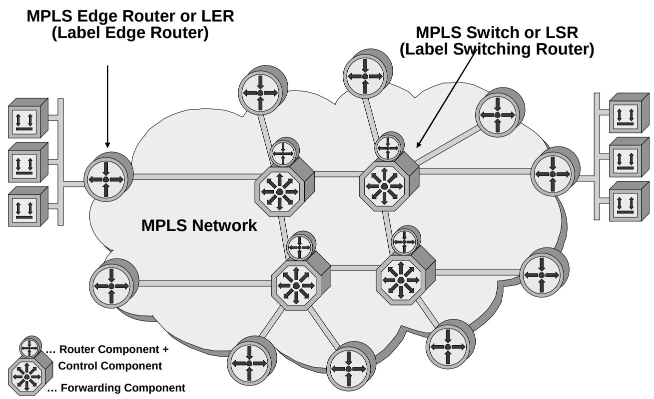
Datenkommunikation 384.081 - SS 2007

Appendix 3 - Multiprotocol Label Switching

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
  - <u>labels</u> used for switching
- **MPLS Network consists of**
  - MPLS Edge Routers and MPLS Switches
- **Edge Routers and Switches**
  - exchange routing information about L3 IP networks
  - exchange forwarding information about the actual usage of labels

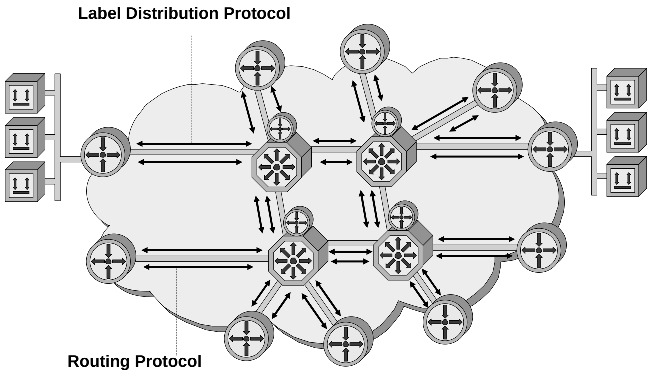
MPLS Network



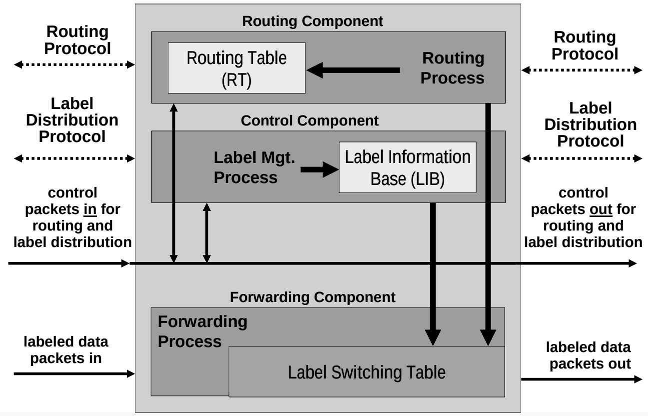
MPLS LSR Internal Components

- **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 (switching table) to perform packet forwarding -> "label swapping"

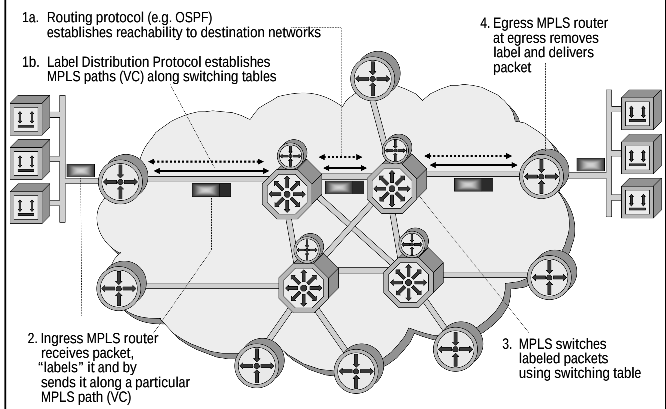
MPLS Control Communication



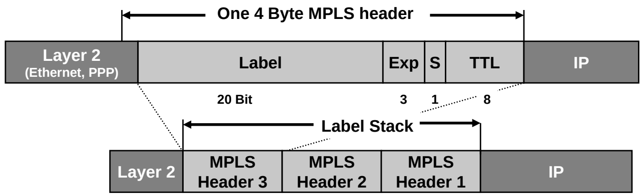
Generic Overview of MPLS LSR Internal Processes and Communication



MPLS Basic Operations



MPLS Header: Frame Mode



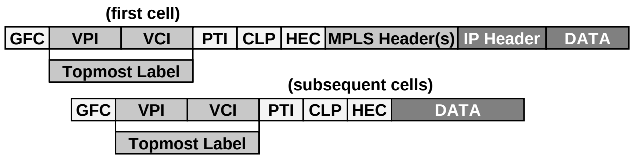
- **"Layer 2.5" can be used over Ethernet, 802.3 or PPP links**
  - note: 2.5 means 32 bit
  - 20-bit MPLS label (Label)
  - 3-bit experimental field (Exp)
    - could be copy of IP Precedence -> MPLS QoS like IP QoS with DiffServ Model based on DSCP
  - 1-bit bottom-of-stack indicator (S)
    - Labels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

MPLS 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):



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 is based on the network layer destination address

Label Binding

- **Two neighboring LSR´s 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
    - 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

 MPLS, v4.5 57

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

 MPLS, v4.5 58

Some FEC Examples for Topology Driven

- **FEC´s 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

 MPLS, v4.5 59

Label Distribution

- **MPLS architecture allows an LSR to distribute bindings to LSR´s 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, v4.5 60

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

Datenkommunikation 384.081 - SS 2007

Appendix 3 - Multiprotocol Label Switching

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

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 <u>egress method</u> 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 LSR´s
- 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 LSR´s
- inherent loop prevention
- slower convergence

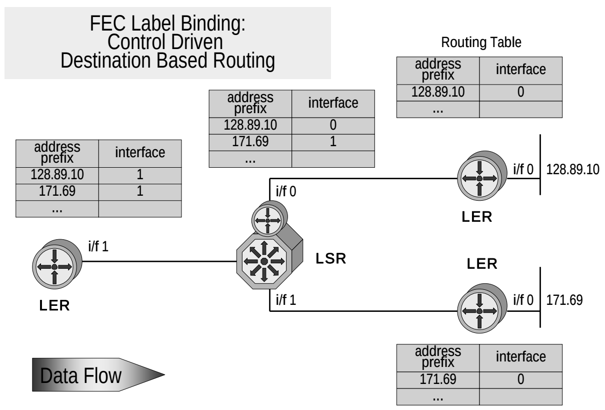
Ordered Control - Ingress

- in case of <u>ingress method</u> 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)

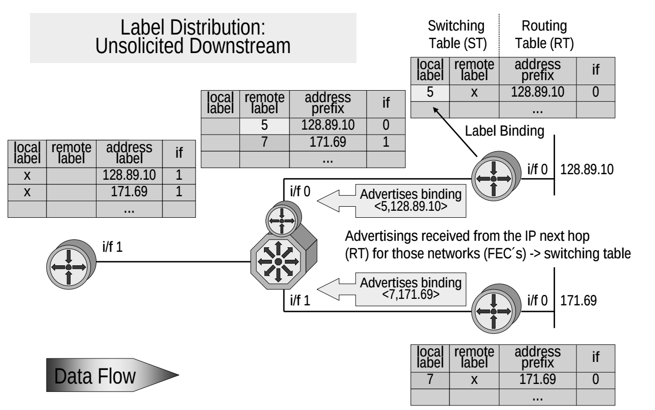
Agenda

- **Review ATM**
- **IP over WAN Problems (Traditional Approach)**
- **MPLS Principles**
- **<u>Label Distribution Methods</u>**
  - <u>Unsolicited Downstream</u>
  - Downstream On Demand
- **RFC's**

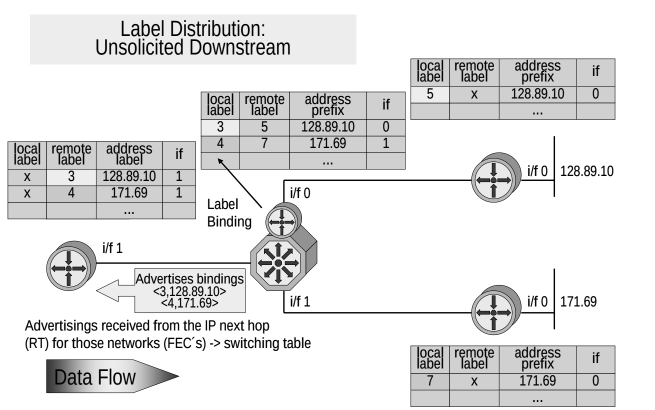
Routing Table Created by Routing Protocol



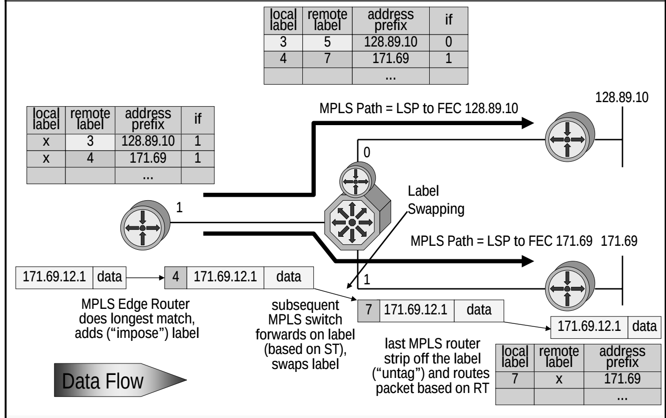
Labels Sent by LDP



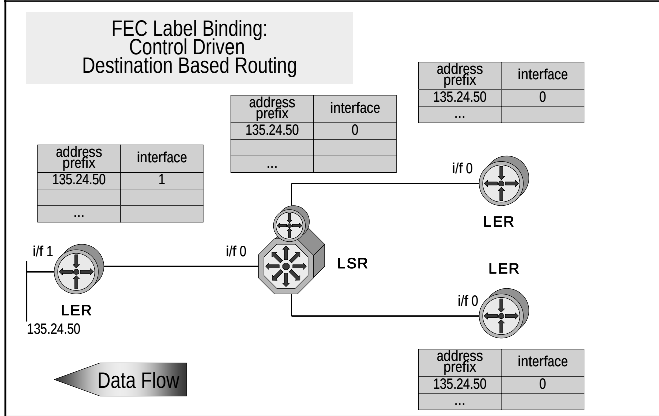
Labels Sent and Switching Table Entry Created by MPLS Switch



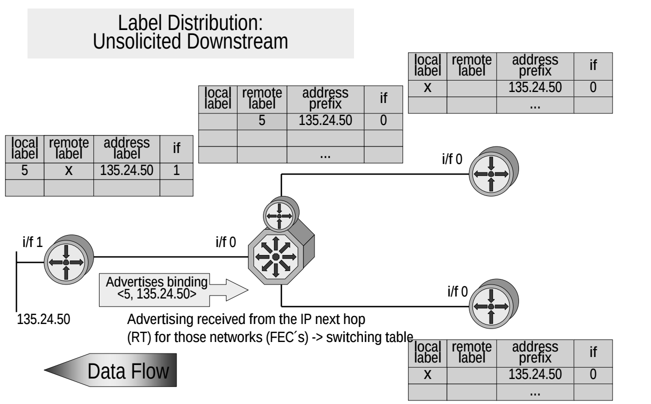
MPLS Switched Packets



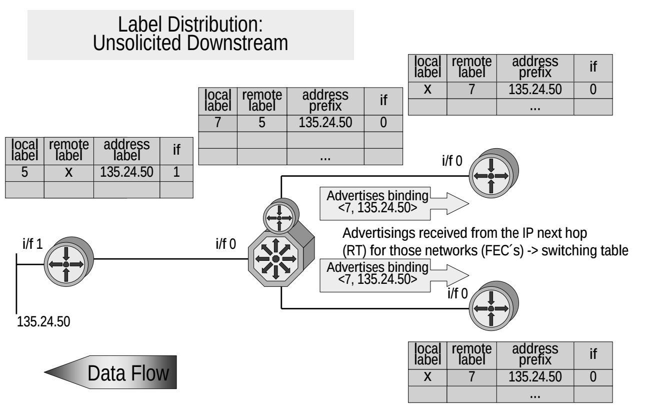
Routing Table Created by Routing Protocol



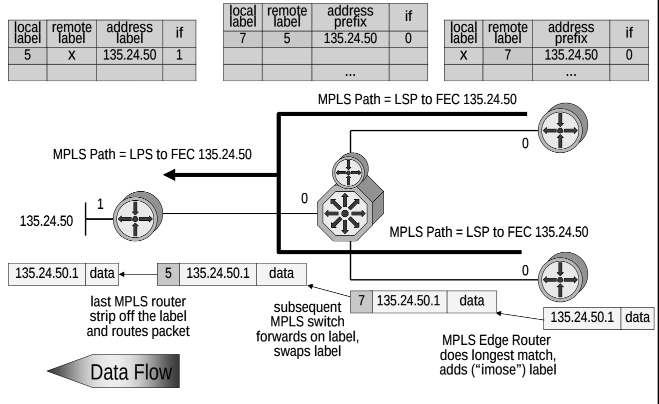
Labels Sent by LDP



Labels Sent and Switching Table Entry Created by MPLS Switch



Label Merging - LSP Merging

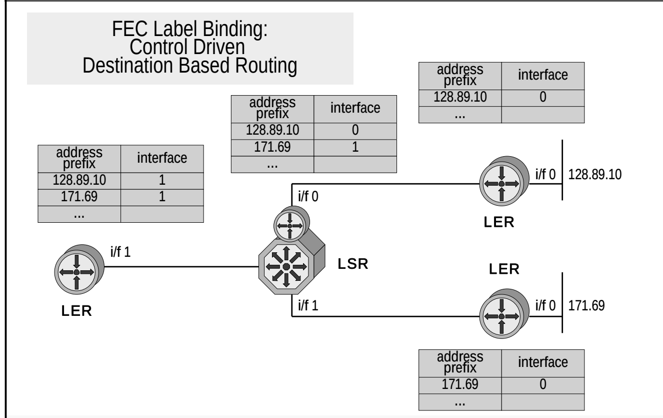


Agenda

- **Review ATM**
- **IP over WAN Problems (Traditional Approach)**
- **MPLS Principles**
- **Label Distribution Methods**
  - Unsolicited Downstream
  - Downstream On Demand
- **RFC's**

© 2008, D.I. Manfred Lindner

Routing Table Created by Routing Protocol



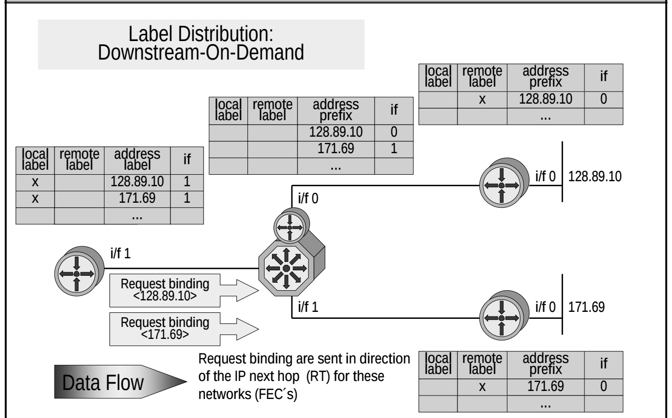
address prefix	interface
128.89.10	1
171.69	1
...	

address prefix	interface
128.89.10	0
171.69	1
...	

address prefix	interface
128.89.10	0
...	

address prefix	interface
171.69	0
...	

Labels Requested by MPLS Edge Routers



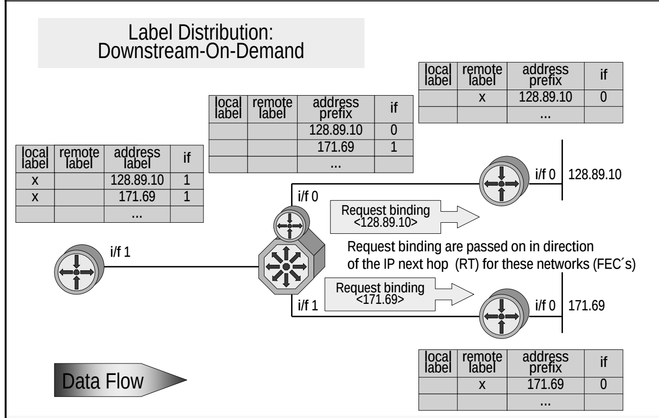
local label	remote label	address label	if
x		128.89.10	1
x		171.69	1
		...	

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

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

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

Labels Requested by MPLS Switch



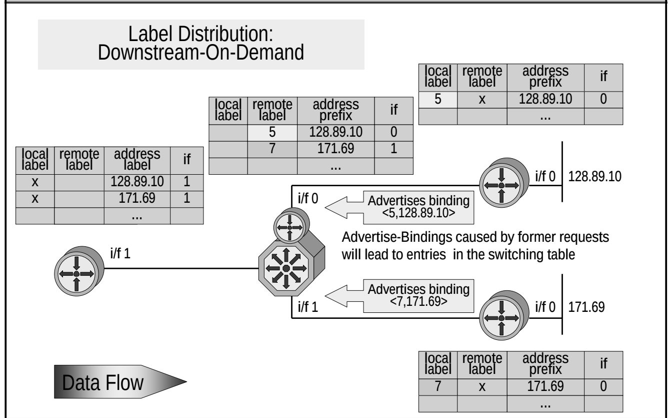
local label	remote label	address label	if
x		128.89.10	1
x		171.69	1
		...	

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

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

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

Labels Allocated by MPLS Edge Router



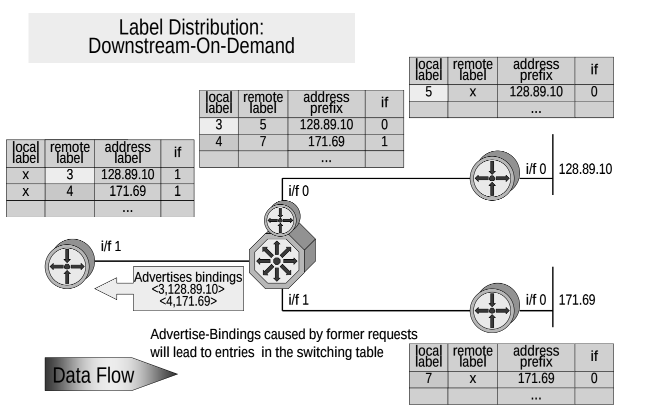
local label	remote label	address label	if
x		128.89.10	1
x		171.69	1
		...	

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

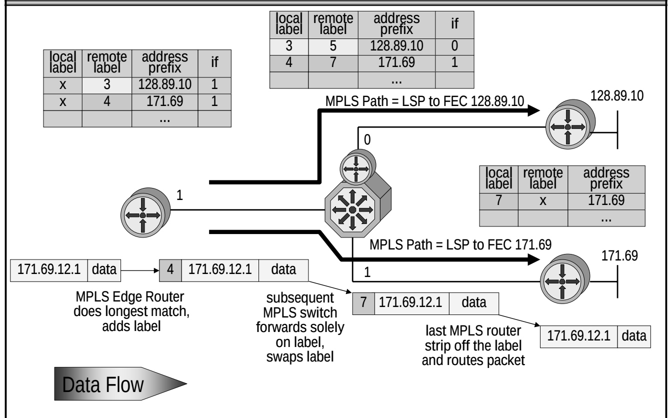
local label	remote label	address prefix	if
5	x	128.89.10	0
		...	

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

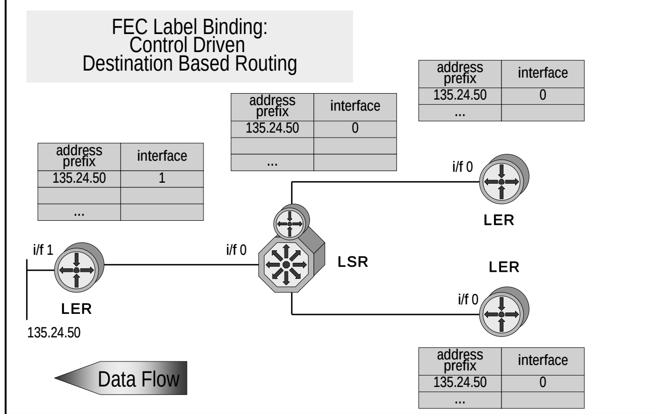
Labels Allocated and Switching Table Built by MPLS Switch



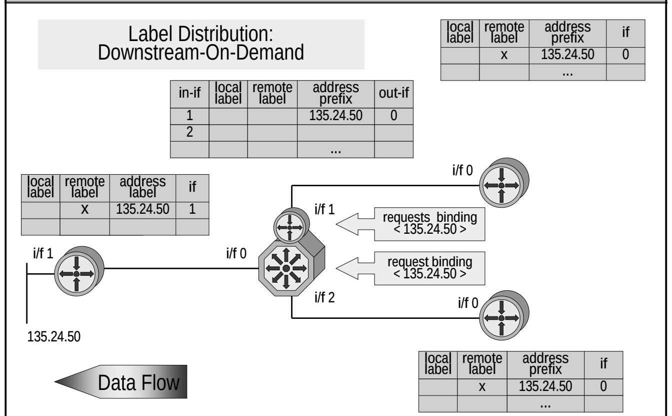
MPLS Switched Packets



Routing Table Created by Routing Protocol

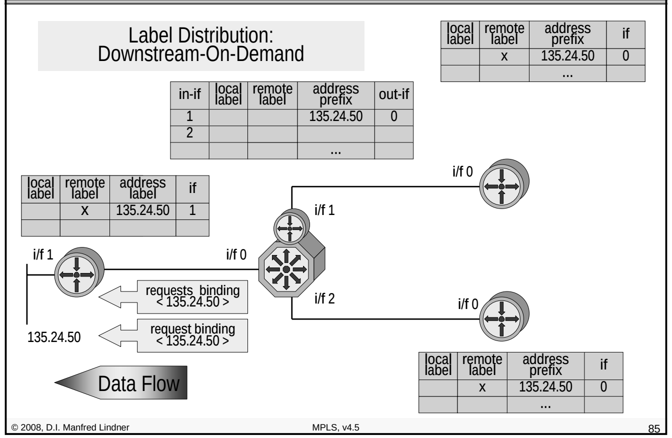


Labels Requested by MPLS Edge Routers

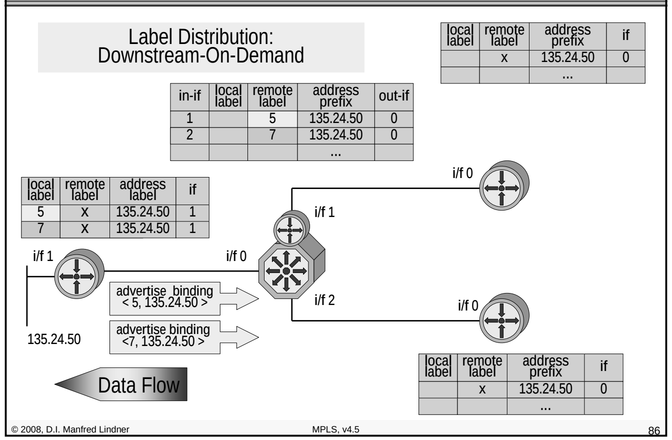


Appendix 3 - Multiprotocol Label Switching

Labels Requested by MPLS Switch

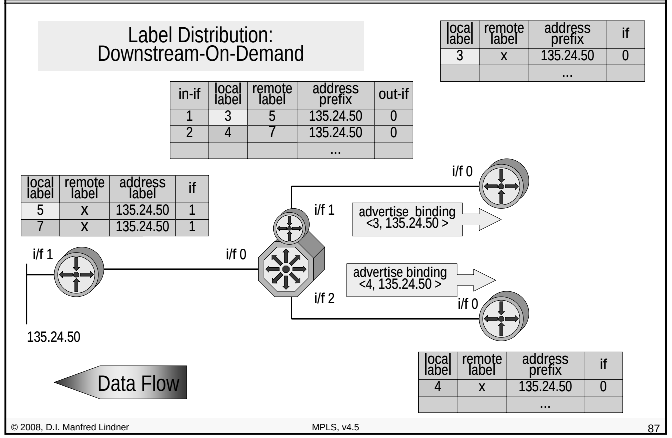


Labels Allocated by MPLS Edge Router

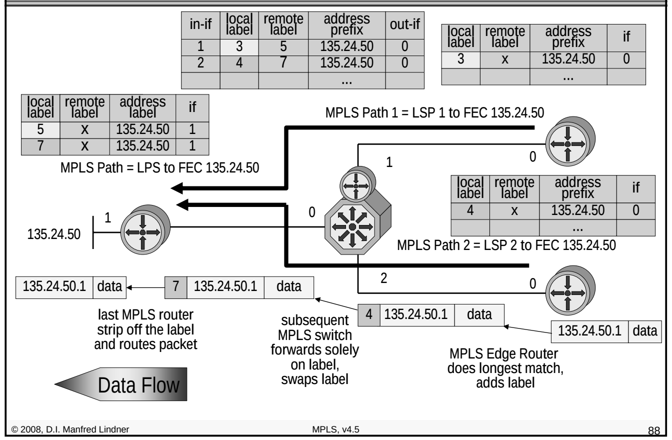


Appendix 3 - Multiprotocol Label Switching

Labels Allocated and Switching Table Built by MPLS Switch



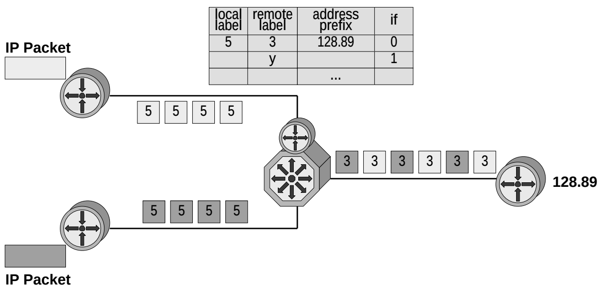
Two Separate LSP´s



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

Label Switching and ATM



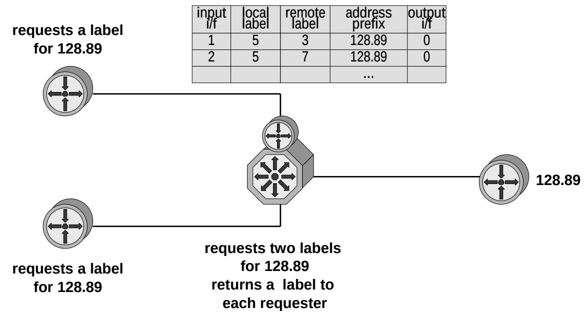
local label	remote label	address prefix	if
5	3	128.89	0
	y		1
		...	

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.

Datenkommunikation 384.081 - SS 2007

Appendix 3 - Multiprotocol Label Switching

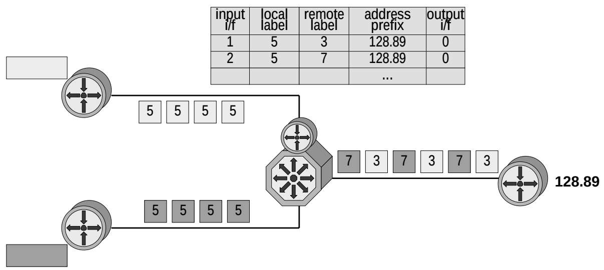
Label Distribution Solution for ATM



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**

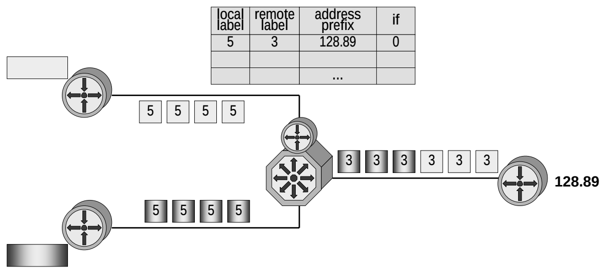
Label Distribution Solution for ATM



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

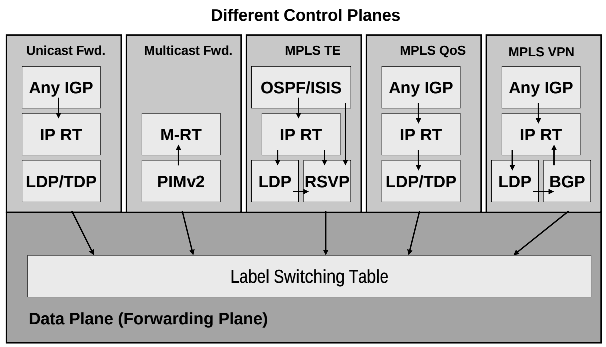


- **ATM switch avoids interleaving of frames**
  - VC Merge technique
  - looking for AAL5 trailers and storing corresponding cells of a frame until AAL5 trailer is seen

Agenda

- **Review ATM**
- **IP over WAN Problems (Traditional Approach)**
- **MPLS Principles**
- **Label Distribution Methods**
- **<u>RFC's</u>**

MPLS Applications and MPLS Control Plane



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)

